Implementasi Algoritma K-Nearest Neighbor (K-NN) dalam Deteksi Dini Penyakit Hepatitis C

Ni Made Rika Padeswari Kusuma[1], L. G. Astuti[2]

Program Studi Informatika, Fakultas MIPA, Universitas Udayana
Jl. Kampus Bukit Jimbaran, Badung, Bali
[1]rikakusumaa23@gmail.com
[2]lg.astuti@unud.ac.id

Abstract

According to the World Health Organization (WHO), Hepatitis is an inflammatory condition that can evolve to Cirrhosis or liver cancer. Hepatitis is a disease that is caused by several types of viruses that attack and cause inflammation and damage the cells of the human liver. Hepatitis C Virus (HCV) is one of the viruses that caused hepatitis and is considered the biggest impact among the other viruses that caused hepatitis. This study uses a classification method with the K-Nearest Neighbor (KNN) algorithm to detect the onset of hepatitis C in patients based on data from the patient's laboratory checks. The classification method with K-Nearest Neighbor (KNN) algorithm is carried out by comparing the neighbors between test data and train data based on the patient's medical history. The tuning parameter is used to determine the number of neighbors or the value of K in K-Nearest Neighbor (KNN) which obtains 92% of accuracy, 92% of precision, and 99% of recall with a 80:20 ratio of training data and test data.

Keywords: *Classification, Machine Learning, Hepatitis C, K-Nearest Neighbor Algorithm, Confusion Matrix*

1. Introduction

Hepatitis merupakan penyakit yang disebabkan oleh beberapa jenis virus yang menyerang dan menyebabkan peradangan serta merusak sel-sel organ hati manusia. Hepatitis dikategorikan ke dalam beberapa golongan, diantaranya Hepatitis A, B, C, D, E [4]. Virus Hepatitis C (HCV) merupakan salah satu virus penyebab hepatitis dan dianggap menimbulkan dampak paling besar diantara virus lainnya yang menjadi penyebab penyakit hepatitis [2]. Berdasarkan data dari *World Health Organization* (WHO) pada tahun 2021, ditunjukkan bahwa sebanyak 1% atau 71 juta orang di seluruh dunia terinfeksi virus hepatitis C (HCV) dimana 399 ribu diantaranya meninggal dunia dikarenakan sirosis hati.

Penyakit hepatitis C apabila tidak ditangani dengan cepat dapat menetap dan berkembang menjadi hepatitis C kronik. Beberapa komplikasi yang dapat terjadi akibat infeksi hepatitis C yaitu sirosis hati dan karsinoma sel hati [1].
Melihat data penderita serta dampak dari penyakit hepatitis tersebut, perlu dilakukannya penanganan untuk menghambat perkembangan penyakit hepatitis C. Salah satu upaya yang dapat dilakukan adalah melakukan *screening* untuk mendeteksi penyakit hepatitis C.

Menurut Prasetyo, dalam dunia kesehatan saat ini rekam medis menyimpan gejala-gejala serta diagnosis penyakit pasien. Dimana hal tersebut dapat berguna bagi para ahli kesehatan untuk dapat dijadikan sebagai bantuan dalam pengambilan keputusan terhadap diagnosis penyakit pasien [6]. Selain itu dalam bidang teknologi, data rekam medis juga dapat dimanfaatkan sebagai deteksi awal sebuah penyakit. Salah satu cara yang dapat dilakukan untuk deteksi awal sebuah penyakit, khususnya penyakit hepatitis C adalah dengan memanfaatkan teknologi *machine learning* yaitu metode klasifikasi.

Dengan teknologi *machine learning* yang sedang berkembang saat ini serta data rekam medis

pasien hendaknya dapat menjadi sebuah solusi bantuan dalam pengambilan keputusan terhadap diagnosis penyakit pasien. Dimana pemanfaatan teknologi *machine learning* dengan menggunakan metode klasifikasi ini dapat dimanfaatkan sebagai pendeteksian awal apakah pasien memiliki kecenderungan penyakit hepatitis C atau tidak. Sehingga pada penelitian ini akan dilakukan proses klasifikasi dengan menggunakan algoritma K-Nearest Neighbor (KNN). Serta dilakukan *parameter tuning* untuk menentukan jumlah tetangga atau nilai K yang menghasilkan tingkat akurasi terbaik. Dataset yang digunakkan pada penelitian ini yaitu Hepatitis C Prediction Dataset yang diambil dari website UCI Machine Learning Repository dengan 14 fitur dan 615 baris data.

2. Reseach Methods

Penelitian yang dilakukan menggunakan salah satu teknik machine learning khususnya supervised learning yaitu metode klasifikasi dengan algoritma K-Nearest Neighbor. Pada penelitian ini, terdapat beberapa tahapan yang dapat dilihat pada diagram alur penelitian berikut.

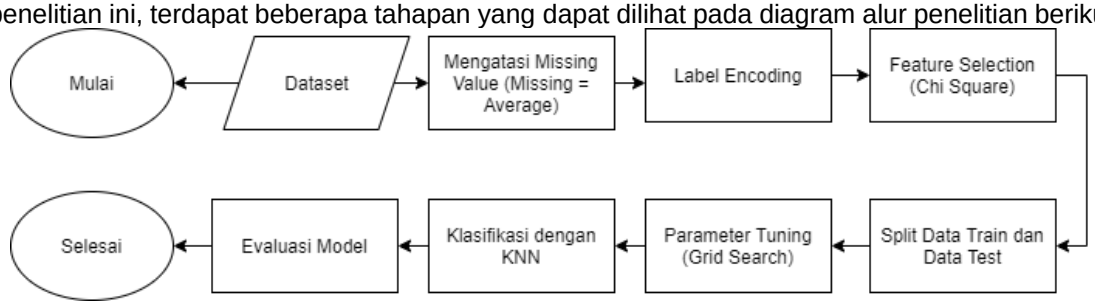

Figure 1. Diagram Alur Penelitian

2.1. Pengumpulan Data

Data pada penelitian ini diambil dari *website* UCI Machine Learning Repository yaitu Hepatitis C Prediction Dataset. Data ini terdiri dari 13 fitur dan 615 baris data yang berisi terkait rekam medis dari pasien yang terdeteksi memiliki penyakit hepatitis C maupun pasien yang terdeteksi sehat. Untuk tiap atribut dari data dapat dilihat pada tabel sebagai berikut.

Table 1. Atribut Dataset

Fitur	Penjelasan	Keterangan
Category	Kategori atau tipe pasien (Healthy Patient, Suspected Patient)	Label
Age	Umur pasien	Atribut
Sex	Jenis kelamin pasien (Female, Male)	Atribut
ALB	Kadar albumin pada darah pasien	Atribut
ALP	Kadar alkaline phosphatase pada darah pasien	Atribut
ALT	Kadar alanine transaminase pada darah pasien	Atribut
AST	Kadar aspartate aminotransferase pada darah pasien	Atribut
BIL	Kadar bilirubin pada darah pasien	Atribut
CHE	Kadar cholinesterase pada darah pasien	Atribut
CHOL	Kadar cholesterol pada darah pasien	Atribut
CREA	Kadar creatine pada darah pasien	Atribut
GGT	Kadar gamma-glutamyl pada darah pasien	Atribut
PROT	Kadar protein pada darah pasien	Atribut

2.2. Preprocessing Data

Sebelum masuk ke tahap pemodelan, perlu dilakukan beberapa tahapan yang salah satunya adalah *preprocessing* data. *Preprocessing* data merupakan tahapan yang dilakukan untuk mempersiapkan data sebelum masuk ke tahapan modeling. Adapun hal yang dilakukan pada saat *preprocessing data* diantaranya sebagai berikut.

a. Mengatasi Missing Value
   Pada dataset yang digunakan, terdapat missing value atau sejumlah data yang hilang pada beberapa fitur yang ada. Data yang hilang tersebut harus diatasi untuk meminimalisir error pada model yang dibangun. Untuk mengatasi hal tersebut, data yang hilang akan diisi dengan nilai rata-rata dari fitur itu sendiri.
b. Label Encoding
   Fitur kategorikal seperti label dan jenis kelamin pada dataset akan diubah ke dalam bentuk numerik dengan menggunakan metode label encoding.
c. Seleksi Fitur
   Untuk meningkatkan efisiensi dan efektifitas model yang dibangun, perlu dilakukan tahap seleksi fitur atau memilih fitur-fitur yang relevan terhadap permasalahan yang dihadapi. Pada tahapan ini digunakan metode chi square untuk memilih fitur terbaik yang akan masuk ke tahap pemodelan.

2.3. Data Preparation

Setelah dilakukan preprocessing data, data akan dibagi menjadi data train dan data set. Dataset dibagi dengan rasio 80:20. Dimana 80% dari dataset akan menjadi data train dan 20% menjadi data test. Setelah dilakukannya tahap splitting data atau membagi data menjadi data train dan data test, akan dilakukan normalisasi pada data dengan menggunakan *min-max scaler*.

2.4. Klasifikasi dengan Algoritma K-Nearest Neighbor (KNN)

Setelah dilakukannya data preprocessing dan preparation, data siap masuk ke tahap modeling dengan menggunakan algoritma K-Nearest Neighbor. Algoritma K-Nearest Neighbor bekerja melakukan klasifikasi dengan melihat jarak terdekat dari suatu objek dengan data pembelajaran. Nilai K pada algoritma K-Nearest Neighbor merupakan jumlah tetangga terdekat dari objek. Dalam tahap pemodelan, nilai K ditentukan dengan melakukan parameter tuning menggunakan grid search. Dimana data akan dilatih dengan menggunakan nilai K yang berbeda-beda untuk menemukan nilai K yang menghasilkan tingkat akurasi tertinggi.

$$d(x,y) = \sum_{i=1}^{m} |x_i - y_i| \quad (1)$$

Keterangan:

d(x,y) = Jarak

x_i = Data training

y_i = Data testing

i = Variabel data

2.5. Evaluasi

Tahap evaluasi dilakukan untuk menilai performa dari model klasifikasi yang telah dibangun. Evaluasi model dilakukan dengan menggunakan confusion matrix. Confusion matrix adalah suatu metode yang digunakan untuk melakukan perhitungan akurasi, presisi, dan recall pada data mining. Confusion matrix digambarkan dengan tabel yang menyatakan jumlah data uji yang benar diklasifikasikan serta jumlah data uji yang salah diklasifikasikan [5].

3. Result and Discussion

Dalam mengimplementasikan metode klasifikasi K-Nearest Neighbor dalam deteksi dini penyakit Hepatitis C, terdapat beberapa tahapan yang dilakukan untuk mencapai akurasi yang optimal pada model yang dibangun.

3.1. Mengatasi Missing Value

Terdapat beberapa fitur pada dataset yang memiliki *missing value*. Beberapa fitur yang terdapat *missing value* diantaranya sebagai berikut.

Table 2. Missing Value pada Dataset

Fitur	Jumlah Missing Value
Category	-
Age	-
Sex	-
ALB	1
ALP	18
ALT	1
AST	-
BIL	-
CHE	-
CHOL	10
CREA	-
GGT	-
PROT	1

Sehingga untuk mengatasinya, data yang hilang pada fitur tersebut diisi dengan nilai rata-rata dari fitur itu sendiri. Nilai rata-rata dari tiap fitur yang memiliki missing value adalah sebagai berikut.

Table 3. Nilai Rata-Rata pada Fitur

Fitur	Nilai Rata-Rata
ALB	41.6
ALP	68.28
ALT	28.45
CHOL	5.36
PROT	72.04

3.2. Label Encoding

Terdapat 2 fitur kategorikal pada dataset yaitu Category dan Sex. Kedua fitur ini akan diubah ke dalam bentuk numerik sehingga dapat digunakan pada saat tahap modeling.

Table 4. Label Encoding pada Fitur Kategorikal

Fitur	Nilai Awal	Hasil Label Encoding
Category	Healthy Patient	0 = Healthy Patient
	Suspected Patient	1 = Suspected Patient
Sex	F	0 = Female (F)
	M	1 = Male (M)

3.3. Seleksi Fitur

Seleksi fitur dilakukan untuk menentukan fitur-fitur yang paling relevan terhadap label guna meningkatkan efisiensi dan efektifitas dari model. Pada penelitian ini, digunakan *chi-square* untuk menentukan fitur terbaik pada dataset. Berikut hasil dari perhitungan *chi-square* fitur-fitur pada dataset terhadap label Category.

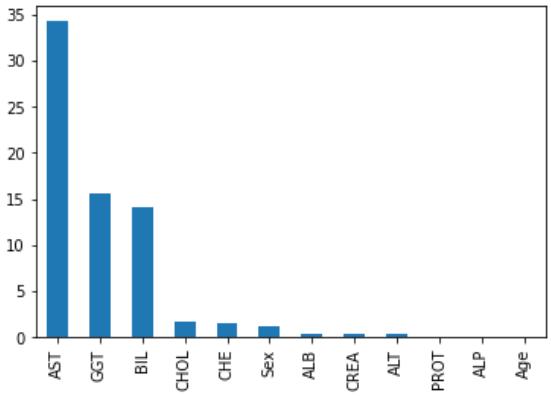

Figure 2. Hasil Perhitungan Chi-Square

Fitur yang dipilih adalah fitur yang memiliki nilai *chi-square* tertinggi. Hal ini dikarenakan pada seleksi fitur, fitur yang dipilih adalah fitur yang paling bergantung dengan label. Dalam perhitungan *chi-square*, apabila kedua fitur tidak bergantung satu sama lain maka frekuensi harapan sangat dekat dengan frekuensi kenyataan sehingga nilai *chi-square* akan rendah. Maka semakin tinggi nilai *chi-square* dapat diartikan bahwa fitur tersebut lebih bergantung pada label dan dapat dipilih untuk menjadi fitur yang digunakan pada tahap pemodelan.
Sehingga dipilih 5 fitur yang memiliki nilai *chi-square* tertinggi atau dapat dikatakan paling relevan. Fitur yang dipilih untuk masuk ke tahap pemodelan diantaranya sebagai berikut.

Table 5. Hasil Seleksi Fitur

Fitur	Penjelasan
AST	Kadar aspartate aminotransferase pada darah pasien
BIL	Kadar bilirubin pada darah pasien
CHE	Kadar cholinesterase pada darah pasien
CHOL	Kadar cholesterol pada darah pasien
GGT	Kadar gamma-glutamyl pada darah pasien

3.4. Data Preparation

Setelah dilakukan *preprocessing* pada data serta seleksi fitur, maka data dipersiapkan sebelum masuk ke tahap modeling. Hal yang dilakukan yaitu membagi data menjadi data latih dan data uji. Dimana data akan dibagi menjadi 80% data latih dan 20% data uji. Sehingga jumlah data untuk pelatihan model yaitu sebanyak 492 data dan 123 data untuk pengujian. Selain itu, dilakukan normalisasi pada data dengan menggunakan *min-max scaler* agar data memiliki rentang nilai yang sama.

3.5. Klasifikasi dengan K-Nearest Neighbor (KNN)

Klasifikasi dengan menggunakan algoritma *K-Nearest Neighbor* (KNN) dilakukan dengan menghitung jarak antara data uji dengan data latih. Prediksi kelas dari data uji adalah kelas aktual terbanyak dari jumlah K data latih yang jaraknya terdekat dengan data uji tersebut [3]. Untuk menentukan parameter yang digunakan dalam klasifikasi dengan algoritma *K-Nearest Neighbor* dilakukan proses *parameter tuning* dengan menggunakan GridSearchCV. Dimana parameter yang akan dicari adalah jumlah K atau jumlah ketetanggaan, pembobotan, dan metrik pengukuran jarak pada K-Nearest Neighbor. Jumlah nilai K yang akan diuji adalah 8, dimana K = {5, 7, 9, 11, 13, 15, 17, 19}. Pembobotan yang akan diuji yaitu *uniform* dan *distance*, dimana akan dilihat apakah pembobotan mempengaruhi model atau tidak. Apabila *uniform* maka tidak ada bobot yang ditambahkan, sedangkan *distance* akan menambahkan bobot lebih banyak

kepada objek yang lebih dekat dibandingkan dengan objek yang lebih jauh. Metrik jarak yang akan diuji yaitu perhitungan jarak dengan metode *Minkowski*, *Euclidean,* dan *Manhattan*. Selanjutnya masing-masing pengujian akan dilakukan sebanyak 3 kali dan kemudian dicari nilai rata-ratanya. Sehingga total percobaan yang dilakukan adalah sebanyak 144 kali dengan parameter yang berbeda-beda.

Table 6. Hasil Pengujian Parameter

Metrik	Nilai K	Pembobotan	Hasil Akurasi Rata-Rata
Minkowski	5	Uniform	0.9491
Minkowski	5	Distance	0.9491
Minkowski	7	Uniform	0.9491
Minkowski	7	Distance	0.9491
Minkowski	9	Uniform	0.9430
Minkowski	9	Distance	0.9471
Minkowski	11	Uniform	0.9410
Minkowski	11	Distance	0.9451
Minkowski	13	Uniform	0.9329
Minkowski	13	Distance	0.9451
Minkowski	15	Uniform	0.9308
Minkowski	15	Distance	0.9430
Minkowski	17	Uniform	0.9288
Minkowski	17	Distance	0.9410
Minkowski	19	Uniform	0.9268
Minkowski	19	Distance	0.9369
Euclidean	5	Uniform	0.9491
Euclidean	5	Distance	0.9491
Euclidean	7	Uniform	0.9491
Euclidean	7	Distance	0.9491
Euclidean	9	Uniform	0.9430
Euclidean	9	Distance	0.9471
Euclidean	11	Uniform	0.9410
Euclidean	11	Distance	0.9451
Euclidean	13	Uniform	0.9329
Euclidean	13	Distance	0.9451
Euclidean	15	Uniform	0.9308
Euclidean	15	Distance	0.9430
Euclidean	17	Uniform	0.9288
Euclidean	17	Distance	0.9410
Euclidean	19	Uniform	0.9268
Euclidean	19	Distance	0.9410
Manhattan	5	Uniform	0.9471
Manhattan	5	Distance	0.9491
Manhattan	7	Uniform	0.9491
Manhattan	7	Distance	0.9491
Manhattan	9	Uniform	0.9491
Manhattan	9	Distance	0.9491
Manhattan	11	Uniform	0.9451
Manhattan	11	Distance	0.9512
Manhattan	13	Uniform	0.9451
Manhattan	13	Distance	0.9512
Manhattan	15	Uniform	0.9369
Manhattan	15	Distance	0.9430
Manhattan	17	Uniform	0.9369
Manhattan	17	Distance	0.9410
Manhattan	19	Uniform	0.9288
Manhattan	19	Distance	0.9390

Berdasarkan tabel di atas, performa klasifikasi dengan algoritma *K-Nearest Neighbor* terbaik yaitu dengan parameter metrik perhitungan jarak menggunakan *Manhattan*, dengan pembobotan *distance*, serta nilai K yaitu 11 dengan rata-rata akurasi sebesar 95%.
Setelah mendapatkan parameter-parameter terbaik, maka model klasifikasi dapat dibangun dan menghasilkan nilai *true positive* sebesar 98 dan *true negative* sebesar 15.

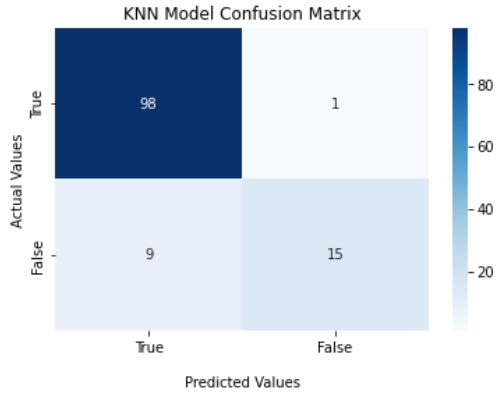

Figure 3. Confusion Matrix Model Klasifikasi

Berdasarkan *confusion matrix* diatas, diperoleh hasil akurasi sebesar 92%, presisi sebesar 92%, dan *recall* sebesar 99%.

4. Conclusion

Berdasarkan hasil penelitian yang dilakukan pada deteksi dini penyakit Hepatitis C dengan metode klasifikasi *K-Nearest Neighbor* dihasilkan akurasi testing sebesar 92%, presisi sebesar 92%, dan recall sebesar 99% dengan parameter metrik perhitungan jarak menggunakan metode Manhattan, pembobotan menggunakan *distance*, serta nilai K yaitu 11. Total percobaan yang dilakukan sebanyak 144 kali sehingga menghasilkan parameter terbaik dengan akurasi *training* sebesar 95%. Sehingga dapat disimpulkan bahwa kombinasi parameter terbaik untuk klasifikasi data penyakit Hepatitis C dengan menggunakan algoritma *K-Nearest Neigbor* (KNN) yaitu metrik perhitungan jarak menggunakan metode Manhattan, pembobotan menggunakan *distance*, serta nilai K = 11.

References

[1] A. Saraswati, TA Larasati, and Suharmanto, “FAKTOR RISIKO TERJADINYA PENYAKIT HEPATITIS C,” Bandar Lampung, May 2022. [Online]. Available: http://jurnal.globalhealthsciencegroup.com/index.php/JPPP

[2] Alhawaris, “Hepatitis C: Epidemiologi, Etiologi, dan Patogenitas,” *Jurnal Sains dan Kesehatan*, vol. 2, no. 2, pp. 139–150, Dec. 2019, doi: 10.25026/jsk.v2i2.132.

[3] D. Kartini, A. Farmadi, Muliadi, D. Turianto Nugrahadi, and Pirjatullah, “Perbandingan Nilai K pada Klasifikasi Pneumonia Anak Balita Menggunakan K-Nearest Neighbor,” 2022.

[4] Darsin and M. F. Sesunan, “PERANCANGAN SISTEM PENDIAGNOSA PENYAKIT HEPATITIS DENGAN METODE CASE BASED REASONING (CBR),” *Jurnal Sistem Informasi dan Sains Teknologi*, vol. 1, no. 2, pp. 1–7, 2019.

[5] M. F. Rahman, M. Ilham Darmawidjadja, and D. Alamsah, “KLASIFIKASI UNTUK DIAGNOSA DIABETES MENGGUNAKAN METODE BAYESIAN REGULARIZATION NEURAL NETWORK (RBNN),” 2017.

[6] W. Dwi Septiani, “ALGORITMA NAÏVE BAYES UNTUK PREDIKSI PENYAKIT HEPATITIS,” Aug. 2022.

Halaman ini sengaja dibiarkan kosong