

# Prediksi Penyakit Liver Dengan Menggunakan Metode *Naïve Bayes* Dan *K-Nearest Neighbour* (KNN)

Astrid Noviriandini<sup>1</sup>, Popon Handayani<sup>2</sup>, Syahriani<sup>3</sup>

Universitas BSI<sup>1</sup>, STMIK Nusa Mandiri<sup>2</sup>, STMIK Nusa Mandiri<sup>3</sup>  
Astrid.asv@bsi.ac.id<sup>1</sup>, Popon.pph@nusamandiri.ac.id<sup>2</sup>, Syahriani.yii@nusamandiri.ac.id<sup>3</sup>

**Abstract**—Keakuratan dalam memprediksi suatu penyakit di bidang kesehatan menjadikannya satu hal yang penting, Karena ini menyangkut kehidupan pasien. Seperti sampel data pada penelitian kami yaitu Liver. Kita tahu bahwa penyakit Liver dapat merusak salah satu organ tubuh manusia yaitu Hati. Ada beberapa factor yang menyebabkannya seperti penggunaan alcohol, merokok dengan berlebihan, ataupun adanya virus. Untuk itulah kami menggunakan metode *Naïve Bayes* dan *K-Nearest Neighbour* untuk mengetahui tingkat keakuratan pada penyakit liver ini. Hasil dari penelitian ini adalah *K-Nearest Neighbour* lebih tinggi nilai keakuratannya dari pada *Naïve Bayes*.

**Key Word** : Liver, *Naïve Bayes*, *K-Nearest Neighbour*

## I. PENDAHULUAN

Kesehatan merupakan salah satu hal terpenting dalam hidup manusia, hal ini menjadi dasar bahwasanya banyak ditemukannya temuan-temuan ilmiah baik berupa temua obat-obatan, alat kesehatan atau penemuan-penemuan baru di bidang kesehatan. Berbicara mengenai kesehatan tentunya berhubungan dengan penyakit, sejak dahulu banyak sekali penyakit-penyakit bermunculan entah itu datangnya dari virus, bakteri, parasite, sel kanker dan lain-lain. Salah satunya adalah organ tubuh hati. Hati penting untuk mencerna makanan, menyingkirkan tubuh kita dari zat beracun dan menyimpan energy tubuh untuk digunakan bila diperlukan. Dalam industry kesehatan dan medis keakuratan prediksi sebuah penyakit sangatlah penting dan memerlukan keputusan yang efektif dalam mengambil suatu analisa dan keakuratan prediksi suatu penyakit yang diderita pasien [1].

Saat ini dalam dunia kedokteran, penentuan penyakit peradangan hati menjadi hal yang tidak mudah dilakukan. Tetapi terdapat catatan rekam medis yang telah menyimpan gejala-gejala penyakit pasien dan diagnosis penyakit peradangan hati. Hal semacam ini tentu sangat berguna bagi para ahli kesehatan. Mereka dapat menggunakan catatan rekam medis yang telah ada sebagai bantuan untuk mengambil keputusan tentang diagnosis penyakit pasien [2].

Dalam mendiagnosis ada atau tidak penyakit liver dapat digunakan acuan dari hasil tes fungsi hati yang dilaksanakan di laboratorium. Tes tersebut antara lain yaitu *transaminase serum*, fosfatase alkali, total bilirubin, bilirubin terkonjugasi, total protein, albumin, serta rasio albumin dan globulin [4].

Penelitian yang dilakukan oleh [3] yang berjudul Analisa Komparasi Algoritma *Naïve Bayes* dan C4.5 Untuk Prediksi Penyakit Liver. Untuk mengukur kinerja kedua algoritma tersebut digunakan metode pengujian *Cross Validation*, dan *Split Percentace*, dan pengukurannya menggunakan *Confusion Matrix*. Algoritma C4.5 memiliki akurasi yang lebih tinggi dengan nilai 69,82% dibandingkan *Naïve Bayes* dengan nilai 63,36%. Dengan demikian algoritma C4.5 dapat memberikan pemecahan dalam mengidentifikasi penyakit liver.

Penelitian yang dilakukan oleh [1] yang berjudul Prediksi Penyakit Liver Dengan Menggunakan Metode *Decision Tree* dan *Neural Network*. Dari hasil pengujian evaluasi dilakukan secara *Confusion Matrix* dan *ROC Curve* membuktikan bahwa pengujian yang dilakukan algoritma C4.5 memiliki nilai akurasi yang lebih tinggi disbanding menggunakan algoritma *Neural Network*. Nilai akurasi untuk model algoritma C4.5 sebesar 75,56% dengan nilai AUC 0,898 dan nilai akurasi algoritma *Neural Network* sebesar 74,17% dengan AUC 0,671.

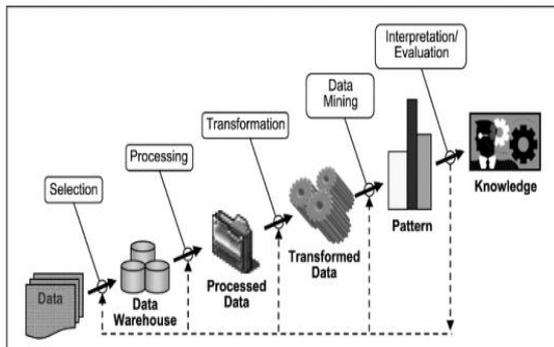
Berdasarkan penelitian terdahulu diatas, dikarekan belum ada yang membandingkan metode *Naïve Bayes* dan KNN maka dari itu penulis ingin membandingkan kedua metode tersebut dan ingin melihat tingkat akurasi dari metode *Naïve Bayes* dan KNN.

## II. KAJIAN LITERATUR

### A. Data Mining

Data Mining adalah proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Teknik dalam Data Mining yaitu bagaimana menelusuri data yang ada untuk membangun sebuah model. Model tersebut digunakan untuk mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan [5].

Data Mining sebenarnya merupakan salah satu bagian proses *Knowledge-discovery in Database* (KDD) yang bertugas untuk mengekstrak pola atau model dari data dengan menggunakan satu algoritma yang spesifik.



**Gambar 1. Proses Knowledge Discovery Database (KDD)**

Sumber : [6]

Keterangan :

#### a. Data Selection

Pemilihan data dari sekumpulan data operasional yang perlu dilakukan sebelum tahap penggalian informasi dalam *Knowledge Discovery Database* (KDD) dimulai. Data hasil seleksi yang digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari data operasional.

#### b. Processing

*Processing* dimana proses pembersihan data mencakup membuang duplikasi data, memeriksa data inkonsisten dan memperbaiki kesalahan data.

#### c. Transformation

*Coding* adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam *Knowledge Discovery Database* (KDD) merupakan proses kreatif dan sangat bergantung pada jenis atau pola informasi yang dicari dalam basis data.

#### d. Proses Mining

Proses *Mining* adalah proses mencari pola atau informasi dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses *Knowledge Discovery Database* (KDD) secara keseluruhan.

#### e. Interpretation (Evaluasi)

Tahapan ini merupakan bagian dari proses *Knowledge Discovery Database* (KDD) yang disebut *Interpretation*. Pola ini melihat apakah ada sesuatu yang baru dan menarik dan dilakukan *iterasi* jika diperlukan. Tahap ini akan diketahui apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

### B. Naïve Bayes

Kata *Naïve*, yang terkesan merendahkan berasal dari asumsi independensi pengaruh nilai suatu atribut dari probabilitas pada kelas yang diberikan terhadap nilai atribut lainnya. Penggunaan teorema *Bayes* pada algoritma *Naïve Bayes* yaitu dengan mengkombinasikan prior *probability* dan probabilitas bersyarat dalam sebuah rumus yang bisa digunakan untuk menghitung probabilitas tiap klasifikasi yang mungkin. Model independence ini menghasilkan pemecahan yang terbaik [7].

Klasifikasi *Bayes* didasarkan pada teorema *Bayes*, yaitu:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Keterangan :

y = data dengan kelas yang belum diketahui

x = hipotesis data y merupakan suatu kelas spesifik

P(x|y) = probabilitas hipotesis x berdasarkan kondisi y

P(x) = probabilitas hipotesis x

P(y|x) = probabilitas hipotesis y berdasarkan kondisi x

P(y) = probabilitas dari y

### C. K-Nearest Neighbour (KNN)

*K-Nearest Neighbour* adalah algoritma klasifikasi yang digunakan untuk menentukan kelompok berdasarkan mayoritas pada k tetangga paling dekat. Jika  $d$  adalah sekumpulan data training, maka ketika data testing disajikan, algoritma akan menghitung jarak antar bagian data dalam  $d$  dengan data testing. Kemudian  $k$  buah data dalam  $d$  yang memiliki jarak terdekat dengan data uji diambil [8].

Himpunan  $k$  merupakan *k-nearest neighbor*. Selanjutnya kategori data uji ditentukan berdasarkan mayoritas kategori dalam tetangga terdekat.

*K-Nearest Neighbour* dengan pencarian jarak *Euclidean*. Seperti dibawah ini rumus KNN dengan pencarian jarak *Euclidean*.

$$d_{Euclidean}(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Dimana :

$x$  =  $x_1, \dots, \dots, x_n$

$y$  =  $y_1, \dots, \dots, y_n$

Keterangan:

$x, y$  = nilai data tes

$n$  = jumlah atribut data

### III. METODE PENELITIAN

Dua pendekatan utama dalam penelitian ini yaitu pendekatan kualitatif dan pendekatan kuantitatif. Dalam penelitian ini metode yang digunakan yaitu metode penelitian kuantitatif.

Tujuan dari penelitian ini adalah melakukan komparasi dan evaluasi model algoritma *Naive Bayes* dan *K-Nearest Neighbour* untuk mengetahui algoritma yang memiliki keakuratan lebih tinggi dalam memprediksi penyakit liver.

Langkah-langkah untuk melakukan penelitian yaitu [9] :

#### 1. Pengumpulan Data

Pada penelitian ini data yang digunakan adalah data penyakit hepatitis yang didapat dari *Machine Learning Repository UCI* (Universitas California Invene) dengan alamat web: <http://archive.ics.uci.edu/ml/>. Data yang diteliti merupakan hasil pemeriksaan terhadap 582 orang dari wilayah Andhra Pradesh, India dengan 10

atribut yaitu *age, gender, TB (Total Bilirubin), DB (Direct Bilirubin), Alkphos (Alkaline Phosphotase), Sgpt\_AA (Almine Aminotransferase), Sgot\_AA (Aspartate Aminotransferase), TP (Total Proteins), ALB (Albumin) dan A/G (Ratio Albumin Globulin Ratio)*.

#### 2. Pengolahan Data Awal

Pengolahan data awal meliputi proses input data ke format yang dibutuhkan, pembersihan data. Dari proses pengolahan data awal diperoleh sebanyak 582 data dengan 415 data dengan kelas positif dan 167 dengan kelas negatif.

#### 3. Metode yang diusulkan

Dalam penelitian ini metode yang diusulkan adalah metode klasifikasi *data mining* algoritma *Naive Bayes* dan KNN.

#### 4. Experimen dan pengujian model

Pada bagian ini pengujian model menggunakan *Cross Validation*.

#### 5. Evaluasi hasil

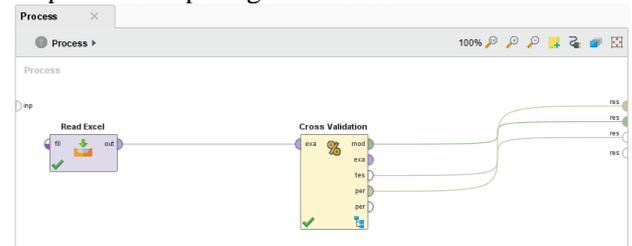
Pada bagian ini dilakukan pengujian terhadap model-model untuk mendapatkan informasi model yang akurat. Evaluasi menggunakan metode *Confusion Matrix* dan *Curva ROC*.

## IV. HASIL DAN PEMBAHASAN

### A. Eksperimen dan Pengujian Model Naive Bayes

Pembuatan model *Naive Bayes* dilakukan pada *dataset* yang terdiri dari 10 atribut yang merupakan atribut dari prediksi penyakit liver. Data kemudian di validasi agar proses pelatihan dapat berjalan dengan cepat dan mampu digunakan untuk melakukan pelatihan.

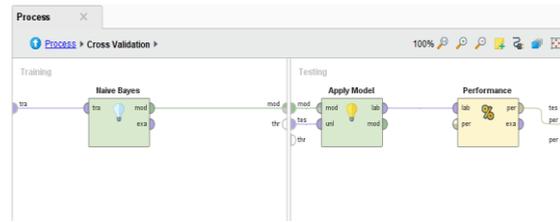
Tahap ini dibuatkan model pengolahan dengan menggunakan perangkat lunak aplikasi Rapidminer seperti gambar dibawah ini:



Gambar 2. Model Proses Naive Bayes

Keterangan :

Gambar 2 tersebut adalah gambar proses pertama yang ada di *software* RapidMiner.



**Gambar 3. Model Validasi Naive Bayes**

Keterangan :  
 Gambar 3 tersebut adalah gambar proses *Cross Validation* yang ada di *software* RapidMiner.

**B. Evaluasi Model Dengan Confusion Matrix**

Model *confusion matrix* akan membentuk *matrix* yang terdiri dari *true positive* atau tupel positif dan *true negative* atau tupel negatif, kemudian masukan data *testing* yang sudah disiapkan ke dalam *confusion matrix* sehingga didapatkan hasil pada tabel dibawah ini:

**Tabel 1. Confusion Matrix Algoritma Naive Bayes Pada Data Testing**

	True 1	True 2	Class Precision
Pred 1	169	7	96.02%
Pred 2	247	160	39.31%
Class Recall	40.62%	95.81%	

Keterangan :  
 Tabel 1 ini adalah hasil *Confusion Matrix* yang mengasilkan *True Positive* (TP) 160, *False Negative* (FN) 169, *False Positive* (FP) 247, *True Negative* (TP) 7.

Berdasarkan tabel diatas dari data *testing* terdapat rincian jumlah *True Positive* (TP) 160, *False Negative* (FN) 169, *False Positive* (FP) 247, *True Negative* (TP) 7. Dari data tersebut maka dapat dihitung nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV*, dan *NPV*. Data hasil olahan dapat diliha pada tabel dibawah ini:

**Tabel 2. Nilai Accuracy, Sensitivity, Specificity, PPV, dan NPV Pada Naive Bayes**

	Nilai
Accuracy	0,5643
Sensitivity	0,4863
Specificity	0,0275
PPV	0,3931
NPV	0,0397

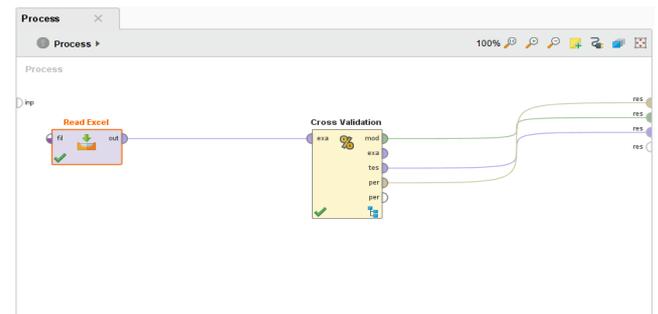
Keterangan :

Tabel 2 ini adalah hasil perhitungan nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV*, dan *NPV* Pada *Naive Bayes*.

**C. Eksperimen dan Pengujian Model KNN**

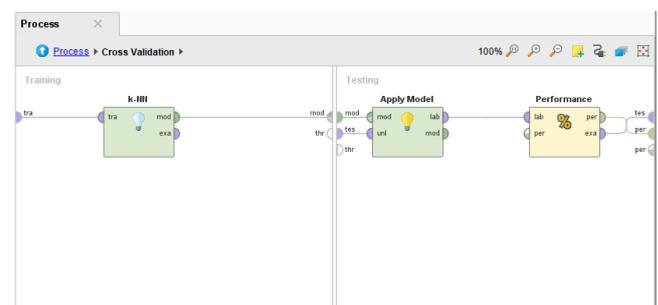
Pembuatan model KNN dilakukan pada *dataset* yang terdiri dari 10 atribut yang merupakan atribut dari prediksi penyakit liver. Data kemudian di validasi agar proses pelatihan dapat berjalan dengan cepat dan mampu digunakan untuk melakukan pelatihan.

Tahap ini dibuatkan model pengolahan dengan menggunakan perangkat lunak aplikasi Rapidminer seperti gambar dibawah ini:



**Gambar 4. Model Prose KNN**

Keterangan :  
 Gambar 4 tersebut adalah gambar proses pertama yang ada di *software* RapidMiner



**Gambar 5. Model Validasi KNN**

Keterangan :  
 Gambar 5 tersebut adalah gambar proses *Cross Validation* yang ada di *software* RapidMiner

**D. Evaluasi Model Dengan Confusion Matrix**

Model *confusion matrix* akan membentuk *matrix* yang terdiri dari *true positive* atau tupel positif dan *true negative* atau tupel negatif, kemudian masukan data *testing* yang sudah disiapkan ke dalam *confusion matrix* sehingga didapatkan hasil pada tabel dibawah ini:

**Tabel 3. Confusion Matrix Algoritma KNN  
 Pada Data Testing**

	True 1	True 2	Class Precision
Pred 1	305	98	75,62%
Pred 2	112	69	38,12%
Class Recall	73,08%	41,32%	

Keterangan :

Tabel 3 ini adalah hasil *Confusion Matrix* yang menghasilkan *True Positive* (TP) 69, *False Negative* (FN) 305, *False Positive* (FP) 112, *True Negative* (TN) 98.

Berdasarkan tabel diatas dari data *testing* terdapat rincian jumlah *True Positive* (TP) 69, *False Negative* (FN) 305, *False Positive* (FP) 112, *True Negative* (TN) 98. Dari data tersebut maka dapat dihitung nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV*, dan *NPV*. Data hasil olahan dapat dilihat pada tabel dibawah ini:

**Tabel 4. Nilai Accuracy, Sensitivity, Specificity, PPV, dan NPV Pada KNN**

	Nilai
<i>Accuracy</i>	0,6421
<i>Sensitivity</i>	0,1844
<i>Specificity</i>	0,4666
<i>PPV</i>	0,3812
<i>NPV</i>	0,2431

Keterangan :

Tabel 4 ini adalah hasil dari perhitungan Nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV*, dan *NPV* Pada KNN.

#### E. Komparasi Model Algoritma Naïve Bayes dengan KNN

Hasil pengujian algoritma *Naïve Bayes* dibandingkan dengan algoritma KNN dapat dilihat pada tabel berikut:

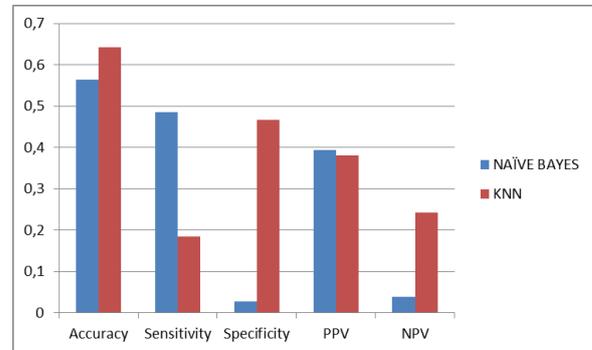
**Tabel 5. Pengujian Algoritma Naïve Bayes dan KNN**

	NAÏVE BAYES	KNN
<i>Accuracy</i>	0,5643	0,6421
<i>Sensitivity</i>	0,4863	0,1844
<i>Specificity</i>	0,0275	0,4666
<i>PPV</i>	0,3931	0,3812
<i>NPV</i>	0,0397	0,2431

Keterangan :

Tabel 5 ini adalah perbandingan nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV*, dan *NPV* terhadap metode *Naïve Bayes* dan KNN.

Berikut perbedaan Algoritma *Naïve Bayes* dibandingkan dengan algoritma KNN dapat dilihat pada gambar berikut:



**Gambar 6. Perbedaan Nilai Accuracy, Sensitivity, Specificity, PPV, dan NPV Algoritma Naïve Bayes dan KNN**

Keterangan :

Gambar 6 ini adalah perbandingan nilai *Accuracy*, *Sensitivity*, *Specificity*, *PPV*, dan *NPV* terhadap metode *Naïve Bayes* dan KNN.

## V. SIMPULAN

Dalam penelitian ini dilakukan analisa dan komparasi dua metode klasifikasi data mining yang memiliki karakteristik yang berbeda. Model yang dihasilkan, dikomparasi untuk mengetahui algoritma yang paling baik dalam penentuan identifikasi penyakit liver. Untuk mengukur kinerja kedua algoritma tersebut digunakan metode pengujian *Cross Validation* dan pengukurannya dengan menggunakan *Confusion Matrix*. Dari hasil pengujian dengan mengukur kinerja dua metode tersebut diketahui bahwa Algoritma *K-Nearest Neighbour* memiliki nilai akurasi sebesar 0,6421 atau 64,21% dibandingkan dengan Algoritma *Naïve Bayes* yang memiliki nilai akurasi sebesar 0,5643 atau 56,43%. Adapun model yang telah terbentuk selanjutnya dapat dikembangkan dan dapat diimplementasikan ke dalam sebuah aplikasi sehingga dapat membantu dan memudahkan bagi para pemegang kepentingan dalam mengambil sebuah keputusan untuk memprediksi penyakit liver.

#### UCAPAN TERIMA KASIH

Kami mengucapkan terima kasih kepada semua penulis baik itu penulis 1, penulis 2 dan penulis 3, yang sudah memberikan kontribusinya pada penelitian ini hingga selesai.

#### DAFTAR PUSTAKA

- [1] P. Handayani, E. Nurlelah, M. Raharjo, and P. M. Ramdani, "PREDIKSI PENYAKIT LIVER DENGAN MENGGUNAKAN METODE," vol. 4, no. 1, pp. 55–59, 2019.
- [2] I. Journal and I. T. Vol, "Penentuan Penyakit Peradangan Hati Dengan Menggunakan Neural Network Backpropagation," vol. 1, no. 1, pp. 27–33, 2016.
- [3] E. Rahmawati, "Vol. XII No. 2, September 2015 Jurnal Techno Nusa Mandiri," vol. XII, no. 2, pp. 125–136, 2015.
- [4] J. Gaussian, "1 , 2 , 3 1," vol. 3, pp. 731–739, 2014.
- [5] T. I. Andini, W. Witanti, and F. Renaldi, "Prediksi Potensi Pemasaran Produk Baru dengan Metode Naïve Bayes Classifier dan Regresi Linear," *Semin. Nas. Apl. Teknol. Inf.*, pp. 27–32, 2016.
- [6] T. Wahyudi, R. E. Indrajit, and Muh. Fauzi, "Pemanfaatan Status Kredit Nasabah Untuk Mengevaluasi Pembiayaan Kpr Pada Bank Muamalat," no. November, pp. 1–2, 2017.
- [7] W. D. Septiani, P. Studi, and M. Informatika, "DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT HEPATITIS," vol. 13, no. 1, pp. 76–84, 2017.
- [8] A. Behori and S. Informasi, "KOMPARASI MODEL NEURAL NETWORK DAN K-NN UNTUK MENGLASIFIKASI KONDISI PENDERITA PENYAKIT HEPATITIS COMPARISON OF NEURAL NETWORK AND K-NN MODEL FOR CLASSIFYING," vol. 3, no. 2, pp. 82–88, 2017.
- [9] D. A. N. T. Komputer, A. Kinerja, C. Algoritma, D. A. N. Naïve, and A. Noviriandini, "( STUDI KASUS : SMK AL-IJTIHAD )," vol. 5, no. 1, pp. 23–28, 2019.