Jurnal EDUKASI ELEKTROMATIKA (JEE)
ISSN: 2747-0784 (p); xxxxxx (e)
Vol 4,No. 1, Juni 2023

CLASSIFICATION FEASIBILITY OF BLOOD THAT CAN BE DONORED USING K-NN WITH K-MEANS OPTIMIZATION

KLASIFIKASI KELAYAKAN DARAH YANG DAPAT DIDONORKAN MENGGUNAKAN K-NN DENGAN OPTIMASI K-MEANS

Yayang Fredyatama 1[1*], Zainal Abidin 2[2], Bijanto 3[3]
Prodi Informatika Sekolah Tinggi Teknik Pati, Indonesia[1,2,3]

e-mail: yayangfredyatama@gmail.com[1], zainal.fsrd@yahoo.id[2], biyantokakoi@ gmail.com[3]

Abstract *Blood is an important component for human body which has various benefits. Blood serves to provide information about the condition of the body, distribute oxygen and nutrients, as well as one of the medical efforts in saving human lives, called blood donation. Blood donors have a huge impact on society and millions of people need blood donation every year. Blood donation has a drawback impact in the form of spreading infectious diseases such as Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome (HIV/AIDS), Hepatitis C, Hepatitis B, Syphilis, Malaria, Dengue Hemorrhagic Fever (DHF), and other life-threatening risks. To avoid transmission of the disease, it is necessary to classify human blood that is eligible for donation. This study classified the eligibility of blood that can be donated using K-NN and K-means optimization. This study succeeded in achieving the highest level of accuracy of 98.91% with an error of 1.09% using k=3. This study also obtained a sensitivity value of 100% with a specificity value of 90.48% and an AUC of 95% which has excellent classification category*.

Keywords*: K-NN, K-Means, Blood Donation, Classification*

Abstrak Darah merupakan komponen penting bagi tubuh manusia yang memiliki berbagai macam manfaat. Darah berfungsi untuk menyediakan informasi mengenai kondisi tubuh, mendistribusikan oksigen dan zat makanan, serta sebagai salah satu upaya medis dalam penyelamatan nyawa manusia yaitu donor darah. Donor darah memiliki pengaruh yang sangat besar bagi masyarakat serta jutaan orang butuh tranfusi darah tiap tahunnya. Mendonorkan darah memiliki dampak negatif berupa penyebaran penyakit menular seperti *Human Immunodeficiency Virus*/*Acquired Immunodeficiency Syndrome* (HIV/AIDS), Hepatitis C, Hepatitis B, Sifilis, Malaria, Demam Berdarah *Dengue* (DBD), dan resiko lainnya yang membahayakan nyawa. Untuk menghindari penularan penyakit tersebut, maka perlu dilakukan klasifikasi darah manusia yang layak untuk didonorkan. Penelitian ini melakukan klasifikasi kelayakan darah yang dapat didonorkan menggunakan K-NN dan optimasi K-*means*. Penelitian ini berhasil mencapai tingkat akurasi tertinggi yaitu 98,91% dengan error sebesar 1,09% menggunakan *k*=3. Penelitian ini juga mendapatkan nilai *sensitifity* 100% dengan nilai *spesificity* sebesar 90,48% dan AUC 95% dengan kategori *excellent classification*

Kata Kunci: K-NN, K-Means, Donor Darah, Klasifikasi

PENDAHULUAN

Darah merupakan wadah untuk mendistribusikan berbagai macam sel ke seluruh tubuh yang terdiri atas cairan kompleks plasma tempat elemen seluler meliputi *eritrosit*, *leukosit*, dan *trombosit* [1]. Darah menjadi materi biologis yang memiliki banyak informasi meliputi kondisi

kesehatan, kondisi akut atau kronik, penyakit keturunan, penyakit yang mengenai kesehatan lingkungan termasuk pembentukan dan akibat yang dihasilkan [2]. Darah berfungsi dalam membawa intisari yang dibutuhkan dalam sel tubuh manusia meliputi oksigen, hasil metabolisme, nutrisi dan elektrolit [2]. Berdasarkan beberapa pernyataan tersebut, maka dapat disimpulkan bahwa jika manusia mengalami kekurangan darah, maka akan berakibat pada kondisi tubuh manusia yang mengalami berbagai macam gangguan.

Donor darah merupakan kegiatan mendistribusikan darah dari satu orang ke sistem peredaran darah orang lainnya [3]. Komponen darah yang dihasilkan dari pendonor sangat berharga bagi pengobatan modern, oleh karena itu pendonor yang sehat merupakan hal yang sangat penting [4]. Dari pernyataan tersebut, dapat disimpulkan bahwa kegiatan donor darah sangat penting dan sangat berharga bagi pengobatan modern dengan menyediakan darah yang sehat.

Melakukan donor darah berpotensi dalam menyebarkan penyakit infeksi menular seperti *Human Immunodeficiency Virus*/*Acquired Immunodeficiency Syndrome* (HIV/AIDS), Hepatitis C, Hepatitis B, Sifilis, Malaria, Demam Berdarah *Dengue* (DBD), dan resiko lainnya yang membahayakan nyawa [5]. Ada berbagai cara penularan penyakit tersebut namun sebagian besar berasal dari sentuhan luka terbuka, hubungan seksual, transfusi darah, obat intravena atau jarum suntik, vertikal darah ibu ke janin melalui infeksi perinatal, intrauterin, dan air susu ibu [6]. Dalam penelitian ini melakukan klasifikasi darah yang layak didonorkan untuk membantu menghindari penyebaran penyakit seperti Hepatitis C, di mana penyakit tersebut dapat menyebabkan hati mengalami komplikasi termasuk *fibrosis*, *sirosis*, kanker hati sehingga menimbulkan masalah kesehatan dunia dengan penderita mencapai 150 – 170 juta jiwa dan perkiraan korban jiwa mencapai 35.000 orang tiap tahun [7]. Berdasarkan pernyataan tersebut, perlu dilakukan klasifikasi kelayakan darah yang layak didonorkan dengan akurat untuk mengantisipasi penyebaran penyakit Hepatitis C dan penyakit lain yang sering menyebar melalui kontak darah.

Pada penelitian ini akan melakukan klasifikasi menggunakan K-NN (*K-Nearest Neighbor*) dan optimasi algoritma K-*Means* untuk menentukan kalayakan darah yang dapat didonorkan. K-*means* akan diaplikasikan pada tahap pemrosesan awal sebagai algoritma untuk melakukan optimasi, dalam kasus ini akan digunakan untuk pengelompokan terlebih dahulu sebelum data diklasifikasi. Setelah tahap pengelompokan selesai, K-NN akan diaplikasikan sebagai metode utama dalam melakukan klasifikasi. K-NN dalam penelitian ini akan menjadi metode utama dalam melakukan klasifikasi yang merupakan metode populer untuk melakukan klasifikasi dalam statistik serta *data mining* karena memiliki kinerja klasifikasi yang signifikan serta implementasi yang bersifat sederhana [8], [9], [10]. Dalam algoritma K-NN akurasi akan semakin tinggi jika jumlah data *training* semakin banyak namun sebaliknya jika jumlah data *training* semakin sedikit akan berakibat sulitnya sistem dalam melakukan klasifikasi [11]. Kelemahan yang lain yaitu apabila parameter k = 1, akan menghasilkan klasifikasi yang terlihat kaku sedangkan bila parameter k terlalu besar menghasilkan klasifikasi menjadi samar [12]. Terkait dengan metode K-NN yang memiliki keunggulan yang signifikan dalam melakukan klasifikasi, maka dari itu K-NN dipilih sebagai metode utama dalam melakukan klasifikasi.

Dalam penggunaannya, K-*means* digunakan sebagai metode *clustering* yang merupakan metode paling dasar dalam melakukan analisa *cluster* [13]. K-*means* juga dalam penggunaannya dapat dijadikan sebagai metode untuk pemrosesan awal ataupun juga dapat dikombinasikan pada metode lain [13]. K-*means* memiliki keunggulan dalam penggunaannya yaitu kesederhanaan implementasi, kemampuan dalam mengklaster data yang besar, dan komputasi yang cepat [14]. K-*means* memiliki kelemahan yaitu dalam menentukan nilai k menggunakan cara manual sehingga nilai k yang terbaik tidak bisa dipastikan serta dapat dipengaruhi oleh jumlah dimensi yang besar [13]. Terlepas dari kelemahannya, K-*means* tetap tidak dapat dikesampingkan terutama sebagai metode untuk pemrosesan awal serta dapat juga digunakan sebagai metode klasifikasi.

Penelitian sebelumnya tentang kelayakan calon pendonor darah menggunakan algoritma *Neural Network* oleh Firdaus, dkk (2020) menghasilkan akurasi sebesar 91.65% [15]. Selanjutnya penelitian oleh Handayani, dkk (2021) mengenai komparasi algoritma C4.5 dan *Naïve Bayes*

untuk menentukan status kelayakan pada donor darah menghasilkan akurasi tertinggi sebesar 93,26% pada algoritma *Naïve Bayes* [16]. Kemudian penelitian dari Mostafa & Hasan (2021) tentang komparasi 3 metode klasifikasi (a*rtificial neural* n*etwork*, *random forest*, *support vector machine*) untuk mendeteksi penyakit hati dan donor darah menghasilkan akurasi sebesar 98.23% menggunakan SVM (*Support Vector Machine)* [17]. Penelitian ini diharapkan dapat menghasilkan akurasi yang tinggi untuk hasil yang semakin akurat. Oleh karena itu, penelitian ini akan mengaplikasikan dua algoritma yaitu K-*means* digunakan sebagai optimasi dan K-NN sebagai metode utama untuk melakukan klasifikasi.

Darah menjadi komponen dalam tubuh manusia yang penting dalam mendistribusikan berbagai macam zat yang dibutuhkan oleh tubuh serta memberikan informasi mengenai kondisi kesehatan tubuh manusia. Oleh karena itu, apabila terdapat faktor yang dapat menimbulkan kekurangan darah pada tubuh manusia dapat berakibat pada berbagai gangguan kesehatan sehingga mengancam nyawa manusia. Salah satu cara mengatasi kekurangan darah pada manusia yaitu dengan cara donor darah, namun donor darah beresiko menularkan berbagai macam penyakit menular seperti Hepatitis C. Sehingga pada penelitian ini akan K-*means* akan diaplikasikan pada tahap pemrosesan awal sebagai algoritma untuk melakukan optimasi, dalam kasus ini akan digunakan untuk pengelompokan terlebih dahulu sebelum data diklasifikasi. Setelah tahap pengelompokan selesai, K-NN akan diaplikasikamn sebagai metode utama dalam melakukan klasifikasi.

METODE PENELITIAN

Pada penelitian ini, data yang digunakan merupakan data yang bersumber dari *website* penyedia data UCI *Machine Learning Repository* yang berjudul HCV *Data Dataset*. Data tersebut terdiri dari 615 *record data* dengan 14 atribut serta terdapat *missing value* (data yang hilang). Pada penelitian ini metode yang digunakan diantaranya meliputi imputasi *mean* untuk mengatasi *missing value*, K-*means clustering* untuk pemrosesan awal sebelum klasifikasi dan digunakan sebagai metode untuk optimasi, serta K-NN yang digunakan sebagai metode untuk melakukan klasifikasi.

A. Metode Imputasi Mean

Imputasi *missing value* telah dipelajari selama beberapa dekade sebagai solusi dasar untuk masalah kumpulan data yang tidak lengkap, khususnya pada beberapa sampel data yang mengandung satu atau lebih nilai atribut yang hilang [18]. Berbagai teknik imputasi telah dikembangkan guna meminimalisir dampak negatif *missing value*, salah satu teknik tersebut ialah metode *mean* yang mengubah *missing value* menggunakan nilai rata-rata dari suatu variabel [19]. Teknik tersebut merupakan teknik yang paling sering digunakan [19]. Teknik tersebut cara kerjanya yaitu mengganti nilai data yang hilang dengan rata-rata dari semua nilai atribut yang sudah diketahui untuk diisikan pada tempat *instance* atribut yang datanya hilang [19]. Rumus dari penanganan *missing value* menggunakan teknik imputasi *mean* dapat dijelaskan pada persamaan sebagai berikut [20]:

$$\bar{x} = \frac{\sum_{i=1}^{n} xi}{n} \quad (1)$$

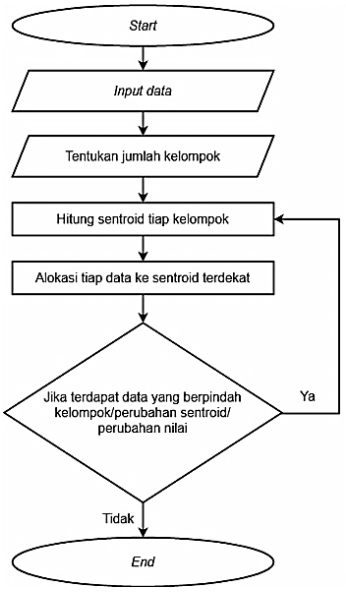
Rumus tersebut dapat dijelaskan bahwa $\bar{x}$ mewakili *record* yang nilainya kosong sedangkan xi mewakili nilai tiap atribut. Simbol n mewakili banyaknya data setiap atribut [20]. Berdasarkan penjelasan tersebut, maka dapat disimpulkan bahwa imputasi *mean* menjumlah semua nilai tiap kolom yang terdapat *missing value* dibagi dengan banyaknya data pada kolom tersebut. Data yang akan digunakan dalam penelitian ini terdapat *missing value* (data yang hilang). Untuk menangani *missing value*, penelitian ini akan menggunakan teknik imputasi *mean* sebagai metode untuk mengganti data yang hilang (replace *missing value*). Teknik imputasi mean merupakan salah satu teknik imputasi untuk mengatasi *missing value* dengan cara mengganti data yang hilang (terdapat *missing value*). Pada penelitian ini, dalam menangani *missing value* dengan cara mengganti data yang kosong diperkirakan dapat menghasilkan kinerja yang lebih baik untuk digunakan sebagai klasifikasi. Hal tersebut dikarenakan data yang terdapat *missing*

value tidak dihilangkan, akan tetapi diperbaiki dengan menerapkan teknik imputasi yaitu imputasi *mean*.

B. Pengelompokkan Menggunakan Algoritma K-Means

Dalam *machine learning*, *clustering* menggunakan K-*means* merupakan metode analisis pengelompokan yang mengarah kepada pembagian partisi *N* objek pengamatan ke dalam *K* kelompok di mana tiap objek pengamatan dimiliki oleh sebuah kelompok dengan rata-rata (*mean*) yang paling dekat [21]. Dalam penggunaannya, *K-means* digunakan sebagai metode *clustering* di mana *k-means* sendiri merupakan metode paling dasar dalam melakukan analisa *cluster* [13]. K-*means* juga dalam penggunaannya dapat dijadikan sebagai metode untuk pemrosesan awal ataupun juga dapat dikombinasikan pada metode lain [13]. Disamping penerapannya sebagai metode untuk melakukan *clustering*, *K-means* juga dapat digunakan sebagai metode untuk melakukan klasifikasi [22]. Jadi, dapat disimpulkan bahwa K-*means* merupakan metode yang digunakan untuk melakukan pengelompokan data serta dapat dikombinasikan untuk pemrosesan awal dan juga dapat digunakan untuk melakukan klasifikasi data.

K-*means* juga dalam penggunaannya dapat dijadikan sebagai metode untuk pemrosesan awal ataupun juga dapat dikombinasikan pada metode lain [13]. Algoritma K-*means* dalam implementasinya memiliki keunggulan dan kekurangan. K-*means* memiliki keunggulan diantaranya kesederhanaan implementasi, relatif cepat, mudah diadaptasi, dan penggunaannya umum digunakan dalam praktek [23]. K-*means* dalam penggunaannya merupakan salah satu algoritma paling penting pada bidang *data mining* [23]. K-*means* dalam penerapannya juga memiliki kelemahan yaitu penentuan nilai *k* secara manual sehingga tidak dapat dipastikan nilai *k* terbaik serta kurang mampu dalam mengolah data dengan dimensi tinggi [13]. Masalah yang paling utama dalam penerapan algoritma K-*means* yaitu menentukan berapa banyak *cluster* yang akan dicari [24]. Artinya tidak ada yang menentukan nilai *k* kecuali analis memiliki pengetahuan tentang prioritas jumlah *cluster* yang mendasari [24]. Langkah pengelompokan pada algoritma K-means disajikan dalam *flowchart* pada gambar 1.



Gambar 1 Flowchart algoritma K-means [21]

Untuk menghitung sentroid data, teknik yang digunakan dalam penelitian ini yaitu *euclidean distance* (jarak *euclidean*). *Euclidean distance* (jarak *euclidean*) merupakan teknik yang paling umum digunakan dalam penentuan jarak [25]. Rumus dari teknik *Euclidean distance* dalam menentukan sentroid data dapat dijabarkan pada persamaan berikut [26]:

Jurnal EDUKASI ELEKTROMATIKA (JEE)
ISSN: 2747-0784 (p); xxxxxx (e)
Vol 4,No. 1, Juni 2023

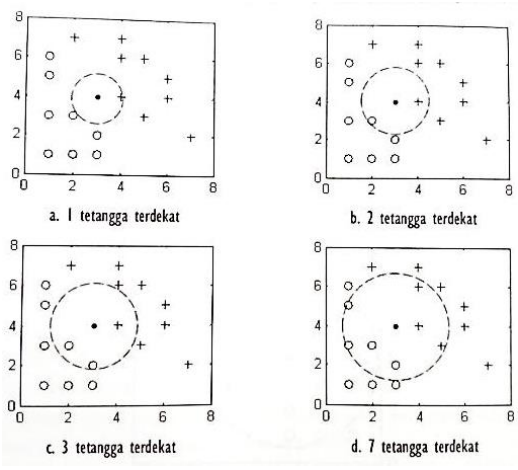
$$Euclidean = \sqrt{\sum_{i=1}^{n}(Pi - Qi)^2} \quad (2)$$

Berdasarkan persamaan tersebut maka dapat dijelaskan bahwa [26]:
Pi = menunjukkan data latih
Qi = menunjukkan data uji
i = menunjukkan data variabel
n = menunjukkan dimensi data.

Pengelompokan menggunakan algoritma K-*means* dalam penelitia ini digunakan sebagai teknik untuk melakukan optimasi data sebelum data diolah menggunakan algoritma K-NN. Sebelum data diklasifikasi dengan K-NN, data akan terlebih dahulu dilakukan proses pengelompokan (*clustering*). Pada lampiran 1 menjelaskan bahwa data yang digunakan pada penelitian ini juga termasuk dalam data yang dapat digunakan untuk tugas pengelompokan (*clustering*). Data yang sudah dikelompokkan akan menghasilkan data baru. Data baru hasil dari pengelompokan mesnggunakan K-*means* akan diklasifikasi dengan tujuan ada peningkatan akurasi dibandingkan pada penelitian sebelumnya yang menggunakan algoritma K-NN.

C. Klasifikasi Menggunakan Algoritma K-NN

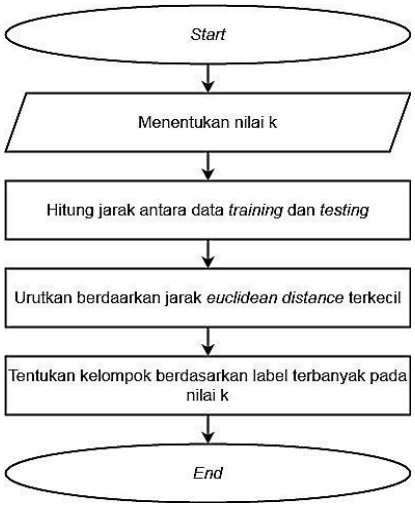
K-NN menjadi salah satu contoh pembelajaran berbasis *instance*, dimana kumpulan data pelatihan disimpan sehingga klasifikasi untuk catatan baru yang tidak terklasifikasi dapat ditentukan dengan membandingkannya pada catatan yang paling mirip dalam kumpulan data pelatihan [24]. K-NN merupakan algoritma yang paling sering digunakan untuk klasifikasi, meskipun juga dapat digunakan untuk estimasi dan prediksi [24]. Untuk melakukan klasifikasi, algoritma K-NN termasuk algoritma yang sederhana serta memiliki fleksibilitas tinggi dalam menyelesaikan permasalahan klasifikasi yang kompleks [10]. K-NN merupakan algoritma yang melakukan proses klasifikasi berdasarkan kedekatan dari jarak suatu data dengan data yang lain [27]. Dalam bidang *machine learning* algoritma termudah ialah K-NN, dalam pengenalan pola K-NN merupakan metode non-parametik untuk klasifikasi dan regresi [28]. Gambar 2 menggambarkan konsep dari ketetanggaan K-NN dari mulai 1 tetangga terdekat, 2 tetangga terdekat, 3 tetangga terdekat, dan 7 tetangga terdekat.



Gambar 2 Ketetanggaan K-NN [21]

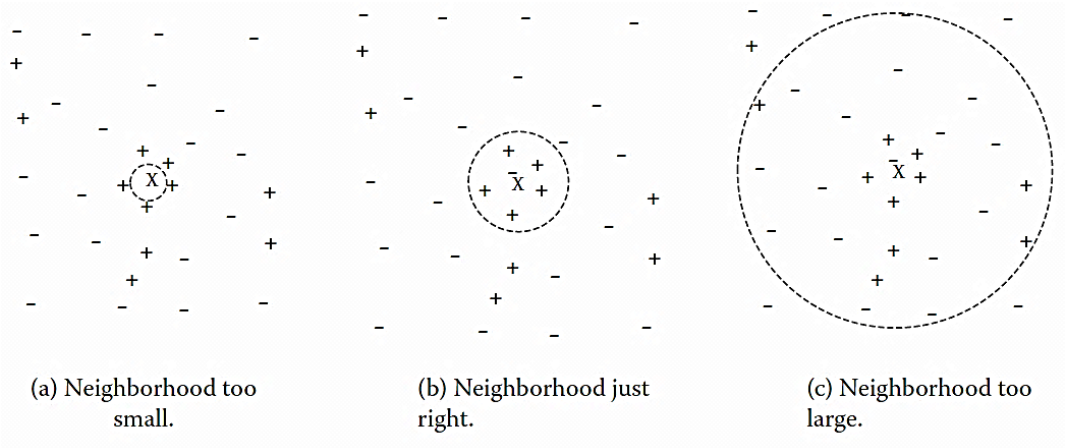
Gambar 2 dapat dijelaskan bahwa nilai *k* pada K-NN berarti k data terdekat dari data uji Pada gambar 2 tanda lingkaran untuk kelas 0, tanda *plus* untuk kelas 1. jika *k* bernilai 1, kelas dari satu data latih sebagai tetangga terdekat (terdekat pertama) dari data uji tersebut akan

diberikan sebagai kelas untuk data uji yaitu kelas 1, jika *k* bernilai 2, akan diambil 2 tetangga terdekat dari data latih. Hal tersebut juga berlaku jika nilai *k* = 3,4,5 dan sebagainya. Jika dalam *k* tetangga ada dua kelas yang berbeda, akan diambil kelas dengan jumlah data terbanyak (voting mayoritas). Pada gambar 2 terlihat bahwa kelas 0 mempunyai jumlah yang lebih banyak daripada kelas 1 sehingga data uji akan dikategorikan ke dalam kelas 0. Jika kelas dengan kata terbanyak ada dua atau lebih, akan diambil kelas dari data dengan jumlah yang sama tersebut acak. Proses klasifikasi menggunakan algoritma K-NN disajikan dalam bentuk *flowchart* pada gambar 2.



Gambar 3 Proses Algoritma K-NN

Dalam penentuan nilai *k* dapat berpengaruh terhadap kinerja algoritma K-NN [23]. Pada gambar 3 menunjukkan objek uji yang tidak berlabel "*x*" dan objek pelatihan dengan kelas "+" atau "−". Jika *k* terlalu kecil, maka hasilnya akan sensitif terhadap titik noise [23]. Di sisi lain, jika *k* terlalu besar, maka lingkungan tersebut dapat memasukkan terlalu banyak titik dari kelas lain [23].



Gambar 4 Klasifikasi K-NN dengan nilai k kecil, medium, dan besar

Pada penelitian ini dalam mencari jarak terdekat menggunakan *Euclidean distance* (jarak *euclidean*) untuk menentukan nilai *k* terdekat untuk melakukan klasifikasi. *Euclidean distance* (jarak *euclidean*) merupakan jarak dari suatu titik masuk dengan garis lurus yang menggunakan metode Teori Pitagoras [9]. Terdapat banyak fungsi untuk menentukan jarak, akan tetapi

algoritma K-NN (*K-Nearest Neighbor*) lebih sering memakai fungsi jarak berupa *euclidean distance* [29]. Rumus *euclidean distance* dapat dilihat pada persamaan berikut [26]:

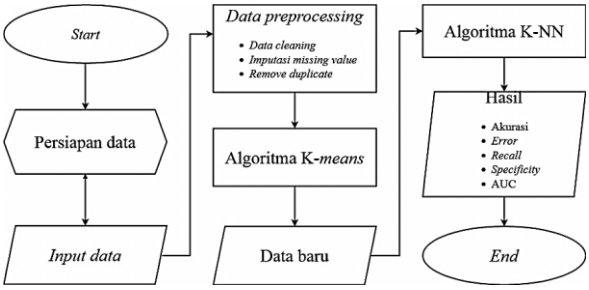
$$Euclidean = \sqrt{\sum_{i=1}^{n}(Pi - Qi)^2} \quad (2)$$

Berdasarkan persamaan tersebut maka dapat dijelaskan bahwa:
Pi = menunjukkan data latih
Qi = menunjukkan data uji
i = menunjukkan data variabel
n = menunjukkan dimensi data.

Algoritma K-NN dalam penelitian ini akan menjadi metode utama dalam mengolah data untuk menghasilkan akurasi, *error*, *recall*, *spesificity*, dan AUC. Pada penelitian ini, klasifikasi menggunakan algoritma K-NN akan mencari akurasi tertinggi berdasarkan nilai k terbaik. Hasil akurasi tertinggi akan digunakan sebagai *output* dari penelitian dan diharapkan akurasi dapat melampaui penelitian sebelumnya.

D. Kerangka Pemikiran

Kerangka pikir menjelaskan tahapan yang dilaksanakan dari awal hingga akhir penelitian di mana *outpuT* dari hasil penelitian yang ingin dicapai peningkatan akurasi dari penelitian sebelumnya. Kerangka pemikiran dalam penelitian ini dapat dilihat pada gambar 5 sebagai berikut:



Gambar 5 Kerangka Pemikiran

a. Persiapan data

Dalam mencari dan mendapatkan data pada penelitian ini memanfaatkan *website* UCI *Machine Learning Repository* sebagai *website* penyedia data publik yang digunakan untuk penelitian. Judul dari *dataset* yang digunakan dalam *website* UCI adalah "*HCV data Data Set*" yang diterbitkan pada tahun 2020. *Dataset* tersebut terdiri dari 615 *instance* dengan 14 atribut serta terdapat 31 *record* dengan *missing value* (data yang hilang).

b. Data *preprocessing*

Setelah data berhasil didapatkan, langkah selanjutnya adalah data akan di *input* untuk tahap pengolahan awal (*data preprocessing*). Langkah *data preprocessing* merupakan hal yang sangat penting sebelum data diolah menggunakan suatu algoritma. Hal tersebut dikarenakan sebagian besar data yang masih mentah dalam *database* merupakan data yang tidak diproses, tidak lengkap, dan terdapat *noise* (gangguan) [24]. Contohnya dalam sebuah data terdapat *field* yang usang atau berlebihan, *missing value*, outlier, data tidak cocok untuk model *data mining*, serta nilai yang tidak konsisten [24]. Tahap *data preprocessing* pada penelitian ini meliputi *data cleaning*, imputasi *missing value*, serta *remove duplicate* yang dapat dijelaskan sebagai berikut:

c. Clustering menggunakan algoritma K-means

Setelah langkah *data preprocessing*, maka data siap untuk diolah menggunakan algoritma. Dalam penelitian ini, setelah langkah data *preprocessing* (meliputi *data cleaning*, imputasi *missing value*, *remove duplicate*) selesai, selanjutnya yaitu melakukan proses *clustering* menggunakan algoritma K-*means*. Algoritma K-*means* dapat dijadikan sebagai metode untuk pemrosesan awal ataupun dapat juga dikombinasikan pada metode lain [13]. Oleh karena itu, pada penelitian ini menggunakan algoritma K-*means* untuk dikombinasikan menggunakan algoritma K-NN dengan tujuan mengoptimalkan data sebelum proses klasifikasi data. Proses *clustering* pada penelitian ini akan menghasilkan data baru. Data baru tersebut selanjutnya akan dilakukan proses klasifikasi menggunakan algoritma K-NN.

d. Klasifikasi menggunakan algoritma K-NN

Tahap klasifikasi menggunakan algoritma K-NN merupakan tahap inti dari penelitian yang dilaksanakan. K-NN termasuk dalam algoritma yang penggunaannya didominasi untuk malakukan klasifikasi [24]. Perbedaan pada jumlah tetangga terdekat (nilai k) yang digunakan dapat berpengaruh pada hasil klasifikasi [30]. Berdasarkan pernyataan tersebut, maka pada penelitian ini akan membandingkan penggunaan nilai k pada algoritma K-NN antara 3,5,7, dan 9 di mana nilai k dengan akurasi tertinggi akan dipilih sebagai *output* dengan harapan akurasi yang mengalami peningkatan dari peneltian sebelumnya. Setelah proses klasifikasi selesai, maka akan keluar hasil dari pengolahan menggunakan algoritma K-NN. *Output* (hasil) pada penelitian ini menggunakan pengukuran akurasi, *error*, *recall*, *specificity*, dan AUC (*Area Under the Curve*).

HASIL DAN PEMBAHASAN

A. Persiapan Data

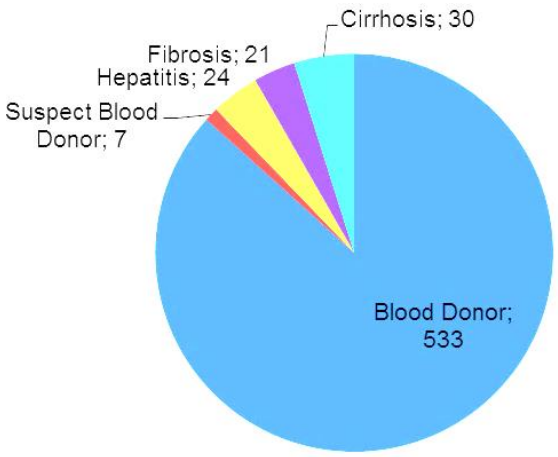
Data didapatkan dari website UCI dengan judul "HCV *data dataset*". Data tersebut terdiri dari 615 *instance* dengan 14 atribut serta terdapat *missing value* (data yang hilang) sebanyak 31 data. Atribut target adalah "*category*" dengan klasifikasi meliputi *blood donors*, Hepatitis C, *Fibrosis*, dan *Cirrhosis*. Data yang digunakan pada penelitian ini dapat dilihat pada tabel 1 sebagai berikut.

Tabel 1 *Sampel data*

Id	*Age*	*Sex*	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	*Category*
1	32	m	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106	12.1	69	0=*Blood donor*
2	32	m	38.5	70.3	18	24.7	3.9	11.17	4.8	74	15.6	76.5	0=*Blood donor*
3	32	m	46.9	74.7	36.2	52.6	6.1	8.84	5.2	86	33.2	79.3	0=*Blood donor*
4	32	m	43.2	52	30.6	22.6	18.9	7.33	4.74	80	33.8	75.7	0=*Blood donor*
...													
534	47	m	22.5	124	79.5	46.7	2.3	6.83	4.3	170	345.6	58.6	0s=*suspect Blood donor*
535	48	m	24.9	116.9	49.2	24.3	4.9	3.44	5.25	29	83	47.8	0s=*suspect Blood donor*
536	49	m	21.6	42.2	9.5	10.6	2.4	3.75	3.01	64	38.9	44.8	0s=*suspect Blood donor*
...													
541	38	m	45	56.3	NA	33.1	7	9.58	6	77.9	18.9	63	1=*Hepatitis*
542	19	m	41	NA	87	67	12	7.55	3.9	62	65	75	1=*Hepatitis*
543	23	m	47	19.1	38.9	164.2	17	7.09	3.2	79.3	90.4	70.1	1=*Hepatitis*
544	25	m	42	38.2	63.3	187.7	14	6	4.28	66.9	40.2	70.5	1=*Hepatitis*
...													
565	29	m	41	43.1	2.4	83.5	6	11.49	5.42	55.2	130	66.5	2=*Fibrosis*
566	40	m	39	43.1	23.8	114.7	11	9.64	4.2	70.9	127.3	81.3	2=*Fibrosis*
567	46	m	45	26.9	23.1	125	17	6.97	4.01	60.5	72.2	73	2=*Fibrosis*
568	48	m	49	45.2	19.3	69.1	30	7.76	4.22	76.7	28.4	72.3	2=*Fibrosis*
...													
612	64	f	24	102.8	2.9	44.4	20	1.54	3.02	63	35.9	71.3	3=*Cirrhosis*
613	64	f	29	87.3	3.5	99	48	1.66	3.63	66.7	64.2	82	3=*Cirrhosis*
614	46	f	33	NA	39	62	20	3.56	4.2	52	50	71	3=*Cirrhosis*
615	59	f	36	NA	100	80	12	9.07	5.3	67	34	68	3=*Cirrhosis*

Berdasarkan tabel 1 dapat dijelaskan bahwa atribut "*category*" merupakan variabel target di mana terdapat 5 jenis variabel target yaitu *blood donor*, *suspect blood donor*, hepatitis, *fibrosis*, dan *cirrhosis*.

Jurnal EDUKASI ELEKTROMATIKA (JEE)
ISSN: 2747-0784 (p); xxxxxx (e)
Vol 4,No. 1, Juni 2023



Gambar 6 Grafik pembagian value pada atribut *category*

Pada gambar 6 menunjukkan bahwa kategori *blood donor* terdapat sebanyak 533 data, suspect blood donor sebanyak 7 data, hepatitis sebanyak 24 data, *fibrosis* sebanyak 21 data, dan *chirrhosis* sebanyak 30 data. Berdasarkan pembagian dari keseluruhan kategori tersebut, maka terdapat total keseluruhan data sebanyak 615 data. Pada atribut "Id" digunakan sebagai index atau penomoran data, atribut "*Age*" menunjukkan umur, dan "*Sex*" melambangkan jenis kelamin. Untuk atribut yang lain dapat dijelaskan pada tabel 2 sebagai berikut [31]*:*

Tabel 2 *Penjelasan atribut data*

No	Nama Atribut	Makna
1	ALB	*Albumin*
2	ALP	*Alkaline Phosphatase*
3	ALT	*Alanine Amino-Transferase*
4	AST	*Aspartate Amino-Transferase*
5	BIL	*Bilirubin*
6	CHE	*Choline Esterase*
7	CHOL	*Cholesterol*
8	CREA	*Creatinine*
9	GGT	*Glutamyl-Transferase*
10	PROT	*Protein*

Sampel data yang digunakan dalam penelitian ini pada tabel 4.1 terdapat *record* yang bernilai NA. NA merupakan *record* yang memiliki kepanjangan dari *Not Available* (nilai tidak tersedia). Artinya data yang memiliki *record* dengan nilai NA adalah data yang memiliki *missing value*.

B. Prepocessing

Sebelum data diolah menggunakan algoritma, data perlu melalui tahap *preprocessing* terlebih dahulu. Pada penelitian ini, tahapan *preprocessing* meliputi *data cleaning*, imputasi *missing value*, dan *remove duplicate*. Berikut adalah penjabaran dari proses *preprocessing* data yang meliputi *data cleaning*, imputasi *missing value*, dan *remove duplicate*.

a. Data cleaning

Pada *record* data yang terdapat *missing value* masih tertulis NA. Pada proses ini atribut yang nilainya masih dalam bentuk huruf akan diubah ke dalam bentuk angka. Untuk data bernilai NA yang melambangkan *missing value* akan dikosongkan untuk proses penanganan *missing value* (setelah proses *data cleaning*). Atribut yang nilainya dalam bentuk huruf adalah "*Sex*" yang melambangkan jenis kelamin serta "*Category*" sebagai atribut yang berfungsi sebagai variabel

target. Atribut "*Sex*" memiliki *value* yaitu "m" yang melambangkan *male* atau jenis kelamin laki-laki, sedangkan "f" melambangkan *female* atau jenis kelamin perempuan. Atribut "*Sex*" nilainya akan dikonversi kedalam bentuk angka di mana "m" dikonversi menjadi "1" kemudian untuk "f" menjadi "0".

Langkah selanjutnya mengonversi nilai dari atribut "*Category*" menjadi dalam bentuk angka. Atribut "*Category*" dengan *value* "0=*Blood donor*" dikonversi menjadi "1", "0s=*suspect Blood donor*" dikonversi menjadi "2", "1=Hepatitis" dikonversi menjadi "3", "2=*Fibrosis*" dikonversi menjadi "4", serta yang terakhir yaitu "3=*Cirrhosis*" dikonversi menjadi "5". Setelah data dikonversi, maka tidak ada lagi atribut yang masih dalam bentuk huruf, sehingga hasil dari data *cleaning* dapat dilihat pada tabel 3. Data hasil proses *cleaning* terdapat kolom yang kosong (*missing value*) untuk selanjutnya kolom yang masih kosong akan diisi dengan data hasil pengolahan menggunakan imputasi *mean*.

Tabel 3 *Sampel data hasil proses cleaning*

Id	*Age*	*Sex*	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	*Category*
1	32	1	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106	12.1	69	1
2	32	1	38.5	70.3	18	24.7	3.9	11.17	4.8	74	15.6	76.5	1
3	32	1	46.9	74.7	36.2	52.6	6.1	8.84	5.2	86	33.2	79.3	1
4	32	1	43.2	52	30.6	22.6	18.9	7.33	4.74	80	33.8	75.7	1
...													
534	47	1	22.5	124	79.5	46.7	2.3	6.83	4.3	170	345.6	58.6	2
535	48	1	24.9	116.9	49.2	24.3	4.9	3.44	5.25	29	83	47.8	2
536	49	1	21.6	42.2	9.5	10.6	2.4	3.75	3.01	64	38.9	44.8	2
...													
541	38	1	45	56.3		33.1	7	9.58	6	77.9	18.9	63	3
542	19	1	41		87	67	12	7.55	3.9	62	65	75	3
543	23	1	47	19.1	38.9	164.2	17	7.09	3.2	79.3	90.4	70.1	3
544	25	1	42	38.2	63.3	187.7	14	6	4.28	66.9	40.2	70.5	3
...													
565	29	1	41	43.1	2.4	83.5	6	11.49	5.42	55.2	130	66.5	4
566	40	1	39	43.1	23.8	114.7	11	9.64	4.2	70.9	127.3	81.3	4
567	46	1	45	26.9	23.1	125	17	6.97	4.01	60.5	72.2	73	4
568	48	1	49	45.2	19.3	69.1	30	7.76	4.22	76.7	28.4	72.3	4
...													
612	64	0	24	102.8	2.9	44.4	20	1.54	3.02	63	35.9	71.3	5
613	64	0	29	87.3	3.5	99	48	1.66	3.63	66.7	64.2	82	5
Id	***Age***	***Sex***	**ALB**	**ALP**	**ALT**	**AST**	**BIL**	**CHE**	**CHOL**	**CREA**	**GGT**	**PROT**	***Category***
614	46	0	33		39	62	20	3.56	4.2	52	50	71	5
615	59	0	36		100	80	12	9.07	5.3	67	34	68	5

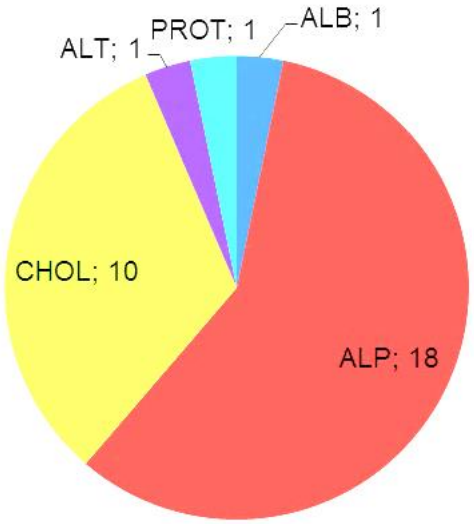
b. Imputasi missing value

Pada tahap ini, metode imputasi *missing value* akan diterapkan untuk mengisi data yang kosong supaya semua data terisi sehingga diharapkan memberikan hasil yang bagus. Teknik imputasi *missing value* yang digunakan dalam penelitian ini adalah imputasi *mean* yang memberikan hasil eksperimen yang baik. Tabel 4 menunjukkan sampel data *missing value*, sedangkan pada gambar 7 menjelaskan total data *missing value*. Hasil dari imputasi *missing value* menggunakan imputasi *mean* dapat dilihat pada data yang dicetak tebal pada tabel 5.

Jurnal EDUKASI ELEKTROMATIKA (JEE)
ISSN: 2747-0784 (p); xxxxxx (e)
Vol 4,No. 1, Juni 2023

Tabel 4 *Sampel dan missing value*

id	*Age*	*Sex*	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	*Category*
122	43	1	48,6	45	10,5	40,5	5,3	7,09	?	63	25,1	70	1
320	32	0	47,4	52,5	19,1	17,1	4,6	10,19	?	63	23	72,2	1
330	33	0	42,4	137,2	14,2	13,1	3,4	8,23	?	48	25,7	74,4	1
414	46	0	42,9	55,1	15,2	29,8	3,6	8,37	?	61	29	71,9	1
425	48	0	45,6	107,2	24,4	39	13,8	9,77	?	88	38	75,1	1
434	48	0	46,8	93,3	10	23,2	4,3	12,41	?	52	23,9	72,4	1
499	57	0	48,4	94,4	2,5	39,6	2,3	8,84	?	82	6,4	76,8	1
541	38	1	45	56,3	?	33,1	7	9,58	6	78	18,9	63	3
542	19	1	41	?	87	67	12	7,55	3,9	62	65	75	3
546	29	1	49	?	53	39	15	8,79	3,6	79	37	90	3
547	30	1	45	?	66	45	14	12,16	6,1	86	43	77	3
569	49	1	39	?	118	62	10	7,28	3,5	72	74	81	4
570	49	1	46	?	114	75	16	10,43	5,2	72	59	82	4
571	50	1	42	?	258	106	15	8,74	4,7	77	80	84	4
...													
615	59	0	36	?	100	80	12	9,07	5,3	67	34	68	5



Gambar 7 Jumlah data *missing value*

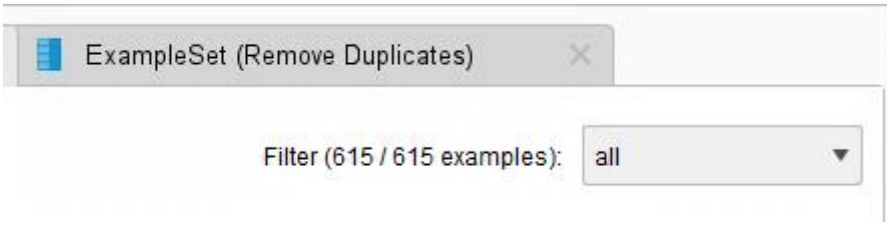
Tabel 5 *Hasil imputasi mean*

id	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	Cat
122	43	1	48,60	45,00	10,50	40,50	5,30	7,09	**5,37**	63,00	25,10	70,00	1
320	32	0	47,40	52,50	19,10	17,10	4,60	10,19	**5,37**	63,00	23,00	72,20	1
330	33	0	42,40	137,20	14,20	13,10	3,40	8,23	**5,37**	48,00	25,70	74,40	1
414	46	0	42,90	55,10	15,20	29,80	3,60	8,37	**5,37**	61,00	29,00	71,90	1
425	48	0	45,60	107,20	24,40	39,00	13,80	9,77	**5,37**	88,00	38,00	75,10	1
434	48	0	46,80	93,30	10,00	23,20	4,30	12,41	**5,37**	52,00	23,90	72,40	1
499	57	0	48,40	94,40	2,50	39,60	2,30	8,84	**5,37**	82,00	6,40	76,80	1
541	38	1	45,00	56,30	**28,45**	33,10	7,00	9,58	6,00	78,00	18,90	63,00	3
542	19	1	41,00	**68,28**	87,00	67,00	12,00	7,55	3,90	62,00	65,00	75,00	3
546	29	1	49,00	**68,28**	53,00	39,00	15,00	8,79	3,60	79,00	37,00	90,00	3
id	**Age**	**Sex**	**ALB**	**ALP**	**ALT**	**AST**	**BIL**	**CHE**	**CHOL**	**CREA**	**GGT**	**PROT**	**Cat**

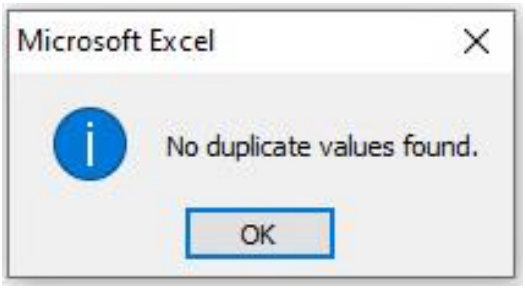
547	30	1	45,00	**68,28**	66,00	45,00	14,00	12,16	6,10	86,00	43,00	77,00	3
569	49	1	39,00	**68,28**	118,00	62,00	10,00	7,28	3,50	72,00	74,00	81,00	4
570	49	1	46,00	**68,28**	114,00	75,00	16,00	10,43	5,20	72,00	59,00	82,00	4
571	50	1	42,00	**68,28**	258,00	106,00	15,00	8,74	4,70	77,00	80,00	84,00	4
...													
615	59	0	36,00	**68,28**	100,00	80,00	12,00	9,07	5,30	67,00	34,00	68,00	5

c. Remove duplicate

Pada tahap *remove duplicate* data akan diperiksa secara menyeluruh dan jika terdapat data yang ganda, maka data ganda tersebut dihapus. Apabila tidak ada data yang ganda, maka keseluruhan data yang didapatkan oleh peneliti dapat diolah semuanya tanpa menghilangkan salah satu *record*. Dalam memeriksa data yang ganda, pada pelitian ini akan menggunakan alat untuk membantu dalam memeriksa data yang ganda. Alat yang digunakan untuk membantu memeriksa data ganda dalam penelitian ini adalah *rapidminer studio* dan *microsoft excel*. Dalam penelitian ini, data yang digunakan tidak terdapat data ganda sehingga keseluruhan data dapat diolah.



Gambar 8 *Remove duplicate* menggunakan *rapidminer*



Gambar 9 *Remove duplicate* menggunakan *microsoft excel*

C. Pengelompokkan menggunakan K-means

Pada tahap ini, data dikelompokkan menggunakan algoritma K-*means* dengan tujuan untuk mengoptimalkan data sebelum proses klasifikasi. Proses ini mengelompokkan data hasil dari *preprocessing* untuk selanjutnya didapatkan data baru hasil dari pengolahan menggunakan algoritma K-*means*. Tabel 6 merupakan data hasil pengelompokan final, dimana data sudah tidak ada lagi yang berpindah kelompok. Artinya, data tersebut merupakan data hasil dari pengolahan menggunakan algoritma K-*means*. Data tersebut selanjutnya akan diolah menggunakan algoritma K-NN untuk menentukan akurasi data, tingkat *error*, *Sensitifity*/*Recall*/TPR, *Specificity,* AUC (*Area Under the Curve*).

Tabel 6 *Sampel data hasil pengelompokan (final)*

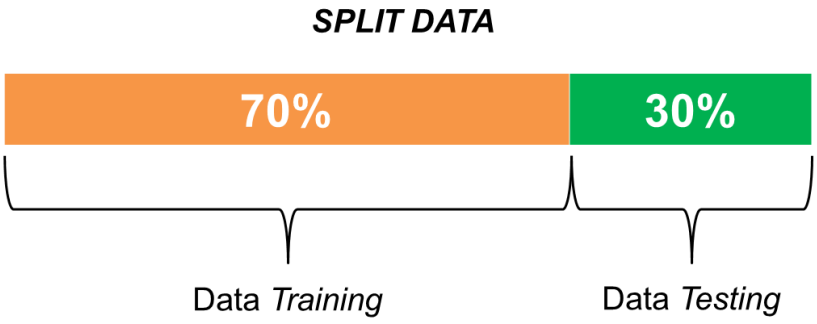
Id	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	*Cluster* ke
1	32	1	38.5	52.5	7.7	22.1	7.5	6.93	3.23	106	12.1	69	*Cluster* 1
2	32	1	38.5	70.3	18	24.7	3.9	11.17	4.8	74	15.6	76.5	*Cluster* 1
3	32	1	46,90	74,70	36,20	52,60	6,10	8,84	5,20	86,00	33,20	79,30	*Cluster* 1
4	32	1	43,20	52,00	30,60	22,60	18,90	7,33	4,74	80,00	33,80	75,70	*Cluster* 1
5	32	1	39,20	74,10	32,60	24,80	9,60	9,15	4,32	76,00	29,90	68,70	*Cluster* 1

...	...	...	...	...	...	...	...	...	...	...	...	...	...
301	65	1	44,70	99,40	31,90	30,50	12,20	7,15	6,31	82,00	38,50	75,70	*Cluster* 1
302	65	1	39,10	45,80	23,10	27,50	6,40	7,00	6,23	73,00	27,10	64,30	*Cluster* 1
303	65	1	43,60	104,00	32,30	34,20	7,70	8,23	4,69	89,00	20,80	75,50	*Cluster* 1
304	66	1	48,40	76,00	31,90	29,60	13,80	8,81	4,17	111,00	26,20	76,70	*Cluster* 1
305	66	1	40,60	79,60	27,00	28,00	10,10	10,88	5,48	76,00	29,80	71,80	*Cluster* 1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
611	62	0	32	416,6	5,9	110,3	50	5,57	6,30	55,7	650,9	68,5	*Cluster* 4
Id	**Age**	**Sex**	**ALB**	**ALP**	**ALT**	**AST**	**BIL**	**CHE**	**CHOL**	**CREA**	**GGT**	**PROT**	***Cluster* ke**
612	64	0	24	102,8	2,9	44,4	20	1,54	3,02	63	35,9	71,3	*Cluster* 1
613	64	0	29	87,3	3,5	99	48	1,66	3,63	66,7	64,2	82	*Cluster* 5
614	46	0	33	68,28	39	62	20	3,56	4,20	52	50	71	*Cluster* 1
615	59	0	36	68,28	100	80	12	9,07	5,30	67	34	68	*Cluster* 5

Cluster 1 dikategorikan sebagai *blood donors*, *cluster* 2 dikategorikan sebagai *suspect blood donor*, *cluster* 3 dikategorikan sebagai Hepatitis C, *cluster* 4 dikategorikan sebagai *Fibrosis*, dan *cluster* 5 dikategorikan sebagai *Cirrhosis*.

D. Split data

Split data membagi data menjadi dua bagian yaitu data *training* dan data *testing*. Rasio dalam penentuan data *training* dan *testing* dalam penelitian ini menggunakan rasio 70:30 % karena dianggap sebagai rasio paling baik untuk pelatihan (*training*) dan validasi model.



Gambar 10 Ilustrasi *split data* (*training* dan *testing*)

Split data dalam peneitian ini bertujuan untuk menyediakan kinerja yang lebih ringan pada sistem karena sistem tidak perlu mengolah keseluruhan data, akan tetapi sistem mengolah data sampling (dalam hal ini data *testing*). Sampel data *testing* dalam penelitian ini dapat dilihat pada tabel 7.

Tabel 7 *Sampel data testing*

Id	Age	Sex	ALB	ALP	ALT	AST	BIL	CHE	CHOL	CREA	GGT	PROT	*Category*
1	32	1	38,50	52,50	7,70	22,10	7,50	6,93	3,23	106,00	12,10	69,00	*Blood donor*
9	32	1	50,90	65,50	23,20	21,20	6,90	8,69	4,10	83,00	13,70	71,30	*Blood donor*
13	33	1	36,30	78,60	23,60	22,00	7,00	8,56	5,38	78,00	19,40	68,70	*Blood donor*
15	33	1	38,70	39,80	22,50	23,00	4,10	4,63	4,97	63,00	15,20	71,90	*Blood donor*
18	33	1	45,20	88,30	32,40	31,20	10,10	9,78	5,51	102,00	48,50	76,50	*Blood donor*
...	...	...	...	...	...	...	...	...	...	...	...	...	...
234	53	1	41,70	45,30	23,20	25,10	10,80	5,68	5,78	119,00	114,90	67,90	Cirrhosis
235	53	1	38,10	82,50	8,00	17,50	2,40	9,13	6,28	103,00	35,80	69,90	*Blood donor*
236	53	1	49,20	71,80	42,80	29,40	6,80	15,10	6,24	107,00	48,30	77,80	*Blood donor*
237	54	1	46,90	74,70	36,20	52,60	6,10	8,84	5,20	86,00	33,20	79,30	*Blood donor*
243	55	1	44,10	60,00	26,30	25,90	5,10	7,23	7,30	88,00	41,60	77,70	*Blood donor*
...	...	...	...	...	...	...	...	...	...	...	...	...	...
610	59	0	39,00	51,30	19,60	285,80	40,00	5,77	4,51	136,00	101,10	70,50	Cirrhosis
611	62	0	32,00	416,60	5,90	110,30	50,00	5,57	6,30	56,00	650,90	68,50	Fibrosis
612	64	0	24,00	102,80	2,90	44,40	20,00	1,54	3,02	63,00	35,90	71,30	*Blood donor*

614	46	0	33,00	68,28	39,00	62,00	20,00	3,56	4,20	52,00	50,00	71,00	*Blood donor*
615	59	0	36,00	68,28	100,00	80,00	12,00	9,07	5,30	67,00	34,00	68,00	Cirrhosis

E. Klasifikasi menggunakan algoritma K-NN

Tahap klasifikasi menggunakan algoritma K-NN dalam penelitian ini merupakan tahap terakhir dan tahap utama dalam mencari akurasi data, *error*, *sensitifity*/*recall*/TPR, *spesificity*, dan AUC (*Area Under the Curve*). Dalam proses perhitungan K-NN perlu penghitung jarak terdekat dari sebuah data. Dalam penelitian ini dalam menentukan jarak terdekat menggunakan rumus *euclidean distance*.

Penelitian ini membandingkan penggunaan nilai k yaitu 3,5,7,9 untuk mencari hasil klasifikasi terbaik. Hasil terbaik akan dipilih dan digunakan sebagai *output* dari penelitian ini. Penelitian ini menggunakan nilai k ganjil guna menghindari jumlah jarak yang sama dalam proses klasifikasi sehingga hasil klasifikasi menjadi samar dalam penentuan prediksi tiap data. Hasil dari klasifikasi menggunakan algoritma K-NN disajikan pada tabel 8.

Tabel 8 *Perbandingan hasil pengujian klasifikasi k=3,5,7,9*

Nilai *k*	Akurasi	*Error*	*Sensitifity*/*Recall*/TPR	*Spesificity*	AUC
3	0,9891	0,0109	1,00	0,9048	0,95
5	0,9783	0,0217	1,00	0,8095	0,90
7	0,9783	0,0217	1,00	0,8095	0,90
9	0,9728	0,0272	1,00	0,7619	0,88

Hasil klasifikasi terbaik pada penelitian ini menggunakan nilai *k*=3 yang menghasilkan akurasi sebesar 0,9891 atau 98,91% dengan tingkat *error* sebesar 0,0109 atau 1,09%. Penggunaan nilai *k*=3 juga menghasilkan *sensitifity*/*recall*/TPR sebesar 1,00 atau 100% dengan nilai *spesificity* sebesar 0,9048 atau 90,48%. Pada penggunaan nilai *k*=3 mendapatkan nilai AUC sebesar 0,95 atau 95%, dimana jika mengacu pada tabel 9 mengenai kriteria dari nilai AUC maka penggunaan *k*=3 tergolong dalam *excellent classification*.

Tabel 9 *Kriteria dari AUC*

Nilai AUC	Kriteria
0.90 - 1.00	*Excellent classification*
0.80 - 0.90	*Good classification*
0.70 - 0.80	*Fair classification*
0.60 - 0.70	*Poor classification*
0.50 - 0.60	*Failure*

Pada penelitian ini mengalami peningkatan akurasi sebesar 7,26% dari penelitian Firdaus, dkk (2020), peningkatan sebesar 5,65% dari penelitian Lisnawanty, dkk (2021), serta peningkatan sebesar 0,68% dari penelitian Mostafa & Hasan (2021). Peningkatan lain juga terdapat pada nilai *spesificity* dimana pada penelitian Mostafa & Hasan (2021) nilai *spesificity* tertinggi sebesar 85,51% yang artinya penelitian ini mengalami kenaikan sebesar 4,97%. Sedangkan pada penelitian Firdaus, dkk (2020) dan Lisnawanty, dkk (2021) tidak menyertakan nilai dari *spesificity*.

KESIMPULAN

Berdasarkan penelitian yang telah dilakukan mengenai klasifikasi kelayakan darah yang dapat didonorkan menggunakan K-NN dengan optimasi K-*means*, maka kesimpulan yang dapat diambil meliputi:

1. Pada penelitian sebelumnya tentang klasifikasi kelayakan darah yang dapat didonorkan oleh Firdaus, dkk (2020) menggunakan *Neural Networks*, selanjutnya oleh Lisnawanty, dkk (2021) menggunakan komparasi 2 metode (C4.5 dan *Naive Bayes*), dan oleh Mostafa & Hasan (2021) dengan komparasi 3 metode klasifikasi (*Artificial Neural Network, Random Forest, Support Vector Machine*). Pada penelitian ini berhasil meningkatkan (*improve*) algoritma K-NN menggunakan algoritma K-*means* untuk melakukan klasifikasi kelayakan darah yang dapat didonorkan dengan hasil yang sangat baik.
2. Metode yang digunakan dalam penelitian ini berhasil mencapai target yang ingin dicapai yaitu peningkatan akurasi dari penelitian sebelumnya mengenai klasifikasi kelayakan darah yang dapat didonorkan.
3. Metode yang diusulkan (dalam hal ini K-NN dan K-*means*) berhasil mendapatkan akurasi tertinggi sebesar 98,91% dengan tingkat error sebesar 01,09%. Metode yang diusulkan berhasil mencapai peningkatan akurasi dari penelitian sebelumnya oleh Firdaus, dkk (2020), penelitian oleh Lisnawanty, dkk (2021), serta dari penelitian Mostafa & Hasan (2021).
4. Metode yang diusulkan juga mengalami kenaikan pada nilai *spesificity* (90,48%) dimana pada penelitian Mostafa & Hasan (2021) nilai *spesificity* tertinggi sebesar 85,51% yang artinya penelitian ini mengalami kenaikan sebesar 4,97%.
5. Tingkat dari *sensitifity* pada penelitian ini mencapai hasil 100%, artinya metode yang digunakan memiliki sensitifitas yang sangat baik.
6. Hasil dari AUC dalam penelitian ini mencapai 95% yang artinya metode yang digunakan untuk melakukan klasifikasi kelayakan darah yang dapat didonorkan tergolong dalam *excellent classification*.

DAFTAR PUSTAKA

[1] K. Fitryadi, “Pengenalan Jenis Golongan Darah Menggunakan Jaringan Syaraf Tiruan Perceptron,” *J. Masy. Inform.*, vol. 7, no. 1, p. 113618, 2017.
[2] N. K. Firani, *Mengenali Sel-Sel Darah dan Kelainan Darah*. Malang: UB Press, 2018.
[3] P. R. Situmorang, W. Y. Sihotang, and L. Novitarum, “Identifikasi Faktor-Faktor yang Mempengaruhi Kelayakan Donor Darah di STIKes Santa Elisabeth Medan Tahun 2019,” *J. Anal. Med. Biosains*, vol. 7, no. 2, p. 122, 2020, doi: 10.32807/jambs.v7i2.195.
[4] A. S. Rigas, O. B. Pedersen, K. Magnussen, C. Erikstrup, and H. Ullum, “Iron deficiency among blood donors: experience from the Danish Blood Donor Study and from the Copenhagen ferritin monitoring scheme,” *Transfus. Med.*, vol. 29, no. S1, pp. 23–27, 2019, doi: 10.1111/tme.12477.
[5] C. R. Lestari and A. A. Saputro, “Gambaran Hasil Pemeriksaan HCV, HIV, dan VDRL Pada Pendonor Unit Donor Darah PMI Kabupaten Kudus,” *Indones. J. Biomed. Sci. Heal.*, vol. 1, no. 1, pp. 11–21, 2021.
[6] T. Ilhami Surya Akbar, S. Rahmayani Siregar, R. Nadia Amris, and A. Penelitian, “Gambaran Hasil Skrining Infeksi Menular Lewat Transfusi Darah (IMLTD) Pendonor di Unit Transfusi Darah (UTD) PMI Kabupaten Aceh Utara Periode 2017-2018,” *Artik. Penelit.*, vol. 70, no. 2, pp. 121–127, 2020.
[7] P. Cacoub and C. Comarmond, “New insights into HCV-related rheumatologic disorders: A review,” *J. Adv. Res.*, vol. 8, no. 2, pp. 89–97, 2017, doi: 10.1016/j.jare.2016.07.005.
[8] S. Zhang, S. Member, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient kNN Classification With Different Numbers of Nearest Neighbors,” pp. 1–12, 2017.
[9] A. F. Pulungan, M. Zarlis, and S. Suwilo, “Analysis of Braycurtis , Canberra and Euclidean Distance in KNN Algorithm,” vol. 4, no. 1, pp. 2017–2020, 2019.
[10] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-elsoud, “Knowledge-Based Systems

A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier," *Knowledge-Based Syst.*, vol. 205, p. 106270, 2020, doi: 10.1016/j.knosys.2020.106270.

[11] W. A. Istiqhfarani, I. Cholissodin, and F. A. Bachtiar, "Klasifikasi Penyakit Dental caries menggunakan Algoritme Modified K- Nearest Neighbor," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 5, pp. 1499–1506, 2020.

[12] S. W. Binabar and Ivandari, "Optimasi Parameter K pada Algoritma KNN untuk Deteksi Penyakit Kanker Payudara," *IC-Tech*, vol. XII, no. 2, pp. 11–18, 2017.

[13] R. J. Kasim, S. Bahri, and S. Amir, "Implementasi Metode K-Means Untuk Clustering Data Penduduk Miskin Dengan Systematic Random Sampling," *Pros. SISFOTEK*, vol. 5, no. 1, pp. 95–101, 2021.

[14] T. Alfina, B. Santosa, and R. Barakbah, "Analisa Perbandingan Metode Hierarchical Clustering , K-means dan Gabungan Keduanya dalam Cluster Data (Studi kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)," vol. 1, 2012.

[15] M. R. Firdaus, A. Latif, and W. Gata, "Klasifikasi Kelayakan Calon Pendonor Darah Menggunakan Neural Network," *Sist. J. Sist. Inf.*, vol. 9, no. 2, p. 362, 2020, doi: 10.32520/stmsi.v9i2.840.

[16] K. Handayani, L. Lisnawanty, A. Latif, M. R. Firdaus, and F. N. Hasan, "Komparasi Algoritma C4.5 Dan Naïve Bayes Dalam Penentuan Status Kelayakan Donor Darah," *Sistemasi*, vol. 10, no. 3, p. 676, 2021, doi: 10.32520/stmsi.v10i3.1440.

[17] F. B. Mostafa and E. Hasan, "Machine Learning Approaches for Inferring Liver Diseases and Detecting Blood Donors from Medical Diagnosis," *arXiv e-prints*, p. 2104.12055, 2021.

[18] W. C. Lin and C. F. Tsai, "*Missing value* imputation: a review and analysis of the literature (2006–2017)," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1487–1509, 2020, doi: 10.1007/s10462-019-09709-4.

[19] E. Acuña and C. Rodriguez, "The Treatment of *Missing value*s and its Effect on Classifier Accuracy," *Classif. Clust. Data Min. Appl.*, no. 1995, pp. 639–647, 2004, doi: 10.1007/978-3-642-17103-1_60.

[20] Isy Karima Fauzia, Budi Arif Dermawan, and Tesa Nur Padilah, "Penerapan K-Means Clustering pada Penyakit Infeksi Saluran Pernapasan Akut (ISPA) di Kabupaten Karawang," *J. Sist. dan Inform.*, vol. 15, no. 1, pp. 81–87, 2020, doi: 10.30864/jsi.v15i1.350.

[21] E. Prasetyo, *Data Mining-Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: ANDI, 2012.

[22] T. Suprawoto, "Klasifikasi Data Mahasiswa Menggunakan Metode K-Means Untuk Menunjang Pemilihan Strategi Pemasaran," *JIKO (Jurnal Inform. dan Komputer)*, vol. 1, no. 1, pp. 12–18, 2016, doi: 10.26798/jiko.2016.v1i1.9.

[23] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining*, vol. 53, no. 9. London: Taylor & Francis Group, 2009.

[24] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: an Introduction to Data Mining*. Hoboken, New Jersey: John Wiley & Sons, 2014.

[25] R. Hidayati, A. Zubair, A. Hidayat Pratama, L. Indana, P. Studi Sistem Informasi, and F. Teknologi Informasi, "Silhouette Coefficient Analysis in 6 Measuring Distances of K-Means Clustering," *Techno.Com*, vol. 20, no. 2, pp. 186–197, 2021.

[26] N. M. Putry, "Komparasi Algoritma Knn Dan Naïve Bayes Untuk Klasifikasi Diagnosis Penyakit Diabetes Mellitus," *EVOLUSI J. Sains dan Manaj.*, vol. 10, no. 1, 2022, doi: 10.31294/evolusi.v10i1.12514.

[27] M. C. Wijaya and A. Prijono, *Pengolahan Citra Digital Menggunakan MatLAB Image Processing Toolbox*. Bandung: Informatika, 2007.

[28] W. Budhiarto, *Machine Learning & Computational Intelligence*. Yogyakarta: ANDI, 2016.

[29] S. Zhang, "Cost-sensitive KNN classification," *Neurocomputing*, vol. 391, no. xxxx, pp. 234–242, 2020, doi: 10.1016/j.neucom.2018.11.101.

[30] I. A. Angreni, S. A. Adisasmita, M. I. Ramli, and S. Hamid, "Pengaruh Nilai K Pada Metode K-Nearest Neighbor (Knn) Terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan," *Rekayasa Sipil*, vol. 7, no. 2, p. 63, 2019, doi: 10.22441/jrs.2018.v07.i2.01.

[31] L. Syafa'ah, Z. Zulfatman, I. Pakaya, and M. Lestandy, "Comparison of Machine Learning Classification Methods in Hepatitis C Virus," *J. Online Inform.*, vol. 6, no. 1, p. 73, 2021, doi: 10.15575/join.v6i1.719.