Cenderawasih : Journal of Statistics and Data Science
E-ISSN: 3026-0981

PEMILIHAN METODE KLASIFIKASI TERBAIK ANTARA *LOGISTIC REGRESSION* DAN *DECISION TREE* PADA DATASET HEPATITIS

DESI RAHMAWATI[1], AGINDA ERSITA MARURUK[2], CAECILIA BINTANG GIRIK ALLO[3]
[1,2,3)]Program Studi Statistika Fakultas MIPA Universitas Cenderawasih Jayapura, Indonesia
e-mail: bintanggirikallo@gmail.com

ABSTRAK

Hepatitis merupakan penyakit peradangan pada hati yang dapat disebabkan oleh berbagai macam penyebab, termasuk infeksi virus atau paparan zat beracun. Penerapan proses *data mining* untuk mengekstrak informasi dari data medis dan klinis. Dengan menggunakan metode ini, kondisi pasien di masa depan dapat diprediksi berdasarkan observasi data pasien lain atau pasien masa lalu. Penelitian ini bertujuan untuk menerapkan proses data mining dan melakukan perbandigan metode klasifikasi yaitu *Logistic Regression* dan *Decision Tree* menggunakan dataset Hepatitis. Berdasarkan hasil perbandingan diperoleh nilai accuracy *Logistic Regression* sebesar 80,207%. Sedangkkan metode *Decision Tree* menghasilkan nilai accuracy sebesar 83,195%. Maka dapat disimpulkan bahwa hasil perbandingan metode terbaik yaitu *Decision Tree*.

Kata Kunci: Hepatitis, Logistic Regression, Decision Tree, Perbandingan Metode Klasifikasi, Data Mining

1. PENDAHULUAN

Hepatitis merupakan penyakit peradangan pada hati yang dapat disebabkan oleh berbagai macam penyebab, termasuk infeksi virus atau paparan zat beracun. Pada hepatitis virus, peradangan hati yang terus-menerus atau berulang, sering dikaitkan dengan alkoholisme kronis yang dapat menyebabkan sirosis yaitu suatu kondisi yang melibatkan penggantian sel-sel hati yang rusak secara permanen dengan jaringan. Jaringan hati mempunyai kemampuan untuk beregenerasi dan dalam keadaan normal dan mengalami pergantian sel secara progresif. Jika sebagian jaringan hati rusak, maka jaringan yang rusak tersebut dapat diganti dengan meningkatkan laju pembelahan sel-sel sehat. Tampaknya ada faktor dalam darah yang bertanggung jawab untuk mengatur proliferasi hepatosit. Meskipun sifat dan mekanisme faktor pengaturan ini masih menjadi misteri, kecepatan pergantian sel hati ada batasnya. Selain hepatosit, beberapa fibroblas (sel jaringan ikat) juga ditemukan di antara lempeng hati dan membentuk jaringan pendukung hati. Jika hati berulang kali terkena zat beracun, seperti alkohol yang menyebabkan sering kali sel-sel hati baru tidak dapat beregenerasi dengan cukup cepat untuk menggantikan sel-sel yang rusak. Fibroblas yang kuat akan memanfaatkan situasi ini dan berkembang biak secara berlebihan. Jaringan ikat tambahan ini mengurangi ruang yang dibutuhkan sel-sel hati untuk beregenerasi.(Wahyu, 2011)

Data mining merupakan suatu proses pengerukan atau pengumpulan informasi penting dari big data. Proses *data mining* sering kali menggunakan metode statistik, matematika, dan bahkan teknologi kecerdasan buatan. Dengan menerapkan proses data mining untuk mengekstrak informasi dari data medis dan klinis. Maka, kondisi pasien di masa depan dapat diprediksi berdasarkan observasi data pasien lain atau pasien masa lalu. Salah satu metode prediksi adalah klasifikasi. Berbagai metode klasifikasi diuji untuk memverifikasi keakuratan hasil prediksi pada data pasien hepatitis.(Khomsah, 2018)

Penelitian ini bertujuan untuk menerapkan proses data mining klasifikasi yaitu regresi logistik dan *decision tree* untuk memprediksi harapan hidup penderita hepatitis kronis dimana, fokus penelitian adalah membandingkan metode klasifikasi terbaik untuk dataset penderita hepatitis.

2. METODE PENELITIAN

2.1 Sumber Data

Dataset yang digunakan adalah data pasien hepatitis yang diunduh dari repository UCI *Machine Learning*. Dataset berisi sejumlah atribut gejala medis beserta identifikasi apakah penderita hepatitis hidup (*live*) atau mati (*die*) jika memiliki gejala medis tersebut. Total data sebanyak 155 record. Atribut yang menunjukkan gejala sejumlah 19 dan 1 atribut kelas keputusan. Atribut kelas keputusan berisi nilai 1 untuk "*die*" dan 2 untuk "*live*".

2.2 Deskripsi Data

2.2.1 Atribut data

Atribut dataset hepatitis ini terdiri dari enam atribut numerik dan empat belas atribut kategorik. Variabel Y terdiri dari 32 data berstatus "Mati" dan 123 data berstatus "Hidup".

Tabel 1. Atribut Y

No	Atribut	Keterangan
1.	Class/Label keputusan	Label yang menunjukan pasien hidup/mati karena gejala yang ditemukan

Tabel 2. Atribut X

No	Atribut	Keterangan
1.	Umur	Umur pasien
2.	Jenis kelamin	Jenis kelamin pasien
3.	Steroid	Riwayat terapi steroid
4.	Antivirals	Riwayat terapi antivirals
5.	Fatigue	Gejala kelehahan akut
6.	Malaise	Gejala malaise(rasa tidak nyaman)
7.	Anorexia	Gejala anorexia(muntah setiap makan)
8.	Liver Big	Hati membesar
9.	Liver Firm	Hati mengeras
10.	Spleen Palpable	Gejala limfa lebih besar dari normal
11.	Spiders	Gejala pembuluh darah upnormal pada kulit(pembuluh darah mengumpul dan menonjol pada permukaan kulit)
12.	Ascites	Penumpukan cairan pada rongga perut
13.	Varices	Pembekakan vena esophagus (varices)
14.	Bilirubin	Kadar bilirubin dalam darah
15.	Alk Phosphate	Kadar alkalin phospate dalam liver
16.	Sgot	Nilai sgot
17.	Albumin	Kadar albumin
18.	Protime	Uji masa protrombhine
19.	Histology	Pemeriksaan dengan histology(biopsy hati)?

2.3 Proses *Pre-Processing*

Pre-processing adalah proses persiapan data sebelum dilakukan analisis atau pemodelan data dengan tujuan untuk memastikan kualitas data yang baik dan memudahkan dalam proses analisis. Beberapa tahapan dalam *pre-processing* data antara lain.

2.3.1 *Cleansing Data*

Proses *cleansing* data adalah proses membersihkan data dari kesalahan, duplikasi, atau ketidaktepatan lainnya dalam sebuah dataset. Proses ini penting dilakukan karena data yang tidak bersih dapat menghasilkan hasil analisis yang tidak akurat. Berikut adalah beberapa langkah yang dapat dilakukan dalam proses *cleansing data*.

- Pengecekan duplikat data guna menghasilkan performa prediksi yang lebih akurat.

- Pengecekkan *missing value* adalah mengecek data yang memiliki nilai "Na". Jika terdapat *missing value* dapat diatasi dengan, setiap data atribut yang memiliki *missing value* bertipe kategorik diisi dengan nilai modusnya dan tipe data numerik diisi dengan nilai mediannya.
- Deteksi *outlier* adalah proses mengidentifikasi nilai yang sangat berbeda dari sebagian besar data lainnya dalam sebuah dataset. Deteksi *outlier* dilakukan menggunakan *boxplot* pada data bertipe numerik dan *countplot* pada data bertipe kategorik.
- Transformasi data dilakukan untuk meningkatkan validitas dan reliabilitas hasil serta untuk memperbaiki karakeristk data dengan menggunakan salah satu metodenya yaitu *Min-Max Normalization*. Metode ini bertujuan untuk mengubah nilai-nilai data ke dalam rentang 0 hingga 1 Dimana, nilai minimum dari setiap atribut akan diubah menjadi 0, sedangkan nilai maksimum akan diubah menjadi 1.

  $x_{new} = \frac{x_{old} - \min(x)}{\max(x) - \min(x)}$ dimana, $\in [0,1]$ ...(1)

2.3.2 *Feature Selection*

Feature Selection adalah proses pemilihan fitur atau variabel yang paling relevan dan signifikan dalam dataset untuk meningkatkan kinerja model dan mengurangi dimensi data**.** Dalam *feature selection* terdapat 2 metode yaitu *filter method* dan *wrapper method*. *Filter method* digunakan untuk mengevaluasi setiap fitur secara bebas dari *classifier* kemudian memberikan peringkat pada fitur setelah mengevaluasi dan mengambil yang unggul.

2.3.3 *Split Data*

Split data adalah metode membagi data menjadi data pelatihan untuk mengembangkan model atau disebut data *training* dan data untuk mengevaluasi kinerja model atau disebut data *testing*. Ada dua teknik untuk melakukan *split* data yaitu *K-fold* dan *Holdout*. *K-fold* yaitu data setiap subset digunakan sebagai data pengujian secara bergantian, sementara subset lainnya digunakan sebagai data pelatihan. *Holdout* yaitu data dibagi menjadi dua subset menjadi data pelatihan dan data pengujian,dengan menggunakan proporsi 80:20 atau 70:30 pada umumnya.

2.3.4 *Imbalance Data*

Ketidakseimbangan data terjadi ketika jumlah sampel pada satu kelas jauh lebih sedikit dibandingkan jumlah sampel pada kelas lain. Ketidakseimbangan data dapat memengaruhi performa model dan memberikan hasil yang bias. Ada beberapa teknik yang dapat digunakan untuk menangani ketidakseimbangan data, seperti *oversampling* dan *undersampling*. Teknik *oversampling* dilakukan dengan menambahkan sampel pada kelas minoritas, sedangkan teknik *undersampling* dilakukan dengan mengurangi sampel pada kelas mayoritas.

2.4 Perbandingan Model

Model klasifikasi yang digunakan dalam penelitian ini terdiri dari 2 model, yaitu *Logistic Regression* dan *Decision Tree*.

2.4.1 *Logistic Regression*

Pada algoritma *Logistic Regression*, independen variabel didefinisikan sebagai persamaan regresi linear $y = \beta 0 + \beta 1x1 + \ldots + \beta pxp$ ...(2)

Algoritma *Logistic Regression* memprediksi probabilitas keanggotaan variabel independen dalam suatu kelas menggunakan model fungsi logit transform atau invers logit dengan persamaan (Gusthvi et al., 2023):

$\hat{p} = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}$ ...(3)

2.4.2 *Decision Tree*

Model Decision Tree dibentuk menyerupai struktur *flowchart*, dimana setiap simpul yang bukan simpul daun merupakan atribut pengujian, setiap cabang mewakili *output* dari pengujian, dan setiap simpul daun (terminal node) menentukan label class. Simpul paling atas dari sebuah pohon adalah node akar.

Decision Tree membedakan antara suatu kelas dengan kelas lainnya berdasarkan tingkat kemurnian kelas (*impurity*) pada suatu simpul. Alat ukur kemurnian kelas (*impurity*) yang umum digunakan adalah GINI Index, Entropy, Misclassification measure, Chi-square measure, G-square measure. Beberapa algoritma yang termasuk kategori Decision Tree antara lain ID3, C4.5, C5.0, CART, CHAID, dan lain-lain.(Gusthvi et al., 2023)

3. HASIL DAN PEMBAHASAN

3.1 Proses Pre-processing Data

3.1.1 *Cleansing Data*

3.1.1.1 Duplicate Data

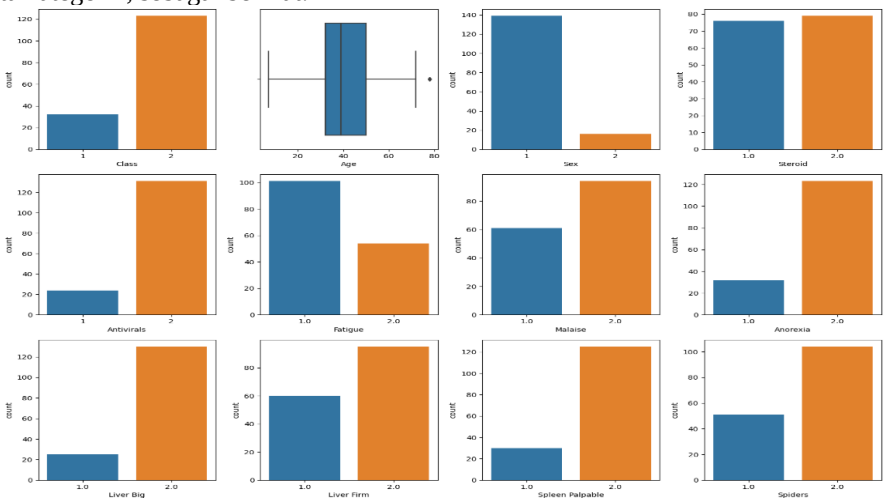
Hasil pengecekan duplikasi data menunjukkan bahwa pada dataset Hepatitis tidak memiliki data yang duplikat maka dapat dilanjutkan ke proses selanjutanya.

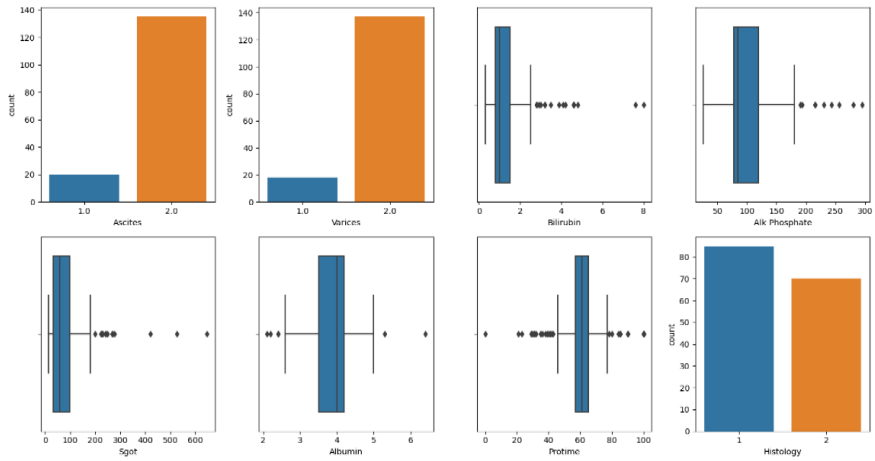
3.1.1.2 Missing Value

Hasil pengecekan *missing value* menunjukkan terdapat *missing value* pada 15 atribut yang terdiri dari 10 data kategorik dan 5 data numerik. Maka diatasi dengan cara mengganti *missing value* pada data numerik menggunakan nilai mediannya dan pada data kategorik menggunakan nilai modusnya. Setelah proses *imputation* maka, dilanjutkan pada proses selanjutnya.

3.1.1.3 Pendeteksian Outlier

Pendeteksian outlier dilakukan menggunakna *boxplot* untuk tipe data numerik dan *countplot* untuk tipe data kategorik, sebagai berikut.





Gambar 1. Hasil *Boxplot* dan *Countplot*

3.1.1.4 Transformasi Data

Transformasi data dilakukan menggunakan *Min-Max Normalization* dengan tujuan untuk menyamakan skala.

3.1.1.5 Feature Selection

Metode *feature selection* yang digunakan dalam penelitian ini yaitu *Filter Method* yang dimana untuk penyeleksian atribut menggunakan nilai korelasinya. Berdasarkan hasil *filter method* diperoleh nilai korelasinya yaitu 0,3 maka atribut yang digunakan ialah atribut yang memiliki nilai korelasi lebih dari sama dengan 0,3. Diperoleh 10 atribut sebagai berikut, Fatigue, Malaise, Spiders, Ascites, Varices, Bilirubin, Albumin, Protime, Histology, dan Class. Setelah diperoleh atribut-atribut hasil feature selection maka dapat dilanjutkan ke proses pemilihan metode klasifikasi terbaik.

3.1.1.6 Split Data

Pada dataset Hepatitis digunakan metode *split Holdout* dengan proporsi 20:80, dimana 20% digunakan sebagai data *testing* dan 80% sebagai data *training*.

3.1.1.7 Imbalance Data

Proses pengecekan *imbalance data* diperoleh data mengalami *imbalance* atau ketidakseimbangan maka untuk mengatasi ketidakseimbangan ini digunakan metode *oversampling* yaitu *Smote*. Metode ini menambahkan sampel pada kelas minoritas dengan menduplikasi data sejumlah data pada kelas mayoritas. Diperoleh jumlah data setelah *Smote* pada masing-masing kelas di atribut y adalah 101 data.

3.2 Perbandingan Metode Klasifikasi

Berdasarkan hasil pemodelan menggunakan Regresi Logistik dengan *K-Fold* sebanyak 5, diperoleh nilai *accuracy* sebesar 80,207%. Sedangkkan pemodelan menggunakan metode *Decision Tree* menghasilkan nilai *accuracy* sebesar 83,195%. Maka, dapat diperoleh metode terbaik yaitu metode yang memiliki nilai *accuracy* tertinggi yaitu *Decision Tree*.

4. SIMPULAN DAN SARAN

Perbandingan metode dilakukan dengan melakukan proses *pre-processing* terlebih dahulu yaitu, *cleansing data*, *feature selection* menggunakan *Filter Method*, *split data* menggunakan proporsi 20:80, mengatasi *imbalance data* dan proses perbandingan metode terbaik antara *Logistic Regression* dan *Decision Tree* yang diperoleh nilai accuracy tertinggi yaitu *Decision Tree* sebesar 83,195%. Maka dapat disimpulkan bahwa hasil perbandingan metode terbaik yaitu *Decision Tree*.

DAFTAR PUSTAKA

Gusthvi, W., Roza, A. A., Bintang, C., & Allo, G. (2023). Perbandingan Metode Klasifikasi Decision Tree, Naive Bayes, K-Nearest-Neighbor, dan Logistic Regression pada Dataset Phishing. In *CENDERAWASIH Journal of Statistics and Data Science* (Vol. 1). https://ejurnal.fmipa.uncen.ac.id/index.php/CJSDS

Khomsah, S. (2018). *Prediksi Harapan Hidup Penderita Hepatitis Kronik Menggunakan Metode-Metode Klasifikasi*.

Wahyu, Y. (2011). *MAKALAH_HEPATITIS*. https://www.academia.edu/18658316/MAKALAH_HEPATITIS