IMPLEMENTASI *DECISION TREE* C4.5 UNTUK KLASIFIKASI CARA KELUAR PASIEN GINJAL KRONIS BERDASARKAN REKAM MEDIS BPJS KESEHATAN DI RSUAM

TUGAS AKHIR

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Program Studi Statistika

Disusun Oleh:

Insani Hasanah

17611074

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2021

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul	:	Implementasi *Decision Tree* C4.5 untuk Klasifikasi Cara Keluar Pasien Ginjal Kronis Berdasarkan Rekam Medis BPJS Kesehatan di RSUAM.
Nama Mahasiswa	:	Insani Hasanah
NIM	:	17611074

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, 26 Januari 2021

Pembimbing

(Dr. Edy Widodo, S.Si., M.Si)

HALAMAN PENGESAHAN

TUGAS AKHIR

IMPLEMENTASI *DECISION TREE* C4.5 UNTUK KLASIFIKASI CARA KELUAR PASIEN GINJAL KRONIS BERDASARKAN REKAM MEDIS BPJS KESEHATAN DI RSUAM

Nama Mahasiswa : Insani Hasanah

NIM : 17611074

TUGAS AKHIR INI TELAH DIUJIKAN

PADA TANGGAL : 16 Februari 2021

Nama Penguji	**Tanda Tangan**
1. Muhammad Muhajir, S.Si., M.Sc.	.....
2. Mujiati Dwi Kartikasari, S.Si., M.Sc.	.....
3. Dr. Edy Widodo, S.Si., M.Si	.....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

(Prof. Riyanto, S.Pd., M.Si., Ph.D.)

KATA PENGANTAR

Assalamu'alaikum Warahmatullaahi Wabarakaatuh

Alhamdulillahi Robbil 'Alamin, puji dan syukur penulis panjatkan kepada Allah SWT, Tuhan semesta alam yang telah melimpahkan rahmat, hidayah serta inayah-Nya kepada penulis, sehingga penulis dapat menyusun dan menyelesaikan tugas akhir yang berjudul **"Implementasi *Decision Tree* C4.5 untuk Klasifikasi Cara Keluar Pasien Ginjal Kronis Berdasarkan Rekam Medis BPJS Kesehatan di RSUAM"** dengan lancar dan sebaik-baiknya sebagai salah satu persyaratan dalam menyelesaikan jenjang strata satu di Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia. Shalawat serta salam tak lupa penulis ucapkan kepada junjungan kita nabi agung Muhammad SAW yang selalu membimbing ke jalan yang penuh berkah ini.

Selama penyusunan tugas akhir, penulis telah banyak mendapatkan bimbingan, bantuan serta dukungan dari berbagai pihak. Untuk itu pada kesempatan kali ini penulis bermaksud menyampaikan ucapan terimakasih yang sebesar-besarnya kepada :

1. Bapak Prof. Fathul Wahid, S.T., M.Sc., Ph.D selaku rektor Universitas Islam Indonesia.
2. Bapak Prof. Riyanto, S.Pd., M.Si., Ph.D. selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia.
3. Bapak Dr. Edy Widodo, S.Si., M.Si, selaku ketua Jurusan Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia dan selaku dosen pembimbing 1 tugas akhir ini yang telah memberikan arahan dan bimbingan selama penyusunan penulisan tugas akhir ini.
4. Keluarga penulis yang selalu memberikan motivasi, dukungan dan mendoakan yang terbaik untuk penulis.
5. Pihak Rumah Sakit Umum Abdoel Moeloek yang sudah membantu selama pengambilan data.

6. Sahabat-sahabat terdekat serta Teman-teman Statistika UII angkatan 2017 yang sudah membantu dan memberikan semangat sehingga penulis dapat menyelesaikan tugas akhir ini.
7. Serta semua pihak baik secara langsung maupun tidak langsung telah membantu penulis dalam menyelesaikan tugas akhir.

Semoga Allah SWT senantiasa melimpahkan rahmat dan ridho-Nya kepada semua pihak yang telah membantu penulis. Demikian penulisan tugas akhir ini, semoga penulisan tugas akhir ini bermanfaat bagi semua pihak. Amin.

Wassalamualaikum Wr.Wb

Yogyakarta, 26 Januari 2021

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan orang lain, kecuali yang diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, 26 Januari 2021

(Insani Hasanah)

INTISARI

IMPLEMENTASI *DECISION TREE* C4.5 UNTUK KLASIFIKASI CARA KELUAR PASIEN GINJAL KRONIS BERDASARKAN REKAM MEDIS BPJS KESEHATAN DI RSUAM

Insani Hasanah

Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Islam Indonesia

Decision tree merupakan salah satu teknik klasifikasi untuk menemukan kumula pola atau fungsi yang mendeskripsikan serta memisahkan kelas data yang satu dengan yang lainnya untuk menyatakan objek tersebut masuk pada kategori tertentu dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Pada penelitian ini dilakukan klasifikasi menggunakan metode *decision tree* C4.5 untuk mengklasifikasi cara keluar pasien ginjal kronis berdasarkan rekam medis BPJS Kesehatan di RSUAM dengan Program R. Dari seluruh data rekam medis tanggal 01 Juni 2020 hingga 31 Oktober 2020 yang berjumlah 358 data dibagi menjadi 287 sebagai data *training* dan 71 sebagai data *testing*. Dari hasil pengujian menghasilkan 11 aturan yang dapat dijadikan pola dalam menentukan pasien yang paling berpotensi meninggal dengan akurasi yang cukup besar yaitu 91% sebelum dilakukan *post-prunning* dan 88% setelah dilakukan *post-purning*. Pada hasil akurasi dengan dan tanpa penerapan metode *re-sampling* untuk mengatasi *imbalance data,* keduanya tidak memiliki perbedaan yang signifikan dengan masing-masing nilai akurasi sebesar 89% ketika menggunakan metode *re-sampling* (*oversampling* dan *undersampling*) dan 91% tanpa menggunakan metode *re-sampling*, namun jika dilihat berdasarkan nilai *specificity*, *sensitivity*, dan *precision*, metode ini mampu meningkatkan ketiga ukuran ketepatan tersebut karena memiliki nilai yang lebih besar dibandingkan sebelum menggunakan metode *re-sampling*.

Kata Kunci: Klasifikasi, *Decision Tree* C4.5, *Imbalance, Sampling*, Rekam Medis, Sembuh, Meninggal

ABSTRACT

IMPLEMENTATION OF C4.5 DECISION TREE FOR CLASSIFICATION OF EXIT CHRONIC RENAL PATIENTS BASED ON MEDICAL RECORD OF BPJS HEALTH IN RSUAM

Insani Hasanah
Department of Statistics, Faculty of Matematics and Natural Sciences
Universitas Islam Indonesia

Decision tree is a classification technique to find the cumulative pattern or function that describes and separates data classes from one another to declare that the object belongs to a certain category by looking at the behavior and attributes of a defined group. In this study, a classification was carried out using the C4.5 decision tree method to classify the way out of chronic kidney patients based on BPJS Kesehatan medical records at RSUAM with the R Program. From all medical record data from 1 June 2020 to 31 October 2020, a total of 358 data is divided into 287 as training data and 71 as testing data. In the accuracy results with and without the application of the re-sampling method to overcome data imbalance, both of them do not have a significant difference with the respective accuracy values of 89% when using the re-sampling method (oversampling and undersampling) and 91% without using the re-sampling method. sampling, but when viewed based on the values of specificity, sensitivity, and precision, this method is able to increase the three measures of accuracy because it has a greater value than before using the re-sampling method.

Keywords: Classification, Decision Tree C4.5, *Imbalance*, *Sampling*, Medical Records, Cured, Died

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

(UNFPA, 2018) memproyeksikan jumlah penduduk Indonesia pada tahun 2020 sebesar 255,6 juta jiwa dengan komposisi 135,3 juta jiwa laki-laki dan 134,3 juta jiwa perempuan. Sementara itu, jumlah kematian (mortalitas) sebesar 1,7 juta jiwa pada 2020 dan terus menanjak hingga 2045 sebesar 3,2 juta jiwa. Tahun ke tahun jumlah kematian akan terus meningkat dan menjadikan keadaan sehat sebagai suatu keadaan yang sulit didapatkan dan dijaga oleh setiap orang. Seiring berjalannya waktu, penyakit-penyakit makin tersebar ke seluruh masyarakat di Indonesia bahkan dunia yang menjadikan rumah sakit semakin disesaki oleh pasien-pasien yang berdatangan.

Rumah Sakit Umum Abdoel Moeloek adalah rumah sakit tipe A yang bertempat di Kota Bandar Lampung Provinsi Lampung dimana rumah sakit ini memiliki alat kesehatan terbaik di Provinsi Lampung. Saat pasien-pasien masuk kedalam rumah sakit ini terdapat dua tipe pendaftaran antara lain dengan menggunakan kartu asuransi (BPJS dan lainnya) dan umum (mandiri). Saat ini, sebagian besar masyarakat Indonesia sudah memiliki kartu BPJS Kesehatan yang berfungsi untuk membiayai pelayanan kesehatan sesuai dengan ketentuan program jaminan sosial, dimana kartu ini dapat digunakan untuk pasien yang melakukan rawat jalan ataupun rawat inap.

Karena Rumah Sakit Umum Abdoel Moeloek adalah rumah sakit tipe A, untuk di rawat inap di rumah sakit ini, perlu beberapa tahap yang harus dilakukan. Dengan itu, hanya dua tipe pasien BPJS Kesehatan yang dapat dirawat di rumah sakit ini antara lain pasien yang dirujuk rumah sakit sebelumnya dan pasien yang masuk IGD, setiap rumah sakit lain yang tidak sanggup/ tidak memungkinkan untuk menyelamatkannya karena minimnya alat kesehatan dan juga faktor lainnya, maka pasien-pasien tersebut akan dirujuk ke Rumah Sakit Umum Abdoel Moeloek. Sehingga pasien-pasien BPJS Kesehatan yang rawat inap di Rumah Sakit Umum Abdoel Moeloek termasuk kedalam pasien gawat darurat. Salah satu penyakit pada pasien gawat daruratnya ialah penyakit ginjal kronis. Menurut hasil

penelitian *Global Burden of Disease* tahun 2010, penyakit ginjal kronis merupakan penyebab kematian peringkat ke-27 di dunia, tahun 1990 dan meningkat menjadi urutan ke-18 pada tahun 2010 (Kementrian Kesehatan Republik Indonesia, 2017).

Penyakit ginjal kronis adalah suatu keadaan klinis yang ditandai dengan penurunan fungsi ginjal yang *irreversible*, pada suatu derajat yang memerlukan terapi pengganti ginjal yang tetap, berupa dialisis atau transplantasi ginjal (Herman, 2016). Sekitar 1 dari 10 populasi global mengalami penyakit ginjal kronis pada tipe/stadium tertentu. Hasil systematic review dan metaanalysis mendapatkan prevalensi global penyakit ginjal kronis sebesar 13,4%. Pada tahun 2013 sebanyak 499.800 penduduk Indonesia menderita penyakit ginjal kronis. Di Indonesia, perawatan penyakit ginjal merupakan ranking kedua pembiayaan terbesar dari BPJS kesehatan setelah penyakit jantung (Kementrian Kesehatan Republik Indonesia, 2017). Berdasarkan data Riset Kesehatan Dasar tahun 2018, prevalensi penyakit ginjal kronis berdasarkan diagnosis dokter penduduk umur ≥ 15 tahun menurut karakteristik di Provinsi Lampung yaitu sebesar 0.39% (Tim Riset Kesehatan Dasar, 2019).

Salah satu analisis statistik ialah analisis klasifikasi. Analisis klasifikasi dilakukan untuk menentukan sebuah *record* data baru ke salah satu dari beberapa kategori yang telah didefinisikan sebelumnya, hal tersebut dapat disebut juga dengan *supervised learning*. Metode-metode yang digunakan untuk menyelesaikan kasus klasifikasi yaitu terdiri dari Pohon Keputusan (*Decision Tree*), Jaringan Syaraf Tiruan, *Naïve Bayes,* Algoritma Genetik, *Rough Sets, Metode Berbasis Aturan,* Analisis Statistik, *kNearest Neighbour, Support Vector Machine, Memory Based Reasoning* (Sumathi, 2006).

Algoritma *decision tree* didasarkan pada pendekatan *divide-and-conquer* untuk klasifikasi suatu masalah. Algoritma tersebut bekerja dari atas ke bawah, mencari pada setiap tahap atribut untuk membaginya ke dalam bagian terbaik *class* tersebut, dan memproses secara rekursif submasalah yang dihasilkan dari pembagian tersebut. Strategi ini menghasilkan sebuah *decision tree* yang dapat diubah menjadi satu set *classification rules* (Witten, 2011). Dalam analisis *decision tree* terdapat tiga jenis algoritma, antara lain CART (*Classification and*

Regression Tree), ID3 yang dibuat tahun 1970 sampai awal tahun 1980 oleh J. Ross Quinlan, seorang peneliti di bidang *machine learning* dan C4.5. Algoritma C4.5 sendiri merupakan pengembangan dari algoritma ID3 yang dibuat oleh Quinlan. Algoritma C4.5 memiliki kelebihan yaitu mudah dimengerti, fleksibel, dan menarik karena dapat divisualkan dalam bentuk gambar *decision tree* (Han, J; Kamber, M, 2006). Beberapa penelitian terdahulu yang menggunakan algoritma C4.5 ini atau juga membandingkan antara beberapa metode klasifikasi yang dimana algoritma C4.5 memiliki hasil tingkat akurasi yang tinggi seperti (Hana, 2020) dengan judul "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5" yang menghasilkan akurasi yang cukup besar yaitu 97,12 %, *precision* sebesar 93,02% %, dan *recall* sebesar 100,00% dan (Andie, 2016) dengan judul "Penerapan Decision Tree Untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa Baru" yang menghasilkan tingkat akurasi sebesar 94.29%.

Berdasarkan hal tersebut, peneliti akan melakukan klasifikasi cara keluar pasien ginjal kronis berdasarkan rekam medis yan diantaranya terdapat tipe cara keluar pasien (sembuh, meninggal), tipe penyakit pasien (tumor ginjal & gagal ginjal ringan, tumor ginjal & gagal ginjal sedang, dan tumor ginjal & gagal ginjal berat), umur pasien (anak-anak, remaja, dewasa, dan lanjut usia), lamanya pasien dirawat, kelas rawat pasien (1,2,3,V), dan jenis kelamin pasien (laki-laki dan perempuan) di rsuam menggunakan *decision tree* c4.5 untuk menentukan klas-klas mana saja yang paling banyak sembuh dan paling banyak meninggal.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, maka permasalahan yang muncul dapat dirumuskan:

1. Bagaimana deskripsi rekam medis pasien ginjal kronis di Rumah Sakit Umum Abdoel Moeloek?
2. Bagaimana hasil klasifikasi cara keluar pasien kronis Rumah Sakit Umum Abdoel Moeloek menggunakan *Decision Tree C4.5*?

3. Bagaimana hasil perbandingan ukuran ketepatan antara sebelum dan setelah mengatasi *imbalance data* berdasarkan metode *Decision Tree C4.5*?

1.3 Batasan Masalah

1. Rekam medis yang digunakan yaitu cara keluar pasien, tipe penyakit gagal ginjal kronis, umur pasien, lama pasien dirawat, kelas rawat pasien, dan jenis kelamin pasien.
2. Software yang digunakan ialah program R untuk menganalisis data menggunakan metode *decision tree c4.5* dan microsoft excel untuk melakukan deskripsi data.
3. Rekam medis pasien yang digunakan yaitu dalam rentang waktu 01 Juni 2020 hingga 31 Oktober 2020.

1.4 Tujuan Penelitian

1. Mengetahui hasil deskripsi dari rekam medis pasien kronis di Rumah Sakit Umum Abdoel Moeloek.
2. Mengetahui hasil klasifikasi cara keluar pasien kronis Rumah Sakit Umum Abdoel Moeloek menggunakan *Decision Tree C4.5*.
3. Mengetahui hasil perbandingan ukuran ketepatan antara sebelum dan setelah mengatasi *imbalance data* berdasarkan metode *Decision Tree C4.5*.

1.5 Manfaat Penelitian

Manfaat dalam proposal ini adalah:

Berdasarkan hasil klasifikasi yang terbentuk, dapat digunakan untuk menentukan pasien ginjal kronis mana yang berpotensi meninggal berdasarkan pola dari 11 aturan yang terbentuk. Untuk penelitian lanjutan, diharapkan penelitian ini mampu menambah pengetahuan terkait metode yang digunakan (*decision tree c4.5*) untuk memecahkan studi kasus rekam medis pasien.

BAB II
TINJAUAN PUSTAKA

Dalam melakukan penelitian ini, peneliti mencari referensi dari beberapa sumber, yang berkaitan dengan judul yang diambil. Berikut beberapa referensi yang berkaitan dengan judul penelitian:

Tabel 2.1 Penelitian-penelitian terdahulu

No	Peneliti	Judul	Metode	Tujuan Penelitian	Kelebihan Penelitian Terdahulu
1	Rian Rafiska, dkk	Analisa Rekam Medis untuk Menentukan Pola Kelompok Penyakit Menggunakan Klasifikasi dengan *Decision Tree*	*Decision Tree* C4.5	Untuk membentuk model klasifikasi pohon keputusan untuk memprediksi loyalitas pelanggan dan melihat variable yang paling berpengaruh	Kelebihan : Memiliki tingkat akurasi hingga mencapai 97.5% serta algoritma C4.5 ini telah berhasil di terapkan dalam menganalisis data rekam medis di RSUD Mayjen H.A. Thalib Kerinci.
2	Fida Maisa Hana	Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma *Decision Tree* C4.5	Decision Tree C4.5	Untuk mengklasifikasi apakah seseorang terkena penyakit diabetes atau tidak	Kelebihan : Menghasilkan akurasi yang cukup besar yaitu 97,12 % Precision sebesar 93,02% %, dan Recall sebesar 100,00%.
3	Yuli Mardi	Data Mining Rekam Medis Untuk Menentukan Penyakit Terbanyak Menggunakan *Decision Tree C4.5*	*Decision Tree* C4.5	Tujuan utama dari penelitian ini adalah bagaimana memanfaatkan data rekam medis yang dianggap sampah tersebut agar dapat memberikan kontribusi yang positif bagi semua pihak baik bagi rumah sakit dalam membuat kebijakan, bagi fasilitas	Kelebihan : Menggunakan *software KNIME*, dimana merupakan pengetahuan baru untuk peneliti

No	Peneliti	Judul	Metode	Tujuan Penelitian	Kelebihan Penelitian Terdahulu
				kesehatan, maupun bagi pemerintah dalam penanganan kesehatan.	
4	Bambang Hermanto dan Azhari SN	Klasifikasi Nilai Kelayakan Calon Debitor Baru Menggunakan *Decision Tree* C4.5	*Decision Tree* C4.5	Untuk menghasilkan sebuah software sebagai pendukung keputusan dalam penilaian kelayakan calon debitur baru menggunakan metoda algoritma C4.5 sehingga bermanfaat bagi manajer perusahaan dalam pembiayaan kendaraan bermotor	Kelebihan : Dalam pembuatan pohon keputasan (generate tree) dan aturan keputusan (generate rules) dibutuhkan waktu yang cukup cepat, yakni tidak lebih dari 15 menit untuk setiap skenario pengujian.
5	Andie	Penerapan *Decision Tree* untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa Baru	*Decision Tree*	Untuk membantu pihak perguruan tinggi swasta, khususnya dalam hal memprediksi jumlah kemungkinan mahasiswa yang akan melakukan registrasi ulang secepat mungkin	Kelebihan : Terdapat perhitungan manual dan menggunakan aplikasi Rapid Miner.

Penelitian oleh (Andie, 2016) yang berjudul "Penerapan Decision Tree Untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa Baru". Penelitian ini dilakukan pada Universitas Islam Kalimantan (UNISKA) Muhammad Arsyad AlBanjary Banjarmasin dengan sampel data yang digunakan sebanyak 10663 data calon mahasiswa baru tahun 2013, 2014 dan 2015. Menariknya, pada penelitian ini terdapat perhitunan manual dan perhitungan menggunakan aplikasi *RapidMiner* yang menghasilkan tingkat akurasi yang berbeda. Tingkat akurasi untuk perhitungan manual (94,29%) lebih tinggi dibandingkan perhitungan menggunakan aplikasi *RapidMiner* yan menghasilkan nilai akurasi sebesar 86,07%.

Sementara itu (Mardi, 2018) melakukan penelitian tentang "*Data Mining* Rekam Medis Untuk Menentukan Penyakit Terbanyak Menggunakan Decision Tree C4.5". Data yan digunakan ialah data yang diperoleh di Rumah Sakit Umum Citra BMC Padang yang berobat pada bulan Januari 2013 sebanyak 21 data pasien yang berobat. Didapatkan total entrophy 2,5061441 dengan jumlah kasus terbanyak terdapat pada BAB XVIII (R00-R99) yaitu sebanyak 8 pasien dari 21, dengan rincian jenis kelamin (perempuan 5 pasien dan laki-laki 3 pasien), usia (tua 5 pasien, muda dan dewasa 1 pasien, bayi dan anak 2 pasien), alamat (Padang Timur 4 pasien, Padang Utara 1 pasien, Lubuk Begalung 2 pasien dan Padang Barat 1 pasien).

(Hana, 2020) melakukan penelitian tentang "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5". Peneliti menggunakan data *Early stage diabetes risk prediction dataset*, dengan variabel sebanyak 17 variabel dengan jumlah data sebanyak 520 data. Kumpulan data ini dibuat dari kuesioner langsung kepada orang-orang yang baru saja menjadi penderita diabetes, atau yang masih nondiabetes tetapi memiliki sedikit atau lebih gejala. Data dikumpulkan dari pasien dengan menggunakan kuesioner langsung dari Sylhet Diabetes Hospital of Sylhet, Bangladesh. Dari 520 data dibagi menjadi 416 sebagai data training dan 104 sebagai data testing. Dari hasil pengujian menghasilkan akurasi yang cukup besar yaitu 97,12 % Precision sebesar 93,02% %, dan Recall sebesar 100,00%.

Berdasarkan penelitian terdahulu yang sudah dijabarkan, terdapat perbedaan terkait penelitian ini, yaitu terkait tempat dan waktu penelitian, variabel, dan penggunaan *software*. Pada penelitian ini bertempat di Rumah Sakit Umum Abdoel Moeloek dan untuk waktu penelitian yaitu 01 Juni 2020 hingga 31 Oktober 2020. Untuk melakukan klasifikasi *decision tree*, mayoritas dari penelitian terdahulu tersebut menggunakan *software Rapidmaner* dan *KNIME* sedangkan untuk penelitian ini, peneliti menggunakan Program R.

BAB III
LANDASAN TEORI

3.1 Statistika Deskriptif

(Hasan, 2001) menjelaskan bahwa stastika deskriptif atau statistik deduktif adalah bagian dari statistik mempelajari cara pengumpulan data dan penyajian data sehingga muda dipahami. Statistik deskriptif hanya berhubungan dengan hal menguraikan atau memberikan keterangan-keterangan mengenai suatu data atau keadaan atau fenomena.

Menurut (Janah, 2014), tabel silang dua dimensi adalah pengelompokan data dari dua variabel dengan melakuka pengelompokan silang dari kedua variabe. Tabel silang dapat disebut juga sebagai tabel kontingensi. Pada tabel silang dua dimensi terdapat keterkaitan antara dua variabel. Contohna seperti terdapat variabel letak suatu wilayah kota dengan tingkat kriminalitas yang ada saling terikat. Tingkat kriminalitas akan semakin tinggi di wilayah yang letaknya sebelah barat atau sama sekali tidak terdapat keterikatan. Semua ini didukung oleh data yang diperoleh dari penelitian. Berikut merupakan gambaran tabel kontingensi dalam bentuk sebuah kerangka:

Tabel 3.1 Tabel Kontingensi

Variabel Terikat (y)	**Variabel Bebas (x)**		**Total y**
	x_1	x_2	
y_1	x_1y_1	x_2y_1	Total y_1
y_2	x_1y_2	x_2y_2	Total y_2
Total x	Total x_1	Total x_2	Total xy

Analisis tabulasi silang/*crosstabs* digunakan untuk menghitung frekuensi dan persentase dua atau lebih variabel dengan melakukan penyilangan variabel-variabel yang dianggap berhubungan sehingga makna hubungan dua variabel dapat mudah dipahami secara deskriptif (Santoso & Tjiptono, 2001). Tujuan dari analisis tabulasi silang adalah untuk mengidentifikasi korelasi antara satu variabel dengan variabel lainnya. Salah satu ciri-ciri dari penggunaan data crosstab adalah

data input yang digunakan yaitu data nominal atau ordinal sehingga akan menghasilkan output yang dapat dijelaskan secara deskriptif (Sarwono, 2009).

a. Hipotesis

   H_0 : Tidak ada hubungan antara baris dan kolom

   H_1 : Ada hubungan antara baris dan kolom

b. Tingkat Signifikasi

   Dengan *confident interval* 95% didapatkkan $\alpha = 5\%$ / $\alpha = 0.05$

c. Daerah Kritis

   Tolak H_0 jika *p-value/ Asymp. Sig (2-sided) Chi-Square* < α

3.2 *Data Mining*

Data Mining merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara yang berbeda dengan sebelumnya, yang dapat dipahami dan bermanfaat bagi pemilik data (Larose, 2005). Menurut (Aprilla, dkk, 2013), kata *mining* merupakan kiasan dari bahasa inggris, *mine*. *Mine* memiliki arti menambang sumber daya yang tersembunyi di dalam tanah, dimana berarti *data mining* merupakan penggalian makna yang tersembunyi dari kumpulan data yang sangat besar. Karena itu *data mining* sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan basis data. *Data mining* dibagi menjadi beberapa kelompok berdasarkan tugas yang dilakukan, yaitu:

a. Klasifikasi

   Klasifikasi yaitu suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan pengelompokkan pada data baru dengan melakukan manipulasi data yang ada dan menggunakan hasilnya untuk memberikan sejumlah aturan. Salah satu contoh yang mudah dan populer adalah dengan menggunakan metode *decision tree* dimana metode ini merupakan salah satu metode klasifikasi yang paling populer karena mudah untuk diinterpretasi. *Decision tree*

adalah model prediksi menggunakan struktur pohon atau struktur berhirarki.

b. Clustering

Clustering digunakan untuk menganalisis pengelompokkan berbeda terhadap data, mirip dengan klasifikasi, namun pengelompokkan belum didefinisikan sebelum dijalankannya *tool data mining*. Biasanya menggunkan metode neural network atau statistik. Clustering membagi item menjadi kelompok-kelompok berdasarkan yang ditemukan tool data mining.

c. Asosiasi

Asosiasi digunakan untuk mengenali kelakuan dari kejadiankejadian khusus atau proses dimana hubungan asosiasi muncul pada setiap kejadian. Salah satu contohnya adalah *Market Basket Analysis*, yaitu salah sati metode asosiasi yang menganalisa kemungkinan pelanggan untuk membeli beberapa item secara bersamaan.

3.3 *Decision Tree*

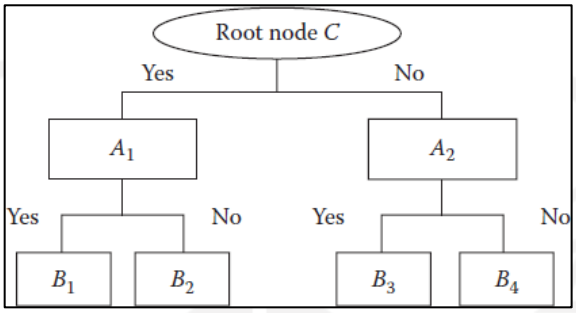
Pohon keputusan atau yang lebih dikenal dengan istilah *Decision Tree* ini merupakan implementasi dari sebuah sistem yang manusia kembangkan dalam mencari dan membuat keputusan untuk masalah-masalah tersebutdengan memperhitungkan berbagai macam faktor yang berkaitan di dalam lingkup masalah tersebut. Secara umum, pohon keputusan adalah suatu gambaran permodelan dari suatu persoalan yang terdiri dari serangkaian keputusan yang mengarah kepada solusi yang dihasilkan (Aprilla, dkk, 2013).

Pada *decision tree* terdapat 3 tipe jenis *node*, yaitu:

a. *Root Node*, merupakan *node* paling atas, pada *node* ini tidak ada *input* dan bisa tidak mempunyai *output* atau mempunyai *output* lebih dari satu. (lihat titik C pada gambar 3.1)
b. *Internal Node* , merupakan *node* percabangan, pada *node* ini hanya terdapat satu *input* dan mempunyai *output* minimal dua. (lihat titik A pada gambar 3.1)

c. *Leaf node atau terminal node* , merupakan *node* akhir, pada *node* ini hanya terdapat satu *input* dan tidak mempunyai *output*. (lihat titik B pada gambar 3.1)

Berikut merupak contoh struktur *decision tree*:



Gambar 3.1 Contoh Struktur *Decision Tree*
Sumber: (Dua & Du, 2011)

3.3.1 *Decision Tree* ID3 dan C4.5

Terdapat beberapa jenis dalam metode *decision tree* antara lain yaitu ID3 atau C4.5. Berikut ini akan dijelaskan model dari *decision tree* tersebut yaitu:

Salah satu jenis metode *decision tree* yang diperkenalkan dan dikembangkan pertama kali oleh Quinlan yaitu ID3 yang merupakan singkatan dari *Iterative Dichotomiser 3* atau *Induction of Decision 3*. Terdapat langkah pembentukan *decision tree* ID3 adalah:

a. *Node* teratas yaitu node tunggal (akar/root) dimana *node* ini merepresentasikan semua data dari data *training*.
b. Sesudah akar *node* dibentuk, maka data pada akar *node* akan diukur dengan nilai *gain* untuk dipilih atribut mana yang akan dijadikan atribut pembaginya.
c. Sebuah cabang dibentuk dari atribut yang dipilih menjadi pembagi dan data akan didistribusikan ke dalam cabang masing-masing.
d. Kemudian dilakukan proses yang sama atau rekursif untuk dapat membentuk sebuah *decision tree*. Ketika sebuah atribut telah dipilih menjadi node pembagi atau cabang, maka atribut tersebut tidak diikutkan lagi dalam penghitungan nilai *gain*.
e. Proses pembagian rekursif akan berhenti jika salah satu dari kondisi dibawah ini terpenuhi :

1. Semua data dari anak cabang sudah termasuk dalam kelas yang sama.
2. Semua atribut telah dipakai, tetapi masih tersisa data dalam kelas yang berbeda. Dalam kasus ini, diambil data yang mewakili kelas yang terbanyak untuk menjadi label kelas pada node daun. Tidak terdapat data pada anak cabang yang baru. Dalam kasus ini, node daun akan dipilih pada cabang sebelumnya dan diambil data yang mewakili kelas terbanyak untuk dijadikan label kelas.

Menurut (Fernitha, 2019) perbedaan utama algoritma C4.5 dan ID3 yaitu:

Tabel 3.2 Perbedaan algoritma C4.5 dan ID3

No	Algoritma C4.5	Algoritma ID3
1	Dapat menangani atribut kontinu dan diskrit	Hanay dapat atribut diskrit
2	Dapat menangani data *training* dengan *missing value*	Tidak dapat menangani data *training* dengan *missing value*
3	Hasil pohon keputusan C4.5 akan terpangkas setelah dibentuk	Hasil pohon keputusan C4.5 tidak terpangkas setelah dibentuk
4	Pemilihan atribut yang dilakukan menggunakan *Gain Ratio*	Pemilihan atribut yang dilakukan dengan mengunakan *Information Gain*

Terdapat beberapa langkah dalam menentukan pohon keputusan menggunakan algoritma *decision tree* C4.5 (Nugraha, dkk, 2016).

1. Menentukan *data training*.
2. Menghitung nilai *Entropy* dengan menggunakan persamaan (1) sebagai berikut:

$$Entropy(\mathrm{S}) = \sum_{i=1}^{n} - p_i * \log_2 p_i \quad (1)$$

dengan,

S : Himpunan *dataset* (kasus)

n : Banyaknya partisi *S*

i : Jumlah partisi *S*

p_i : Proporsi dari S_i terhadap $S = \frac{S_i}{S}$ $(\frac{Himpunan\ (Yes)}{Himpunan/Total\ Kasus})$

S_i : Partisi S ke $-i$

Misal terdapat sebuah kasus dengan variabel independen memiliki 14 kasus dengan *yes* sebanyak 9 kasus dan *no* sebanyak 5 kasus. Salah satu variabel independennya terdapat variabel *wind* dengan dua tipe yaitu *weak* dan *strong*. Diketahui bahwa:

Tabel 3.3 Total *Wind*

	Total	**No**	**Yes**
Weak	8	2	6
Strong	6	3	3

Hasil perhitungan *entropy* untuk partisi *weak* dan *strong* yaitu:

$$Entropy(\mathrm{S}) = \sum_{i=1}^{n} - p_i * \log_2 p_i$$

$$= \left(-\frac{2}{8} * log_2\left(\frac{2}{8}\right)\right) + \left(-\frac{6}{8} * log_2\left(\frac{6}{8}\right)\right) = 0.81$$

$$Entropy(\mathrm{S}) = \sum_{i=1}^{n} - p_i * \log_2 p_i$$

$$= \left(-\frac{3}{6} * log_2\left(\frac{3}{6}\right)\right) + \left(-\frac{3}{6} * log_2\left(\frac{3}{6}\right)\right) = 1$$

3. Hitung *Gain*, untuk menentukan atribut sebagai akar dengan melihat nilai *gain* tertinggi dari berbagai atribut. Untuk menghitung *gain* dapat menggunakan persamaan 2 (Nugraha, dkk, 2016) sebagai berikut:

$$Gain(S, A) = Entropy(\mathrm{S}) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (2)$$

dengan,

S : Himpunan kasus

A : Atribut

n : Jumlah partisi atribut A

$|S_i|$: Jumlah Kasus pada partisi ke-i

$|S|$: Jumlah kasus S /Himpunan kasus

$Entropy(S_i)$: *Entrophy* untuk kasus-kasus pada partisi ke-i

Diketahui nilai *entrophy* untuk seluruh himpunan kasus sebesar 0.94. Hasil perhitungan *gain* untuk variabel *wind* yaitu:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

$$= 0.94 - \left(\left(\frac{8}{14} * 0.81\right) + \left(\frac{6}{14} * 1\right)\right) = 0.05$$

4. Menentukan *root node*.
5. Proses partisipasi pohon keputusan akan berhenti saat semua cabang dalam node N mendapat kelas yang sama.
6. Melakukan pengujian model dengan data dibagi menjadi 2 bagian yaitu, sebanyak 80% data menjadi data *training* yang digunakan untuk membangun struktur (pola) pohon keputusan melalui metode *decision tree* C4.5. Sedangkan 20% data akan digunakan sebagai data uji *(data testing)*.
7. Menganalisa hasil klasifikasi. Menggunakan *confusion matrix*. *Confusion matrix* merupakan sebuah tabel yang terdiri dari banyaknya baris data uji yang diprediksi benar dan tidak benar oleh model klasifikasi.

Tabel 3.4 *Confusion Matrix*

	Actual Class	
Predicted Class	***Class* = 1**	***Class* = 0**
***Class* = 1**	TP	FP
***Class* = 0**	FN	TN

Menurutu (Fernitha, 2019), pada saat membangun pohon keputusan, banyaknya cabang mungkin karena adanya *noise* atau *outlier*pada data training. Pemangkasan pohon (*pruning*) dapat dilakukan untuk mengenali dan menghilangkan cabang tersebut agar pohon lebih kecil dan lebih mudah dipahami. Selain itu, pemangkasan pohon juga perlu dilakukan karena dalam teknik klasifikasi yang akan dijalankan nantinya akan mengeluarkan pola atau *rule* yang dibentuk berdasarkan struktur pohon, jadi jika struktur pohon tidak teratur atau kurang sederhana, maka *rule* yang dihasilkan pun akan rumit untuk diimplementasikan. Ada dua metode yang dapat digunakan untuk melakukan pemangkasan pohon keputusan, yaitu :

1. *Prepruning*

   Prepruning yaitu melakukan pemangkasan *subtree* lebih awal, yakni dengan memutuskan untuk tidak lebih jauh mempartisi data *training*. Pada

pendekatan *prepruning*, sebuah pohon dipangkas dengan cara menghentikan pembangunannya jika partisi yang akan dibuat dianggap tidak signifikan. Keuntungan dari *prepruning* yaitu lebih hemat waktu dalam proses pembentukan pohon keputusan.

2. *Postpruning*

   Postpruning yaitu menyederhanakan pohon dengan cara memangkas beberapa cabang *subtree* setelah pohon selesai dibangun.

   Langkah-langkah pemangkasan pohon :

   1. Hitung *Pessimistic error rate parent.*
   2. Hitung *Pessimistic error rate child*.
   3. Jika *Pessimistic error rate child>parent*, maka lakukan pemangkasan.
   4. Jika *Pessimistic error rate child<parent*, maka lanjutkan *split*.

   Untuk menghitung *Pessimistic error rate* digunakan rumus dari persamaan (3).

$$e = \frac{r + \frac{z^2}{2n} + z\sqrt{\frac{r}{n} - \frac{r^2}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \quad (3)$$

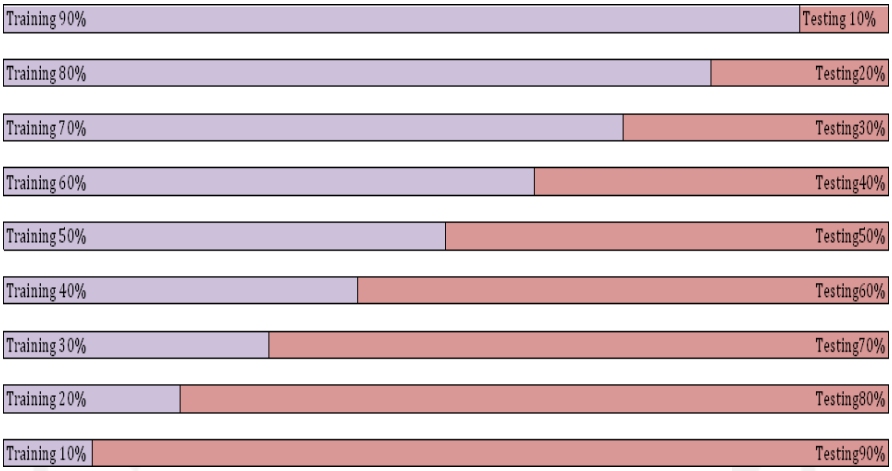
   Jika c = 25% (*default* untuk C4.5) maka z =0.69 (dari distribusi normal)

   r = nilai perbandingan *error rate*

   n = total kasus

3.4 *Split Validation*

Split Validation merupakan teknik validasi yang membagi data menjadi dua bagian, yaitu data *training* dan data *testing*. Dengan menggunakan *split validation* akan dilakukan proses *training* berdasarkan *splitratio* yang telah ditentukan sebelumnya, kemudian sisa dari *split ratio* data *training* akan dianggap sebagai data *testing*. Data *training* merupakan data latih yang akan dipakai dalam melakukan pembelajaran sedangkan data *testing* merupakan data yang belum pernah dipakai sebagai pembelajaran dan akan digunakan sebagai data pengujian kebenaran hasil pembelajaran.



Gambar 3.2 Ilustrasi *split validation*

3.5 *Imbalance Data*

Pengatasan *imbalance* data dalam melakukan penyebaran secara signifikan yaitu dengan melakukan metode *sampling* (Hulse & Khoshgoftaar, 2009). Beberapa teknik untuk mengatasi *class imbalance* seperti *oversampling* cenderung mengurangi jumlah pemangkasan yang terjadi, sedangkan *under-sampling* sering membuat pemangkasan yang tidak perlu (Drummond & Holte, 2003).

Metode *oversampling* adalah metode yang digunakan untuk menangani kelas minoritas dengan melakukan random kelas selama proses pengambilan sampel. Proses pengambilan sampel dengan metode oversampling ini yaitu dengan menduplikasi kelas positif dan dilakukan penyeimbangkan kelas secara acak. Sedangkan untuk metode *undersampling*, metode ini dilakukan dengan menghitung selisih kelas mayoritas dan kelas minoritas. Kemudian dilakukan perulangan sebanyak selisih kelas mayoritas dengan kelas minoritas. Selama proses perulangan dilakukan penghapusan terhadap kelas mayoritas sehingga didapatkan jumlah yang sama dengan kelas minoritas (Ganganwar, 2012).

3.6 Gagal Ginjal Kronik

Gagal ginjal merupakan suatu keadaan klinis yang ditandai dengan penurunan fungsi ginjal yang ireversibel pada suatu derajat dimana memerlukan terapi pengganti ginjal yang tetap, berupa dialisis 14 atau transplantasi ginjal (Rahman, dkk, 2013).

3.7 Klasifikasi Penyakit Gagal Ginjal Kronik

Menurut (Corwin, Bahasa, & Pendit, 2001)**,** gagal ginjal kronis selalu berkaitan dengan penurunan progresif *GFR (Glomerulo Filtration Rate).* Stadium-stadium gagal ginjal kronis yaitu berdasarkan pada tingkat *GFR* yang tersisa antara lain:

a. Penurunan cadangan ginjal, yang terjadi apabila *GFR* turun 50% dari normal.
b. Insufisiensi ginjal, yang terjadi apabila *GFR* turun menjadi 20-35% dari normal. Nefron-nefron yang tersisa sangat rentan mengalami kerusakan sendiri karena beratnya beban yang mereka terima.
c. Gagal ginjal, yang terjadi apabila *GFR* kurang dari 20% normal. Semakin banyak nefron yang mati.
d. Penyakit ginjal stadium-akhir, yang terjadi apabila *GFR* menjadi kurang dari 5% dari normal. Hanya sedikit nefron fungsional yang tersisa. Di seluruh ginjal ditemukan jaringan parut dan atrofi tubulus. Klasifikasi gagal ginjal kronis berdasarkan derajat *(stage)* LFG (Laju Filtrasi *Glomerolus*) dimana nilai normalnya adalah 125 ml/min/1,73 m2 .

3.8 Penyebab atau Etiologi Gagal Ginjal Kronis

Beberapa penyebab penyakit ginjal kronis adalah sebagai berikut:

a. Glomerulonefritis

Glomerulonefritis adalah inflamasi nefron, terutama pada *glomerulus*. *Glomerulonefritis* terbagi menjadi dua, yaitu *glomerulonefritis* akut dan *glomerulonefritis* kronis. *Glomerulonefritis* akut seringkali terjadi akibat respon imun terhadap toksin bakteri tertentu (kelompok *streptokokus* beta A) (Sloane, 2004).

b. *Pielonefritis* kronis

Pielonefritis adalah inflamasi ginjal dan pelvis ginjal akibat infeksi bakteri. Inflamasi dapat berawal di traktus urinaria bawah (kandung kemih) dan menyebar ke ureter, atau karena infeksi yang dibawa darah dan limfe ke ginjal. Obstruksi *kaktus urinaria* terjadi akibat pembesaran

kelenjar prostat, batu ginjal, atau defek kongenital yang memicu terjadinya pielonefritis (Sloane, 2004).

c. Batu ginjal

Batu ginjal atau kalkuli urinaria terbentuk dari pengendapan garam kalsium, magnesium, asam urat, atau sistein. Batu-batu kecil dapat mengalir bersama urine, batu yang lebih besar akan tersangkut dalam ureter dan menyebabkan rasa nyeri yang tajam (kolik ginjal) yang menyebar dari ginjal ke selangkangan (Sloane, 2004).

d. Penyakit polikistik ginjal

Penyakit ginjal polikistik ditandai dengan kista *multiple*, bilateral, dan berekspansi yang lambat laun mengganggu dan menghancurkan parenkim ginjal normal akibat penekanan (Price & Wilson, 2012)

3.9 Rekam Medis

Berdasarkan Permenkes 269 tahun 2008, rekam medis adalah berkas yang berisi catatan dan dokumen tentang pasien yang didalamnya terdapat identitas pasien, pemeriksaan, pengobatan, tindakan medis lain pada sarana pelayanan kesehatan untuk rawat jalan dan rawat inap baik dikelola pemerintah maupun swasta.

Menurut Depkes RI tahun 2002, rekam medis adalah keterangan yang tertulis maupun terekam tentang identitas, *anamnese*, penentuan fisik laboratorium, diagnose segala pelayanan dan tindakan medis yang diberikan kepada pasien, serta pengobatan baik rawat inap, rawat jalan, maupun yang mendapatkan pelayanan gawat darurat.

3.10 Tujuan Rekam Medis

Tujuan rekam Medis berdasarkan Hatta (1985) terdiri dari beberapa aspek diantaranya aspek administrasi, legal, finansial, riset, edukasi dan dokumentasi, yang dijelaskan sebagai berikut:

1. Aspek administrasi yaitu suatu berkas rekam medis yang mempunyai nilai administrasi karena isinya meyangkut tindakan berdasarkan wewenang dan

tanggung jawab sebagai tenaga medis dan paramedis dalam mencapai tujuan pelayanan kesehatan.

2. Aspek Medis yaitu suatu berkas rekam Medis mempunyai nilai Medis, karena catatan tersebut dipergunakan sebagai dasar untuk merencanakan pengobatan /perawatan yang harus diberikan seorang pasien.
3. Aspek Hukum yaitu suatu berkas rekam medis mempunyai nilai hukum karena isinya menyangkut masalah adanya jaminan kepastian hukum atas dasar keadilan, dalam rangka usaha menegakkan hukum serta penyediaan bahan bukti untuk menegakkan keadilan.
4. Aspek keuangan yaitu suatu berkas rekam medis mempunyai nilai uang karena isinya menyangkut data dan informasi yang dapat digunakan dalam menghitung biaya pengobatan/tindakan dan perawatan.
5. Aspek penelitian yaitu suatu berkas rekam medis mempunyai nilai penelitian, karena isinya menyangkut data/informasi yang dapat dipergunakan dalam penelitian dan pengembangan ilmu pengetahuan di bidang kesehatan.
6. Aspek pendidikan yaitu suatu berkas rekam medis mempunyai nilai pendidikan, karena isinya menyangkut data/informasi tentang perkembangan/ kronologis dan kegiatan pelayanan medis yang diberikan kepada pasien. Informasi tersebut dapat dipergunakan sebagai bahan/referensi pengajaran di bidang profesi kesehatan.
7. Aspek dokumentasi yaitu suatu berkas reka medis mempunyai nilai dokumentasi, karena isinya menyangkut sumber ingatan yang harus didokumentasikan dan dipakai sebagai bahan pertanggung jawaban dan laporan sarana pelayanan kesehatan.

3.11 Manfaat Rekam Medis

Manfaat rekam medis berdasarkan Permenkes Nomor 269/MenKes/Per/III/2008, tentang Rekam Medis adalah sebagai berikut:

1. Pengobatan. Rekam medis bermanfaat sebagai dasar dan petunjuk untuk merencanakan dan menganalisis penyakit serta merencanakan pengobatan, perawatan dan tindakan medis yang harus diberikan kepada pasien

2. Peningkatan Kualitas Pelayanan. Membuat Rekam Medis bagi penyelenggaraan praktik kedokteran dengan jelas dan lengkap akan meningkatkan kualitas pelayanan untuk melindungi tenaga medis dan untuk pencapaian kesehatan masyarakat yang optimal.
3. Pendidikan dan Penelitian. Rekam medis yang merupakan informasi perkembangan kronologis penyakit, pelayanan medis, pengobatan dan tindakan medis, bermanfaat untuk bahan informasi bagi perkembangan pengajaran dan penelitian di bidang profesi kedokteran dan kedokteran gigi.
4. Pembiayaan Berkas rekam medis dapat dijadikan petunjuk dan bahan untuk menetapkan pembiayaan dalam pelayanan kesehatan pada sarana kesehatan. Catatan tersebut dapat dipakai sebagai bukti pembiayaan kepada pasien.
5. Statistik Kesehatan Rekam medis dapat digunakan sebagai bahan statistik kesehatan, khususnya untuk mempelajari perkembangan kesehatan masyarakat dan untuk menentukan jumlah penderita pada penyakit-penyakit tertentu.
6. Pembuktian Masalah Hukum, Disiplin dan Etik Rekam medis merupakan alat bukti tertulis utama, sehingga bermanfaat dalam penyelesaian masalah hukum, disiplin dan etik.

BAB IV
METODOLOGI PENELITIAN

4.1 Populasi Penelitian

Data yang digunakan adalah data sekunder yaitu didapatkan dari bagian BPJS Kesehatan terkait data pasien pengguna BPJS Kesehatan pada tanggal 01 Juni 2020 hingga 31 Oktober 2020.

4.2 Tempat Penelitian

Penelitian ini dilakukan di Rumah Sakit Umum Daerah Dr.H. Abdul Moeloek Provinsi Lampung yaitu pada bagian BPJS Kesehatan.

4.3 Variabel Penelitian

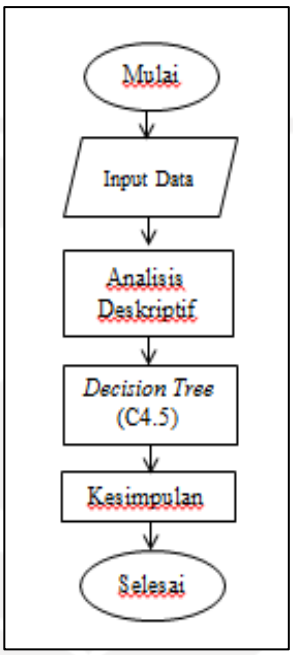
Berikut merupakan penjelasan operasional variabel peneletian yang digunakan.

Tabel 4.1 Definisi operasional variabel penelitian

	Variabel	Satuan	Definisi
Dependen	Y = CaraKeluar		Kondisi pasien ketika keluar dari rumah sakit. (Sembuh/Meninggal)
Independen	X_1= KelasRawat		Kelas perawatan pasien selama di rumah sakit. (1,2,3,V)
	X_2 = LamaRawat	Hari	Lama waktu pasien dirawat hingga keluar.
	X_3 = TipePenyakit		Tipe penyakit ginjal kronis (Ringan/Sedang/Berat).
	X_4 = JenisKelamin		Jenis kelamin pasien(Laki-laki/Perempuan)
	X_5= Umur		Umur pasien yang dirawat. (Anak-anak/Remaja/Dewasa/Lanjut Usia)

4.4 Metode Analisis Data

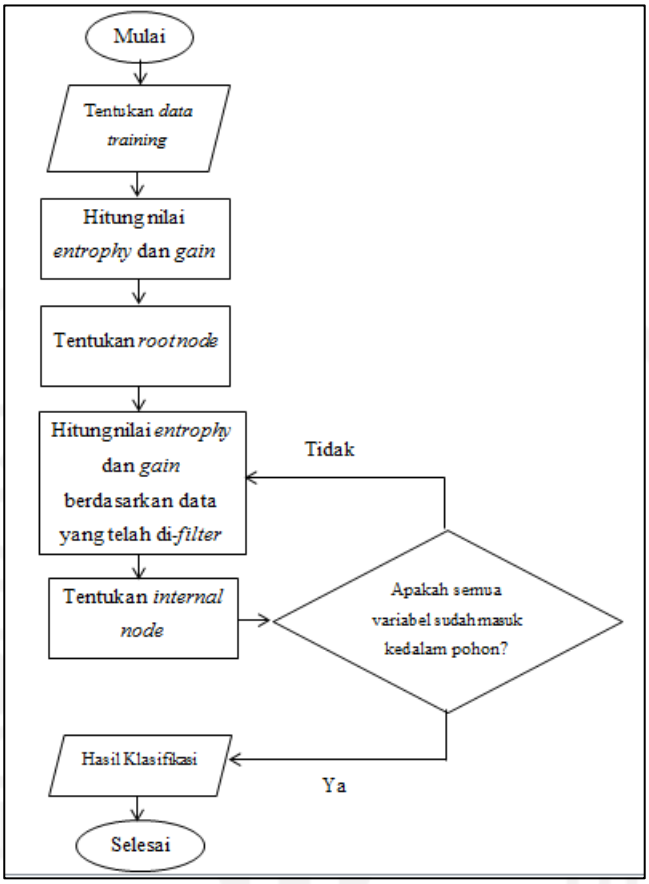
Untuk teknik pengolahan dan analisis data dalam penelitian ini, digambarkan dalam bentuk *flowchart* berikut:



Gambar 4.1 *Flowchart* penelitian

Tahapan analisis:

1. Analisis deskriptif : untuk menggambarkan karakteristik atau deskripsi dari setiap variabel penelitian. Analisis deskriptif akan digambarkan dalam bentuk *pie chart* dan tabel kontingensi.
2. Decision tree C4.5 : dijelaskan pada subbab 3.3.



Gambar 4.2 *Flowchart decision tree* C4.5

Tahapan analisis metode *decision tree* dan algoritma C4.5 secara matematis adalah sebagai berikut:

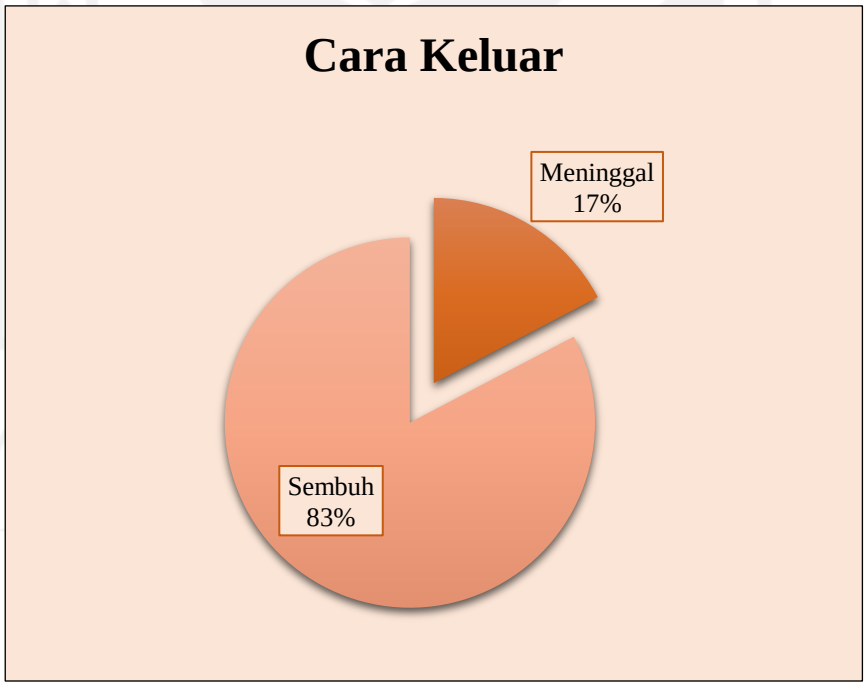
1. Menentukan *data training*.
2. Menghitung nilai *Entropy* dengan menggunakan persamaan (1).
3. Menghitung nilai *Gain* dengan menggunakan persamaan (2).
4. Menentukan *root node*.
5. Proses partisipasi pohon keputusan akan berhenti saat semua cabang dalam node N mendapat kelas yang sama.
6. Melakukan pengujian model untuk membangun struktur (pola) pohon keputusan melalui metode *decision tree* C4.5.
7. Menganalisa hasil klasifikasi menggunakan *confusion matrix*

BAB V
HASIL DAN PEMBAHASAN

Pada bab ini akan dijabarkan terkait hasil-hasil yang telah didapatkan dari hasil klasifikasi menggunakan metode *decision tree* c4.5. Pembahasan ini meliputi analisis deskriptif, klasifikasi menggunakan metode *decision tree* c4.5 dengan program R dan manual sebelum mengatasi *imbalance data* dan setelah mengatasi *imbalance data*.

5.1 Analisis Deskriptif

Penelitian ini terdapat lima variabel dengan satu variabel sebagai variabel dependen yaitu CaraKeluar dan empat variabel lainnya sebagai variabel independen yaitu TipePenyakit, KelasRawat, Umur, dan JenisKelamin. Berikut merupakan deskriptif untuk kelima variabel:



Gambar 5.1 Distribusi pasien berdasarkan cara keluar

Berdasarkan gambar 5.1 yang menunjukkan deskriptif dari variabel cara keluar (n = 358), mayoritas pasien keluar karena sembuh (83%) dan sekitar 17% pasien yang keluar karena meninggal.

Tabel 5.1 Tabel kontingensi variabel CaraKeluar, TipePenyakit dan KelasRawat

				CaraKeluar		Total
				Sembuh	Meninggal	
TipePenyakit	Berat	KelasRawat		25	9	**34**
			1	2		**2**
			2	2		**2**
			3	21	9	**30**
	Sedang			226	33	**259**
			1	44	6	**50**
			2	29	3	**32**
			3	152	23	**175**
			V	1	1	**2**
	Ringan			45	20	**65**
			1	13	1	**14**
			2	7	2	**9**
			3	25	17	**42**
Total				**296**	**62**	**358**

Pada tabel 5.1 menunjukkan tabel kontingensi untuk variabel CaraKeluar, Tipe Penyakit, dan KelasRawat. Dapat dilihat dari 358 pasien paling banyak pasien yang menderita penyakit ginjal kronis dengan cara keluar pasien karena sembuh yaitu pada tipe penyakit sedang, kelas pasien selama dirawat tipe 3 sebanyak 152 pasien. Begitu juga untuk pasien yang menderita penyakit ginjal kronis dengan cara keluar pasien karena meninggal, dari 62 pasien paling banyak yang meninggal yaitu pada tipe penyakit sedang, kelas pasien selama dirawat tipe 23 sebanyak 19 pasien.

Tabel 5.2 Tabel kontingensi variabel CaraKeluar, JenisKelamin dan KelasRawat

				CaraKeluar		Total
				Sembuh	Meninggal	
JenisKelamin	Laki - Laki	KelasRawat		148	28	**176**
			1	35	2	**37**
			2	15	4	**19**
			3	97	21	**118**
			V	1	1	**2**
	Perempuan			148	34	**182**

	1	24	5	**29**
	2	23	1	**24**
	3	101	28	**129**
Total		**296**	**62**	**358**

Berdasarkan tabel 5.2, dari 358 total pasien yang dirawat, sebagian besar pasien yang menderita penyakit ginjal kronis yaitu pada pasien dengan jenis kelamin perempuan, kelas pasien selama dirawat tipe 3, dan cara keluar pasien karena sembuh sebanyak 101 pasien. Kemudian disusul oleh jenis kelamin laki-laki, kelas pasien selama dirawat tipe 3, dan cara keluar pasien karena sembuh sebanyak 97 pasien yang dimana berarti pasien dengan jenis kelamin perempuan lebih banyak yang sembuh dibandingkan pasien dengan jenis kelamin laki-laki. Sedangkan pada variabel yang sama pula, namun dengan cara keluar karena meninggal, dapat dilihat bahwa lebih banyak pasien yang meninggal dengan jenis kelamin perempuan sebanyak 28 pasien dan kelas rawat tipe 3 dibandingkan dengan jenis kelamin laki-laki dan kelas rawat tipe 3 sebanyak 21 pasien.

Tabel 5.3 Tabel kontingensi variabel CaraKeluar, JenisKelamin dan Umur

			CaraKeluar		**Total**
			Sembuh	**Meninggal**	
JenisKelamin	**Laki - Laki**		148	28	**176**
		Umur			
		Dewasa	104	20	**124**
		Lanjut Usia	41	7	**48**
		Remaja	3	1	**4**
	Perempuan		148	34	**182**
		Anak-Anak	1		**1**
		Dewasa	112	22	**134**
		Lanjut Usia	32	11	**43**
		Remaja	3	1	**4**
Total			**296**	**62**	**358**

Tabel 5.3 menunjukkan bahwa dari 358 pasien yang dirawat, sebagian besar pasien yang menderita penyakit ginjal kronis yaitu pada tipe kategori umur dewasa dengan jenis kelamin laki-laki sebanyak 104 pasien dan perempuan sebanyak 112 pasien dan cara keluar pasien karena sembuh. Pada tipe kategori umur yang sama namun cara keluar pasien karena meninggal, dapat dilihat bahwa

pasien yang menderita ginjal kronis dengan jenis kelamin laki-laki sebanyak 20 pasien dan 22 pasien untuk jenis kelamin perempuan.

Tabel 5.4 Tabel kontingensi variabel CaraKeluar, TipePenyakit, dan Umur

				CaraKeluar		Total
				Sembuh	Meninggal	
TipePenyakit	Berat			25	9	**34**
		Umur	Dewasa	21	8	**29**
			Lanjut Usia	4	1	**5**
	Sedang			226	33	**259**
			Anak-Anak	1		**1**
			Dewasa	163	20	**183**
			Lanjut Usia	56	11	**67**
			Remaja	6	2	**8**
	Ringan			45	20	**65**
			Dewasa	32	14	**46**
			Lanjut Usia	13	6	**19**
Total				**296**	**62**	**358**

Pada tabel 5.4 menunjukkan tabel kontingensi untuk variabel CaraKeluar, Umur, dan Tipe Penyakit. Dapat dilihat dari 358 pasien yang dirawat, paling banyak pasien yang menderita penyakit ginjal kronis yaitu pada pasien tipe kategori umur dewasa, tipe penyakit sedang, dan cara keluar karena sembuh sebanyak 163 pasien. Pada variabel yang sama namun untuk cara keluar pasien karena meninggal ada sebanyak 20 pasien.

Tabel 5.5 Tabel kontingensi variabel CaraKeluar, TipePenyakit dan JenisKelamin

				CaraKeluar		Total
				Sembuh	Meninggal	
TipePenyakit	Berat			25	9	**34**
		JenisKelamin	Laki - Laki	13	4	**17**
			Perempuan	12	5	**17**
	Sedang			226	33	**259**
			Laki - Laki	106	13	**119**
			Perempuan	120	20	**140**
	Ringan			45	20	**65**
			Laki - Laki	29	11	**40**
			Perempuan	16	9	**25**
Total				**296**	**62**	**358**

Tabel 5.5 menunjukkan bahwa dari 358 pasien yang menderita penyakit ginjal kronis yaitu pada tipe penyakit sedang sebanyak 226 pasien yang diantaranya dengan cara keluar pasien karena sembuh dengan jenis kelamin perempuan sebanyak 120 pasien dan jenis kelamin laki-laki sebanyak 106 pasien. Sedangkan pasien dengan cara keluar karena meninggal yang berjenis kelamin perempuan terdapat 20 pasien dan 13 pasien berjenis kelamin laki-laki.

Dibawah ini merupakan uji hipotesis untuk mengetahui apakah terdapat hubungan antara masing-masing variabel independen dengan variabel dependen.

a. Hipotesis

   H_0 : Tidak ada hubungan antara variabel CaraKeluar dengan KelasRawat

   H_1 : Ada hubungan antara variabel CaraKeluar dengan KelasRawat

   H_0 : Tidak ada hubungan antara variabel CaraKeluar dengan LamaRawat

   H_1 : Ada hubungan antara variabel CaraKeluar dengan LamaRawat

   H_0 : Tidak ada hubungan antara variabel CaraKeluar dengan TipePenyakit

   H_1 : Ada hubungan antara variabel CaraKeluar dengan TipePenyakit

   H_0 : Tidak ada hubungan antara variabel CaraKeluar dengan JenisKelamin

   H_1 : Ada hubungan antara variabel CaraKeluar dengan JenisKelamin

   H_0 : Tidak ada hubungan antara variabel CaraKeluar dengan Umur

   H_1 : Ada hubungan antara variabel CaraKeluar dengan Umur

b. Tingkat Signifikasi

   Dengan *confident interval* 95% didapatkkan $\alpha = 5\%$ / $\alpha = 0.05$

c. Daerah Kritis

   Tolak H_0 jika *p-value* < α

d. Statistik Uji

Tabel 5.6 Hasil statistik uji antara variabel CaraKeluar dengan variabel KelasRawat, LamaRawat, TipePenyakit, JenisKelamin, dan Umur

Variabel Independen	***P-Value***
KelasRawat	0.131
LamaRawat	0.000
TipePenyakit	0.001
JenisKelamin	0.488
Umur	0.773

e. Keputusan

Tabel 5.7 Hasil keputusan antara variabel CaraKeluar dengan variabel KelasRawat, LamaRawat, TipePenyakit, JenisKelamin, dan Umur

Variabel Independen	*P-Value*		*Alpha (α)*	Keputusan
KelasRawat	0.131	>	0.05	Gagal Tolak Ho
LamaRawat	0.000	<		Tolak Ho
TipePenyakit	0.001	<		Tolak Ho
JenisKelamin	0.488	>		Gagal Tolak Ho
Umur	0.773	>		Gagal Tolak Ho

f. Kesimpulan

Tabel 5.8 Kesimpulan antara variabel CaraKeluar dengan variabel KelasRawat, LamaRawat, TipePenyakit, JenisKelamin, dan Umur

Variabel Independen	Kesimpulan
KelasRawat	Tidak ada hubungan antara variabel CaraKeluar dengan KelasRawat.
LamaRawat	Ada hubungan antara variabel CaraKeluar dengan LamaRawat.
TipePenyakit	Ada hubungan antara variabel CaraKeluar dengan TipePenyakit.
JenisKelamin	Tidak ada hubungan antara variabel CaraKeluar dengan JenisKelamin.
Umur	Tidak ada hubungan antara variabel CaraKeluar dengan Umur.

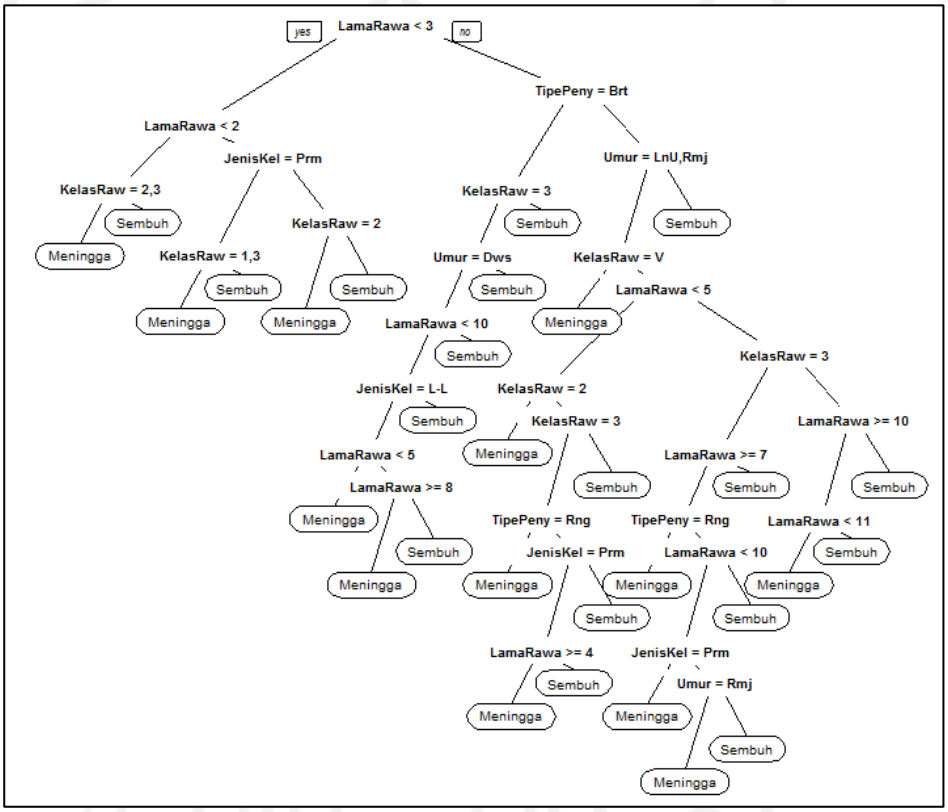
5.2 *Decision Tree* C4.5

Klasifikasi dengan menggunakan beberapa metode *machine learning* bertujuan untuk mendapatkan target kelas yang akurat. Namun ketika pada saat melakukan proses klasifikasi terkadang muncul permasalahan dimana salah satu kelas memiliki jumlah data yang jauh lebih kecil pada *data training*-nya atau dapat disebut *imbalance data*. Salah satu cara populer untuk mengatasi *imbalance data* yaitu dengan melakukan metode *sampling* dengan cara menyeimbangkan data. Terdapat tiga metode *sampling* yaitu *oversampling*, *undersampling*, dan *combine*. Dibawah ini terdapat penjelasan secara rinci untuk hasil klasifikasi menggunakan metode *decision tree* c4.5 tanpa mengatasi *imbalance data* (sebelum dilakukan metode *sampling*) dan terdapat perbandingan pengukuran data hasil klasifikasi sebelum dan setelah dilakukan *metode sampling*.

5.2.1 *Decision Tree C4.5* Menggunakan Program R

Pada saat membuat pohon keputusan, banyak cabang yang terdapat *noise* sehingga terjadi adanya *overfitting*. *Overfitting* dapat diatasi dengan melakukan pemangkasan pohon/*post-prunning*. Dibawah ini terdapat hasil pohon keputusan sebelum dan setelah melakukan *post-prunning* menggunakan Program R dengan sebuah data *train* dan data *test* dimana terbagi kerangka data yang dimiliki menjadi 287 data untuk data *train* dan 71 data untuk data *test.*

Dibawah ini adalah hasil pembentukan pohon keputusan sebelum melakukan proses *post-prunning*.



Gambar 5.2 Hasil pohon keputusan sebelum *post-prunning*

Pembentukan pohon dilakukan dengan menghitung nilai *enthropy* dan *gain* kemudian dibentuk *root note*, pada gambar 5.1 didapatkan variabel LamaRawat < 3 Hari sebagai *root note* kemudian diikuti dengan variabel lain sebagai note selanjutnya. Berikut merupakan evaluasi dengan menganalisa hasil klasifikasi menggunakan *confusion matrix* :

Tabel 5.9 Hasil *confusion matrix* data training

	Data Aktual

Hasil Prediksi	Aktual : Meninggal	Aktual : Sembuh
Prediksi : Meninggal	TP = 32	FP = 9
Prediksi : Sembuh	FN = 18	TN = 228

Berdasarkan output diatas dapat diketahui bahwa:

- Frekuensi prediksi cara keluar karena meninggal dan data aktual sesuai atau benar karena meninggal adalah sebanyak 32. Keadaan seperti ini disebut juga sebagai *true positive.*
- Frekuensi prediksi cara keluar karena sembuh namun data aktual menunjukkan cara keluar karena meninggal sebanyak 18. Kasus seperti ini disebut sebagai *false positive.*
- Frekuensi prediksi cara keluar karena meninggal namun data aktual menunjukkan cara keluar karena sembuh sebanyak 9. Keadaan seperti ini disebut sebagai *false negative*.
- Frekuensi prediksi cara keluar karena sembuh dan data aktual sesuai atau benar karena sembuh adalah sebanyak 228. Keadaan seperti ini disebut juga sebagai *true negative.*

Perhitungan dari *confusion matrix* data *training*:

- *Accuracy* : $\frac{(\text{TP} + \text{TN})}{\text{Total}} = \frac{(32+228)}{287} = 0.91 = 91\%$.

  Yang berarti tingkat keakuratan klasifikasi sebesar 91%.
- *Misclassification Rate* : $\frac{(\text{FP} + \text{FN})}{\text{Total}} = \frac{(9+18)}{287} = 0.09 = 9\%$
- *True Positive Rate/Recall/Sensitivity* : $\frac{\text{TP}}{Actual\ yes} = \frac{\text{TP}}{(\text{TP} + \text{FN})} = \frac{32}{32+18} = 0.64 = 64\%$

  Yang berarti jika sebenarnya *no*/meninggal, maka tingkat seberapa sering diprediksi *no*/meninggal sebesar 64%.
- *False Positive Rate* : $\frac{\text{FP}}{Actual\ no} = \frac{\text{FP}}{(\text{TN}+\text{FP})} = \frac{9}{228+9} = 0.04 = 4\%$

  Yang berarti jika sebenarnya yes/sembuh, maka tingkat seberapa sering diprediksi *no*/meninggal sebesar 4%.
- *False Neative Rate* : $\frac{\text{FN}}{Actual\ yes} = \frac{\text{FN}}{(\text{TP}+\text{FN})} = \frac{18}{32+18} = 0.36 = 36\%$

Yang berarti jika sebenarnya *no*/meninggal, maka tingkat seberapa sering diprediksi yes/sembuh sebesar 36%.

- *True Negative Rate/ Specificity* : $\frac{\text{TN}}{Actual\ no} = \frac{\text{TN}}{(\text{TN+FP})} = \frac{228}{228+9} = 0.96 = 96\%$

  Yang berarti jika sebenarnya yes/sembuh, maka tingkat seberapa sering diprediksi yes/sembuh sebesar 96%.

- *Precision* : $\frac{\text{TP}}{\text{Prediksi yes/sembuh}} = \frac{\text{TP}}{(\text{TP+FP})} = \frac{32}{32+9} = 0.78 = 78\%$

  Yang berarti jika memprediksi *no*/meninggal, maka tingkat kebenaran sebesar 78%.

- *Prevelance* : $\frac{Actual\ yes}{\text{Total}} = \frac{(\text{TP+FN})}{\text{Total}} = \frac{32+18}{287} = 0.17 = 17\%$

  Yang berarti seberapa sering kondisi *yes*/sembuh benar-benar terjadi.

Pada data test didapatkan dengan cara pengambilan sampel secara acak tanpa pengembalian, hal ini dilakukan untuk mengetahui peramalan dari hasil klasifikasi. Setelah mendapatkan hasilnya, dibawah ini terdapat hasil peramalan menggunakan data test yaitu sebesar 71 data. Untuk peramalan hasil akhirnya akan berupa sembuh atau meninggal dan terdapat tampilan tabel klasifikasi untuk mengetahui jumlah error yang dihasilkan menggunakan data test.

Tabel 5.10 Hasil *confusion matrix* data uji/test

	Data Aktual	
Hasil Prediksi	**Aktual : Meninggal**	**Aktual : Sembuh**
Prediksi : Meninggal	TP = 5	FP = 4
Prediksi : Sembuh	FN = 7	TN = 55

Berdasarkan output diatas dapat diketahui bahwa:

- Frekuensi prediksi cara keluar karena meninggal dan data aktual sesuai atau benar karena meninggal adalah sebanyak 5. Keadaan seperti ini disebut juga sebagai *true positive.*
- Frekuensi prediksi cara keluar karena sembuh namun data aktual menunjukkan cara keluar karena meninggal sebanyak 7. Kasus seperti ini disebut sebagai *false positive.*
- Frekuensi prediksi cara keluar karena meninggal namun data aktual menunjukkan cara keluar karena sembuh sebanyak 4. Keadaan seperti ini disebut sebagai *false negative*.

- Frekuensi prediksi cara keluar karena sembuh dan data aktual sesuai atau benar karena sembuh adalah sebanyak 55. Keadaan seperti ini disebut juga sebagai *true negative*.

Dibawah ini terdapat tabel perbandingan *confusion matrix* data *training* dan uji/*test*:

Tabel 5.11 Perbandingan dari *confusion matrix* data *training* dan uji/*test*

	Training	***Testing***
Accuracy	91%	85%
Misclassification Rate	9%	15%
True Positive Rate/ Recall/Sensitivity	64%	42%
False Positive Rate	4%	7%
False Negative Rate	36%	58%
True Negative Rate/ Specificity	96%	93%
Precision	78%	56%
Prevelance	17%	17%

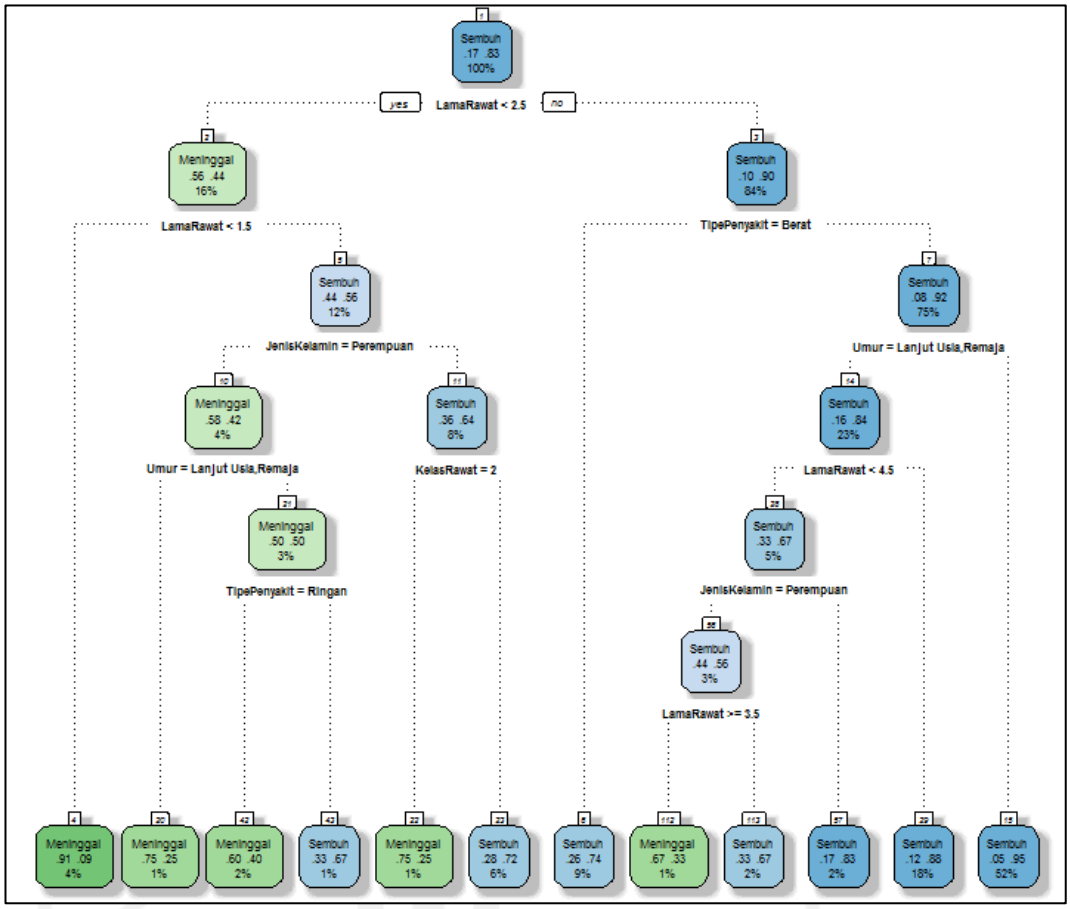
Kemudian dilakukan metode *post-prunning* untuk memangkas pohon.

Berikut hasil pembentukan pohon keputusan setelah melakukan proses *post-prunning* :

Tabel 5.12 Ringkasan dari model *decision tree*

Node Number	**Total Obser-vation**	**Predicted Class**	**Expected Loss**	**P(node)**	**Class count**	**Probabilitas** $\boldsymbol{P = \frac{n(kelas)}{n(total)}}$
1	287	Sembuh	0.174	1	50	0.174
					237	0.826
2	45	Meninggal	0.444	0.157	25	0.556
					20	0.444
3	242	Sembuh	0.103	0.843	25	0.103
					217	0.897
4	11	Meninggal	0.091	0.038	10	0.909
					1	0.091
5	34	Sembuh	0.441	0.118	15	0.441
					19	0.559
6	27	Sembuh	0.259	0.094	7	0.259
					20	0.741
7	215	Sembuh	0.084	0.749	18	0.084
					197	0.916
10	12	Meninggal	0.417	0.042	7	0.583
					5	0.417
11	22	Sembuh	0.364	0.077	8	0.364
					14	0.636
14	67	Sembuh	0.164	0.233	11	0.164

Node Number	Total Obser-vation	Predicted Class	Expected Loss	P(node)	Class count	Probabilitas $P = \frac{n(kelas)}{n(total)}$
					56	0.836
15	148	Sembuh	0.047	0.516	7	0.047
					141	0.953
20	4	Meninggal	0.25	0.014	3	0.750
					1	0.250
21	8	Meninggal	0.5	0.028	4	0.5
					4	0.5
22	4	Meninggal	0.25	0.014	3	0.750
					1	0.250
23	18	Sembuh	0.278	0.063	5	0.278
					13	0.722
28	15	Sembuh	0.333	0.052	5	0.333
					10	0.667
29	52	Sembuh	0.115	0.181	6	0.115
					46	0.885
42	5	Meninggal	0.4	0.017	3	0.600
					2	0.400
43	3	Sembuh	0.333	0.010	1	0.333
					2	0.667
56	9	Sembuh	0.444	0.031	4	0.444
					5	0.556
57	6	Sembuh	0.167	0.021	1	0.167
					5	0.833
112	3	Meninggal	0.333	0.105	2	0.667
					1	0.333
113	6	Sembuh	0.333	0.021	2	0.333
					4	0.667



Gambar 5.3 Hasil pohon keputusan setelah *post-prunning*

Interpretasi dari hasil *decision tree*:

1. Jika lama rawatnya < 3 hari maka akan lanjut ke langkah dua dengan 16% dari seluruh data dan jika lama rawatnya >= 3 hari maka akan lanjut ke langkah tiga dengan 84% dari seluruh data.
2. Jika lama rawat < 2 hari maka langsung dinyatakan meninggal, jika lama rawat >= 2 hari maka akan lanjut ke langkah empat. 4% dari data dinyatakan meninggal dan jumlah data yang diproses ke langkah empat yaitu 12% dari data.
3. Jika tipe penyakit berat maka langsung dinyatakan sembuh (terdapat 9% dari data) dan jika tipe penyakit sedang dan ringan akan lanjut ke langkah lima dengan 75% dari data untuk data yang akan diproses kelangkah lima.

4. Jika jenis kelaminnya perempuan maka akan lanjut ke langkah enam dengan 4% dari seluruh data dan jika jenis kelamin laki-laki maka akan lanjut ke langkah tujuh dengan 8% dari seluruh data.
5. Jika lanjut usia dan remaja maka akan lanjut ke langkah delapan (terdapat 23% dari data), sedangkan jika anak-anak dan dewasa langsung dinyatakan sembuh dengan 52% dari data.
6. Jika lanjut usia dan remaja maka langsung dinyatakan meninggal dengan 1% dari data dan jika anak-anak dan dewasa maka akan lanjut kelangkah sembilan dengan 3% dari data.
7. Jika kelas rawat 2 maka langsung dinyatakan meninggal dengan 1% dari data dan jika kelas rawat 1,3,4, dan V maka langsung dinyatakan sembuh dengan 6% dari data.
8. Jika lama rawat < 5 hari maka akan lanjut kelangkah sepuluh, tetapi jika lama rawat >= 5 hari maka langsung dinyatakan sembuh (terdapat 18% dari data).
9. Jika tipe penyakit ringan langsung dinyatakan meninggal dengan 2% dari data dan jika berat dan sedang maka akan dinyatakan sembuh dengan 1% dari data.
10. Jika jenis kelamin perempuan maka akan lanjut ke langkah sebelas dan jika jenis kelamin laki-laki langsung dinyatakan sembuh dengan 2% dari data.
11. Jika lama rawat >= 4 hari maka langsung dinyatakan meninggal dengan 1% dari data dan jika lama rawat < 4 hari maka langsung dinyatakan sembuh (terdapat 2 % dari data).

Tabel 5.13 Hasil *confusion matrix* data training

	Data Aktual	
Hasil Prediksi	**Aktual : Meninggal**	**Aktual : Sembuh**
Prediksi : Meninggal	TP = 21	FP = 6
Prediksi : Sembuh	FN = 29	TN = 231

Berdasarkan output diatas dapat diketahui bahwa:

- Frekuensi prediksi cara keluar karena meninggal dan data aktual sesuai atau benar karena meninggal adalah sebanyak 21. Keadaan seperti ini disebut juga sebagai *true positive.*
- Frekuensi prediksi cara keluar karena sembuh namun data aktual menunjukkan cara keluar karena meninggal sebanyak 29. Kasus seperti ini disebut sebagai *false positive.*
- Frekuensi prediksi cara keluar karena meninggal namun data aktual menunjukkan cara keluar karena sembuh sebanyak 6. Keadaan seperti ini disebut sebagai *false negative*.
- Frekuensi prediksi cara keluar karena sembuh dan data aktual sesuai atau benar karena sembuh adalah sebanyak 231. Keadaan seperti ini disebut juga sebagai *true negative.*

Perhitungan dari *confusion matrix* data *training*:

- *Accuracy* : $\frac{(\mathrm{TP}+\mathrm{TN})}{\mathrm{Total}} = \frac{(21+231)}{287} = 0.88 = 88\%$.

  Yang berarti tingkat keakuratan klasifikasi sebesar 88%.
- *Misclassification Rate* : $\frac{(\mathrm{FP}+\mathrm{FN})}{\mathrm{Total}} = \frac{(29+6)}{287} = 0.12 = 12\%$
- *True Positive Rate/Recall/Sensitivity* : $\frac{\mathrm{TP}}{Actual\ yes} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} = \frac{21}{21+29} = 0.42 = 42\%$

  Yang berarti jika sebenarnya *no*/meninggal, maka tingkat seberapa sering diprediksi *no*/meninggal sebesar 42%.
- *False Positive Rate* : $\frac{\mathrm{FP}}{Actual\ no} = \frac{\mathrm{FP}}{(\mathrm{TN}+\mathrm{FP})} = \frac{6}{231+6} = 0.03 = 3\%$

  Yang berarti jika sebenarnya yes/sembuh, maka tingkat seberapa sering diprediksi *no*/meninggal sebesar 3%.
- *False Neative Rate* : $\frac{\mathrm{FN}}{Actual\ yes} = \frac{\mathrm{FN}}{(\mathrm{TP}+\mathrm{FN})} = \frac{29}{21+29} = 0.58 = 58\%$

  Yang berarti jika sebenarnya *no*/meninggal, maka tingkat seberapa sering diprediksi yes/sembuh sebesar 58%.
- *True Negative Rate/ Specificity* : $\frac{\mathrm{TN}}{Actual\ no} = \frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FP})} = \frac{231}{231+6} = 0.97 = 97\%$

  Yang berarti jika sebenarnya yes/sembuh, maka tingkat seberapa sering diprediksi yes/sembuh sebesar 97%.

- *Precision* : $\frac{\text{TP}}{\text{Prediksi yes/sembuh}} = \frac{\text{TP}}{(\text{TP+FP})} = \frac{21}{21+6} = 0.78 = 78\%$

  Yang berarti jika memprediksi *no*/meninggal, maka tingkat kebenaran sebesar 78%.

- *Prevelance* : $\frac{Actual\ yes}{\text{Total}} = \frac{(\text{TP+FN})}{\text{Total}} = \frac{21+29}{287} = 0.17 = 17\%$

  Yang berarti seberapa sering kondisi *yes*/sembuh benar-benar terjadi.

Klasifikasi hasil data testing:

Tabel 5.14 Hasil *confusion matrix* data uji/test

	Data Aktual	
Hasil Prediksi	**Aktual : Meninggal**	**Aktual : Sembuh**
Prediksi : Meninggal	TP = 5	FP = 1
Prediksi : Sembuh	FN = 7	TN = 58

Berdasarkan output diatas dapat diketahui bahwa:

- Frekuensi prediksi cara keluar karena meninggal dan data aktual sesuai atau benar karena meninggal adalah sebanyak 5. Keadaan seperti ini disebut juga sebagai *true positive.*
- Frekuensi prediksi cara keluar karena sembuh namun data aktual menunjukkan cara keluar karena meninggal sebanyak 1. Kasus seperti ini disebut sebagai *false positive.*
- Frekuensi prediksi cara keluar karena meninggal namun data aktual menunjukkan cara keluar karena sembuh sebanyak 7. Keadaan seperti ini disebut sebagai *false negative*.
- Frekuensi prediksi cara keluar karena sembuh dan data aktual sesuai atau benar karena sembuh adalah sebanyak 58. Keadaan seperti ini disebut juga sebagai *true negative.*

Dibawah ini terdapat tabel perbandingan *confusion matrix* data *training* dan uji/*testing*:

Tabel 5.15 Perbandingan dari *confusion matrix* data *training* dan uji/*test*

	Training	***Testing***
Accuracy	88%	89%
Misclassification Rate	12%	11%
True Positive Rate/ Recall/Sensitivity	42%	42%
False Positive Rate	3%	2%
False Negative Rate	58%	58%

True Negative Rate/ Specificity	97%	98%
Precision	78%	83%
Prevelance	17%	17%

Jika dilihat dari tabel 5.8 dan 5.12, lebih tinggi nilai akurasi sebelum post-prunning sebesar 3% untuk training, dan sebaliknya untuk testing, lebih tinggi nilai akurasi setelah post-prunning sebesar 4%.

5.2.2 *Decision Tree C4.5* Perhitungan Manual

1. Menentukan node 1, pertama-tama hitung nilai entropy semua variabel. Dibawah ini terdapat contoh perhitungan *entropy* dari salah dua variabel pada data rekam medis yaitu variabel TipePenyakit dan LamaRawat.

a. Seluruh Data

Diketahui :

Tabel 5.16 Total Seluruh Data

Total	**Total**	**Sembuh**	**Meninggal**
	287	237	50

$$Entropy(\mathrm{S}) = \sum_{i=1}^{n} - p_i * \log_2 p_i$$

$$= \left(-\frac{237}{287} * log_2\left(\frac{237}{287}\right)\right) + \left(-\frac{50}{287} * log_2\left(\frac{50}{287}\right)\right) = 0.67$$

b. Tipe Penyakit

Diketahui bahwa variabel TipePenyakit termasuk kedalam variabel kategori. Berikut contoh perhitungan *entropy* variabel TipePenyakit :

Diketahui :

Tabel 5.17 Total Tipe Penyakit

Total	**Total**	**Sembuh**	**Meninggal**
TipePenyakit			
Tipe Berat	29	21	8
Tipe Sedang	203	176	27
Tipe Ringan	55	40	15

- **Tipe Berat**

$$Entropy(\mathrm{S}) = \sum_{i=1}^{n} - p_i * \log_2 p_i$$

$$= \left(-\frac{21}{29} * log_2\left(\frac{21}{29}\right)\right) + \left(-\frac{8}{29} * log_2\left(\frac{8}{29}\right)\right) = 0.85$$

- **Tipe Sedang**

$$Entropy(\mathrm{S}) = \sum_{i=1}^{n} - p_i * \log_2 p_i$$

$$= \left(-\frac{176}{203} * log_2\left(\frac{176}{203}\right)\right) + \left(-\frac{27}{203} * log_2\left(\frac{27}{203}\right)\right) = 0.57$$

- **Tipe Ringan**

$$Entropy(\mathrm{S}) = \sum_{i=1}^{n} - p_i * \log_2 p_i$$

$$= \left(-\frac{40}{55} * log_2\left(\frac{40}{55}\right)\right) + \left(-\frac{15}{55} * log_2\left(\frac{15}{55}\right)\right) = 0.85$$

c. LamaRawat

Diketahui bahwa variabel LamaRawat termasuk kedalam variabel numerik. Dalam algoritma *decision tree* C4.5 perlu mengelompokkan data diskrit/kontinu dengan mengurutkan nilai dari terkecil – terbesar (secara *asscending*). Pada variabel Lamarawat, terdapat 13 kelompok yang akan menjadi partisi, yaitu pada lama rawat 2 hari, 3 hari hingga 26 hari. Data terkecil yaitu 1 hari dan data terbesar yaitu 16 hari. Perpartisi akan terdapat dua kelompok, sebagaimana contoh untuk partisi 2 hari yaitu < 2 hari dan >=2 hari. 1 hari tidak dimasukkan dalam partisi karena tidak ada data yang < 1 hari sehingga partisi dimulai dari 2 hari.

Dibawah ini terdapat perhitungan *entropy* variabel LamaRawat untuk partisi 3 hari.

Diketahui:

Tabel 5.18 Total Lama Rawat

	Total	**Sembuh**	**Meninggal**
<3	45	20	25
>=3	242	217	25

- **< 3 Hari**

$$Entropy(\mathrm{S}) = \sum_{i=1}^{n} - p_i * \log_2 p_i$$

$$= \left(-\frac{20}{45} * log_2\left(\frac{20}{45}\right)\right) + \left(-\frac{25}{45} * log_2\left(\frac{25}{45}\right)\right) = 0.99$$

- **>= 3 Hari**

$$Entropy(\mathrm{S}) = \sum_{i=1}^{n} - p_i * \log_2 p_i$$

$$= \left(-\frac{217}{242} * log_2\left(\frac{217}{242}\right)\right) + \left(-\frac{25}{242} * log_2\left(\frac{25}{242}\right)\right) = 0.48$$

2. Hitung *Gain.*

a. Tipe Penyakit

Berikut perhitungan *gain* variabel TipePenyakit:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

$$= 0.67 - \left(\left(\frac{29}{287} * 0.85\right) + \left(\frac{203}{287} * 0.57\right) + \left(\frac{55}{287} * 0.85\right)\right) = 0.02$$

b. LamaRawat Partisi 3 Hari

Berikut perhitungan *gain* variabel LamaRawat Partisi 3 Hari:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

$$= 0.67 - \left(\left(\frac{45}{287} * 0.99\right) + \left(\frac{242}{287} * 0.48\right)\right) = 0.11$$

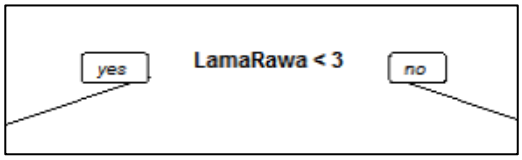
Berikut merupakan hasil perhitungan untuk Node 1.

Tabel 5.19 Hasil perhitungan *node 1*

		Total	Sembuh	Meninggal	Entropy	Gain
Node 1	**Total**	287	237	50	0.67	
	TipePenyakit					
	Berat	29	21	8	0.85	0.02
	Sedang	203	176	27	0.57	
	Ringan	55	40	15	0.85	
	LamaRawat					
	<2	11	1	10	0.44	0.08
	>=2	276	236	40	0.60	
	<3	45	20	25	0.99	0.11
	>=3	242	217	25	0.48	
	<4	94	62	32	0.93	0.06
	>=4	193	175	18	0.45	
	<5	134	98	36	0.84	0.04

	>=5	153	139	14	0.44	
	<6	198	158	40	0.73	0.01
	>=6	89	79	10	0.51	
	<7	224	181	43	0.71	0.01
	>=7	63	56	7	0.50	
	<8	248	204	44	0.67	0.00
	>=8	39	33	6	0.62	
	<9	261	216	45	0.66	0.00
	>=9	26	21	5	0.71	
	<10	270	222	48	0.68	0.00
	>=10	17	15	2	0.52	
	<11	277	228	49	0.67	0.00
	>=11	10	9	1	0.47	
	<12	283	233	50	0.67	-
	>=12	4	4		-	
	<13	285	235	50	0.67	-
	>=13	2	2		-	
	<16	286	236	50	0.67	-
	>=16	1	1		-	
	KelasRawat					
	1	49	45	4	0.41	0.01
	2	34	29	5	0.60	
	3	202	162	40	0.72	
	V	2	1	1	1.00	
	Umur					
	Anak-Anak	1	1		-	0.00
	Dewasa	203	171	32	0.63	
	Lanjut Usia	75	59	16	0.75	
	Remaja	8	6	2	0.81	
	JenisKelamin					
	Laki - Laki	142	119	23	0.64	0.00
	Perempuan	145	118	27	0.69	

Setelah mendapatkan nilai *gain* dari seluruh partisi, didapatkan bahwa variabel LamaRawat partisi 3 hari memiliki nilai *gain* terbesar dari seluruh variabel yang ada yaitu 0.11. Sehingga variabel ini akan menjadi Node 1.



Gambar 5.4 *Output* Node 1

Jumlah kasus variabel LamaRawat (<3 hari) adalah 25 yang meninggal dan 20 yang sembuh sehingga banyaknya pasien yang meninggal dengan LamaRawat < 3 Hari lebih besar dibandingkan yang sembuh.

Perhitungan untuk node berikutnya sama dengan perhitungan untuk menentukan Node 1 namun terdapat perbedaan pada data. Untuk perhitungan selanjutnya yaitu mem-*filter* data berdasarkan nilai variabel pada node 1 (LamaRawat 3 Hari). Pada akar *yes,* menggunakan 45 data dimana berisi data pasien yang dirawat selama < 3 hari dan pada akar *no*, menggunakan 242 data dimana berisi data pasien yang dirawat selama >= 3 hari. Proses partisipasi pohon keputusan akan berhenti saat semua cabang dalam node N mendapat kelas yang sama. Dapat dikatakan memiliki kelas yang sama yaitu dengan dilihat apakah dalam variabel tersebut tidak ada pasien yang meninggal atau tidak ada pasien yang sembuh.

5.2.3 Pengatasan *Imbalance Data*

Sebagaimana bahwa *imbalance data* merupakan kondisi dimana salah satu atau lebih dari kelas yg ada memiliki jumlah sampel yang cukup timpang diantara kelas lainnya. Sehingga perlu dilakukan pengatasan terhadap *imbalance data* yaitu dengan melakukan beberapa metode. Terdapat tiga metode *sampling* untuk mengatasi *imbalance data* yaitu over-*sampling, under-sampling* dan *combine*.

Pada metode *oversampling*, dapat dilihat kelas mana yang memiliki sampel terbanyak. Diketahui kelas sembuh memiliki 237 sampel/pasien dan meninggal sebanyak 50 sampel/pasien. Untuk menentukan *sampling* yang akan digunakan yaitu dengan mengkalikan kelas yang memiliki sampel terbesar dan banyaknya kelas (2 kelas). Sedangkan untuk metode *undersampling,* dapat dilihat kelas mana yang memiliki sampel terkecil dan untuk metode *combine*, menggunakan seluruh sampel.

Dibawah ini merupakan perbandingan antar metode *sampling* berdasarkan hasil pengukuran akurasi menggunakan *decision tree* c4.5 sebagai algoritma klasifikasi.

Tabel 5.20 Perbandingan dari *confusion matrix* data *training* dan uji/*test*

	Without sampling	***Oversampling***	***Undersampling***	***Combine***
Accuracy	91%	89%	86%	89%
Misclassification Rate	9%	11%	14%	11%
True Positive Rate/ Recall/Sensitivity	64%	88%	82%	84%
False Positive Rate	4%	9%	10%	6%
False Negative Rate	36%	12%	18%	16%
True Negative Rate/ Specificity	96%	91%	90%	94%
Precision	78%	91%	90%	94%
Prevelance	17%	50%	59	51%

Berdasarkan nilai akurasi pada tabel 5.20 dapat dilihat bahwa tidak ada perbedaan yang signifikan antara nilai akurasi sebelum mengatasi *imbalance data* sebesar 91% dan setelah mengatasi *imbalance data* sebesar 89% untuk *oversampling* dan *combine*. Sedangkan untuk nilai *sensitivity*, *specificity*, dan *precision* lebih besar setelah menerapkan metode *re-sampling* dibandingkan sebelum menerapkan metode *re-sampling*. Untuk nilai *sensitivity* terbesar yaitu ketika setelah mengatasi *imbalance data* dengan menggunakan metode *oversampling* sebesar 88% dan nilai *specificity* terbesar juga ketika setelah mengatasi *imbalance data* dengan menggunakan metode *combine* sebesar 94%, dan nilai *precision* setelah mengatasi *imbalance data* jauh lebih besar dibandingkan dengan sebelum mengatasi *imbalance data* dengan masing-masing memiliki nilai sebesar 94% ketika menggunakan metode *combine* dan 78% sebelum menggunakan metode *re-sampling*.

Jika dilihat berdasarkan akurasi, perbandingan akurasi antara sebelum dan setelah penerapan metode *re-sampling* tidak memiliki perbedaan yang signifikan yang menunjukkan bahwa penerapan metode *re-sampling* untuk penelitian ini tidak memiliki pengaruh untuk meningkatkan kinerja akurasi pada dataset rekam medis yang memiliki kelas tidak seimbang, namun dengan menggunakan metode ini dapat meningkatkan nilai *specificity*, *sensitivity*, dan *precision* karena memiliki nilai yang lebih besar dibandingkan sebelum menerapkan metode *re-sampling*.

BAB VI
PENUTUP

6.1 Kesimpulan

Berdasarkan analisis yang telah dilakukan, maka dapat disimpulkan bahwa:

1. Metode *decision tree* c4.5 yang digunakan dalam klasifikasi cara keluar pasien Rumah Sakit Umum Abdoel Moeloek menghasilkan 11 aturan/*rule*. 11 aturan tersebut dapat dijadikan pola dalam menentukan pasien yang sembuh/meninggal.
2. Metode *decision tree* c4.5 dapat digunakan untuk mengetahui hasil klasifikasi cara keluar pasien Rumah Sakit Umum Abdoel Moeloek dengan tingkat keakuratan sebesar 91% sebelum *post-prunning* dan 88% setelah *post-prunning* dan berdasarkan evaluasi yang dilakukan dapat diketahui bahwa proses pembentukan pohon menggunakan teknik *prunning* tidak selalu memiliki akurasi yang lebih besar.
3. Penggunaan metode *re-sampling* tidak memiliki pengaruh untuk meningkatkan nilai akurasi dengan hasil yang menunjukkan bahwa nilai akurasi antara sebelum dan setelah penerapan metode *re-sampling* tidak memiliki perbedaan yang signifikan dengan masing-masing memiliki nilai akurasi sebesar 91% untuk sebelum dan 89% setelah menggunakan metode *oversampling* dan *undersampling*, namun jika dilihat berdasarkan nilai *specificity*, *sensitivity*, dan *precision*, metode ini mampu meningkatkan ketiga ukuran ketepatan tersebut karena memiliki nilai yang lebih besar dibandingkan sebelum menerapkan metode *re-sampling*.

6.2 Saran

Berdasarkan hasil analisis untuk penelitian ini, terdapat beberapa saran untuk penelitian selanjutnya antara lain:

1. Menambahkan variabel pada data rekam medis pasien ginjal kronis seperti hal-hal yang menjadi pendukung adanya penyakit ginjal kronis (dapat dilihat berdasarkan kadar hemoglobin, hipertensi, kolesterol, dan lainnya).
2. Menggunakan metode *decision tree* terbaru yaitu *decision tree c5.0*.
3. Menggunakan metode lain untuk mengatasi *imbalance data* seperti SMOTE (*Synthetic Minority Oversampling Tehnique*) dan lainnya.

DAFTAR PUSTAKA

Andie. (2016). Penerapan Decision Tree Untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa Baru.

Aprilla, D., Bakoro, D. A., Ambarwati, L., & Simri, I. W. (2013). *Belajar Data Mining dengan Rapidmaner.* Jakarta.

Bambang, H., & Azhari, S. (2017). Klasifikasi Nilai Kelayakan Calon Debitur Baru Menggunakan Decision Tree C4.5.

Corwin, E. J., Bahasa, a., & Pendit, B. U. (2001). *Buku Saku Patofisiologi.* Jakarta: EGC.

Drummond, C., & Holte, R. (2003). Class Imbalance, and Cost Sensitivity: Why Under-Sampling beats Over-Sampling. Canada, Ottawa, Ontario: Department of Computing Science, University of Alberta.

Dua, S., & Du, X. (2011). *Data Mining and Machine Learning in Cybersecurity.* USA: Taylor & Francis Group.

Fernitha, V. (2019). *Penerapan Algoritma C4.5 untuk Deteksi Penyakit Kanker Serviks.* Yogyakarta: Universitas Sanata Dharma.

Ganganwar, V. (2012). An Overview of Classification Algorithms for Imbalanced Datasets. *International Journal of Emerging Technology and Advanced Engineering Vol.2 Issue 4* , 42-47.

Gorunescu. (2011). *Data Mining: Concepts, Models, and Techniques.* Springer.

Han, J; Kamber, M. (2006). *Data Mining Concepts and Teqniques – 2nd Ed.* San Fransisco: Elsevier Inc.

Hana, F. M. (2020). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5.

Hasan, I. (2001). *Pokok-Pokok Materi Statistik 1 (Statistik Deskriptif).* Jakarta: PT Bumi Aksara.

Herman, I. (2016). *Hubungan Lama Hemodialisis dengan Fungsi Kognitif Pasien Penyakit Ginjal Kronis yang Menjalani Hemodialisis di RSUD Abdul Moeloek Bandar Lampung.* Universitas Lampung.

Hulse, J., & Khoshgoftaar, T. (2009). Knowledge discovery from imbalanced and noisy data. Elsevier.

Janah, L. M. (2014). Hubungan Antar Variabel : Tabel Silang .

Kementrian Kesehatan Republik Indonesia. (2017). *Info Pusat Data dan Informasi Kementerian Kesehatan RI, Situasi Penyakit Ginjal Kronis.* Jakarta: Kementrian Kesehatan Republik Indonesia.

Larose, D. T. (2005). *Discovering Knowledge in Data, 2nd Edition.* New Jersey: John Willey & Sons, Inc.

Mardi, Y. (2018). Data Mining Rekam Medis Untuk Menentukan Penyakit Terbanyak Menggunakan Decision Tree C4.5 .

Nugraha, P. S., Ariwibawa, I., Priyana, I. P., & G, I. (2016). Penerapan Metode Decision Tree (Data Mining) Untuk Memprediksi Tingkat Kelulusan Siswa SMPN 1 Kintamani . Denpasar-Bali.

Price, S. A., & Wilson, L. (2012). Patofisiologi: konsep klinis proses prosespenyakit, 6 ed. (H. e. Hartanto, Penyunt.) vol. 1.

Rafiska, R., Defit, S., & Nurcahyo, G. W. (2018). Analisis Rekam Medis untuk Menentukan Pola Kelompok Penyakit Menggunakan Algoritma C4.5.

Rahman, A., & Dwi, E. (2013). Hubungan Antara Hemodialisis dan Kualitas Hidup Pasien di RSUD Ulin Banjarmasin. Banjarmasin: Universitas Banjarmasin.

Santoso, S., & Tjiptono, F. (2001). Riset Pemasaran Konsep dan Aplikasi dengan SPSS. Jakarta: Elex Media Komputindo.

Sarwono, J. (2009). Statistik Itu Mudah: Panduan Lengkap untuk Belajar Komputasi Statistik Menggunakan SPSS 16. Yogyakarta: CV. Andi Offset.

Sloane, E. (2004). *Anatomi dan Fisiologi Untuk Pemula.* Jakarta: EGC.

Sumathi, S. (2006). *Introduction to Data Mining and Its Applications.* Germany: Springer Verlag berlin Heidelberg.

Tim Riset Kesehatan Dasar. (2019). *Riset Kesehatan Dasar 2018.* Jakarta: Badan Penelitian dan Pengembangan Kesehatan.

UNFPA, B. d. (2018). *Proyeksi Penduduk Indonesia 2015-2045.* Jakarta.

Witten, I. H. (2011). *Data Mining Practical Machine Learning Tools and Techniques (3rd ed).* USA: Elsevier.

LAMPIRAN

Lampiran 1 Data Rekam Medis Pasien Rumah Sakit Abdoel Moeloek Provinsi Lampung

No	CaraKeluar	KelasRawat	LamaRawat	TipePenyakit	JenisKelamin	Umur
1	Sembuh	2	3	Sedang	Perempuan	Anak-Anak
2	Sembuh	3	5	Sedang	Laki - Laki	Remaja
3	Sembuh	3	5	Sedang	Laki - Laki	Remaja
4	Sembuh	3	11	Sedang	Perempuan	Remaja
5	Sembuh	2	8	Sedang	Perempuan	Remaja
6	Sembuh	3	8	Sedang	Perempuan	Remaja
7	Sembuh	2	11	Sedang	Laki - Laki	Remaja
8	Meninggal	3	9	Sedang	Laki - Laki	Remaja
9	Meninggal	3	2	Sedang	Perempuan	Remaja
10	Sembuh	3	3	Sedang	Laki - Laki	Dewasa
11	Sembuh	3	7	Sedang	Perempuan	Dewasa
12	Sembuh	3	4	Sedang	Laki - Laki	Dewasa
13	Sembuh	3	5	Sedang	Laki - Laki	Dewasa
14	Sembuh	1	4	Sedang	Laki - Laki	Dewasa
15	Sembuh	3	6	Sedang	Laki - Laki	Dewasa
16	Meninggal	3	9	Berat	Laki - Laki	Dewasa
17	Meninggal	3	5	Berat	Perempuan	Dewasa
18	Sembuh	1	5	Sedang	Perempuan	Dewasa
19	Meninggal	3	2	Ringan	Laki - Laki	Dewasa
20	Meninggal	3	1	Sedang	Perempuan	Dewasa
21	Sembuh	3	2	Sedang	Laki - Laki	Dewasa
22	Sembuh	3	2	Ringan	Laki - Laki	Dewasa
23	Sembuh	3	3	Sedang	Perempuan	Dewasa
24	Sembuh	3	5	Sedang	Perempuan	Dewasa
25	Sembuh	1	6	Berat	Laki - Laki	Dewasa
26	Meninggal	3	1	Ringan	Perempuan	Dewasa
27	Sembuh	2	2	Ringan	Laki - Laki	Dewasa
28	Sembuh	3	3	Ringan	Laki - Laki	Dewasa
...	...	...	...	...	...	
358	Sembuh	3	8	Sedang	Laki - Laki	Lanjut Usia

Lampiran 2 Program R

```
pasien =read.csv(file.choose(),header=TRUE,sep=",")
```

```
View(pasien)
dim(pasien)
sapply(pasien, class)
table(pasien$CaraKeluar)
str(pasien)
suppressMessages(library(rattle))
library(pastecs)
stat.desc(pasien)
library(rpart)
library(rpart.plot)
#pasien.df = data.frame(factor(pasien$CaraKeluar),
factor(pasien$KodePenyakit),pasien$LamaRawat)
#colnames(pasien.df)<- c("CaraKeluar","KodePenyakit")
#View(pasien.df)
library(caret)
set.seed(12345)
train.set =read.csv(file.choose(),header=TRUE,sep=",")
test.set =read.csv(file.choose(),header=TRUE,sep=",")
#index = createDataPartition(y=pasien$CaraKeluar, p=0.80, list=FALSE)
#train.set = pasien[index,]
#write.csv(train.set,"E://dataskrpsitrainfix8.csv")
#View(train.set)
#test.set = pasien[-index,]
#write.csv(test.set,"E://dataskrpsitestfix8.csv")
#View(test.set)
library(rattle)
dim(train.set)
dim(test.set)
# fit the model
fit <- rpart(CaraKeluar~., data =train.set, method = 'class',control =
rpart.control(minsplit = 1,  minbucket = 1, cp = 0.001))
fit <- rpart(CaraKeluar~., data =train.set, method = 'class',control =
```

```
rpart.control(minsplit = 10,  minbucket = 1, cp = 0.001))
prp(fit)
fancyRpartPlot(fit)
fit$variable.importance
barplot(fit$variable.importance)
summary(fit)
# prediksi testing
pasien.pred = predict(fit, newdata = test.set, type = "class")
# Confusion matrix
table(pasien.pred, test.set$CaraKeluar)
confusionMatrix(data = pasien.pred, reference = test.set$CaraKeluar)
# prediksi training
pasien.train = predict(fit, newdata = train.set, type = "class")
# Confusion matrix
table(pasien.train, train.set$CaraKeluar)
confusionMatrix(data = pasien.train, reference = train.set$CaraKeluar)


################################imbalance
data##############################
library(ROSE)
library(caret)
################oversampling#####################
overtrain <- ovun.sample(CaraKeluar~., data =train.set, method = 'over', N =
474)$data
table(overtrain$CaraKeluar)
fit <- rpart(CaraKeluar~., data =overtrain, method = 'class',control =
rpart.control(minsplit = 1,  minbucket = 1, cp = 0.001))
fit <- rpart(CaraKeluar~., data =overtrain, method = 'class',control =
rpart.control(minsplit = 10,  minbucket = 1, cp = 0.001))
prp(fit)
fancyRpartPlot(fit)
fit$variable.importance
barplot(fit$variable.importance)
```

```
summary(fit)
# prediksi training
pasien.train = predict(fit, newdata = overtrain, type = "class")
# Confusion matrix
table(pasien.train, overtrain$CaraKeluar)
confusionMatrix(data = pasien.train, reference = overtrain$CaraKeluar)


################undersampling#####################
undertrain <- ovun.sample(CaraKeluar~., data =train.set, method = 'under', N
= 100)$data
table(undertrain$CaraKeluar)
fit <- rpart(CaraKeluar~., data =undertrain, method = 'class',control =
rpart.control(minsplit = 5,  minbucket = 1, cp = 0.001))
fit <- rpart(CaraKeluar~., data =undertrain, method = 'class',control =
rpart.control(minsplit = 10,  minbucket = 1, cp = 0.001))
prp(fit)
fancyRpartPlot(fit)
fit$variable.importance
barplot(fit$variable.importance)
summary(fit)
# prediksi training
pasien.train = predict(fit, newdata = undertrain, type = "class")
# Confusion matrix
table(pasien.train, undertrain$CaraKeluar)
confusionMatrix(data = pasien.train, reference =undertrain$CaraKeluar)


################combine#####################
both <- ovun.sample(CaraKeluar~., data=train.set, method = 'both', p=0.49,
seed = 232, N = 574)$data
table(both$CaraKeluar)
fit <- rpart(CaraKeluar~., data =both, method = 'class',control =
rpart.control(minsplit = 1,  minbucket = 1, cp = 0.001))
fit <- rpart(CaraKeluar~., data =both, method = 'class',control =
```

```
rpart.control(minsplit = 10,  minbucket = 1, cp = 0.001))
prp(fit)
fancyRpartPlot(fit)
fit$variable.importance
barplot(fit$variable.importance)
summary(fit)
# prediksi training
pasien.train = predict(fit, newdata = both, type = "class")
# Confusion matrix
table(pasien.train, both$CaraKeluar)
confusionMatrix(data = pasien.train, reference =both$CaraKeluar)
```