# PENERAPAN ALOGARITMA *K-NEAREST NEIGHBOR* UNTUK KLASIFIKASI PENYAKIT LIVER

Ega Yusni Habibie[1], Hardian Oktavianto[2], Qurrota A'yun[3]
Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember
egayusni001@gmail.com, hardian@unmuhjember.ac.id, qurotta.ayun@unmuhjember.ac.id

**ABSTRAK**

Liver merupakan organ yang sagat penting dalam tubuh manusia, Penyakit liver disebabkan oleh berbagai faktor yang merusak hati. Permasalahan yang terjadi yaitu sulitnya mengenali penyakit liver sejak dini, bahkan saat penyakit liver ini sudah menyebar masih sulit untuk diteksi. Tujuan dari penelitian tugas akhir untuk mengetahiu berapa tingkat akurasi, presisi dan recall. Pada diagnosa penyakit liver menggunakan alogaritma *K-Nearst Neighbor*. Alogaritma *K-Nearest Neighbor* merupakan sebuah metode untuk melakukan klasifikasi terhadap objek yang berdasarkan dari data pembelajaran yang jarakanya paling dekat dengan objek tersebut. Pada penelitian ini diterapkan metode untuk klasifikasi penyakit liver menggunakan alogaritma *K-Nearest Neighbor* karena dianggap cukup fleksibel. Data yang di gunakan adalah data Indian Liver Patients Dataset (ILDP) yang diambil dari website *UCI Machine Learning* . Tujuan dari penlitian ini untuk mengatahui berapa tingkat akurasi, presisi dan recall pada diagnosa penyakit liver menggunakan K-Nearest Neighbor. Hasil yang didapatkan dari pengujian alogaritma *K-Nearst Neighbor* dengan nilai akurasi tertinggi sebesar 67,20% , presisi tertinggi sebesar 59,36%. Jadi, Alogaritma *K-Nearest Neighbor,* cukup akurat di atas 50% untuk mengklasifikasi data pasien penyakit liver

**Kata Kunci** : Klasifikasi, Penyakit Liver, *K-Nearest Neighbor*

***ABSTRACT***

*The liver is a cery important organ in the human body. Liver disease is caused by various factors that damage the liver. The problem that occurs us the difficulty in recognizing liver sisease from an early age, even when the liver disease has spread, it is still difficult to detect. The purpose of this final project research is to determine the level of accuracy, precision and recall. in diagnosing liver disease using the K-Nearst Neighbor algorithm. The K-Nearst Neighbor algorithm is a method for classifying objects based on learning data that is closest to the object.In this study, a method for classification of liver disease was applied using the K-Nears Neighbor algorithm because it was considered quite flexible. The data used is the Indian Liver Patients Dataset (ILDP) which is taken from the UCI Machine Learning website. The purpose of this study was to determine the level of accuracy, precision and recall in diagnosing liver disease using the K-Nearst Neighbor algorithm. The results obtained from testing the K-Nearst Neighbor algorithm with the highest accuracy value of 67.20%, the highest precision of 59.36%, and the highest recall of 58.57%. So, the K-Nearest Neighbor Algorithm, is quite accurate above 50% for classifying liver disease patient data.*

*Keywords : Classification, Liver Disease, K-Nearest Neighbor*

## 1. PENDAHULUAN

Masalah yang ditimbulkan oleh penyakit liver adalah susah mengenali penyakit liver sejak dini, bahkan ketika penyakit tersebut sudah menyebar. Diagnosa penyakit liver yang lebih awal dapat meningkatkan tindak kelangsungan hidup pasien. Dengan perkembangan teknologi saat ini diagnosa penyakit dapat menggunakan metode data mining. Salah satu pengembangan dari data mining adalah klasifiakasi. Metode klasifikasi dapat melakukan pembelajaran dengan memetakan suatu item dan ke dalam kelas berdasadarkan kealas data yang telah didefinisikan sebelumnya (Agarwal, 2014).

Metode yang digunakan untuk membangun model klasifikasi dalam penyakit liver yaitu K-Nearst Neighbor (KNN). Karena pada penelitian ini sebelumnya yang dilakukan oleh (Prahudaya & Harjoko, 2017) dengan judul Metode Klasifikasi Mutu Jambu Biji menggunakna KNN berdasarkan fitur dan tekstur dari penelitian tersebut dapat disimpulkan bahwa n nilai akurasi tertinggi yaitu 91.25%. Pada penelitian selanjutnya yang dilakukan oleh (Reza Noviansyah et al., 2018)

## 2. PENELITIAN TERKAIT

### A. Penyakit Liver

Penyakit liver merupakan peradangan pada hati yang disebabkan oleh bakteri, virus atau bahan – bahan beracun sehingga membuat hati tidak dapat melakukan fungsinya dengan baik (Musyaffa & Rifai, 2018).

### B. *Data Mining*

*Data mining* merupakan metode yang di gunakan dalam pengolahan data berskala besar oleh karena itu *Data Mining* memili peranan yang sangat pentingdalam berbagai aspek bidang kehidupaan yaitu dalam bidang industri, bidang cuaca,bidang keuangan, ilmu dan teknologi. Dalam *Data Mining* terdapat metodeyang dapat di gunakan seperti metode klasfikasi, clustering, regresi, seleksi variable, dan market basket analisis (Nielza Atthina, 2014).

### C. *Klasifikasi*

Menurut (Prasetyo, 2013)ditulis dalam bukunya yang berjudul "Data Mining Konsep dani Aplikas Menggunakan Matlab" menjelaskan bahwa : "Klasifikasi dapat didefinisikan sebagai pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi targetif yang memetakan setiap set atribut (fitur) x kesatu dari sejumlah label kelas yang tersedia" (Nofriansyah, 2014).

### D. *K-Nearest Neighbor* (KNN)

Nilai k atau jumlah tetangga terdekat pada metode ini bergantung pada data yang digunakan. Nilai k yang tinggi akan mengurangi noise pada klasifikasi, namun akan membuat batasan antara setiap klasifikasi menjadi semakin kabur. Fungsi jarak yang umumnya digunakan adalah jarak Euclidean dengan menggunakan rumus sebagai berikut :

$$d\,(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

keterangan :

$x = x_1, x_2, ..., x_m$ adalah instance data uji

$y = y_1, y_2, ..., y_m$ adalah instance data latih

$x_i - y_i$ = kuadrat selisih data uji dan data lati

## F. *K-Fold Cross Validation*

Pengujian menggunakan k-fold cross validation. Cross validation adalah bentuk sederhana dari teknik statistik. Jumlah fold standar untuk memprediksi tingkat error dari data adalah dengan menggunakan 10-fold cross validation (Diamant & Witten, 2011).

## G. Confusion Matrix

Confusion matrix memberikan keputusan yang diperoleh dalam traning dan testing, confusion matrix memberikan penilaian performance klasifikasi berdasarkan objek dengan benar atau salah (Gorunescu, 2011).

## H. Rapid Miner Studio

Rapid Miner ditulis dengan munggunakan bahasa java sehingga dapat bekerja di semua sistem operasi. Rapid Miner sebelumnya bernama YALE (Yet Another Learning Environment), dimana versi awalnyai mulai dikembangkan pada tahun 2001 oleh RalfKl nkenberg, Ingo Mierswa,dan Simon Fischer di Artificial Intelligence Unit dari University of Dortmund. Rapid Miner didistribusikan di bawah lisensi AGPL (GNU Affero General Public License) versi 3. Hingga saat initelah ribuan aplikasi yang dikembangkan mengunakan Rapid.

## I. Tahapan Penelitian


Gambar 1. Diagram Alur Penelitian

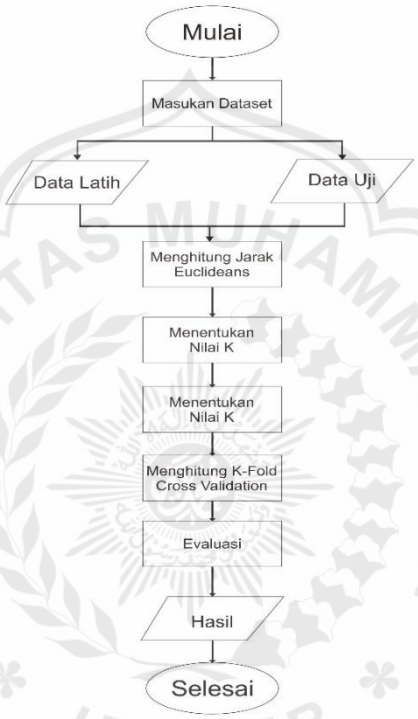
## J. Studi Literatul

Studi literatur adalah serangkaian kegiatan yang berkenaan dengan metode pengumpulan data, membaca, mencatat, dan mencari referensi teori yang relevan dengan Kasus atau permasalahan permasalahan yang ditemukan serta mengelolah bahan penelitian. 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 yang akan memperkuat argumentasi-argumentasi yang ada. Maka penulis perlu mempelajari beberapa literatur yang digunakan, kemudian literatur tersebut diseleksi untuk dapat ditentukan literatur mana yang akan digunakan dalam penelitian.

## K. Pengumpulan Data

Pada pengumpulan data diperoleh dari Website *UCI Machine Learning Repository* (https://archive.ics.uci.edu/ml/index.php) yaitu berupa dataset ILPD (Indian Liver Patient Datatse) tahun 2012. Dataset tersebut diolah menggunakan algoritma *K-Nearest Neighbor*. Dan dataset ini terdapat beberapa atribut, yaitu Age, Gender, Total pada Bilirubin, Direct Bilirubin, ALK (Alkaline Phosphotase) ,SGPT (Alamine Aminoteransferase), SGOT (Aspartate Aminotransferase), TP (Total Protein), ALB (Albumin), A/G (Ratio Albumin and Globulin Ratio), Selector (Class). Dengan jumlah data sebanyak 583.

## L. Implementasi Modifed K-Nearst Neighbor




Gambar 2. Flowchart KNN

# 3. HASIL DAN PEMBAHASAN

Data yang digunakan berasaldari situs website UCI Machine Learning. Data yang diambil dari situs tersebut adalah ILPD (Indian Liver Patient Dataset) tahun 2012. Pengumpulan data didapatkan sebanyak 583 data dan 11 atribut.

## A. *Pre-* Processing Data

### 1. Skenario dan Hasil Pengujian

Pre-processing pada data ILPD (Indian Liver Patient Dataset) yaitu menormalisasi data. Normalisasi ini bertujuan untuk terjadinya penyebaran data supaya nilai dari masing-masing atribut atau variable tidak terlalu jauh nilai data pada setiap atribut/variable akan di ubah pada rentang 0-1. Hasil normalisasi data ditunjukan pada tabel 4.2

$$X^* = \frac{X - min(X)}{\max(X) - \min(X)}$$

Keterangan :
Normalisasi untuk atribut/variable Total Bilirubin (TB) :
Nilai x untuk data yang pertama = 0,7
Nilai min (x) untuk data yang terendah = 0,4
Nilai min (x) untuk data yang tertinggi = 75
$x = \frac{0,7-0,4}{75-0,4}$ = 0,004021

| PA (AGE) | GENDER (PG) | TB | DB | ALK | SGPT | SGOT | TP | ALB | A/G | CLASS |
|---|---|---|---|---|---|---|---|---|---|---|
| 0,709 | 1 | 0,004 | 0 | 0,060 | 0,003 | 0,002 | 0,594 | 0,522 | 0,24 | 1 |
| 0,674 | 0 | 0,141 | 0,276 | 0,312 | 0,027 | 0,018 | 0,696 | 0,5 | 0,176 | 1 |
| 0,674 | 0 | 0,092 | 0,204 | 0,208 | 0,025 | 0,012 | 0,623 | 0,522 | 0,236 | 1 |
| 0,628 | 0 | 0,008 | 0,015 | 0,058 | 0,002 | 0,002 | 0,594 | 0,543 | 0,28 | 1 |
| 0,791 | 0 | 0,047 | 0,097 | 0,064 | 0,008 | 0,010 | 0,667 | 0,326 | 0,04 | 1 |
| 0,488 | 0 | 0,019 | 0,031 | 0,071 | 0,004 | 0,001 | 0,710 | 0,761 | 0,4 | 1 |
| 0,256 | 1 | 0,007 | 0,005 | 0,044 | 0,003 | 0,001 | 0,623 | 0,565 | 0,28 | 1 |
| 0,291 | 1 | 0,007 | 0,010 | 0,068 | 0,002 | 0,001 | 0,579 | 0,587 | 0,32 | 1 |
| 0,151 | 0 | 0,007 | 0,010 | 0,068 | 0,006 | 0,002 | 0,681 | 0,695 | 0,36 | 2 |
| 0,593 | 0 | 0,004 | 0,005 | 0,111 | 0,022 | 0,010 | 0,594 | 0,543 | 0,28 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 0,395 | 0 | 0,008 | 0,010 | 0,075 | 0,005 | 0,003 | 0,667 | 0,761 | 0,48 | 2 |

Tabel 1. Normalisasi Data

## 2. Hasil Klasifikasi

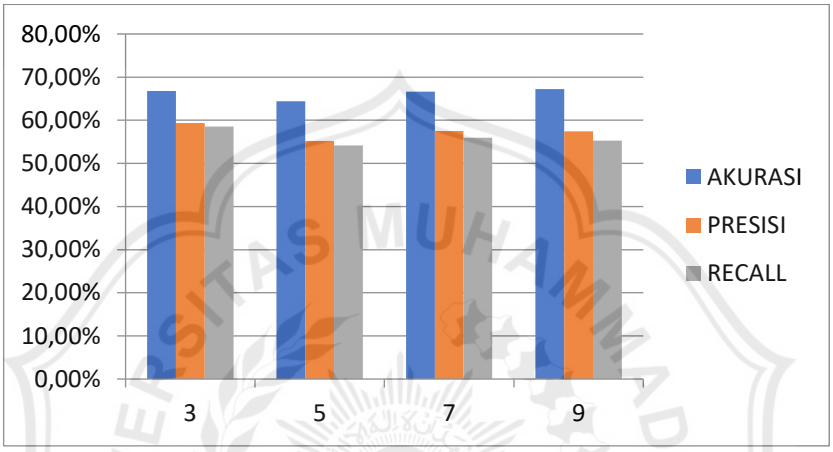
Pada proses klasifikasi dapat dilakukan yaitu dengan menggunakan algoritma *K-Nearest Neighbor* (KNN). Data ILPD (Indian Liver Patient Dataset) sebelumnya telah melalui pre-processing maka selanjutnya adalah proses *K-Nearest Neighbor* (KNN). Klasifikasi ini mengetahui nilai akurasi, presisi, dan recall.

| | | ***Predicted Class*** | |
|---|---|---|---|
| | | Liver | Non Liver |
| ***Actual Class*** | Liver | 277 | 73 |
| | Non Liver | 93 | 57 |

*Tabel 2. Confuxion Matrix pada Algorritma KNN menggunakan* Kfold *2*

| NILAI K | AKURASI | PRESISI | RECALL |
|---|---|---|---|
| 3 | 66,80% | 59,36% | 58,57% |
| 5 | 64,40% | 55,17% | 54,38% |
| 7 | 66,60% | 57,53% | 55,59% |
| 9 | 67,20% | 57,45% | 55,24% |

Tabel 3. Nilai Akurasi, Presisi dan Recall menggunakan Kfold *2*




Gambar 3. Hasil Pengujian pada Kfold *2*

Berdasarkan grafik 4.1, pada algoritma KNN didapatkan hasil terbesar dengan akurasi sebesar 67,20%, presisi sebesar 59,36%, dan *recall* sebesar 58,57%.

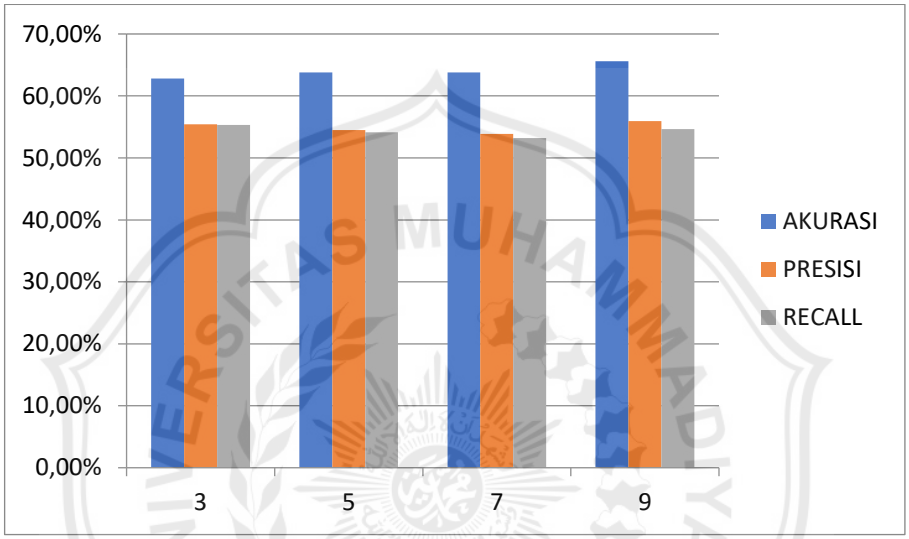
### 3. Pengujian Kfold 4

| | | ***Predicted Class*** | |
|---|---|---|---|
| | | Liver | Non Liver |
| ***Actual Class*** | Liver | 259 | 91 |
| | Non Liver | 95 | 55 |

Tabel 4. Confuxion Matrix pada Algoritma KNN menggunakan Kfold 4

| NILAI K | AKURASI | PRESISI | RECALL |
|---|---|---|---|
| 3 | 62,80% | 55,42% | 55,34% |
| 5 | 63,80% | 54,5% | 54,15% |
| 7 | 63,80% | 53,88% | 53,19% |
| 9 | 65,60% | 55,95% | 54,67% |

Tabel 5. Nilai Akurasi, Presisi dan Recall pada Kfold 4




Gambar 4. Hasil Pengujian pada Kfold 4

Berdasarkan grafik 4.2, pada algoritma KNN didapatkan hasil terbesar dengan akurasi sebesar 65,60%, presisi sebesar 55,95%, dan *recall* sebesar 55,34%.

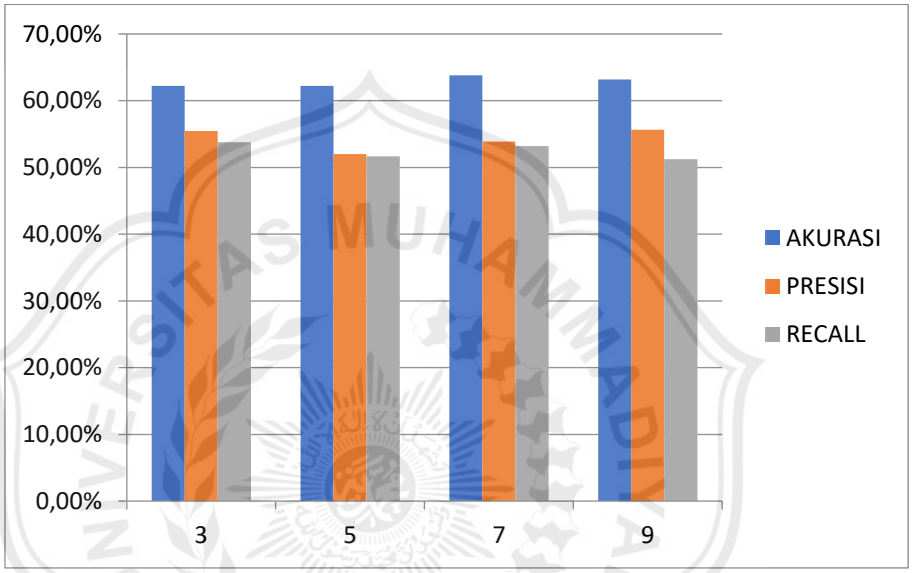
## 4. Pengujian Kfold 5

| | | ***Predicted Class*** | |
|---|---|---|---|
| | | Liver | Non Liver |
| ***Actual Class*** | Liver | 262 | 88 |
| | Non Liver | 101 | 49 |

Tabel 6. Confuxion Matrix pada Algoritma KNN menggunakan Kfold 5

| NILAI K | AKURASI | PRESISI | RECALL |
|---|---|---|---|
| 3 | 62,20% | 55,48% | 53,77% |
| 5 | 62,20% | 51,98% | 51,67% |
| 7 | 63,80% | 53,88% | 53,19% |
| 9 | 63,20% | 55,65% | 51,24% |

Tabel 7. Nilai Akurasi, Presisi dan Recall pada Kfold 5




Gambar 5. Hasil Pengujian pada Kfold 5

Berdasarkan grafik 4.3, pada algoritma KNN didapatkan hasil terbesar dengan akurasi sebesar 63,80%, presisi sebesar 55,48%, dan *recall* sebesar 53,77%.

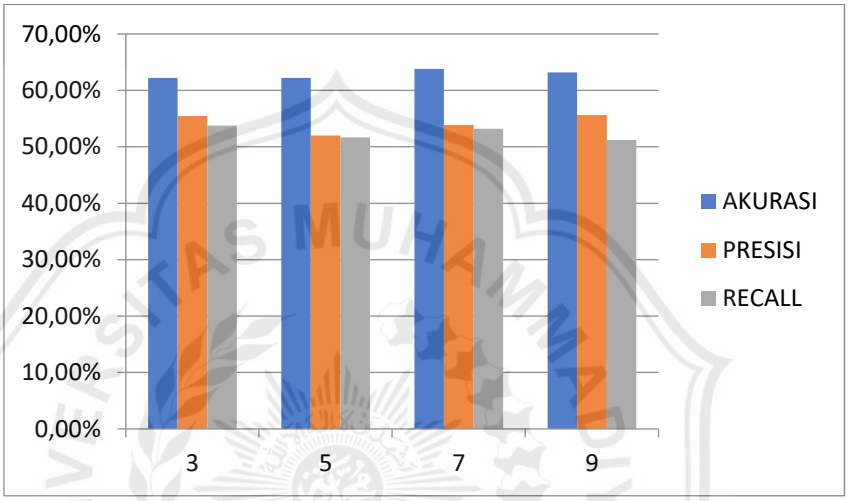
## 5. Pengujian Kfold 10

| | | ***Predicted Class*** | |
|---|---|---|---|
| | | Liver | Non Liver |
| ***Actual Class*** | Liver | 266 | 84 |
| | Non Liver | 100 | 50 |

Tabel 8. Confuxion Matrix pada Algoritma KNN menggunakan K-fold 10

| NILAI K | AKURASI | PRESISI | RECALL |
|---|---|---|---|
| 3 | 63,20% | 55% | 54,67% |
| 5 | 64,20% | 56,79% | 54,62% |
| 7 | 64,80% | 54,66% | 53,81% |
| 9 | 62,20% | 49,94% | 49,95% |

Tabel 9. Nilai Akurasi, Presisi dan Recall pada Kfold 10




Gambar 6. Hasil Pengujian pada Kfold 10

Berdasarkan grafik 4.4,pada algoritma KNN didapatkan hasil terbesar dengan akurasi sebesar 63,20%, presisi sebesar 56,79%, dan *recall* sebesar 54,67%.

## 3. KESIMPULAN DAN SARAN

### 3.1 Kesimpulan

Berdasar kanpenelitian yang telah dilakukan,dapat diambil kesimpulan sebagai berikut:

1. Hasil akurasi paling tinggi dalam klasifikasi Penyakit Liver didapatkan hasil sebesar 67,20% pada Kfold 2 dengan nilai K = 9.
2. Hasil presisi paling tinggi dalam klasifikasi Penyakit Liver didapatkan hasil sebesar 59,36% pada Kfold 2 dengan nilai K = 3.
3. Hasil recall paling tinggi dalam klasifikasi Penyakit Liver didapatkan hasil sebesar 58,57% pada Kfold 2 dengan nilai K = 3.

Jadi, pada Algoritma *K-Nearest Neighbor* nilai akurasi, presisi, dan recall ada pada *k-fold 2* dan untuk nilai $k$ ada pada $k = 3$ dan $k = 9$. Dan Algoritma K-Nearest Neighbor, cukup akurat diatas 50% untuk mengklasifikasi data pasien penyakit liver.

### 3.2 Saran

Berdasarkan pada penelitian yang sudah dilakukan, adapun beberapa saran yang bisa dikembangkan pada penelitian berikutnya adalah seperti bertikut :

1. Untuk Penelitian selanjutnya bisa menggunakan algoritma yang lebih baru contohnya SVM dan Random Forest.
2. Untuk Penelitian berikutnya dapat menggunakan nilai K dan *Kfold* lebih beragam.

## 4. DAFTAR PUSTAKA


Agarwal, S. (2014). Data mining: Data mining concepts and techniques. *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Diamant, H., & Witten, T. A. (2011). Compression induced folding of a sheet: An integrable system. *Physical Review Letters*. https://doi.org/10.1103/PhysRevLett.107.164302

Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques*.

Prahudaya,T., & Harjoko A. (2017). Metode Klasifikasi Mutu Jambu Biji menggunakan KNN berdasarkan Fitur Warna dan Tekstur. Ju*rnal Tekno Sains.*

Musyaffa, N., & Rifai, B. (2018). Model Support Vector Machine Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Liver. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*. https://doi.org/https://doi.org/10.33480/jitk.v3i2

Nielza Atthina, D. (2014). Klasterisasi Data Kesehatan Penduduk untuk Menentukan Rentang Derajat Kesehatan Daerah dengan Metode K-Means. *Seminar Nasional Aplikasi Teknologi Infromasi (SNATI)*, *1*(Klustering), B-52-B-59.

Nofriansyah, D. (2014). *Konsep Data Mining Vs Sistem Pendukung Keputusan*. Deepublish.

Reza Noviansyah, M., Rismawan, T., & Marisa Midyanti, D. (2018). Penerapan Data Mining Menggunakan Metode K-Nearest Neighbor Untuk Klasifikasi Indeks Cuaca Kebakaran Berdasarkan Data Aws (Automatic Weather Station) (Studi Kasus: Kabupaten Kubu Raya). *Jurnal Coding, Sistem Komputer Untan*.