PERBANDINGAN KINERJA ALGORITMA *GAUSSIAN NAIVE BAYES* DAN *K-NEAREST NEIGHBOR* (KNN) UNTUK MENGKLASIFIKASI PENYAKIT HEPATITIS C VIRUS (HCV)

Kukuh Yudha Raharja[1], Hardian Oktavianto[2], Reni Umilasari[3]
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember
Email : yudharaharja0@gmail.com[1], hardian@unmuhjember.ac.id[2],
Reni.umilasari@gmail.com[3].

ABSTRAK

Hepatitis adalah penyakit peradangan pada hati, penyakit Hepatitis merupakanpenyakit yang menular sehingga menjadi beban kesehatan bagi masyarakat karena penularannya yang begitu mudah. Jenis penyakit Hepatitis yang paling berbahaya adalah Hepatitis C yang disebabkan oleh Virus RNA. Pada penelitian ini dilakukan klasifikasi terhadap penyakit Hepatitis C Virus pada pasien di Jerman dikarenakan kasus hepatits di jerman termasuk tinggi dan di Jerman belum menerapkan stategi nasional untuk mengurangi beban Hepatitis C, Metode klasifikasi pada penelitian ini adalah membandingkan antara algoritma *Gaussian Naive Bayes* dan *K-Nearest Neighbor*. Hasil yang didapatkan dari penelitian ini adalah pada algoritma *Gaussian Naive Bayes* nilai akurasi sebesar 90,98%, presisi sebesar 69,91%, dan recall sebesar 61,57% sedangkan pada algoritma *K-Nearest Neighbor* (KNN) didapatkan nilai akurasi sebesar 91,80%, presisi sebesar 68,96%, dan recall sebesar 51,85%. Untuk akurasi, algoritma *K-Nearest Neighbor* (KNN) mendapatkan hasil lebih baik dibandingkan *Gaussian Naive Bayes*. Sedangkan untuk presisi, algoritma *Gaussian Naive Bayes* mendapatkan hasil lebih baik dibandingkan *K-Nearest Neighbor* (KNN). Dan untuk recall, algoritma *Gaussian Naive Bayes* mendapatkan hasil lebih baik dibandingkan *K-Nearest Neighbor* (KNN).

Kata Kunci: Klasifikasi penyakit, Hepatitis C Virus (HCV), *Gaussian Naive Bayes*, *K-Nearest Neighbor.*

.

ABSTRACT

Hepatitis is an inflammatory disease of the liver, it is an infectious disease that becomes health burden for the community because of its easy to get infection. The most dangerous type of hepatitis is Hepatitis C caused by RNA virus. In this study, classification of Hepatitis C Virus disease in patient at Germany because cases of hepatitis in Germany were high and Germany had not implemented a national strategy to reduce the burden of Hepatitis C Virus, The classification method in this study is comparing between algorithm of the Gaussian Naive Bayes and K-Nearest Neighbor (KNN). The results obtained from this study are in Gaussian Naive Bayes algorithm accuracy value of 90.98%, precision by 69.91%, and recall by 61.57% while in K-Nearest Neighbor (KNN) algorithm obtained accuracy value of 91.80%, precision of 68.96%, and recall of 51.85%. For accuracy, K-Nearest Neighbor (KNN) algorithm gets better results than Gaussian Naive Bayes. As for precision, Gaussian Naive Bayes algorithm gets better results than K-Nearest Neighbor (KNN). And for recalls, Gaussian Naive Bayes algorithms get better results than K-Nearest Neighbor (KNN).

Keywords*: Disease classification, Hepatitis C Virus (HCV), Gaussian Naive Bayes, K-Nearest Neighbor.*

1. PENDAHULUAN

Hepatitis adalah penyakit peradangan pada hati, terjadi karena adanya toxin yang ada pada hati. Penyakit ini dapat menyerang semua orang tak terkecuali orang memiliki kekebalan tubuh yang tinggi. Penyakit *Hepatitis* atau *Liver* merupakan penyakit yang menular yang menjadi masalah kesehatan yang besar bagi masyarakat karena penularannya yang mudah. Peyakit *Hepatitis* atau *Liver* yang disebabkan oleh virus yang langsung menyerang pada organ hati. Penyakit ini dibagi menjadi lima jenis, yaitu, Hepatitis A, B, C, D, dan E (Kudesia & Wreghitt, 2009). Menurut WHO pada tahun 2017, jenis penyakit Hepatitis yang paling berbahaya adalah Hepatitis C (Kudesia & Wreghitt, 2009).

Pada tahun 2014 diperkirakan terdapat 28 juta penduduk Indonesia terserang penyakit hepatitis dan 14 juta diantaranya berpotensi untuk menjadi kronis dan dari yang kronis tersebut berpotensi menjadi kanker hati.Maka diperlukan penanganan yang khusus untuk penyakit Hepatitis C, dengan melakukan diagnosa dini terhadap penyakit Hepatitis C. Cara untuk melakukan diagnosa dini terhadap penyakit Hepatitis C adalah memanfaatkan teknologi informasi. Salah satunya menggunakan metode *data mining. Data mining* yaitu pemrosesan suatu data yang menggunakan cara statistik, matematik, kecerdaan tiruan dan *machine learning* untuk mengidentifikasi informasi pengetahuan potensial dan berguna yang tersimpan dalam database besar (Riadi, 2017).

2. PENELITIAN TERKAIT

A. Hepatitis C Virus

Hepatitis merupakan peradangan yang terjadi pada hati, yang disebabkan oleh suatu proses infeksi, seperti virus, bakteri, jamur, dan parasit,atau bukan infeksi, seperti alkohol, obat, autoimun, dan metabolik. Prosesperadangan yang terjadi pada hatidapat bersifat akut maupun kronik. Padakeadaan kronik, terjadi peradangan dalam kurun waktu minimal enam bulandengan tingkat kerusakan yang bervariasi(Blachier et al., 2013).

B. Data Mining

Data mining merupakan proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Turban et al., 2003). Secara sederhana data mining merupakan penemuan informasi baru dengan mencari pola dan aturan tertentu dari sejumlah data yang sangan besar (Davies et al., 2004).

C. Klasifikasi

Klasifikasi merupakan proses menemukan model dan fungsi yang menjelaskan atau membedakan kelas-kelas atau konsep, dengan tujuan agar model yang diperoleh dapat digunakan untuk mengetahui kelas atau objek yang memiliki label kelas yang tidak diketahui. Model yang diturunkan didasarkan pada analisis dari training data. Fungsi dari klasifikasi adalah untuk mengklasifikasi suatu target kelas ke dalam kategori yang dipilih. Ada banyak metode untuk membangun model klasifikasi seperti *Naive Bayesian classification, support vector machine* (SVM) dan *K-Nearest Neighbor Classification* (Crivat, 2011)

D. *Naive Bayes*

Algoritma *Naive Bayes* adalah algoritma yang terdapat pada teknik data mining klasifikasi. *Naive Bayes* merupakan teknik pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuan inggris yaitu Thomas Bayes, *Naive Bayes* memprediksi peluang yang akan dating/dimasa depan berdasarkan pengalaman dimasa sebelumnya, sehingga dikenal dengan Teorema Bayes. Teorema dikombinasikan dengan Naive yaitu mengasumsikan kondisi antar atribut saling bebas. Persamaan dari Teorema Bayes adalah :

E. $P(H|X) = \frac{P(X|H).P(H)}{P(x)}$ (2.1)

F. Keterangan :

G. X : Data kelas yang belum diketahui

H. H : Data X merupakan suatu kelas spesifik

I. P(H|X): Probabilitas H berdasarkan kondisi X

J. P(X) : Probabilitas H

K. P(X|H): Probabilitas X berdasarkan kondisi pada H

L. P(X) : Probabilitas X

E. *Gaussian Naive Bayes*

Ketika dihadapkan dengan data kontinu maka akan dilukukan pendistribusian *Gausssian Naive Bayes* dibawah bentuk distribusi *Gaussian Naive Bayes*:

1. Baca data latih
2. Hitung nilai probabilitas pada data kontinu dengan cara :
   a. Cari nilai mean dan standar deviasi dari masing-masing parameter yang merupakan data kontinu.
   b. Cari nilai probabilistik dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut. Menggunakan cara dibawah :

   $f(w) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(w-\mu)^2}{\sigma^2}}$ (2.2)
3. Mendapatkan nilai mean, standar deviasi dan probabilitas

F. Mean Dan Standar Deviasi

Mean yaitu nilai rata-rata yang didapatkan dari hasil penjumlahan seluruh nilai dari masing-masing data, lalu dibagi dengan banyak nya data yang ada. Adapun rumus mean dapat dilihat pada persamaan 2.3.

$$x = \frac{1}{2}\sum_{i=1}^{n} x1 \quad (2.3)$$

Standar deviasi adalah simpangan baku, seperti varian dan standar deviasi yang juga merupakan suatu ukuran dispersi. Standar deviasi paling banyak dipakai karena standar deviasi mempunyai satuan ukuran data asalnya. Adapun rumus standara deviasi dapat dilihat pada persamaan 2.4.

$$\sigma = \sqrt{\frac{\Sigma_{i=1}^{n} (x_i - \mu)^2}{n-1}} \quad (2.4)$$

G. Algoritma K-Nearest Neighbor (KNN)

Algoritma *K-Nearest Neighbor* (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut (Widiarsana et al., 2011). Algoritma *K-Nearest Neighbor* (KNN) adalah metode yang digunakan untuk mengelompokkan objek berdasarkan contoh pelatihan terdekat. Algoritma *K-Nearest Neighbor* dilakukan dengan mencari kelompok objek dalam data latih yang paling dekat dengan objek pada data testing (Saxena et al., 2014)

H. Phyton

Python adalah bahasa pemrograman *interpreatif* multiguna dengan filosofi perancangan yang berfokus pada tingkat keterbacaan kode. *Python* diklaim sebagai bahasa yang menggabungkan kapabilitas, kemampuan, dengan sintakis kode yang jelas dan dilengkapi dengan fungsionalitas pustaka standar serta komprehensif.

I. Tahapan Penelitian

Gambar 1. Diagram Alur Penelitian

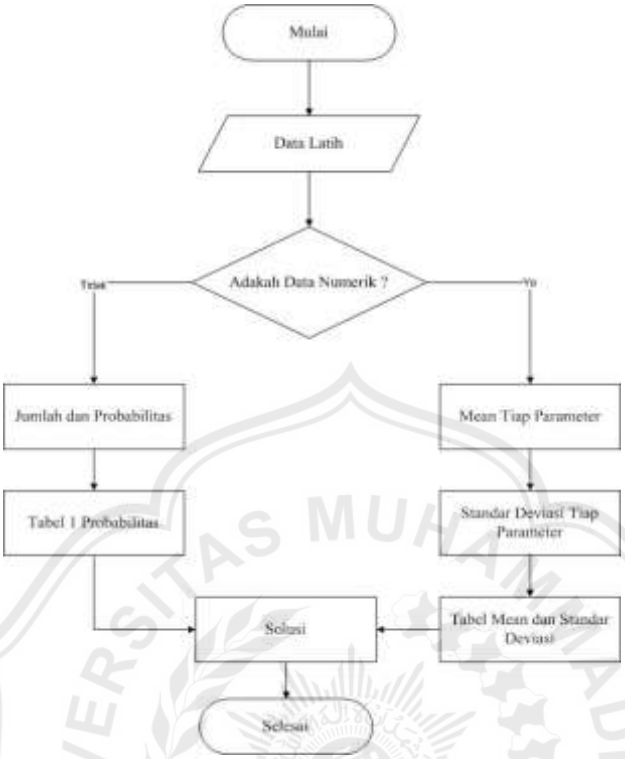
J. Studi Literatul

Penulis mencari referensi teori yang relevan dengan kasus atau permasalahan yang ditemukan. Studi literatur yang dilakukan oleh penelitian ini yaitu dengan melakukan pencarian terhadap jurnal-jurnal dari penelitian sebelumnya yang nantinya dapat mendukung dan dapat dijadikan rujukan atau referensi yangmemperkuat argumentasi-argumentasi yang ada. Maka penulis perlu mempelajari beberapa literatur yang digunakan, kemudian literatur tersebut diseleksi untuk dapat ditentukan literatur mana yang digunakan dalam penelitian

K. Pengumpulan Data

Pada penelitian ini data yang digunakan merupakan data yang diperoleh dari media internet pada website *UCI Machine Learning Repository* tentang penyakit Hepatitis C Virus Germany Pasien pada tahun 2020, dataset ini akan diolah menggunakan algoritma *Gaussian Naive Bayes* dan *K-Nearest Neighbor* . Dataset ini terdapat beberapa atribut yaitu, Age, Sex, ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT, dan Category.

L. *Implementasi Algoritma Gaussian Naive Bayes*



Gambar 2. Flowchart Gaussian Naive Bayes

M. Implementasi *K-Nearest Neighbor* (KNN)

Gambar 3. Flowchart KNN

3. HASIL DAN PEMBAHASAN

Data yang digunakan berasal dari situs website *UCI Machine Learning.* Data yang diambil dari situs tersebutadalah Hepatitis C virus (HCV) Jerman pasien pada tahun 2020. Pengumpulan data didapatkansebanyak 608 data dan 13 atribut.

A. Hasil Klasifikasi Gaussian Naive Bayes

Berikut *matrix* beserta perhitungan akurasi, presisi, dan recall pada *Gaussian Naive Bayes.* Akurasi adalah menggabarkan seberapa akurat sebuah algoritma dalam melakukan klasifikasi sehingga akurasi dapat dikatakan rasio prediksi benar (positif dan negatif) dengan kesuluruhan data.

Tabel 1. Counfunsion Matrix Akurasi Pada Algoritma Gussian Naive Bayes

		PREDICT CLASS			
		Blood Donor	Hepatitis	Fibrosis	Cirrhosis
ACTUAL CLASS	Blood Donor	104	1	3	0
	Hepatitis	3	3	0	0
	Fibrosis	1	1	2	0
	Cirrhosis	0	1	1	2

Tabel 2. Counfunsion Matrix Presisi Pada Algoritma Gaussian Naive Bayes

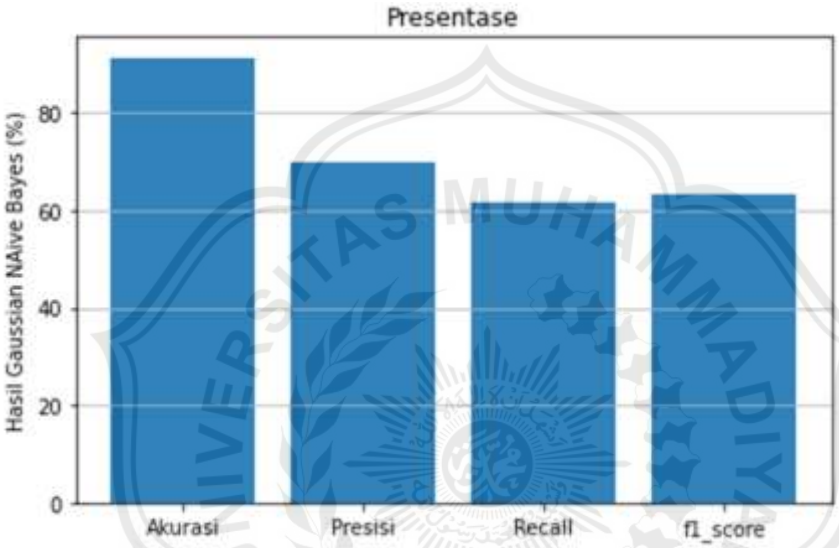
		PREDICT CLASS			
		Blood Donor	Hepatitis	Fibrosis	Cirrhosis
ACTUAL CLASS	Blood Donor	104	1	3	0
	Hepatitis	3	3	0	0
	Fibrosis	1	1	2	0
	Cirrhosis	0	1	1	2

Tabel 3. Counfunsion Matrix Presisi Pada Algoritma Gaussian Naive Bayes

		PREDICT CLASS			
		Blood Donor	Hepatitis	Fibrosis	Cirrhosis
ACTUAL CLASS	Blood Donor	104	1	3	0
	Hepatitis	3	3	0	0
	Fibrosis	1	1	2	0
	Cirrhosis	0	1	1	2

Tabel 4. Counfunsion Matrix Recall Pada Algoritma Guassian Naive Bayes

		PREDICT CLASS			
		Blood Donor	Hepatitis	Fibrosis	Cirrhosis
ACTUAL CLASS	Blood Donor	104	1	3	0
	Hepatitis	3	3	0	0
	Fibrosis	1	1	2	0
	Cirrhosis	0	1	1	2

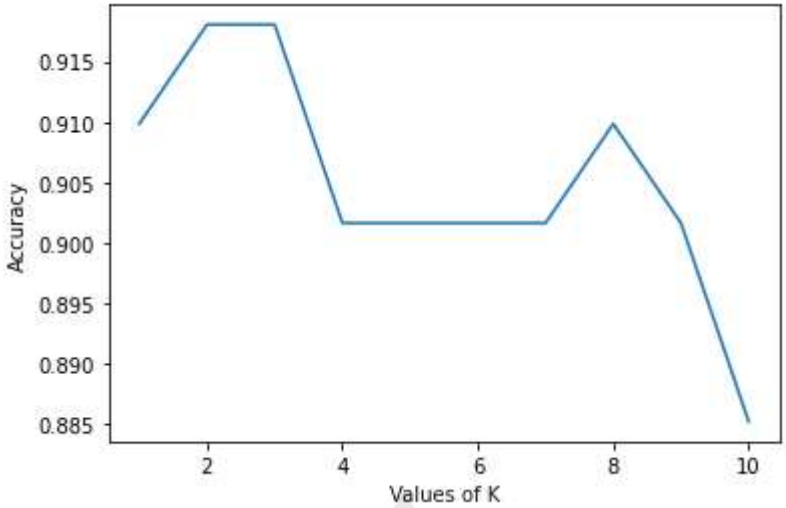


Gambar 4. Grafik Gaussian naïve bayes

B. Hasil Klasifikasi K-*Nearest Neighbor* (KNN)

Tabel 5. Counfusion Matrix Akurasi pada Algoritma K-Nearest Neighbor

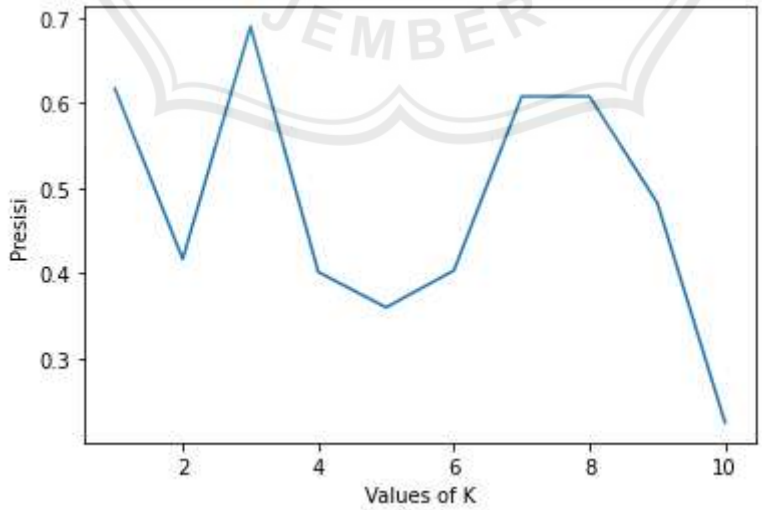
		PREDICT CLASS			
		Blood Donor	Hepatitis	Fibrosis	Cirrhosis
ACTUAL CLASS	Blood Donor	107	1	0	0
	Hepatitis	2	2	2	0
	Fibrosis	0	2	2	0
	Cirrhosis	1	2	0	1



Gambar 5. *Source code* Input Grafik Akurasi.

Tabel 6. Counfusion Matrix Presisi Pada Algoritma K-Nearest Neighbor

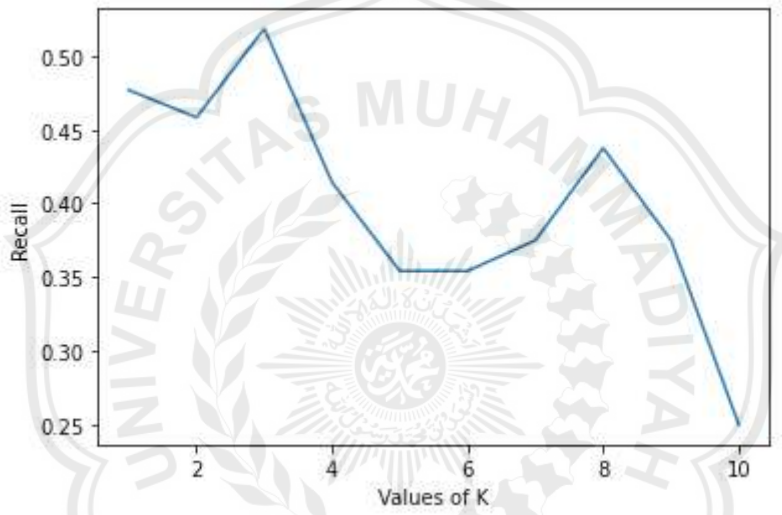
		PREDICT CLASS			
		Blood Donor	Hepatitis	Fibrosis	Cirrhosis
ACTUAL CLASS	Blood Donor	107	1	0	0
	Hepatitis	2	2	2	0
	Fibrosis	0	2	2	0
	Cirrhosis	1	2	0	1



Gambar 6. Grafik Presisi

Tabel 7. Counfusion Matrix Recall Pada Algoritma K-Nearest Neighbor

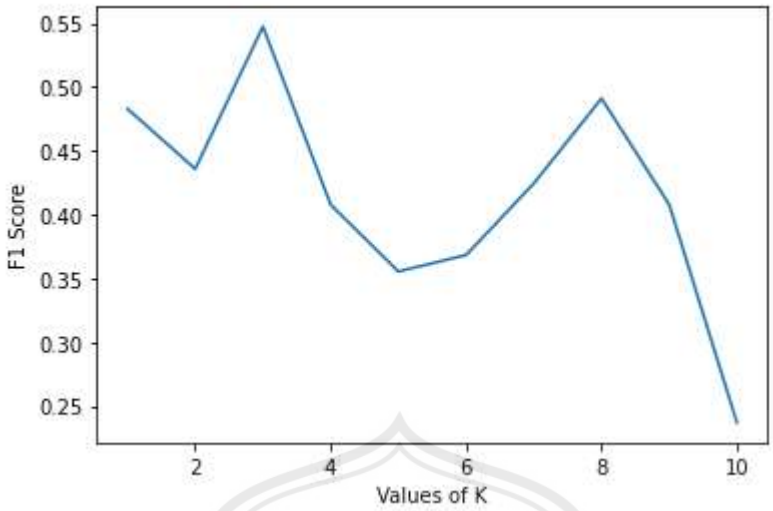
		PREDICT CLASS			
		Blood Donor	Hepatitis	Fibrosis	Cirrhosis
ACTUAL CLASS	Blood Donor	107	1	0	0
	Hepatitis	2	2	2	0
	Fibrosis	0	2	2	0
	Cirrhosis	1	2	0	1



Gambar 7. Grafik Recall

Tabel 8. *Confusion Matrix f1-score*

		PREDICT CLASS			
		Blood Donor	Hepatitis	Fibrosis	Cirrhosis
ACTUAL CLASS	Blood Donor	107	1	0	0
	Hepatitis	2	2	2	0
	Fibrosis	0	2	2	0
	Cirrhosis	1	2	0	1



Gambar 8. Grafik *f1-score*

Tabel 9. Nilai Akurasi, Presisi, Recall, F1-*Score* Pada Percobaan k

K	Akurasi	Presisi	Recall	F1 Score
1	90,98%	61,68%	47,69%	48,29%
2	91,80%	41,69%	45,83%	43,57%
3	91,80%	68,96%	51,85%	54,73%
4	90,16%	40,17%	41,44%	40,77%
5	90,16%	36,01%	35,42%	35,53%
6	90,16%	40,35%	35,42%	36,82%
7	90,16%	60,78%	37,50%	42,44%
8	90,98%	60,78%	43,75%	49,10%
9	90,16%	48,28%	37,50%	40,77%
10	88,25%	22,50%	25,00%	23,68%

Tabel 10. Perbandingan Akurasi, Presisi, Recall, Dan F1-Score

	Akurasi	Presisi	Recall	F1 Score
Gaussian Naive Bayes	90,98%	69,91%	61,57%	63,24%
K-Nearest Neighbor	91,80%	68,91%	51,85%	54,73%

Berdasarkan Tabel hasil akurasi, presisi, recall, dan f1-score untuk dataset Hepatitis CVirus (HCV) dengan algoritma *Gaussian Naive Bayes* didapatkan dengan hasil akurasi sebesar 90,98%, presisi sebesar 69,91%, dan recall sebesar 61,57% sedangkan pada algoritma *K-Nearest Neighbor* (KNN) didapatkan dengan hasil akurasi sebesar 91,80%, presisi sebesar 68,96%, dan recall sebesar 51,85%. Sehingga dalam melakukan prediksi penyakit Hepatitis C Virus (HCV) bahwa Algoritma *K-Nearest Neighbor* (KNN) lebih baik dari Algoritma *Gaussian Naive Bayes* karena memiliki nilai akurasi yang lebih tinggi sebesar 91,80% sedangkan jika dilihat dari tingkat ketepatan antara informasi yang diminta dengan jawaban yang diberikan Algortima *Gaussian Naive Bayes* lebih baik dari Algoritma *K-Nearest Neighbor* (KNN) karena memiliki nilai presisi sebesar 69,91% dan nilai recall sebesar 61,57%, f1 score dibutuhkan untuk menetukan nilai presisi dan recall yang baik dari setiap algoritma hasil f1 score pada klasifikasi penyakit Hepatitis C Virus (HCV), algoritma *Gaussian Naive Bayes* mendapatkan hasil f1 score lebih baik dibandingkan dengan *K-Nearest Neighbor* (KNN) yaitu sebesar 63,24% untuk algoritma *Gaussian Naive Bayes* dan 54,73% untuk *K-Nearest Neighbor*. Semakin besar nilai f1 score nya maka semakin baik nilai presisi dan recall.

3. KESIMPULAN DAN SARAN

3.1 Kesimpulan

hasil penelitian yang telah dilakukan dapat dilihat sebagai berikut.

1. Hasil akurasi yang paling tinggi pada klasifikasi Hepatitis C virus (HCV) diantara algoritma *Gaussian Naive Bayes* dan *K-Nearest Neigbor* (KNN) adalah pada algoritma *K-Nearest Neighbor* (KNN) yaitu sebesar 91,80%.
2. Hasil presisi yang paling tinggi pada klasifikasi Hepatitis C virus (HCV) diantara algoritma *Gaussian Naive Bayes* dan *K-Nearest Neigbor* adalah pada algoritma *Gaussian Naive Bayes* yaitu sebesar 85,71%, sehingga dalam memberikan informasi prediksi pasien terkena hepatitis yang pada kenyataanya pasien tidak terkena hepatitis Algoritma *Gaussian Naive Bayes* lebih baik dibandingkan *K-Nearest Neigbor* (KNN).
3. Hasil recall yang paling tinggi pada klasifikasi Hepatitis C virus (HCV) diantara algoritma *Gaussian Naive Bayes* dan *K-Nearest Neigbor* adalah pada algoritma *Gaussian Naive Bayes* yaitu sebesar 61,57%, sehingga dalam memberikan informasi prediksi pasien tidak terkena hepatitis yang pada kenyataanya pasien terkena hepatitis Algoritma *Gaussian Naive Bayes* lebih baik dibandingkan *K-Nearest Neigbor* (KNN).
4. Hasil F1 Score paling tinggi pada klasifikasi Hepatitis C Virus (HCV) diantara algoritma *Gaussian Naive Bayes* dan K-Nearest Neighbor adalah pada algoritma *Gussian Naive Bayes* yaitu sebesar 63,24% sehingga nilai rata-rata antara presisi dan recall algoritma *Gaussian Naive Bayes* lebih baik dibandingkan *K-Nearest Neighbor.*

3.2 Saran

Berdasarkan hasil penelitian yang telah dijabarkan sebelumnya,adapun saran yang diberikan untuk penelitian selanjutnya adalah sebagai berikut.

1. Agar menggunakan data Hepatitis C Virus (HCV) dalam kurun waktu terbaru untuk dapat meningkatkan hasil yang didapatkan.
2. Bisa menerapkan atau membangun pembuatan sistem klasifikasi penyakit Hepatitis C Virus (HCV) dengan penelitian ini sebagai acuan.

4. DAFTAR PUSTAKA

Kudesia, G., & Wreghitt, T. (2009). *Clinical and Diagnostic Virology*. United States of America by Cambridge University Press.

Riadi, M. (2017). *Pengertian, Fungsi, Proses dan Tahapan Data Mining*. Kajianpustaka.Com.

Blachier, M., Leleu, H., Peck-Radosavljevic, M., Valla, D. C., & Roudot-Thoraval, F. (2013). The burden of liver disease in Europe: A review of available epidemiological data. In *Journal of Hepatology*. https://doi.org/10.1016/j.jhep.2012.12.005

Davies, C. E., Moss, D., & Hill, M. O. (2004). EUNIS Habitat Classification Revised 2004. *Technology*.

Crivat, J. M. Z. T. B. (2011). Microsoft Data Mining with SQL Server ® ® 2008. In *Molecular ecology resources*.

Saxena, K., Khan, Z., & Singh, S. (2014). Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm. *International Journal of Computer Science Trends and Technology (IJCST)*