PREDIKSI KELANGSUNGAN HIDUP PENDERITA HEPATITIS DENGAN MENGGUNAKAN NAIVE BAYES, K-NN DAN SVM

Muhamad Arif Rahman[1], Rosmala Dwi[2]
Universitas Nahdlatul Ulama Lampung
Lampung, Indonesia
e-mail: [1]arifrahman47@gmail.com , [2]rosmaladwi65@gmail.com

Abstrak

Hepatitis adalah salah satu penyakit yang mengancam kesehatan bukan hanya di negara Indonesia akan tetapi juga dunia. Melihat dari bahayanya penyakit hepatitis yang mengancam kesehatan, maka diperlukan penanggulangan yang tepat untuk menangani penyakit hepatitis. Salah satu cara menangani penyakit hepatitis bisa dengan memanfaatkan data mining. Data mining dapat dimanfaatkan untuk memprediksi penyakit hepatitis berdasarkan data-data masa lalu. Pada penelitian ini akan memanfaatkan data mining untuk memprediksi kelangsungan hidup penderita hepatitis. Pada penelitian ini juga akan dilakukan perbandingan algoritma Naïve Bayes, K-NN dan SVM. Berdasarkan hasil penelitian yang telah dilakukan akurasi tertinggi terdapat pada algoritma Naïve Bayes dengan nilai akurasi sebesar 93.55%.

Kata kunci: *hepatitis, prediksi, naïve bayes, k-nn, svm.*

Abstract

Hepatitis is a disease that threatens health not only in Indonesia but also in the world. Seeing from the dangers of hepatitis that threatens health, it is necessary to take appropriate measures to deal with hepatitis. One way to deal with hepatitis can be by utilizing data mining. Data mining can be used to predict hepatitis disease based on past data. In this study, data mining will be used to predict the survival of hepatitis sufferers. This research will also compare the Naïve Bayes algorithm, K-NN and SVM. Based on the results of research that has been done, the highest accuracy is found in the Naïve Bayes algorithm with an accuracy value of 93.55%.

Keywords*: hepatitis, predict, naïve bayes, k-nn, svm.*

1. Pendahuluan

Hepatitis merupakan penyakit peradangan hati karena berbagai sebab diantaranya seperti bakteri, virus, penyakit autoimun, obat-obatan, lemak berlebihan, alkohol dan zat berbahaya lainnya. Penyakit hepatitis terdiri dari beberapa golongan diantaranya A, B, C, D, dan E. Di negara Indonesia, menurut riset kesehatan pada tahun 2013 dari 100 penduduk terdapat 10 orang menderita hepatitis, penduduk Indonesia yang menderita penyakit hepatitis, terkena hepatitis B atau C. Dalam sebuah wawancara dengan Ketua PB Perhimpunan Peneliti Hati Indonesia DR. dr. Irsan Hasan. Sp. PD-KGEH.FINASIM mengatakan bahwa 1 dari 4 penderita hepatitis akan meninggal karena kanker hati atau gagal hati, sehingga diperlukan perawatan dan penangan yang tepat bagi pasien pengidap hepatitis. Dengan memanfaatkan teknik data mining, dapat digali informasi mengenai data-data pasien hepatitis di masa lalu melalui catatan rekam medik, sehingga kondisi pasien hepatitis di masa depan dapat diprediksi. Terdapat berbagai macam metode pada data mining yang bisa digunakan untuk prediksi kelangsungan hidup pasien hepatitis.

Berikut ini merupakan beberapa penelitian yang telah dilakukan oleh peneliti terdahulu, penelitian pertama dilakukan oleh Siti Khomsah melakukan penelitian tentang prediksi kelangsungan hidup penderita hepatitis kronik dengan metode-metode klasifikasi. Penelitian ini bertujuan membandingkan kinerja dari empat metode yaitu k-nn, naïve bayes, decision tree dan random forest. Pengujian pada penelitian ini menggunakan k-fold cross validation dengan nilai k=10. Hasil penelitian menunjukan bahwa metode yang memilki akurasi tertinggi yaitu random forest dengan nilai akurasi sebesar 79.35%.

Metode yang menghasilkan akurasi terendah yaitu k-nn dengan nilai akurasi sebesar 79.31%. selain menguji akurasi, penelitian ini juga menguji kinerja dari masing-masing metode yaitu naïve bayes, decision tree, random forest dengan nilai AUC 0.8 yang berarti classifier level good, sedangkan untuk metode k-nn memiliki nilai AUC 0.7 yang berarti nilai classifier level fair [1].

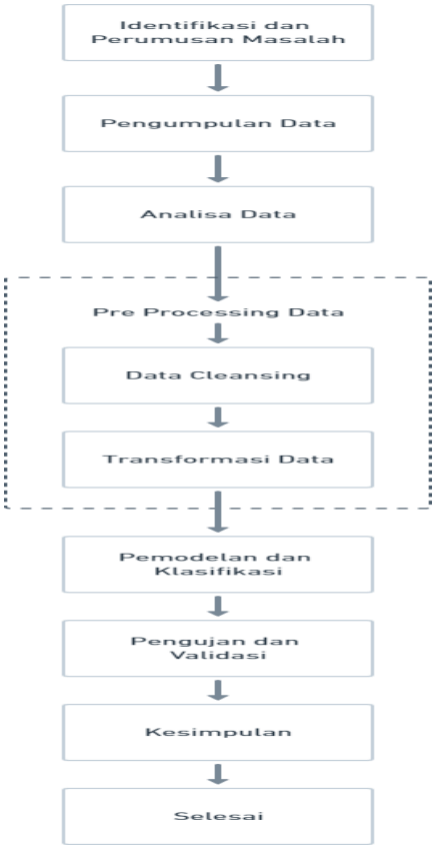
Penelitian selanjutnya dilakukan oleh Wisti Dwi Septiani dengan tujuan mengkomparasi algoritma C.45 dan Naïve Bayes untuk memprediksi kelangsungan hidup penderita hepatitis. Hasil penelititan menunjukan metode yang memiliki nilai akurasi tertinggi yaitu naïve bayes dengan nilai akurasi sebesar 83.71% dan terendah metode C.45 dengan nilai akurasi sebesar 77.29%. Metode Naïve bayes lebih baik dalam memprediksi jika dilihat dari segi akurasi, namun dari segi nilai AUC metode C.45 lebih baik dari pada Naïve Bayes karena memiliki nilai lebih tinggi yaitu 0.846 sedangkan naïve bayes memiliki nilai AUC sebesar 0.812 [2]. Penelitian selanjutnya dilakukan oleh Nurajizah dengan tujuan memprediksi penyakit jantung menggunakan metode Support Vector Machine berbasis particle swarm optimization dan menghasilkan akurasi sebesar 88.61% dan nilai AUC sebesar 0.899. Pada penelitian tersebut algoritma SVM menunjukan kinerja yang sangat baik. Dengan kinerja yang sangat baik, SVM merupakan salah satu metode yang akan digunakan dalam penelitian ini [3].

Penelitian lain juga dilakukan oleh Sulastri tentang prediksi penyakit hepatitis dengan menggunakan tiga algoritma yang di komparasi yaitu K-NN, Naïve Bayes dan Neural Network dan memperoleh hasil Naïve Bayes dengan nilai akurasi 76.62%, K-NN dengan nilai akurasi 93% dan Neural Network Dengan Nilai Akurasi 82.97%. Dari penelitian tersebut metode K-NN memperoleh nilai akurasi tertinggi, di banding dengan metode lainnya [4].

Berdasarkan studi literatur di atas pada penelitian ini akan dilakukan komparasi metode Naïve bayes, K-NN dan SVM dan berfokus pada pencarian metode yang dapat menghasilkan akurasi tertinggi untuk memprediksi kelangsungan hidup penderita hepatitis. Hasil dari penelitian ini adalah metode dengan akurasi tertinggi yang dapat digunakan untuk prediksi kelangsungan hidup penderita hepatitis.

2. Metode Penelitian

Berikut ini merupakan tahapan penelitian yang dijelaskan pada Gambar 1:



Gambar 1. Tahapan Penelitian

2.1 Data

Pada penelitian ini data pasien hepatitis didapatkan dari repository UCI yang dapat di unduh melalui link archive.ics.uci.edu/ml/datasets/Hepatitis. Dataset terdiri dari 20 atribut yang berisi 19 atribut mengenai profil pasien, gejala medis, pemeriksaan fisik, tes lab dan pengobatan yang sudah dilakukan dan 1 atribut keputusan. Atribut keputusan berupa Meninggal yang berjumlah 32 dan Hidup yang berjumlah 123. Berikut merupakan contoh data yang di tunjukan pada tabel 1.

Tabel 1. Atribut pada Dataset

No	Atribut	Domain Nilai	Keterangan
1	Keputusan	Meninggal, Hidup	Menunjukkan pasien hidup/ meninggal karena atribut yg ada
2	Umur	Numerik	Umur pasien
3	Jenis Kelamin	Laki-laki, Perempuan	Jenis kelamin pasien
4	Steroid	Tidak, Ya	Pasien menjalani terapi steroid ?
5	Antiviral	Tidak, Ya	Pasien menjalani terapi antiviral ?
6	Fatigue	Tidak, Ya	Pasien mengalami kelelahan akut ?
7	Malaise	Tidak, Ya	Pasien mengalami malaise ?
8	Anorexia	Tidak, Ya	Pasien mengalami anorexia ?
9	Liver Big	Tidak, Ya	Pembesaran liver ?
10	Liver Firm	Tidak, Ya	Liver pasien mengeras ?
11	Spleen Palpable	Tidak, Ya	Kondisi limfa lebih besar dari normal?
12	Spiders	Tidak, Ya	Kondisi pembuluh darah upnormal pada kulit?
13	Ascites	Tidak, Ya	Terjadi penumpukan cairan pada rongga perut?
14	Varices	Tidak, Ya	Terjadi pembekakan pembuluh darah vena?
15	Bilirubin	Numerik	Nilai kadar bilirubin
16	Alk Phosphate	Numerik	Nilai kadar alkain phosphate dalam liver
17	SGOT	Numerik	Nilai SGOT
18	Albumin	Numerik	Nilai kadar albumin
19	Protime	Numerik	Uji masa protrombin
20	Histology	Tidak, Ya	Pasien menjalani pemeriksaan histology?

2.2 Preprocessing

2.2.1 Cleansing Data

Pada dataset yang null bernilai ”?” dalam proses ini akan di rubah nilainya menjadi 0.

2.2.2 Transformasi data

Untuk memudahkan proses klasifikasi, maka nilai-nilai pada atribut diseragamkan menjadi tipe data numerik. Sehingga data yang memiliki tipe binominal akan diubah menjadi numerik.

Tabel 2. Transformasi data

No	Atribut	Domain Nilai	Keterangan
1	Keputusan	Meninggal, Hidup	Meninggal = 1, Hidup = 2
2	Jenis Kelamin	Laki-laki, Perempuan	Laki-laki = 1, Perempuan = 2
3	Steroid	Tidak, Ya	Tidak = 1, Ya = 2
4	Antiviral	Tidak, Ya	Tidak = 1, Ya = 2
5	Fatigue	Tidak, Ya	Tidak = 1, Ya = 2
6	Malaise	Tidak, Ya	Tidak = 1, Ya = 2
7	Anorexia	Tidak, Ya	Tidak = 1, Ya = 2
8	Liver Big	Tidak, Ya	Tidak = 1, Ya = 2
9	Liver Firm	Tidak, Ya	Tidak = 1, Ya = 2
10	Spleen Palpable	Tidak, Ya	Tidak = 1, Ya = 2
11	Spiders	Tidak, Ya	Tidak = 1, Ya = 2
12	Ascites	Tidak, Ya	Tidak = 1, Ya = 2
13	Varices	Tidak, Ya	Tidak = 1, Ya = 2
14	Histology	Tidak, Ya	Tidak = 1, Ya = 2

2.3 Klasifikasi

Proses klasifikasi dilakukan dengan membandingkan tiga metode yaitu Naïve Bayes, K-NN dan SVM.

2.3.1 Naïve Bayes

Merupakan suatu sistem klasifikasi statistic ntuk memprediksi probabilitas keanggotaan pada suatu class. Metode ini didasari oleh teori Bayes yang memiliki sistem klasifikasi seperti decision tree dan neural network. Metode naïve bayes memiliki akurasi dan kecepatan yang tinggi pada saat diaplikasikan kedalam database dengan data yang besar [5]. Berikut merupakan rumus algoritma naïve bayes yang di tunjukan pada rumus 1.

$$(X|H) = \frac{P(H)\,P(H)}{P(X)} \qquad (1)$$

Keterangan:

X : Data kelas yang belum diketahui

H : Hipotesis data X merupakan suatu kelas spesifik.

P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probability)

P(H) : Probabilitas hipótesis H (prior probability)

P(X|H) : Probabilitas X berdasar kondisi pada hipotesis

H P(X) : Probabilitas dari X

2.3.2 K-NN

Merupakan algoritma klasifikasi pada sekumpulan data yang telah terdefinisi sebelumnya . Algoritma ini termasuk kelompok supervised learning, yaitu hasil nilai klasifikasi baru yang dihasilkan berdasarkan banyaknya data-data yang memiliki kedekatan jarak dari kategori yang ada pada K-NN. Kelas baru dari suatu data akan dipilih berdasarkan kelompok kelas yang paling dekat dengan jarak vektornya. Algoritma ini menggunakan klasifikasi dengan nilai ketetanggaan sebagai nilai prediksi. Sehingga jumlah data terdekat (k) dapat ditentukan oleh user, misalnya ditentukan k=3, maka setiap data testing dilakukan perhitungan jarak dengan data trainingnya dan dipilih 3 data training yang memiliki jarak paling dekat ke data testing. Kemudian diperiksa outputnya dan ditentukan mana yang frekuensinya paling banyak. Tujuan dari algoritma K-NN adalah untuk mengklasifikasikan data baru berdasarkan data training sebelumnya. Jarak antar titik dapat dihitung dengan rumus Euclidean Distance [6]. Berikut merupakan rumus algolitma K-NN yang di tunjukkan pada rumus 2.

$$dist = \sqrt{\sum^{n}_{i=1} (pk - qk)^2} \qquad (2)$$

Keterangan:

n : Jumlah dimensi (attribute)

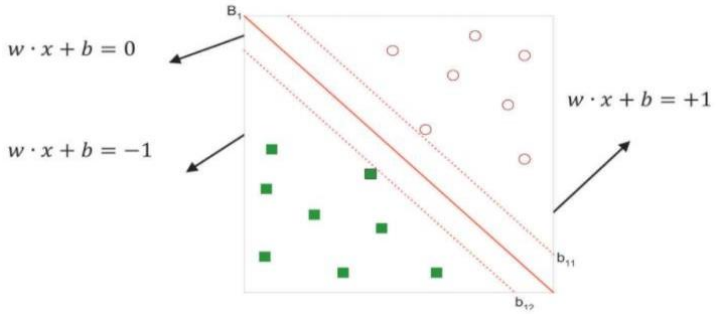
pk : Atribut ke-k / Objek data p

qk : Atribut ke-k / Objek data q

Kelebihan dari algoritma K-NN adalah sangat nonlinear, mudah dipahami, memiliki konsistensi yang kuat dan tangguh terhadap data uji yang noise. Akan tetapi algoritma ini juga memiliki kekurangan yaitu memerlukan parameter k.

2.3.3 SVM

Merupakan teknik prediksi, klasifikasi maupun regresi yang berkembang sejak tahun 1960-an. Metode ini banyak digunakan pada banyak aplikasi seperti bioinformatika, pengenalan tulisan tangan, dan sebagainya. Implementasi dari SVM ini memerlukan *training* dan *testing.* Sehingga tergolong dalam *supervised learning.* Konsep dasar SVM adalah untuk memaksimalkan *margin*, yaitu jarak yang memisahkan antarkelas data dengan mencari *hyperlane* terbaik. Berikut merupakan contoh ilustrasi margin yang di tunjukkan pada gambar 1.



Gambar 1 Hyperline Linier

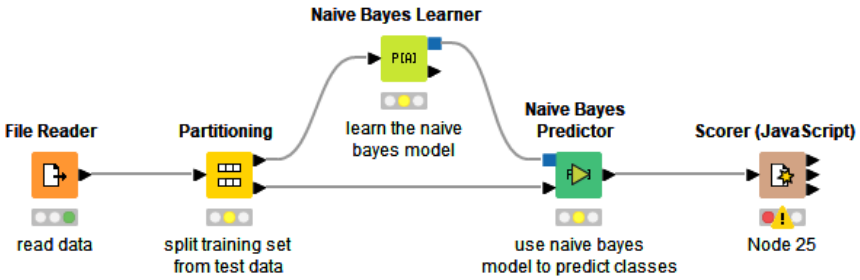
Untuk menghitung margin digunakan rumus 2 dan 3 sebagai berikut [7]:

$$Margin = \frac{2}{\|w\|} = \frac{2}{\sqrt{w_1^2 + w_2^2}} \quad (3)$$

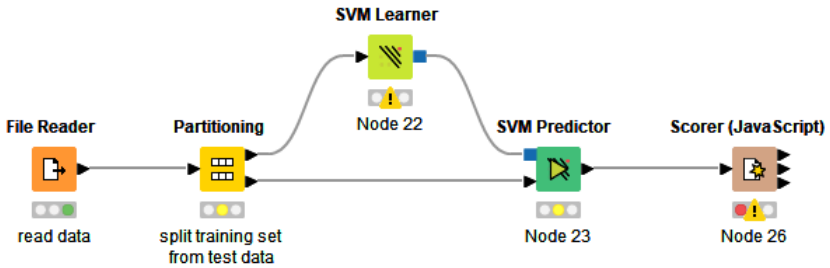
$$f(x) = \begin{cases} 1 \ if \ w \cdot x + b \geq 1 \\ -1 \ if \ w \cdot x + b \leq -1 \end{cases} \quad (4)$$

3. Hasil dan Pembahasan

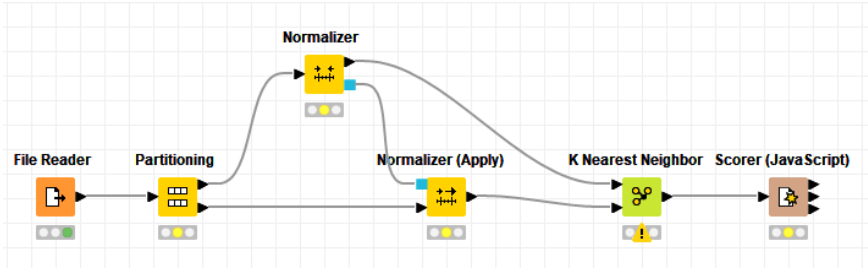
Pada penelitian kali ini menggunakan aplikasi KNIME untuk proses klasifikasi. Workflow pada masing masing algoritma dapat dilihat pada gambar 2 untuk Naïve Bayes, gambar 3 untuk SVM dan gambar 4 untuk K-NN. Seluruh data training pada algoritma menggunakan 80% data training dan 20% data testing. Data training digunakan untuk membangun model sedangkan data testing digunakan untuk menguji model yang telah dibangun sebelumnya. Hasil pengujian pada ketiga algoritma ini menggunakan *confusion matrix*.



Gambar 2. Workflow Naïve Bayes



Gambar 3. SVM



Gambar 4. K-Nearest Neighbor

Kemudian, Pengujian dengan menggunakan confusion matrix pada masing masing algoritma dapat dilihat pada Tabel 3 untuk Naïve Bayes, Tabel 4 untuk SVM dan Tabel 5 untuk K-NN.

Tabel 3. Pengujian dengan Naïve Bayes

	Meninggal (Predicted)	Hidup (Predicted)		Akurasi
Meninggal (Actual)	5	1	83,33 %	93,55%
Hidup (Actual)	1	24	96,00%	
	83,33 %	96,00%		

Tabel 4. Pengujian dengan SVM

	Meninggal (Predicted)	Hidup (Predicted)		Akurasi
Meninggal (Actual)	2	1	66,67 %	90,91%
Hidup (Actual)	1	18	96,00%	
	66,67 %	94,74%		

Tabel 5. Pengujian dengan K-NN

	Meninggal (Predicted)	Hidup (Predicted)		Akurasi
Meninggal (Actual)	2	1	66,67 %	87,50%
Hidup (Actual)	1	12	92,31%	
	66,67 %	92,31%		

Setelah itu, pada Tabel 6 terdapat matriks perbandingan akurasi pada hasil pengujian ketiga algoritma. Hasil pengujian ini menunjukkan nilai akurasi tertinggi didapatkan oleh Naïve Bayes dengan 93,55% kemudian disusul dengan SVM dengan 90,91% dan yang terakhir adalah K-NN dengan 87,50%.

Tabel 6. Matriks Kalkulasi Algoritma

Algoritma	Akurasi
Naïve Bayes	93,55%
SVM	90,91%
K-NN	87,50%

4. Kesimpulan

Ketiga algoritma ini memiliki tingkat akurasi yang cukup baik, karena nilai akurasinya diatas dari 80%, akan tetapi pada penelitian berikutnya bisa dikembangkan lagi agar nilai akurasi pada ketiga algoritma ini bisa diatas 90%. Dan juga, bisa menggunakan data sampel yang lebih banyak dan lebih bervariasi lagi.

Daftar Pustaka

[1] S. Khomsah, "Prediksi Harapan Hidup Penderita Hepatitis Kronik Menggunakan Metode-Metode Klasifikasi," *Semin. Nas. Inform. Medis*, pp. 38–45, 2018.

[2] W. D. Septiani, "KOMPARASI METODE KLASIFIKASI DATA MINING ALGORITMA C4.5 DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT HEPATITIS," *None*, 2017, doi: 10.33480/pilar.v13i1.149.

[3] S. Nurajizah, "PENERAPAN METODE SUPPORT VECTOR MACHINE BERBASIS PARTICLE SWARM OPTIMIZATION UNTUK PREDIKSI PENYAKIT JANTUNG," *J. Techno Nusa Mandiri*, 2013.

[4] S. Sulastri, K. Hadiono, and M. T. Anwar, "ANALISIS PERBANDINGAN KLASIFIKASI PREDIKSI PENYAKIT HEPATITIS DENGAN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR, NAÏVE BAYES DAN NEURAL NETWORK," *Dinamik*, 2020, doi: 10.35315/dinamik.v24i2.7867.

[5] Kusrini and E. T. Luthfi, *Algoritma Data Mining - Kusrini, Emha taufiq luthfi, Universitas Amikom - Google Buku*. Yogyakarta: ANDI, 2009.

[6] D. Wanto, Anjar, "Data Mining : Algoritma dan Implementasi - Books," *Yayasan kita menulis*, 2020. .

[7] I. Werdiningsih, B. Nuqoba, and Muhammadun, "Data Mining Menggunakan Android, Weka, dan Spss," 2020. .