# Klasifikasi Pasien Penyakit Liver di Andhra Pradesh, India Menggunakan Metode Regresi Logistik, *Random Forest*, dan *Extra Trees*

Anita Syafira Irfan [1)], Hasri Wiji Aqsari [2)], Indah Maryana [3)],
Santi Wulan P [4)], Kartika Fithriasari [5)] , Irhamah [6)]
Departemen Statistika, Fakultas Sains dan Analitika Data
Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
*e-mail*: anitasyafirairfan@gmail.com [1)], hasriwiji@gmail.com [2)], indahmaryanaa@gmail.com [3)],
santi_wp@statistika.its.ac.id [2)], kartika_f@statistika.its.ac.id [3)] , irhamah@statistika.its.ac.id [4)]

***Abstrak*— Diagnosis sejak dini pada pasien dengan penyakit liver akan meningkatkan *survival rate* dari pasien tersebut. Menganalisis enzim-enzim yang terdapat pada sampel darah pasien akan membantu proses diagnosis penyakit liver. Oleh karena itu dibutuhkan prasarana diagnosis yang dapat mempercepat proses tersebut dan ini dapat dituntaskan dengan analisis secara statistik. Analisis statistik yang digunakan adalah klasifikasi di mana metode yang digunakan adalah regresi logistik, *random forest,* dan *extra trees*. Metode regresi logistik digunakan karena respon dari data yang digunakan adalah dikotomus atau data biner, yang hanya memuat kategori memiliki penyakit liver dan tidak memiliki penyakit liver. Metode *random forest* digunakan karena cocok untuk pengklasifikasian data dengan akurasi yang tinggi. Metode *extra trees* digunakan karena metode ini menggunakan algoritma *top-down* yang cocok dalam mendiagnosis suatu penyakit. Karakteristik dalam data ini dijelaskan dengan visualisasi dat yang menunjukkan bahwa pasien laki-laki lebih banyak yang menderita penyakit liver dibandingkan dengan yang perempuan. Ciri-ciri pasien yang sakit liver adalah memiliki *total bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate amminotransderase* yang tinggi, albumin, albumin *and globulin ratio* yang tinggi, dan juga sudah berusia tua. Hasil penelitian ini adalah metode regresi logistik akan memiliki hasil klasifikasi yang cukup baik apabila menggunakan metode pembagian data *repeated holdout* daripada *k-fold CV*. Sedangkan klasifikasi dengan menggunakan metode *random forest* akan memiliki hasil klasifikasi yang cukup baik apabila menggunakan metode pembagian data *k-fold CV* daripada *repeated holdout*. Dan untuk klasifikasi dengan menggunakan metode *extra trees* akan memiliki hasil klasifikasi yang cukup baik apabila menggunakan metode pembagian data *repeated holdout*. Masing-masing model untuk mengklasifikasi pasien penyakit liver di Andhra Pradesh, India menggunakan regresi logistik, *random forest,* dan *extra trees* menghasilkan akurasi secara berurutan adalah sebesar 77,71%, 81,03%, dan 78,29%.**

***Kata Kunci – Extra trees*, Klasifikasi, Penyakit liver, *Random forest*, Regresi logistik**

## I. PENDAHULUAN

Organ terbesar dalam perut adalah liver atau hati yang berbentuk triangular. Liver memiliki dua bagian yaitu bagian kanan dan kiri di mana liver ini merupakan organ tunggal yang memiliki fungsi yang esensial untuk tubuh manusia. Organ ini merupakan organ utama untuk mengontrol glukosa, menyeimbangkan nutrisi, lemak, vitamin, kolesterol, dan hormon. Diagnosis sejak dini pada pasien dengan penyakit liver akan meningkatkan *survival rate* dari pasien tersebut. Menganalisis enzim-enzim yang terdapat pada sampel darah pasien akan membantu proses diagnosis penyakit liver. Oleh karena itu dibutuhkan analisis secara statistik yang dapat menjadi prasarana dalam mendiagnosis apakah pasien tersebut terkena penyakit liver atau tidak. Analisis statistik yang dapat dilakukan adalah analisis klasifikasi karena diagnosis yang dimaksudkan akan berupa klasifikasi berdasarkan beberapa faktor yang diamati, dalam hal ini merupakan enzim yang terkandung pada sampel darah pasien, yang akan mengklasifikasikan pasien ke dalam kategori memiliki penyakit liver dan tidak memiliki penyakit liver. Pada penelitian ini, data yang digunakan adalah data pasien di Andhra Pradesh, India di mana terdapat sebanyak 583 pasien. Metode klasifikasi yang digunakan adalah regresi logistik, *random forest*, dan *extra trees*. Metode regresi logistik digunakan karena respon dari data yang digunakan adalah dikotomus atau data biner, yang hanya memuat kategori memiliki penyakit liver dan tidak memiliki penyakit liver. Metode *random forest* digunakan karena cocok untuk pengklasifikasian data dengan akurasi yang tinggi. Metode *extra trees* digunakan karena metode ini menggunakan algoritma *top-down* yang cocok dalam mendiagnosis suatu penyakit. Sebelum melakukan analisis klasifikasi, dilakukan gambaran secara deskritif dari data menggunakan visualisasi data *bar chart*, *box plot*, dan *correlogram* sehingga dapat dilihat bagaimana karakteristik dari data pasien di Andhra Pradesh ini. Prosedur klasifikasi menggunakan metode regresi logistik, *random forest*, dan *extra trees* memanfaatkan metode *repeated holdout* dan *k-fold cross validation* dalam pembagian data *training* dan *testing*-nya agar pembagian datanya dilakukan berulang guna mendapat pembagian data yang baik dan menghasilkan ketepatan klasifikasi yang tinggi nantinya. Di mana ukuran ketepatan klasifikasi pada penelitian ini akan dilihat berdasarkan nilai akurasi model, *sentivity*, *specificity*, *precision*, F1-*score*, AUC, dan ROC *curve*. Pada akhir berdasarkan ukuran-ukuran tersebut akan menuju ke keputusan model terbaik dalam mengklasifikasi pasien penyakit liver di Andhra Pradesh, India.

Laporan ini bertujuan untuk mengklasifikasi pasien liver di Andhra Pradesh, India menggunakan metode klasifikasi regresi logistik, *random forest*, dan *extra trees* sehingga dapat dijadikan prasarana dalam melakukan diagnosis. Diharapkan hasil dari penelitian ini dapat memenuhi tujuan tersebut.

## II. TINJAUAN PUSTAKA

### A. *Statistika Deskriptif*

Statistika deskriptif adalah metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna dengan menyusun tabel, grafik, diagram, dan besaran – besaran lain di majalah dan di koran [1]

#### *1. Rata-Rata*

Rata-rata hitung (*mean*) adalah nilai rata-rata dari data-data yang ada. Rata-rata hitung dari populasi diberi simbol μ. Rata-rata hitung dari sampel diberi symbol [1]. Rata-rata hitung (*mean*) untuk data tunggal.

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

Di mana,

$\bar{x}$ : rata-rata hitung (*mean*)
$x_i$ : nilai data ke-i di mana i=1,2,…,n
$n$ : jumlah sampel

#### *2. Varians dan Standar Deviasi*

Standar deviasi dan varians salah satu teknik statistik yang digunakan untuk menjelaskan homogenitas kelompok. Varians merupakan jumlah kuadrat semua deviasi nilai-nilai individual terhadap rata-rata kelompok. Sedangkan akar dari varians disebut dengan standar deviasi atau simpangan baku [1].

Standar deviasi dan varians simpangan baku merupakan variasi sebaran data. Semakin kecil nilai sebarannya berarti variasi nilai data makin sama Jika sebarannya bernilai 0, maka nilai semua datanya adalah sama. Semakin besar nilai sebarannya berarti data semakin bervariasi. Varians adalah salah satu ukuran dispersi atau ukuran variasi. Varians diberi simbol $\sigma^2$ untuk populasi dan untuk $s^2$ sampel, sedangkan standar deviasi diberi simbol σ untuk populasi dan s untuk sampel. Rumus untuk mencari varian adalah sebagai berikut.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad (2)$$

Untuk standar deviasi adalah akar dari *s* atau sama seperti rumus di atas hanya saja tinggal di akar kan.

Keterangan :

$\bar{x}$ : rata-rata hitung (*mean*)
$x_i$ : nilai data ke-i di mana i=1,2,…,n
$n$ : jumlah sampel

#### *3. Minimum*

Minimum adalah suatu nilai fungsi objek, yang menghasilkan nilai terendah pada daerah himpunan penyelesaian.

#### *4. Maksimum*

Maksimum adalah suatu nilai fungsi objek, yang menghasilkan nilai tertinggi pada daerah himpunan penyelesaian.

### B. *Preprocessing Data*

*Preprocessing* data merupakan tahapan pertama dan merupakan tahapan yang penting dalam *data mining* atau *data analysis* [2]. Pada umunya data pada dunia nyata yang masih mentah atau biasa disebut data primer memiliki kekurangan diantaranta tidak lengkap, banyak *noise*, dan juga tidak konsisten. Oleh karenanya tahapan ini sangat penting untuk memastikan data sumber diolah sehingga menghasilkan dataset yang siap dipakai pada tahapan selanjutnya. Ada tiga tahapan yaitu *data cleaning, data transformation*, dan *feature selection*.

1. *Data cleaning*

*Data cleaning* yaitu tahapan awal pada *preprocessing* data merupakan tahapan yang berusaha memperbaiki kualitas agar menjadi lebih baik seperti mengisi nilai yang hilang, menghaluskan data yang memiliki *noise*, hingga memperbaiki data yang tidak konsisten.

2. *Data transformation*

*Data transformation* dibutuhkan dalam implementasi *data mining* khususnya pada saat *preprocessing* data. Pada tahapan ini data diubah atau dikonsolidasikan sehingga proses penambangan yang dihasilkan dapat lebih efisien dan pola yang ditemukan dapat lebih mudah dipahami.

3. *Feature selection*

*Feature selection* adalah suatu proses yang mencoba untuk menemukan subhimpunan dari himpunan fitur yang tersedia untuk meningkatkan aplikasi dari suatu algoritma pembelajaran [3]. *Feature selection* digunakan dibanyak area aplikasi sebagai alat untuk menghilangkan fitur yang tidak relevan dan atau fitur berlebihan. Sebuah fitur dikatakan tidak relevan jika memberikan sedikit informasi, sedangkan sebuah fitur dikatakan berlebihan jika informasi yang diberikan adalah informasi yang terkandung dalam fitur lain (tidak memberikan informasi baru).

### C. *Visualisasi Data*

Bidang visualisasi difokuskan pada penciptaan gambar yang menyampaikan infornasi penting tentang data yang mendasari [4]. Visualisasi data dilihat oleh banyak bidang ilmu sebagai komunikasi visual modern. Visualisasi data tidak berada di bawah bidang manapun, melainkan interpretasi di antara banyak bidang misalnya, terkadang dilihat sebagai cabang modern dari statistik deskriptif oleh beberapa orang, tetapi juga sebagai dasar alat pengembangan oleh yang lain. Visualisasi data mengikutkan pembuatan dan kajian dari representasi visual dari data, artinya informasi yang telah diabstraksikan dalam bentuk skematis, termasuk atribut atau variabel dari unit informasi.Berikut merupakan beberapa visualisasi data yang sering digunakan.

1. Diagram Batang (*Bar Chart*)

*Bar chart* atau diagram batang merupakan diagram yang digunakan untuk menggambarkan perkembangan nilai suatu objek penelitian dalam kurun waktu tertentu. Diagram batang menunjukkan keterangan – keterangan dengan batang – batang tegak atau mendatar secara vertikal dan sama lebar dengan batang – batang terpisah.

2. *Correlogram*

*Correlogram* adalah suatu grafik mengenai korelasi data. Warna dari *correlogram* sangat berguna karena merupakan perwakilan dari hampir semua data. Warna *correlogram* juga dapat diubah menjadi seperti histogram berdimensi tiga yang pada dimensi satu dan dua menunjukkan warna dari setiap pasangan pixel dan dimensi tiga merupakan jarak spasialnya. Warna correlogram adalah varians dari histogram yang dihitung untuk korelasi spasial[9]

3. *Box plot*

*Box plot* merupakan cara paling mudah untuk menggambarkan distribusi data kategoris sesuai dengan

jangkauan kuartilnya. *Box plot* mempunyai garis vertikal yang memanjang keluar dari kotak yang dikenal dengan istilah *whisker*. *Whisker* inilah yang menggambarkan variabilitas data di luar batas kuartil atas dan kuartil bawah (atau yang dikenal sebagai pencilan). Maka dari itu, terkadang *box plot* juga dikenal sebagai diagram *box-whisker*. Seluruh pencilan dalam *box plot* divisualisasikan sebagai titik - titik individual di luar kotak.

*D. Algoritma Membagi data*

Banyak metode yang dapat digunakan untuk membagi data menjadi data *training* dan *testing* diantaranya yaitu:

1. *Repeated Holdout*

Metode *holdout* adalah metode paling sederhana untuk membagi data menjadi dua set data [5]. Tetapi metode ini tidak terlalu baik jika data tidak terlalu besar karena akan menghasilkan permasalahan klasifikasi. Untuk menghasilkan klasifikasi digunakan *repeated holdout*, yaitu dengan mengulang pembagian data sebanyak *n* kali sehingga menghasilkan sebanyak *n* akurasi dan di rata-rata. Metode *repeated holdout* terbukti efektif untuk menurunkan *bias.*

2. *K-Fold Cross Validation*

Metode ini membagi dataset menjadi k bagian dengan ukuran yang sama, dimana k adalah parameter dari metode k-fold cross-validation [5] . Untuk setiap pergantian k akan digunakan untu evaluasi, sedangkan untuk bagian k-1 yang lain digunakan untuk model pembelajaran. Jika k yang digunakan kecil, maka kemungkinan menghasilkan bias besar, begitu pula sebaliknya.

*E. Analisis Klasifikasi*

Metode klasifikasi yang digunakan pada penelitian kali ini ada yaitu regresi logistik, *extra trees*, dan *random forest*

1. Regresi logistik

Regresi logistik biner merupakan suatu metode statistika yang digunakan untuk pemodelan terbaik yang menggambarkan hubungan antara variabel respon (y) yang bersifat biner atau dikotomus dengan variabel prediktor (x) yang bersifat kualitatif, kuantitatif ataupun kombinasi keduanya [6] [7]. Variabel respon y terdiri dari 2 kategori yaitu "sukses" dan "gagal" yang dinotasikan dengan y=1 (sukses) dan y=0 (gagal). Dalam keadaan demikian, variabel y mengikuti distribusi Bernoulli untuk setiap observasi, untuk n pengamatan maka mengikuti distribusi binomial dengan *p* adalah banyaknya variabel prediktor. Bentuk model regresi logistik dengan *p* variabel prediktor adalah sebagai berikut.

$$\pi(x) = \frac{exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}{1+exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)} \quad (3)$$

Transformasi persamaan 4 pada regresi logistik disebut dengan *logit transformation* yang didefinisikan sebagai berikut.

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_0 X_1 + \ldots + \beta_p X_p \quad (4)$$

Pada regresi logistik, vaariabel respon diekspresikan sebagai $y = \pi(x) + \varepsilon$ dimana $\varepsilon$ mempunyai salah satu dari kemungkinan dua nilai yaitu $\varepsilon = 1 - \pi(x)$ dengan peluang $\pi(x)$ jika y = 1 dan $\varepsilon = -\pi(x)$ dengan peluang 1 - $\pi(x)$ jika y = 0 dan mengikuti distribusi binomial dengan rataan nol dan varians $(\pi(x))(1-\pi(x))$.

2. *Extra Trees*

Klasifikasi dengan metode *extra trees* atau yang disebut juga sebagai *"Extremly randomized trees*" merupakan varian pengembangan dari *decision tree* acak pada berbagai sub bagian dataset dan menghitung rata-ratanya untuk meningkatkan akurasi prediksi dan pengendalian *over vitting*. Berbeda dengan *Random forest* dimana pada setiap tahapannya, sample dan keputusan diambil secara acak dan bukan diambil dari yang terbaik [7]. *Extra trees* membangun grup *decision tree* sesuai dengan prosedur top down.

3. *Random Forest*

*Random forest* adalah pengklasifikasian yang terdiri dari kumpulan pengklasifikasian pohon terstruktur dimana masing-masing pohon melemparkan unit suara untuk kelas paling populer di input *x* [8], dengan kata lain *random forest* terdiri dari sekumpulan *decision tree*, dimana kumpulan *decision tree* tersebut digunakan untuk mengklasifikasi data ke suatu kelas.

*F. Evaluasi Model*

Evaluasi model dilakukan berdasarkan *confussion matrix* dan ROC *Curve*

1. *Confusion Matrix*

*Confusion matrix* merupakan metode yang menggunakan tabel matriks seperti pada Tabel 1. Jika data set terdiri dari dua kelas, maka kelas 1 sebagai positif dan yang lain negatif [9].

**Tabel 1** Model *Confussion Matrix*

| | | True Class | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted Class | Positive | True Positives (TP) | False Negative (FN) |
| | Negative | False Positive (FP) | True Negatives (TN) |

*True positive* adalah jumlah *recored* positif yang diklasifikasikan sebagai positif, *false positives* adalah jumlah *record* negatif yang diklasifikasikan sebagai positif. *False Negatives* adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* adalah jumlah *record* negatif yang diklasifikasikan sebagai negatif. Berdasarkan *confussion* matrik tersebut dapat dihitung nilai-nilai *sensitivity* (*recall*), *specificity*, *precision*, dan *accuracy* [9]. Persamaan-persamaan berikut adalah cara menghitungnya.

$$\begin{aligned} accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ sensitifity &= \frac{TP}{TP+FN} \\ specificity &= \frac{TN}{TN+FP} \\ precission &= \frac{TP}{TP+FP} \end{aligned} \quad (5)$$

2. *ROC Curve*

ROC (*Receiver Operating Characteristics*) *curve* adalah pengujian berdasarkan performanya. ROC mengekpresikan *confussion matrix*. Nilai dari ROC *curve* hanya terdiri dari 0 sampai 1, semakin nilai ROC *curve* mendekati 1, maka akan semakin baik seperti diperlihatkan pada Tabel 2 [10].

**Tabel 2** Kriteria Nilai ROC *Curve*

| Range | Keterangan |
|---|---|
| 0.9-1.00 | *Excellent classification* |
| 0.8-0.9 | *Good classification* |
| 0.7-0.8 | *Fair classification* |
| 0.6-0.7 | *Poor classification* |
| 0.5-0.6 | *Failure* |

*G. Liver*

Penyakit liver adalah istilah yang digunakan untuk setiap gangguan pada liver atau hati yang menyebabkan organ ini tidak dapat berfungsi dengan baik. Hati merupakan organ yang dapat melakukan regenerasi dengan cepat untuk mengganti sel-selnya yang rusak. Akan tetapi, jika sel-sel yang rusak cukup banyak, fungsi dan kerja hati dapat terganggu. Biasanya, fungsi hati akan mulai terlihat penurunannya ketika kerusakan sel-sel hati mencapai 75%. Penurunan fungsi hati umumnya terjadi secara bertahap. Tahapan kerusakan yang terjadi akan mengikuti perkembangan penyakit yang mendasarinya dan seberapa besar kerusakan jaringan hati yang dialami [11].

## III. METODOLOGI PENELITIAN

*A. Sumber Data*

Dalam laporan penelitian ini, data yang digunakan adalah data sekunder yang diunduh dari laman https://www.kaggle.com yang berjudul "*Indian Liver Patient Dataset*". Pada dataset ini berisikan tentang observasi pada pasien penderita penyakit liver di Andhra Pradesh, India. Di mana observasi yang dimaksud merupakan faktor-faktor yang dilihat pada setiap pasien penderita penyakit liver. Dataset ini memuat 583 data pasien dengan 416 di antaranya memiliki riwayat liver dan 167 tidak memiliki.

*B. Variabel Penelitian*

Variabel penelitian yang digunakan pada penelitian ini adalah sebagaimana Tabel 3.

**Tabel 3.** Variabel Penelitian

| Variabel | Skala Data | Keterangan |
|---|---|---|
| *Age* | Rasio | Usia pasien |
| *Gender* | Nominal | Jenis kelamin pasien |
| *Tot_bilirubin* | Rasio | Kadar bilirubin total pada pasien |
| *Direct_bilirubin* | Rasio | Kadar bilirubin *direct* (terkonjugasi) pada pasien |
| *Alkphos* | Rasio | Kadar alkaline phosphotase pada pasien |
| *Sgpt* | Rasio | Kadar alamine aminotransferase pada pasien |
| *Sgot* | Rasio | Kadar aspartate aminotransferasi pada pasien |
| *Tot_proteins* | Rasio | Kadar protein total pada pasien |
| *Albumin* | Rasio | Kadar albumin pada pasien |
| *Ag_ratio* | Rasio | Rasio albumin dan globulin |
| *Is_patient* | Nominal | 1 : Pasien dengan riwayat liver<br>2 : Pasien tidak memiliki riwayat liver |

*C. Langkah Analisis*

Beberapa langkah analisis yang digunakan dalam penelitian ini adalah sebagai berikut.

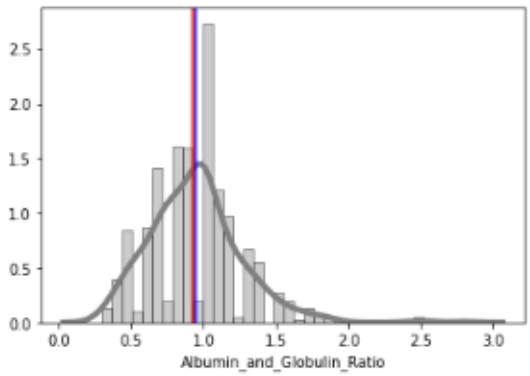
1. Melakukan pengumpulan data *Indian Liver Patient.*
2. Melakukan *preprocessing* data yaitu mendeteksi serta mengatasi data *outlier, missing value,* dan melakukan *feature selection*.
3. Melakukan eksplorasi data menggunakan visualisasi data pada tiap variabel dari data *Indian Liver Patient*.
4. Melakukan analisis klasifikasi berdasarkan metode regresi logistik, *random forest,* dan *extra trees* dengan masing-masing metode menggunakan pembagian data *training-testing* berdasarkan *repeated holdout* dan *k-fold cross validation*.
5. Melakukan penarikan kesimpulan dan saran untuk penelitian selanjutnya.

## IV. HASIL DAN PEMBAHASAN

*A. Preprocessing Data*

*Preprocessing* yang telah dilakukan pada data menghasilkan beberapa keadaan diantaranya adalah sebagai berikut :

1. Membaca data, menghasilkan pengetahuan yaitu data memiliki 11 variabel. Terdapat 2 variabel kategorik yaitu variabel *gender* dan *dataset.* Sedangkan 9 lainnya adalah variabel numerik yaitu *age, total bilirubin, direct bilirubin, alkaline phosphatase, alamine aminotransferase, aspartate aminotransferase, total protiens, albumin, albumin and globulin ratio*.
2. Memeriksa data apakah ada yang *missing value* atau tidak. Hasilnya adalah terdapat *missing value* di variabel *albumin and globulin ratio*.
3. Membuat *density plot* dari variabel *albumin and globulin ratio* untuk penentuan pengisian *missing value*. Berikut adalah gambar dari *density plot* menggunakan garis *mean* dan *median.* Garis *mean* diwakili dengan garis lurus berwarna biru sedangkan garis *median* diwakili dengan garis lurus berwarna merah.




**Gambar 1** Density plot

Berdasarkan gambar 1 tersebut didapatkan keputusan bahwa *missing value* diisi menggunakan *mean* karena mean lebih berada di tengah distribusi daripada *median.*

*B. Feature Selection*

*Feature selection* dibagi menjadi dua metode yaitu *correlation threshold* dan *variance threshold.* Pada metode pertama, *correlation threshold,* diatur korelasi sama dengan 0.9. Artinya variabel yang memiliki nilai korelasi 0.9 atau lebih akan dikeluarkan dari *dataset* karena dianggap salah satu variabel sudah mewakili variabel yang lain. Berdasarkan

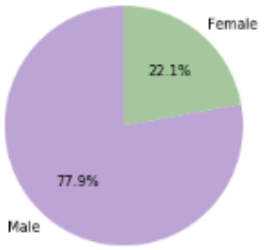
metode *correlation threshold* tidak ada variabel yang dikeluarkan dari *dataset* karena masing-masing dari variabel tidak memiliki nilai korelasi tinggi atau lebih dari 0.9.

Pada metode kedua *variance threshold* diatur korelasi samam dengan 0.1. Artinya variabel yang memiliki nilai varian sama dengan atau lebih kecil dari 0.1 dikeluarkan dari *dataset* karena dianggap data memiliki nilai yang hampir sama sehingga dinilai tidak berguna untuk analisis lanjutan. Berdasarkan metode *variance threshold* didapatkan basil bahwa variabel *albumin and globulin ratio* dikeluarkan dari model karena memiliki nilai varian sebesar 0.09.

*C. Visualisasi Data*

Setelah di *preprocessing*, selanjutnya data divisualisasikan, sehingga dapat memberikan informasi yang bermanfaat. Visualisasi data yang dihasilkan diantaranya adalah sebagai berikut.

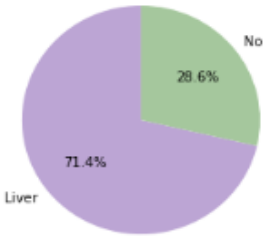
1. *Pie chart* untuk jenis kelamin




**Gambar 2** Pie chart jenis kelamin

Informasi yang didapatkan dari gambar 2 adalah lebih banyak laki-laki yang menderita penyakit liver yaitu dengan proporsi 77.9% sedangkan untuk perempuan memiliki proporsi sebesar 22.1%

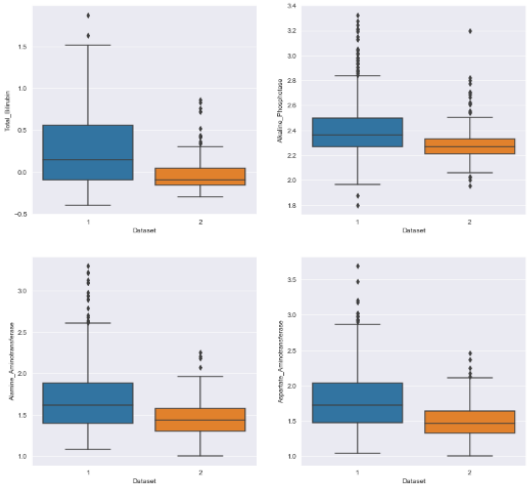
2. *Pie chart* untuk pasien




**Gambar 3** Pie chart pasien

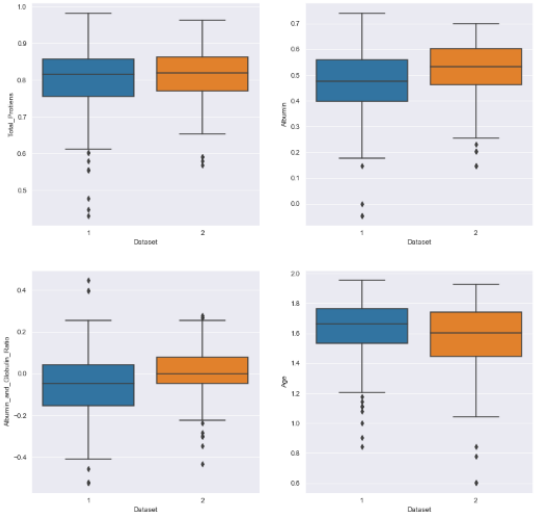
Informasi yang didapatkan dari gambar 3 adalah lebih banyak pasien yang menderita penyakit liver yaitu dengan proporsi sebesar 71.4% sedangkan untuk proporsi yang tidak penyakit liver adalah sebesar 28.6%.

3. *Boxplot* setiap variabel



**Gambar 4** Boxplot 4 variabel pertama

*Boxplot* digunakan untuk melihat ciri-ciri pasien yang menderita penyakit liver, dengan cara membandingkan *boxplot* antara pasien yang sakit liver dan tidak. Berdasarkan gambar 4 didapatkan informasi bahwa ciri-ciri pasien yang sakit liver adalah memiliki *total bilirubin, alkaline phosphatase, alamine aminotransferase* dan *aspartate amminotransderase* yang tinggi.



**Gambar 5** Boxplot 4 variabel kedua

Berdasarkan gambr 5 didapatkan informasi bahwa ciri-ciri lain dari pasien yang sakit liver adalah memiliki *albumin* dan *albumin and globulin ratio* yang tinggi. Selain itu juga terdapat informasi bahwa pasien yang sakit liver rata-rata berumur sudah tua. Variabel *total protiens* tidak dapat digunakan untuk mengetahui ciri-ciri pasien sakit liver karena berdasarkan *boxplot* tidak menunjukkan perbedaan yang signifikan.

*D. Analisis Klasifikasi*

Selanjutnya adalah melakukan analisis klasifikasi untuk data liver di India dengan menggunakan metode klasifikasi yaitu regresi logistik, *Extra Trees*, dan *Support Vector Machine* (SVM). Sebelum dilakukan analisis data terlebih dahulu dibagi menjadi dua data yaitu data *training* dan data *testing* dengan menggunakan metode *repeated holdout* dan *K-Fold CV* agar dapat membandingkan metode mana yang terbaik dengan melihat nilai kebaikan model yaitu nilai akurasi, *sensitivity*, *specificity*, AUC ROC, dan ROC *curve.*

1. Regresi Logistik

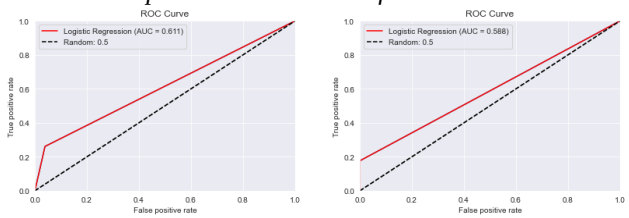
Hasil analisis klasifikasi dengan menggunakan metode klasifikasi regresi logistik dan metode pembagian data *repeated holdout* dan *K-Fold CV* didapatkan kesimpulan kebaikan model yang dapat dilihat pada tabel berikut.

**Tabel 4** Kebaikan model metode regresi logistik

| | ***Repeated Holdout*** | ***K-Fold CV*** |
|---|---|---|
| *Accuracy* | 77,71% | 75,86% |
| *Sensitivity* | 78,48% | 74,55% |
| *Specificity* | 70,59% | 100,00% |
| F-Score | 79,23% | 77,18% |
| AUC | 61,11% | 58,,82% |

Pada tabel 2 dapat dilihat bahwa dengan membandingkan hasil pembagian data menjadi data *training* dan *testing* metode *repeated holdout* menghasilkan nilai *accuracy*, *sensitivity*, *f-score* dan AUC yang lebih besar daripada metode *K-Fold CV* sedangkan *specificity* memiliki nilai yg lebih tinggi pada metode *K-Fold CV* maka dapat disimpulkan bahwa dengan menggunakan metode regresi logistik metode pembagian data yang baik adalah *repeated holdout*. Sehingga

dapat dikatakan bahwa dengan menggunakan metode regresi logistik dengan metode pembagian data menggunakan *repeated holdout* didapatkan sebesar 77,71% data telah diprediksi dengan tepat, sedangkan dari nilai sensitivitas diketahui bahwa 78,48% dari data yang terkena liver dapat diprediksi dengan benar dan dari nilai spesifisitas diketahui bahwa sebanyak 70,59% dari data yang tidak terkena liver dapat diprediksi dengan benar. Nilai AUC sebesar 61,11% menunjukkan bahwa data belum terprediksi secara baik. Berikut merupakan grafik ROC untuk pembagian data berdasarkan *repeated holdout* dan *k-fold CV*




**Gambar 6** Grafik ROC Regresi Logistik (a)*repeated holdout* (b) *k-fold CV*

Gambar 6 menampilkan grafik ROC dari kedua metode pembagian data *testing training* dimana dari gambar tersebut diketahui bahwa metode *repeated holdout* lebih baik karena memiliki kurva ROC yang lebih jauh dari garis *random* dibandingnya dengan metode *k-fold CV* dan juga dilihat dari nilai AUC metode *repeated holdout* yang lebih besar daripada *k-fold CV.*

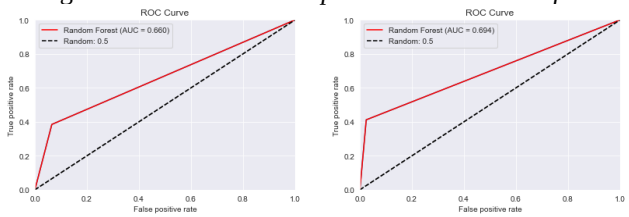
2. *Random Forest*

Hasil analisis klasifikasi dengan menggunakan metode klasifikasi *extra trees* dan metode pembagian data *repeated holdout* dan *K-Fold CV* didapatkan kesimpulan kebaikan model yang dapat dilihat pada tabel berikut.

**Tabel 5** Kebaikan Model Metode *Random Forest*

| | ***Repeated Holdout*** | ***K-Fold CV*** |
|---|---|---|
| *Accuracy* | 77,14% | 81,03% |
| *Sensitivity* | 78,23% | 80,00% |
| *Specificity* | 71,43% | 87,50% |
| F-Score | 79,11% | 80,00% |
| AUC | 65,98% | 69,37% |

Berdasarkan tabel 3 dapat disimpulkan bahwa metode pembagian data terbaik adalah *k-fold CV* karena memiliki nilai semua nilai kebaikan model yang lebih tinggi dari pada *repeated holdout.* Maka dapat disimpulkan bahwa dengan menggunakan metode klasifikasi *random forest* dengan metode pembagian data menggunakan metode *k-fold CV* didapatkan bahwa 81,03% data telah diprediksi dengan tepat dimana nilai senstifitas diketahui bahwa 80,00% dari data pasien yang terkena liver diprediksi dengan benar dan dari nilai spesifisitas diketahui bahwa sebanyak 87,50% dari data pasien yang tidak terkena liver diprediksi dengan benar. Nilai AUC sebesar 69,37% menunjukkan bahwa data belum terprediksi secara baik. Berikut merupakan grafik ROC untuk pembagian data berdasarkan *repeated holdout* dan *k-fold CV*




**Gambar 7** Grafik ROC *Random Forest* (a)*repeated holdout* (b) *k-fold CV*

Gambar 7 menunjukkan bahwa metode *k-fold CV* lebih baik karena memiliki kurva ROC yang lebih jauh dari garis *random* dibandingnya dengan metode *repeated holdout* dan juga dilihat dari nilai AUC metode *k-fold CV* yang lebih besar daripada nilai AUC metode *repeated holdout.*

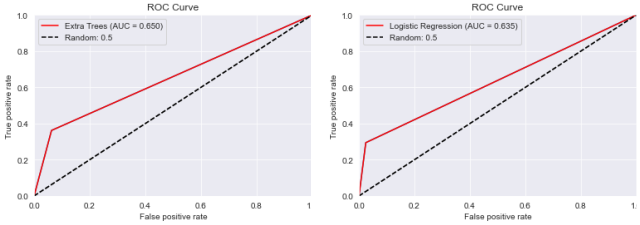
3. *Extra Trees*

Hasil analisis klasifikasi dengan menggunakan metode klasifikasi *extra trees* dan metode pembagian data *repeated holdout* dan *K-Fold CV* didapatkan kesimpulan kebaikan model yang dapat dilihat pada tabel berikut.

**Tabel 6** Kebaikan model metode *extra trees*

| | ***Repeated Holdout*** | ***K-Fold CV*** |
|---|---|---|
| *Accuracy* | 78,29% | 77,59% |
| *Sensitivity* | 80,00% | 76,92% |
| *Specificity* | 68,00% | 83,33% |
| F-Score | 80,00% | 78,43% |
| AUC | 64,96% | 63,49% |

Pada tabel 4 dapat disimpulkan bahwa metode *repeated holdout* menghasilkan nilai *accuraccy*, *specificity*, *precision*, *f-score*, dan AUC lebih besar dari metode *k-fold CV* sedangkan nilai *sensitifity* lebih besar pada saat menggunakan metode *k-fold CV* maka dapat disimpulkan bahwa dengan menggunakan metode klasifikasi *extra trees* metode pembagian data yang baik adalah *repeated holdout.* Sehingga dapat dikatakan bahwa sebesar 78,29% data telah diprediksi dengan tepat, Sedangkan dari nilai sensitivitas diketahui bahwa sebanyak 80,00% dari data pasien yang terkena liver diprediksi dengan benar dan dari nilai spesifisitas diketahui bahwa sebanyak 68,00% dari data pasien yang tidak terkena liver diprediksi dengan benar. Nilai AUC sebesar 64,96% menunjukkan bahwa data belum terprediksi secara baik. Berikut merupakan grafik ROC untuk pembagian data berdasarkan *repeated holdout* dan *k-fold CV.*




Gambar 8 Grafik ROC *Extra Trees* (a)repeated holdout (b) k-fold CV

Gambar 8 menunjukkan bahwa metode *repeated holdout* lebih baik karena memiliki kurva ROC yang lebih jauh dari garis *random* dibandingnya dengan metode *k-fold CV* dan juga dilihat dari nilai AUC metode *repeated holdout* yang lebih besar daripada nilai AUC metode *k-fold CV.*

## V. KESIMPULAN DAN SARAN

Kesimpulan yang didapatkan dari hasil analisis diatas yaitu terdapat *missing value* di variabel *albumin and globulin ratio* yang kemudian diatasi dengan menggunakan *mean* karena dari hasil density plot menggambarkan bahwa *mean* lebih berada di tengah plot daripada median. Dilakukan *feature selection* dan menghasilkan dengan menggunakan metode *variance threshold* didapatkan basil bahwa variabel *albumin and globulin ratio* dikeluarkan dari model. Diketahui pula bahwa penderita penyakit liver sebagian besar adalah laki-laki dengan ciri-ciri pasien yang sakit liver adalah memiliki *total bilirubin, alkaline phosphatase, alamine* aminotransferase, *aspartate amminotransderase, albumin* dan *albumin and globulin ratio* yang tinggi dengan pasien liver rata-rata berumur sudah tua.

Analisis klasifikasi dengan menggunakan metode regresi logistik, *random forest*, dan *extra trees* didapat kesimpulan bahwa klasifikasi dengan menggunakan metode

regresi logistik akan memeliki hasil klasifikasi yang cukup baik apabila menggunakan metode pembagian data *repeated holdout* dari pada *k-fold CV*. Sedangkan klasifikasi dengan menggunakan metode *random forest* akan memiliki hasil klasifikasi yang cukup baik apabila menggunakan metode pembagian data *k-fold CV* dari pada *repeated holdout*. Dan untuk klasifikasi dengan menggunakan metode *extra trees* akan memiliki hasil klasifikasi yang cukup baik apabila menggunakan metode pembagian data *repeated holdout*.

Saran untuk penelitian selanjutnya yaitu untuk lebih memperhatikan *imbalanced* data agar pada saat dilakukan klasifikasi hasil akurasi klasifikasi yang didapat lebih tinggi dan juga disarankan untuk mencoba lebih banyak lagi metode klasifikasi.

## DAFTAR PUSTAKA


[1] R. E. Walpole, Pengantar Statistika Edisi ke-3, Jakarta: PT. Gramedia Pustaka Utama, 1995.

[2] J. K. Han and J. Pei, Data Mining Concepts and Techniques, Morgan Kaufmann, 2012.

[3] M. Dash and H. Liu, Feature selection for Classification, Intelligent Data Analysis, 1997.

[4] H. C and J. C, The Visualization Handbook, USA: Elsevier Inc., 2005.

[5] A. Karahoca, Advances in Data mining Knowledge Discovery and Applications, Rijeka: Intech, 2012.

[6] D. L. Hosmer and S. Lemeshow, Applied Logistic Regression, New York: John Willey and Sons, Inc, 2000.

[7] G. P, E. D and L. Wehenkel, "Extremely Randomized Trees," 2006. [Online]. [Accessed Oktober 2005].

[8] L. Breimen, Machine Learning, Berkeley: University of California, 2001.

[9] Olson, David and S. Yong, Pengantra Ilmu Penggalian Data Bisnis, Jakarta: Salemba Empat, 2008.

[10] F. Gorunescu, Data Mining Concepts, Model and Techniques, Springer: s.1., 2011.

[11] Alodokter, "Penyakit Liver," 2020. [Online]. Available: https://www.alodokter.com/penyakit-liver.