e-ISBN 978-623-93343-3-8
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 April 2021

# Klasifikasi Penyakit Liver dengan Menggunakan Metode Decision Tree

Endah Patimah[1], Ballya Vicky Haekal[2], Desta Sandya Prasvita, S.Komp., M.Kom.[3]
[1,2,3]Program Studi Informatika, Universitas Pembangunan Nasional Veteran Jakarta
[1,2,3]Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Jawa Barat 12450
email: [1]endahp@upnvj.ac.id, [2]vballya12@gmail.com, [3]desta.sandya@gmail.com

**Abstrak.** Penyakit liver merupakan penyakit yang berbahaya. Sehingga, penanganan pasien pada tahap awal sangatlah penting, sehingga kami melakukan klasifikasi dataset liver dengan menggunakan metode *decision tree*. Di mana dataset yang digunakan yaitu ILPD (Indian Liver Patient Dataset). Tahap pertama yang dilakukan adalah penghapusan data yang duplikat, lalu pembagian data, normalisasi, dan tahap terakhir yaitu pemodelan dengan menggunakan decision tree. Untuk pembagian data dengan metode *holdout validation* dengan nilai test size 0.2 dan *k-fold cross validation* dengan nilai test size 0.6, lalu dinormalisasi menggunakan *min-max* atau *standar scaler*. Hasil yang didapatkan dari penelitian ini dengan metode *k-fold cross validation* dan *min-max* hasil akurasinya adalah 0.7, lalu menggunakan *k-fold cross validation* dan *standar scaler* hasil akurasinya sebesar 0.7333, jika menggunakan *holdout validation* dan *min-max* hasil akurasinya sebesar 0.5342, dan jika menggunakan *holdout validation* dan *standar scaler* hasilnya 0.6027. Sehingga dapat disimpulkan bahwa akurasi yang terbesar adalah dengan menggunakan *k-fold cross validation* dan *standar scaler*.

**Kata kunci:** liver, *decision tree*, klasifikasi.

## 1. Pendahuluan

Liver atau hati merupakan satu-satunya organ dalam tubuh yang dapat melakukan regenerasi yaitu kemampuan untuk mengganti sel yang rusak. Liver memiliki beberapa fungsi di antaranya adalah metabolisme lemak, di mana hati menghasilkan empedu dan kolesterol yang nantinya akan mencerna lemak yang ada dalam tubuh. Lalu, hati juga berfungsi sebagai metabolisme protein, di mana hati akan menghasilkan asam amino untuk menyusun protein yang nantinya akan melawan infeksi serta membersihkan ammonia [1].

Dikarenakan fungsi yang beragam itu, tidak menjamin liver terus sehat. Terdapat beberapa hal yang dapat membuat fungsi liver terganggu, diantaranya dari infeksi yang disebabkan oleh parasit atau virus, bisa juga karena kebiasaan mengonsumsi alkohol dalam waktu yang lama. Alkohol sangat berbahaya bagi hati, karena alkohol bersifat toksik untuk sel-sel hati, oleh karena itu ketika terjadi penyaringan terhadap alkohol oleh hati akan membuat sel-sel hati mengalami kematian [2].

Menurut Santosa klasifikasi yaitu peramalan yang di mana outputnya merupakan nilai diskrit, yang tujuannya untuk mendapatkan suatu keputusan yang akurat dari kelas suatu data [3]. Contoh metode klasifikasi diantaranya ada *decision tree, naïve bayes, neural network* dan lain-lain. Dari banyaknya algoritma klasifikasi, algoritma yang akan digunakan pada penelitian ini yaitu *decision tree*.

Seperti penelitian sebelumnya, yaitu penelitian yang dilakukan oleh Popon Handayani, dkk, dengan menggunakan *decision tree* dengan algoritma C4.5 dan *neural network*, data yang digunakan memiliki 11 atribut yang terdiri dari 10 atribut dan 1 class. Dari hasil penilitian ini didapatkan nilai akurasi *decision tree* menggunakan algoritma C4.5 yaitu sebesar 75,56% dengan nilai AUC 0,898 dan nilai akurasi menggunakan algoritma *neural network* sebesar 74,17% dengan AUC 0,671. Dari hasil tersebut dapat disimpulkan bahwa untuk memprediksi penyakit liver lebih akurat dengan menggunakan *decision tree* dari pada menggunakan *neural network* [4].

Oleh karena itu, dalam penelitian kali ini, kita akan membuat klasifikasi pada penyakit liver, data yang akan digunakan adalah berasal dari UCI Machine Learning Repository yaitu dataset ILPD (*Indian Liver Patient Dataset*). Di mana pada dataset terdapat dua kelas, yaitu kelas 1 dan kelas 2 yang terletak pada kolom class.

## 2. Penelitian Terdahulu

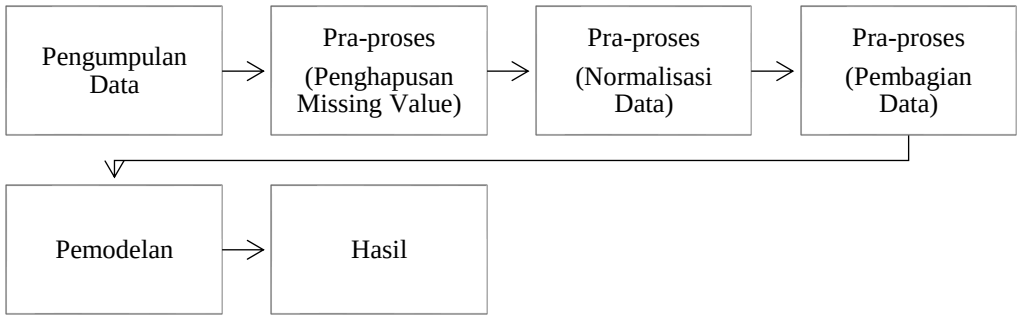
Penelitian yang menggunakan dataset ILPD (*Indian Liver Patient Dataset*) sudah banyak dilakukan, diantaranya adalah sebagai berikut.

1. Popon Handayani, dkk. Melakukan penelitian dengan metode *decision tree* dengan algoritma C4.5 dan neural network, data yang digunakan memiliki 11 atribut yang terdiri dari 10 atribut dan 1 class. Dari hasil penilitian ini didapatkan nilai akurasi *decision tree* menggunakan algoritma C4.5 yaitu sebesar 75,56% dengan nilai AUC 0,898 dan nilai akurasi menggunakan algoritma *neural network* sebesar 74,17% dengan AUC 0,671. Dari hasil tersebut dapat disimpulkan bahwa untuk memprediksi penyakit liver lebih akurat dengan menggunakan *decision tree*[4].
2. Intan Setiawati, dkk. Melakukan penelitian dengan metode *decision tree* dengan melakukan pengolahan dataset ILPD menggunakan bantuan aplikasi yaitu aplikasi Rapidminer versi 7.4. Dari 583 data yang diproses, 433 data digunakan sebagai data *training*, 150 data digunakan sebagai data *testing*. Penelitian ini menunjukkan bahwa hanya 2 atribut (SPGT_AA dan Age) diantara 10 atribut pada dataset ILPD (*Indian Liver Patient Dataset*) yang paling berpengaruh dalam penentuan klasifikasi penyakit liver dan hasil akurasi sebesar 72.67% [5].

## 3. Fokus Pengerjaan

Pada penelitian ini kami menggunakan pemodelan data dengan teknik *decision tree,* dengan menambahkan proses pra-proses terlebih dahulu sebelum melakukan pemodelan. Contoh proses pra-proses yang dilakukan adalah dengan menghapus data missing value, normalisasi dan pembagian data. Normalisasi adalah salah satu pra-proses yang digunakan untuk membuat penskalaan nilai atribut dari data [6]. Normalisasi yang digunakan yaitu *min-max* dan *standar scaler*. *Min- max* adalah normalisasi dengan melakukan transformasi linier terhadap data asli, sedangkan *standar scaler* adalah normalisasi berdasarkan nilai rata-rata dan deviasi standar dari data [7]. Lalu, Pembagian data dilakukan untuk membagi dataset menjadi data training dan data testing. Pada penelitian ini pembagian data menggunakan *k-fold cross validation* dan *holdout validation validation*. Teknik yang digunakan dalam *K-fold cross validation* ini adalah membagi dataset menjadi sebanyak K partisi secara acak, lalu akan dilakukan sebanyak k kali percobaan, di mana masing-masing percobaan menggunakan data partisi ke k sebagai data testing dan sisanya sebagai data training [8]. Lalu, *Holdout validation* adalah pembagian dataset di mana data akan dibagi menjadi data testing dan data training, misalnya 0.2 maka data tersebut 20% digunakan untuk data testing dan sisanya untuk data training [9]. Untuk data yang digunakan sama dengan penelitian-penelitian sebelumnya, yaitu data ILPD (*Indian Liver Patient Dataset*). *Class* yang akan digunakan yaitu kolom 'class' yang terdiri dari dua *class*, yaitu *class* 1 dan *class* 2.

## 4. Metode Penelitian




**Gambar 1.** Metode Penelitian

1. Pengumpulan Data
   Data diperoleh dari website UCI yaitu dataset ILPD (*Indian Liver Patient Dataset*).
2. Penghapusan Missing Value
   Setelah data diperoleh, data akan diolah dengan menghapus missing value yang terdapat pada data.
3. Normalisasi Data
   Setelah melakukan penghapusan missing value, lalu langkah selanjutnya adalah melakukan normalisasi dengan menggunakan *standar scaler* atau *min-max*.
4. Pembagian Data
   Setelah melakukan normalisasi data, maka langkah selanjutnya adalah pembagian data dengan menggunakan *k-fold cross validation* atau *holdout*.
5. Pemodelan
   Setelah pembagian data, selanjutnya adalah melakukan pemodelan dengan menggunakan *decision tree*.
6. Hasil

Setelah melakukan pemodelan data, kita akan mengetahui akurasi dari data.

## 5. Hasil dan Pembahasan

### 5.1. Dataset

Penelitian dilakukan dengan menggunakan dataset ILPD (*Indian Liver Patient Dataset*) yang terdiri dari 583 record, dataset ini tidak memiliki missing value namun memiliki duplikat sebanyak 221 record [10]. Dataset ini memiliki 11 atribut, berikut aalah atribut-atribut yang dirincikan pada tabel di bawah ini.

**Tabel 1. Variabel dataset**

| No | Variabel | Keterangan |
|---|---|---|
| 1 | Age | Umur dari pasien. |
| 2 | Gender | Jenis kelamin dari pasien. |
| 3 | TB | Total Bilirubin. |
| 4 | DB | Direct Bilirubin. |
| 5 | Alkphos | Alkaline Phosphotase. |
| 6 | SGPT | Alamine Aminotransferase. |
| 7 | SGOT | Aspartate Aminotransferase |
| 8 | TP | Total Protiens |
| 9 | ALB | Albumin |
| 10 | A/G Ratio | Ratio Albumin dan Globulin Ratio |
| 11 | Class | Terdapat dua kelas, yaitu kelas 1 dan kelas 2 |

### 5.2. Metodologi yang digunakan

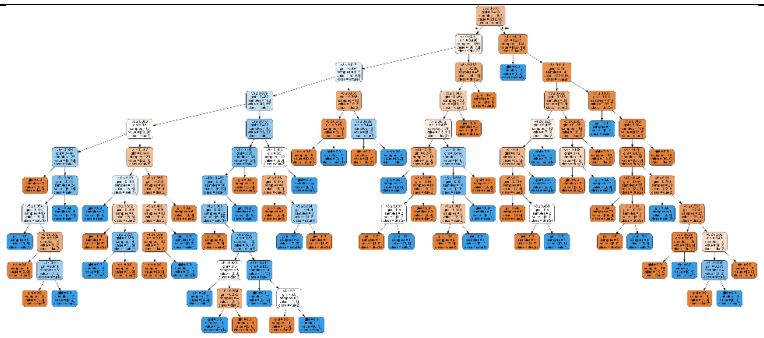
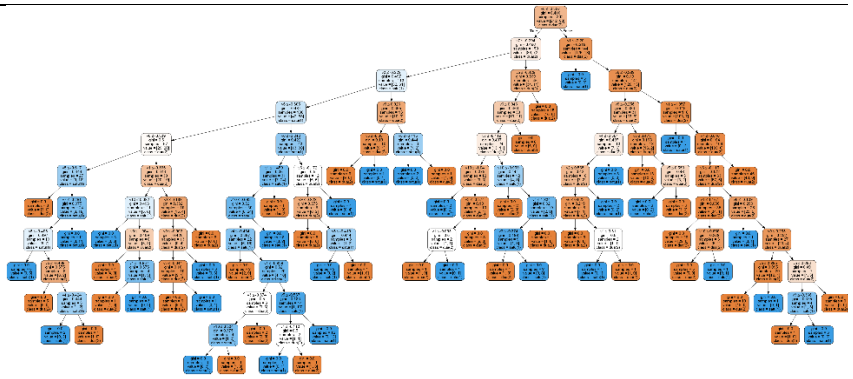
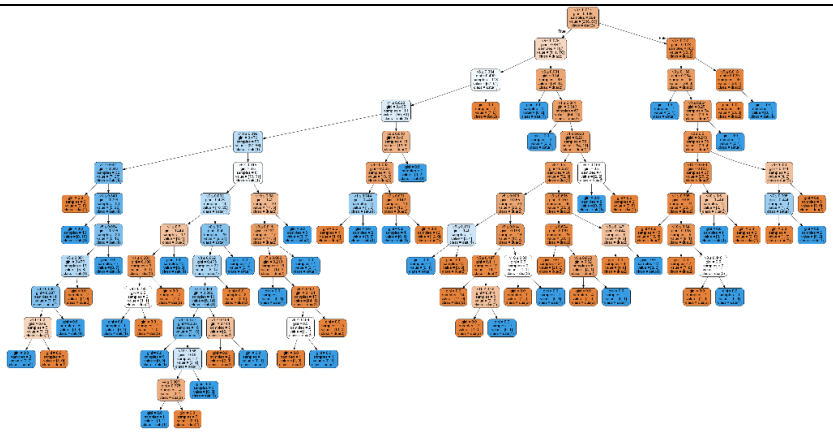
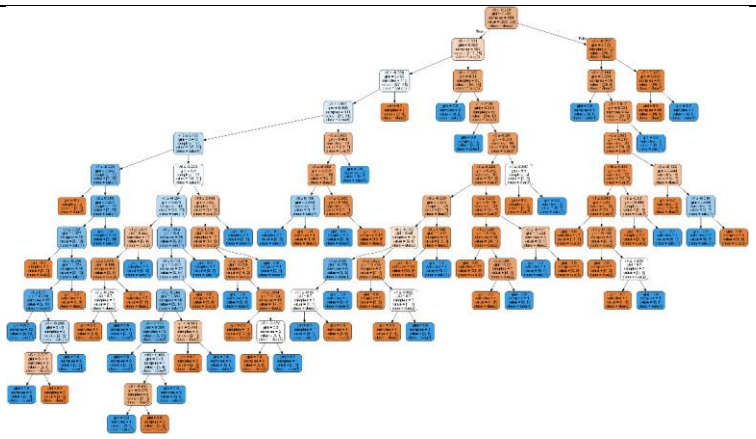
Dalam penelitian ini hal yang pertama kali dilakukan adalah membaca dataset liver, kemudian dataset tersebut dievaluasi apakah terdapat missing value atau data yang duplikat pada dataset. Setelah melakukan evaluasi pada dataset langkah selanjutnya adalah melakukan normalisasi data dengan menggunakan *min-max* atau *standar scaler*. Lalu lakukan pembagian data dengan menggunakan *k-fold cross validation* di mana nilai untuk k adalah 6 atau *holdout validation* dengan nilai test-sizenya 0.2, lalu langkah terakhir adalah melakukan pemodelan data dengan menggunakan *decision tree*.

### 5.3. Hasil Penelitian

Setelah melakukan proses pengolahan data seperti yang telah dijelaskan pada metodologi penelitian, yaitu dengan melakukan penghapusan data yang duplikat, lalu selanjutnya melakukan normalisasi dengan menggunakan *min-max* atau *standar scaler*, dan dilanjutkan dengan pembagian data, didapatkan hasil akurasi dari penelitian dapat dilihat pada tabel di bawah ini.

**Tabel 2. Hasil Penelitian**

| Metode | Akurasi | Gambar |
|---|---|---|

| | | |
|---|---|---|
| *K-fold cross validation* Min max | 0.7 |  |
| *K-fold cross validation Standar scaler* | 0,7333 |  |
| *Holdout validation* Min max | 0.5342 |  |
| *Holdout validation Standar scaler* | 0.6027 |  |

## 6. Kesimpulan dan Saran

Hasil pengolahan pada dataset liver menunjukkan bahwa setiap pengujian menghasilkan nilai akurasi yang berbeda. Ketika mengatur nilai untuk *k-fold cross validation* dan *holdout validation* akan membuat pengaruh pada setiap nilai akurasi. Jika melakukan penghapusan data yang duplikat dengan metode *k-fold cross validation* dan *min-max* hasil akurasinya adalah 0.7, lalu menggunakan *k-fold cross validation* dan *standar scaler* hasilnya akurasinya sebesar 0.7333, jika menggunakan *holdout validation* dan *min-max* hasil akurasinya sebesar 0.5342, dan jika menggunakan *holdout validation* dan *standar scaler* hasilnya 0.6027. Sehingga dapat disimpulkan bahwa akurasi yang terbesar adalah dengan menggunakan *k-fold cross validation* untuk pembagian datanya dan untuk normalisasi dengan menggunakan *standar scaler*.

Saran yang dapat kami berikan adalah diharapkan pada penelitian selanjutnya dapat menggunakan metode *feed forward backpropagation* atau metode regresi untuk mendapatkan nilai akurasi lainnya, sehingga dapat mengetahui metode mana yang lebih baik untuk klasifikasi pada dataset liver.

## 7. Daftar Pustaka

[1] Pujiyanta, Ardi, et. all. 2012. "*Sistem Pakar Penentuan Jenis Penyakit Hati dengan Metode Inferensi Fuzzy Tsukamoto (Studi Kasus di RS PKU Muhammadiyah Yogyakarta)"* dalam Jurnal Informatika : Vol 6, No. 1. Yogyakarta: Universitas Ahmad Dahlan Yogyakarta

[2] Conreng, Dicky, et. al. 2014. "*Hubungan Konsumsi Alkohol dengan Gangguan Fungsi Hati Pada Subjek Pria Dewasa Muda Di Kelurahan Tateli dan Tateli Atas Manado*" dalam Jurnal e-Clinic (eCl): Vol. 2, No.2. Manado: Universitas Sam Ratulangi.

[3] Noviandi. 2018. *Modul Kuliah Data Mining*. Universitas Esa Unggul.

[4] Handayani, Popon, et. all. 2019. "*Prediksi Penyakit Liver dengan Menggunakan Metode Decision tree dan Neural Network*" dalam CESS (Journal of Computer Engineering System and Science): Vol. 4, No. 1. Jakarta: STMIK Nusa Mandiri Jakarta.

[5] Setiawati, Intan, et. all. 2019. "*Implementasi Decision tree untuk Mendiagnosis Penyakit Liver*" dalam JOISM: JURNAL OF INFORMATION SYSTEM MANAGEMENT: Vol 1, No 1. Yogyakarta: Universitas Teknologi Yogyakarta.

[6] Wiley, John, et. all. 2015. *Data Science & Big Data Analytics*. Kanada: EMC Education Service.

[7] Nasution, Darnisa Azzahra, et, all. "*Perbandingan Normalisasi Data Untuk Klasifikasi Wine Menggunakan Algoritma KNN*" dalam Journal of Computer Engineering System and Science: Vol. 4, No 1. Jakarta: Universitas Pembangunan Nasional Veteran Jakarta.

[8] Raju, K Srujan, et. all. 2018. "*Support Vector Machine with K-Fold Cross Validation Model for Software Fault Prediction"* dalam International Journal of Pure and Applied Mathematics: Vol 118, No.20.

[9] Giarno, et, all. 2019. *"Suitable Proportion Sample of Holdout validation Validation for Spatial Rainfall Interpolation in Surrounding the Makassar Strait"* dalam Universitas Muhamadiyah Surajrta Online Journals: Vol 32, No. 2.

[10] UCI Machine Learning Repository. *https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)*, diakses pada 2 Juni 2020.