



**KEMENTERIAN KOMUNIKASI DAN INFORMATIKA  
BADAN PENELITIAN DAN PENGEMBANGAN SUMBER DAYA MANUSIA  
PUSAT PENGEMBANGAN PROFESI DAN SERTIFIKASI**

PELATIHAN DAN SERTIFIKASI ASSOCIATE DATA SCIENTIST  
Makassar 20 – 24 Oktober 2024

**LAPORAN TUGAS PELATIHAN**

**KELOMPOK : II**

**Anggota :**

- 1. Andi Muhammad Alfian**
- 2. Andi Sri Mulyani**
- 3. Budi Prasetyo**
- 4. Cristin Sherina Miracle**

## **1. DAFTAR ISI**

<b>DAFTAR ISI</b>	<b>1</b>
<b>1. MENGUMPULKAN DATA</b>	<b>2</b>
1.1 MENENTUKAN KEBUTUHAN DATA	2
1.2 MENGAMBIL DATA	
1.3 MENGINTEGRASIKAN DATA	
<b>2. MENELAAH DATA</b>	<b>2</b>
2.1 MENGANALISIS TIPE DAN RELASI DATA STATISTIK	2
2.2 MENGANALISIS KARAKTERISTIK DATA STATISTIK	2
2.3 MEMBUAT LAPORAN TELAAH DATA STATISTIK	2
<b>3. MEMVALIDASI DATA</b>	<b>2</b>
3.1 MELAKUKAN PENGECEKAN KELENGKAPAN DATA	2
3.2 MEMBUAT REKOMENDASI KELENGKAPAN DATA	2
<b>4. MENENTUKAN OBJEK DATA</b>	<b>3</b>
4.1 MEMUTUSKAN KRITERIA DAN TEKNIK PEMILIHAN DATA	3
4.2 MENENTUKAN ATRIBUT (KOLOM) DAN RECORDS (BARIS) DATA	3
<b>5. MEMBERSIHKAN DATA</b>	<b>3</b>
5.1 MELAKUKAN PEMBERSIHKAN DATA KOTOR	3
5.2 MEMBUAT LAPORAN DAN REKOMENDASI HASIL PEMBERSIHKAN DATA	3
<b>6. MENGKONSTRUKSI DATA</b>	<b>3</b>
6.1 MENGANALISIS TEKNIK TRANSFORMASI DATA	3
6.2 MELAKUKAN TRANSFORMASI DATA	3
6.3 MEMBUAT DOKUMENTASI KONSTRUKSI DATA	3
<b>7. MENENTUKAN LABEL DATA</b>	<b>3</b>
7.1 MELAKUKAN PELABELAN DATA	3
7.2 MEMBUAT LAPORAN HASIL PELABELAN DATA	3
<b>8. MEMBANGUN MODEL</b>	<b>4</b>
8.1 MENYIAPKAN PARAMETER MODEL	4
8.2 MENGGUNAKAN TOOLS PEMODELAN	4
<b>9. MENGEVALUASI HASIL PEMODELAN</b>	<b>4</b>
9.1 MENGGUNAKAN MODEL PADA DATA	4
9.2 MENILAI HASIL PEMODELAN	4

Isilah setiap tahapan dari proses Data Science berikut dengan hasil pengolahan data dan interpretasinya sesuai dengan konteks permasalahan yang diselesaikan.

Sebagai input data, silahkan memilih data dari link berikut:  
<https://drive.google.com/drive/folders/1uQyJxRHqf8PdPfyc4GQwZHMWFhyjz7LI?usp=sharing>

## 1. Mengumpulkan Data

### 1.1 Menentukan Kebutuhan Data

Kebutuhan data yang diperlukan adalah data hasil pemeriksaan laboratorium pasien rumah sakit, yang memiliki fitur (atribut) terkait dengan kondisi kesehatan pasien, serta label target yang menunjukkan apakah pasien tersebut memerlukan rawat inap (rwi) atau rawat jalan (rwj).

Row No.	Hematokrit	Hemoglobin	Jumlah Eritr...	Jumlah Lek...	Jumlah Tro...	MCH	MCHC	MCV	SOURCE
1	45	15.200	5.810	5.900	219	26.200	33.800	77.500	rwj
2	45	15.300	5.490	6.200	312	27.900	34	82	rwj
3	47.500	16.600	5.840	7.400	244	28.400	34.900	81.300	rwj
4	45.400	15	5.330	6.700	342	28.100	33	85.200	rwj
5	44.900	15.800	5.190	7	328	30.400	35.200	86.500	rwj
6	46.400	16.400	5.240	15.600	241	31.300	35.300	88.500	rwj
7	44.900	15.700	5.260	7.400	183	29.800	35	85.400	rwj
8	35.900	11.900	4.250	21.100	343	28	33.100	84.500	rwj
9	42	14.300	4.960	14.300	216	28.800	34	84.700	rwj
10	43.200	14.100	4.900	13.200	228	28.800	32.600	88.200	rwj
11	38.400	13.200	4.590	4.100	166	28.800	34.400	83.700	rwj
12	38.200	12.600	4.450	14.400	221	28.300	33	85.800	rwj
13	41.500	13.900	4.790	8.300	335	29	33.500	86.600	rwj
14	35.900	12.500	3.990	6.500	169	31.300	34.800	90	rwj

Dataset yang digunakan memiliki 4517 sampel atau baris data. Setiap sampel mewakili satu pasien, dengan informasi hasil pemeriksaan laboratorium yang terdiri dari 7 atribut (fitur). Atribut ini menggambarkan hasil berbagai tes medis, seperti tekanan darah, kadar gula, atau parameter medis lainnya, yang bisa mempengaruhi keputusan apakah pasien akan menjalani rawat inap (rwi) atau rawat jalan (rwj). Kolom ke-8 dari dataset ini adalah kelas (label), yang menunjukkan apakah pasien membutuhkan rawat inap atau rawat jalan. Dari keseluruhan 4518 sampel ini, model klasifikasi akan dilatih untuk memprediksi kelas tersebut berdasarkan atribut yang tersedia.

Adapun atribut dalam dataset yaitu :

- Hematokrit: Persentase volume darah yang terdiri dari sel darah merah. Ini menunjukkan seberapa banyak sel darah merah yang ada dalam darah.
- Hemoglobin: Protein dalam sel darah merah yang mengangkut oksigen dari paru-paru ke seluruh tubuh.
- Jumlah Eritrosit: Jumlah sel darah merah dalam darah, yang penting untuk membawa oksigen ke jaringan tubuh.
- Jumlah Lekosit: Jumlah sel darah putih dalam darah, yang berfungsi melawan infeksi dan menjaga sistem kekebalan tubuh.

- Jumlah Trombosit: Jumlah platelet dalam darah, yang penting untuk pembekuan darah dan penyembuhan luka.
- MCH (Mean Corpuscular Hemoglobin): Rata-rata jumlah hemoglobin dalam satu sel darah merah.
- MCHC (Mean Corpuscular Hemoglobin Concentration): Rata-rata konsentrasi hemoglobin dalam volume sel darah merah tertentu.
- MCV (Mean Corpuscular Volume): Rata-rata ukuran sel darah merah, yang dapat membantu mendiagnosis jenis anemia.
- Source: Asal atau sumber data, bisa merujuk pada institusi, laboratorium, atau metode pengumpulan data.

## 1.2 Mengambil Data

Import Data - Format your columns.

Format your columns.

Date format: Enter value... ▾  Replace errors with missing values ⓘ

	Hematokrit * real	Hemoglobin * real	Jumlah Erit... * real	Jumlah Le... * real	Jumlah Tro... * integer	MCH * real
1	45.000	15.200	5.810	5.900	219	26.200
2	45.000	15.300	5.490	6.200	312	27.900
3	47.500	16.600	5.840	7.400	244	28.400
4	45.400	15.000	5.330	6.700	342	28.100
5	44.900	15.800	5.190	7.000	328	30.400
6	46.400	16.400	5.240	15.600	241	31.300
7	44.900	15.700	5.260	7.400	183	29.800
8	35.900	11.900	4.250	21.100	343	28.000
9	42.000	14.300	4.960	14.300	216	28.800
10	43.200	14.100	4.900	13.200	228	28.800
11	38.400	13.200	4.590	4.100	166	28.800

no problems.

Data diambil dari dataset yang telah disediakan, yang berisi 7 atribut hasil pemeriksaan laboratorium dan 1 atribut target (kelas) yang menunjukkan status rawat jalan atau rawat inap. Proses mengambil data melibatkan langkah-langkah teknis untuk mengimpor dataset ke dalam alat analisis. Di sini, data hasil pemeriksaan laboratorium pasien diimporkan ke RapidMiner, dan setiap kolom diperiksa untuk memastikan data siap digunakan dalam proses klasifikasi lebih lanjut.

### 1.3 Mengintegrasikan Data

Karena dataset sudah dalam satu file lengkap dan semua atribut sudah termasuk, tidak ada kebutuhan untuk mengintegrasikan data dari sumber lain. Data yang ada langsung diolah lebih lanjut untuk proses berikutnya.

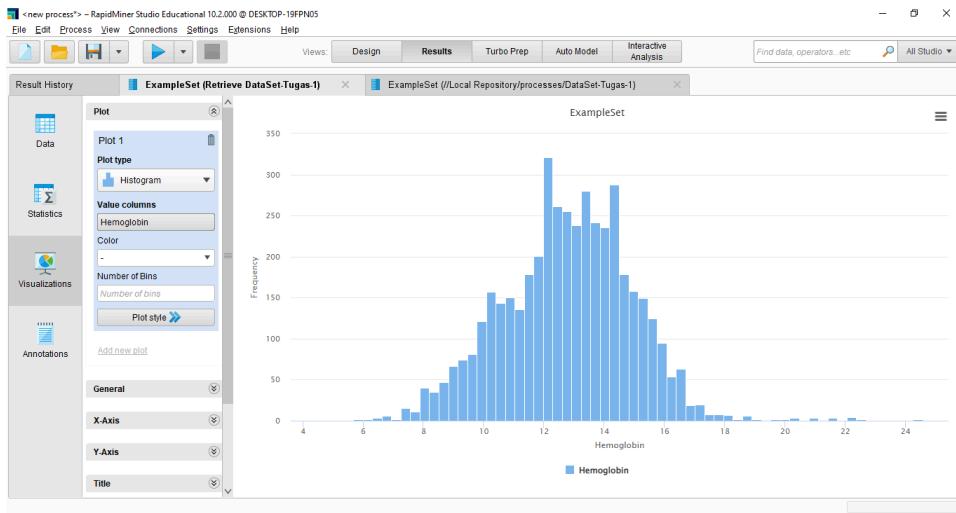
## 2. Menelaah Data

### 2.1 Menganalisis Tipe dan Relasi Data Statistik

Name	Type	Missing	Statistics
Hematokrit	Real	0	Min: 13.700 Max: 71.300 Average: 38.358
Hemoglobin	Real	0	Min: 3.800 Max: 24.900 Average: 12.803
Jumlah Eritrosit	Real	0	Min: 1.480 Max: 7.860 Average: 4.552
Jumlah Lekosit	Real	1	Min: 1.100 Max: 76.600 Average: 8.915
Jumlah Trombosit	Integer	0	Min: 8 Max: 1183 Average: 258.350
MCH	Real	0	Min: 14.900 Max: 40.800 Average: 28.295
MCHC	Real	0	Min: 26 Max: 39 Average: 33.350
MCV	Real	0	Min: 54 Max: 115.600 Average: 84.764

- Tipe Data: kolom berupa numerik atau kategorikal. Kolom 8 harus terdeteksi sebagai label (rwi/rwj).
- Relasi Atribut: Menggunakan operator Correlation Matrix untuk melihat hubungan antar atribut numerik. Korelasi mendekati +1 atau -1 menunjukkan hubungan kuat, sedangkan mendekati 0 menunjukkan tidak ada hubungan linier signifikan.

## 2.2 Menganalisis Karakteristik Data Statistik



Histogram yang ditampilkan menunjukkan distribusi frekuensi dari variabel numerik "Hemoglobin". Setiap batang pada histogram mewakili rentang nilai hemoglobin tertentu dan tinggi batang menunjukkan jumlah observasi (data) yang jatuh dalam rentang tersebut.

Menganalisis karakteristik data statistik melibatkan pengumpulan dan interpretasi informasi dari dataset untuk memahami pola, tren, dan struktur yang ada. Proses ini biasanya dimulai dengan analisis deskriptif, seperti menghitung ukuran pemusatan (mean, median, modus) dan ukuran dispersi (rentang, varians, deviasi standar), yang memberikan gambaran umum tentang distribusi data. Selain itu, analisis frekuensi digunakan untuk melihat seberapa sering nilai tertentu muncul. Dengan menggunakan visualisasi seperti histogram, boxplot, dan scatter plot, analis dapat lebih mudah mengidentifikasi outlier dan hubungan antar variabel. Tujuan dari analisis ini adalah untuk memberikan wawasan yang berguna untuk pengambilan keputusan dan untuk mempersiapkan data sebelum dilakukan analisis lebih lanjut.

## 2.3 Membuat Laporan Telaah Data Statistik

Attribut...	Hemato...	Hemogl...	Jumlah ...	Jumlah ...	Jumlah ...	MCH	MCHC	MCV	SOURCE
Hematok...	1	0.978	0.857	-0.091	0.005	0.159	0.137	0.129	-0.216
Hemogl...	0.978	1	0.806	-0.074	-0.031	0.274	0.334	0.181	-0.198
Jumlah ...	0.857	0.806	1	-0.131	0.026	-0.331	-0.033	-0.385	-0.207
Jumlah ...	-0.091	-0.074	-0.131	1	0.282	0.081	0.043	0.076	0.167
Jumlah ...	0.005	-0.031	0.026	0.282	1	-0.114	-0.185	-0.058	-0.234
MCH	0.159	0.274	-0.331	0.081	-0.114	1	0.597	0.937	0.020
MCHC	0.137	0.334	-0.033	0.043	-0.185	0.597	1	0.282	0.029
MCV	0.129	0.181	-0.385	0.076	-0.058	0.937	0.282	1	0.010
SOURCE	-0.216	-0.198	-0.207	0.167	-0.234	0.020	0.029	0.010	1

Hubungan korelasi yang mendekati 1 adalah indikator yang sangat baik mengenai kekuatan hubungan antara dua variabel. Hubungan korelasi antara variabel target dan variabel prediktor yang mendekati 1 menunjukkan hubungan yang sangat kuat dan positif. Artinya, ketika nilai salah satu variabel meningkat, nilai variabel lainnya juga cenderung meningkat secara signifikan.

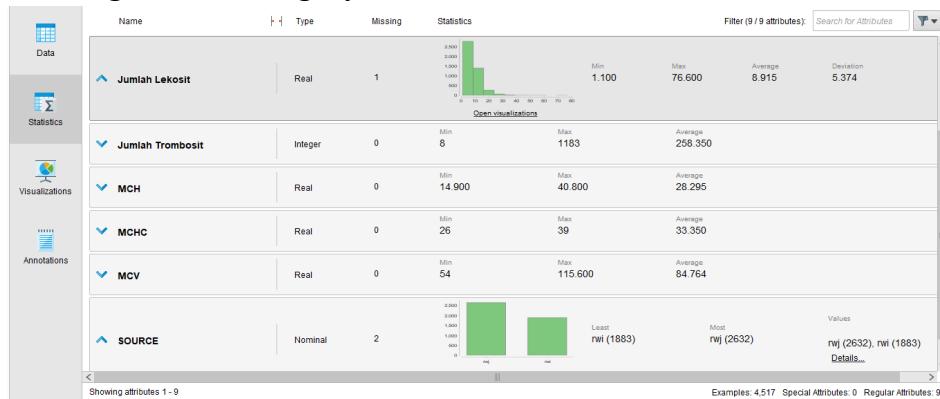
Korelasi antar variabel adalah ukuran statistik yang menunjukkan sejauh mana dua variabel saling terkait. Korelasi ini dinyatakan dengan koefisien korelasi, yang bernilai antara -1 hingga 1. Nilai positif menunjukkan hubungan searah, di mana peningkatan pada satu variabel diikuti oleh peningkatan pada variabel lain, sedangkan nilai negatif menunjukkan hubungan berlawanan. Korelasi bernilai 0 menunjukkan tidak ada hubungan linear antara dua variabel. Korelasi membantu dalam memahami apakah suatu variabel dapat digunakan untuk memprediksi variabel lain, yang penting dalam analisis data, terutama untuk model prediksi.

Variabel yang paling berkorelasi dengan variabel target adalah yang memiliki koefisien korelasi tertinggi (positif atau negatif) terhadap target, karena ini menunjukkan hubungan paling kuat antara variabel tersebut dan target dalam memprediksi hasil.

x1	-0,216		0,216	x2		0,234	x1	Jumlah Trombosit
x2	-0,198		0,198	x4		0,216	x2	Hemaktorit
x3	-0,207		0,207	x3		0,207	x3	Jumlah Eritrosit
x4	0,167		0,167	x5		0,198	x4	Hemoglobin
x5	-0,234		0,234	x1		0,167	x5	Jumlah Lekosit
x6	0,02		0,02	x7		0,029	x6	MCHC
x7	0,029		0,029	x6		0,02	x7	MCH
x8	0,01		0,01	x8		0,01	x8	MCV

### 3. Memvalidasi Data

#### 3.1 Melakukan Pengecekan Kelengkapan Data



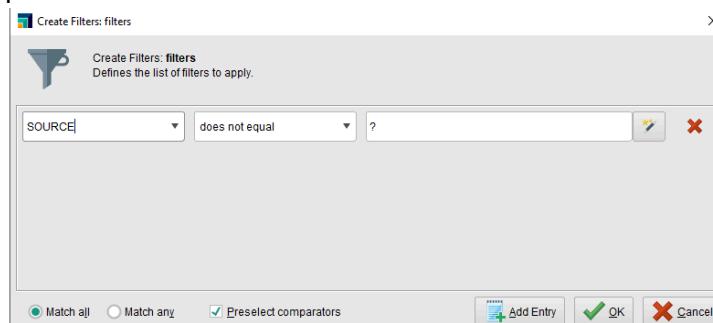
Pengecekan data adalah proses penting untuk memastikan kualitas, kelengkapan, konsistensi, dan validitas data sebelum analisis dilakukan. Proses ini meliputi identifikasi nilai yang hilang, pemeriksaan konsistensi antar kolom, serta validasi rentang dan tipe data. Selain itu, pengecekan duplikasi juga dilakukan untuk menghindari bias.

Operator:

- Missing Value: Operator ini bisa digunakan untuk mendeteksi dan menangani data yang hilang. Anda bisa menggunakannya untuk mengidentifikasi kolom mana yang memiliki nilai kosong.
- Replace Missing Values: Operator ini digunakan untuk mengisi nilai yang hilang dengan metode yang sesuai (misalnya rata-rata, nilai yang sering muncul, atau nilai tertentu).
- Filter Examples: Digunakan untuk menghapus baris (examples) dengan missing values.
- Type Conversion: Digunakan untuk mengonversi tipe data yang salah, misalnya mengubah kolom teks menjadi kolom numerik jika diperlukan.

#### 3.2 Membuat Rekomendasi Kelengkapan Data

Proses pembuatan rekomendasi kelengkapan data melibatkan beberapa tahap, dan salah satu tahapan awal, yaitu pemfilteran data.



Pada gambar, terlihat pengguna sedang membuat filter dengan kriteria "SOURCE does not equal ?". Ini berarti data yang memiliki nilai "?" pada kolom "SOURCE" akan disaring atau tidak dimasukkan dalam analisis selanjutnya.

Rekomendasi kelengkapan data merupakan langkah penting dalam analisis data. Dengan melakukan pembersihan dan pengolahan data yang baik, kita dapat meningkatkan kualitas analisis dan mendapatkan hasil yang lebih akurat.

## 4. Menentukan Objek Data

### 4.1 Memutuskan Kriteria dan Teknik Pemilihan Data

Pada poin ini kriteria yang kami pakai adalah secara relevansi dan kelengkapannya. Dimana dari segi relevansi sudah memadai dengan tersedianya atribut-atribut seperti Hematokrit, Hemoglobin, Jumlah Eritrosit, Jumlah Lekosit, Jumlah Trombosit, MCH, MCHC, dan MCV yang berdasarkan laporan hasil lab pengujian darah atribut tersebut yang digunakan untuk menentukan hasil analisa kondisi pasien.

Kemudian dari segi kelengkapan datanya sendiri sudah memadai, hanya saja semakin banyak jumlah data yang dimiliki maka hasil akhirnya akan lebih optimal.

### 4.2 Menentukan Atribut (Kolom) dan Records (Baris) Data

Setelah kriteria ditetapkan, langkah selanjutnya adalah menentukan atribut (kolom) dan records (baris) dalam dataset. Atribut mewakili variabel yang akan dianalisis, sedangkan records adalah entri individual dalam dataset. Penting untuk memastikan bahwa atribut yang dipilih relevan dan dapat memberikan informasi yang dibutuhkan untuk analisis lebih lanjut.

Disini kami mengambil semua atribut dan records yang terdapat dalam dataset, karena menurut kami semua atribut yang tersedia sangat dibutuhkan dalam pemprosesan namun dikarenakan terdapat beberapa missing value maka kami akan melakukan penanganan missing value yang akan di jelaskan di poin selanjutnya.

Name	Type	Missing	Statistics	Average
Hematokrit	Real	0	Min: 13.700 Max: 71.300	38.358
Hemoglobin	Real	0	Min: 3.800 Max: 24.900	12.803
Jumlah Eritrosit	Real	0	Min: 1.480 Max: 7.860	4.552
Jumlah Lekosit	Real	1	Min: 1.100 Max: 76.600	8.915
Jumlah Trombosit	Integer	0	Min: 8 Max: 1163	258.350
MCH	Real	0	Min: 14.900 Max: 40.800	28.295
MCHC	Real	0	Min: 26 Max: 39	33.350
MCV	Real	0	Min: 54 Max: 115.600	84.764
SOURCE	Binominal	2	Negative: rwj Positive: rwi	rwj (2632), rwi (1883)

Gambar di atas merupakan gambar yang ditampilkan dari Rapidminer yang menunjukkan atribut atribut yang akan digunakan. Adapun dapat kita lihat bahwa terdapat missing value pada atribut "Jumlah Lekosit" dan atribut "Source" sehingga tindakan yang akan kami lakukan selanjutnya adalah mengganti record dari atribut "Jumlah Lekosit" menjadi rata-rata atribut "jumlah lekosit", alasannya atribut ini memiliki tipe data numerikal sehingga kita masih bisa mempertahankan jumlah record pada atribut "jumlah lekosit".

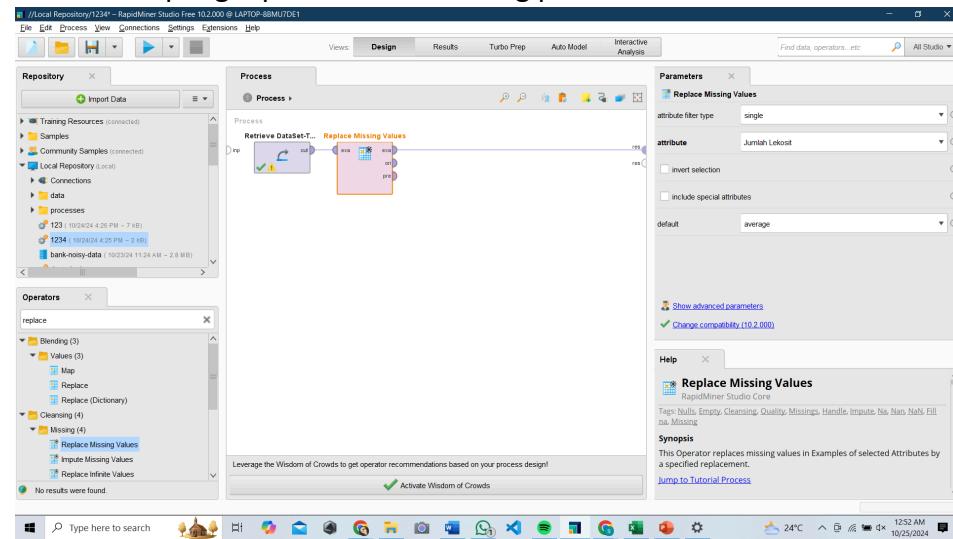
Sedangkan pada atribut "source" akan dilakukan penghapusan data dikarenakan tipe datanya bersifat nominal dengan alasan tidak mempengaruhi hasil.

## 5. Membersihkan Data

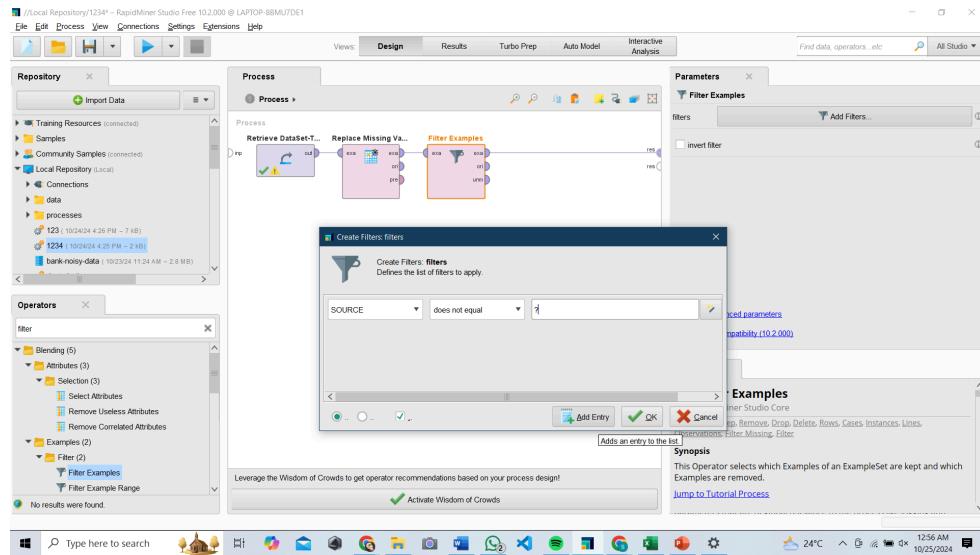
### 5.1 Melakukan Pembersihan Data Kotor

Pembersihan data adalah proses mengidentifikasi dan memperbaiki kesalahan dalam dataset, seperti duplikasi, inkonsistensi, dan nilai yang hilang. Langkah ini sangat penting karena data yang bersih meningkatkan akurasi analisis dan hasil yang diperoleh. Proses ini meliputi penghapusan data ganda, perbaikan kesalahan format, dan penanganan outlier.

Pada data pasien terdapat atribut yang memiliki record kosong yaitu atribut "Jumlah Lekosit" dan atribut "source" yang dimana seperti yang dibahas di poin sebelumnya, kami akan melakukan penggantian record kosong menjadi nilai rata-rata atribut pada atribut "jumlah lekosit" dan penghapusan record kosong pada atribut source.



Menambahkan operator Replace Missing Value dan pada parameter kita atur jumlah lekosit sebagai atribut yang akan di ganti dengan memilih average sebagai nilai penggantinya.



Kemudian untuk penanganan record kosong pada atribut "source" ditambahkan operator Filter Examples yang pada parameter filernya kita atur atribut "source", memilih "does not equal", dan menetapkan karakter "?" yang mewakili record kosong. Jadi, pada atribut "source" kita akan mengambil semua record kecuali "?", sehingga record kosong yang diwakili oleh karakter "?" tidak akan digunakan. Langkah ini merupakan langkah yang dapat ditempuh pada penanganan record kosong pada tipe data nominal

## 5.2 Membuat Laporan dan Rekomendasi Hasil Pembersihan Data

Setelah pembersihan dilakukan, maka hasilnya jumlah record data akan berkurang dan akan lebih optimal dalam penentuan hasil

Row No.	Hematocrit	Hemoglobin	Jumlah Erit...	Jumlah Leko...	Jumlah Tro...	MCH	MCHC	MCV	SOURCE
64	38.200	12.500	4.600	3.600	143	27.200	32.700	83	nwj
65	43	15	5.120	4.500	236	29.300	34.900	84	nwj
66	32.800	10.800	3.600	14.400	281	30	32.900	91.100	nwj
67	44.400	14.600	5.180	3.900	276	28.200	32.900	85.700	nwj
68	38.100	12.900	4.160	10.000	233	31	33.900	91.600	nwj
69	38.600	12.700	4.310	6.200	380	29.500	32.900	89.600	nwj
70	37.900	11.900	5.400	11.600	359	21.800	31.400	69.400	nwj
71	41.400	13.600	5.130	5.700	343	26.500	32.900	80.700	?
72	34.500	10.800	4.370	7.800	509	24.700	31.300	78.900	nwj
73	47.600	16.200	5.470	4.500	158	29.600	34	87	nwj
74	33.700	10.800	3.650	6.900	359	29.600	32	92.300	nwj
75	35.200	11.600	4.200	?	380	27.600	33	83.800	nwj
76	40	13.400	4.430	6.300	437	30.200	33.500	90.300	nwj
77	42.900	14.400	5.030	4.400	244	28.600	33.600	85.500	nwj
78	33.500	10.800	3.550	3.200	144	30.400	32.200	94.400	nwj
79	41.500	13.300	4.830	2.900	217	27.500	32	85.900	nwj
80	35	11.400	3.970	7.700	314	28.700	32.600	88.200	nwj
81	44.900	14.900	5.950	6.300	179	25	33.200	75.500	nwj

sebelum diproses dapat dilihat jumlah data dalam data set sebanya 4,517 data dan pada row ke 75 terdapat record kosong pada atribut "jumlah lekosit"

RapidMiner Studio interface showing the 'Replace Missing Values' operator applied to the 'ExampleSet'. The 'Data' tab displays a table with 78 rows and 9 columns: Row No., Jumlah Lek..., Hematokrit, Hemoglobin, Jumlah Erit..., Jumlah Tro..., MCH, MCHC, MCV, and SOURCE. The 'SOURCE' column shows 'rwj' for most rows except row 75 which shows 'average'.

kemudian setelah ditambahkan operator Replace Missing Value, jumlah data tetap sama namun pada baris 75 kolom atribut "jumlah lekosit" telah mengalami perubahan record yang berisi average atribut "jumlah lekosit".

RapidMiner Studio interface showing the 'Filter Examples' operator applied to the 'ExampleSet'. The 'Data' tab displays a table with 75 rows and 9 columns: Row No., Jumlah Lek..., Hematokrit, Hemoglobin, Jumlah Erit..., Jumlah Tro..., MCH, MCHC, MCV, and SOURCE. The 'SOURCE' column shows 'rwj' for most rows except row 75 which shows 'average'.

kemudian setelah kita menambahkan kembali operator baru yaitu Filter Examples, jumlah data berkurang 2 (dikarenakan record kosong pada atribut "source" berjumlah 2) dan nilai average pada baris 75 juga berubah diakibatkan oleh penghapusan 2 record tadi

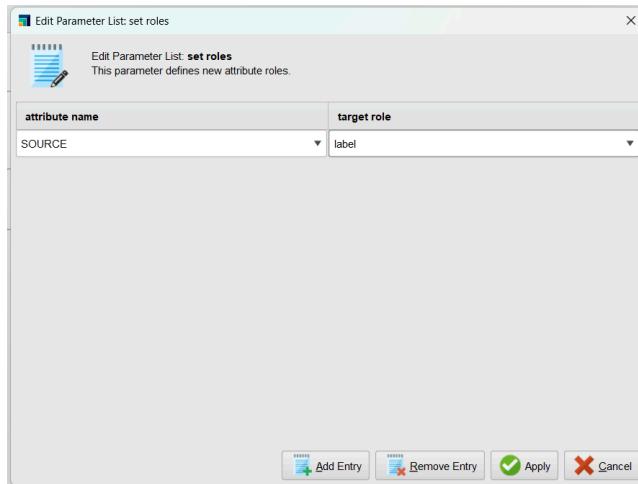
## 6. Mengkonstruksi Data

Pada tahap ini kami menghilangkan proses konstruksi data karena variabel target telah bernilai nominal, yang apabila dilakukan proses konstruksi data salah satunya normalisasi maka hasil yang didapatkan akan bernilai sama

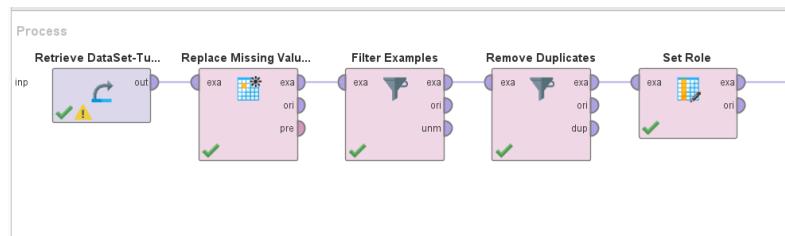
## 7. Menentukan Label Data

### 7.1 Melakukan Pelabelan Data

Pada dataset ini, data yang dijadikan sebagai inputan data yaitu variabel source (Variabel target) dijadikan label data dengan menggunakan operator set role.



### 7.2 Membuat Laporan Hasil Pelabelan Data



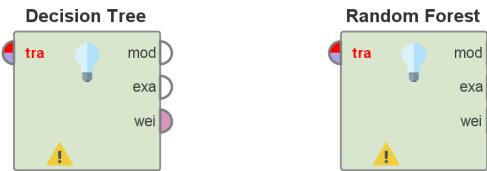
Pada gambar diatas, tahapan proses operator set role yang digunakan untuk pelabelan data pada variabel sorce yang dijadikan sebagai target. Dan menghasilkan data dibawah ini.

Row No.	SOURCE	Jumlah Leko...	Hematokrit	Hemoglobin	Jumlah Eritr...	Jumlah Tro...	MCH	MCHC	MCV
1	rwj	5.900	45	15.200	5.810	219	26.200	33.800	77.500
2	rwj	6.200	45	15.300	5.490	312	27.900	34	82
3	rwj	7.400	47.500	16.600	5.840	244	28.400	34.900	81.300
4	rwj	6.700	45.400	15	5.330	342	28.100	33	85.200
5	rwj	7	44.900	15.800	5.190	328	30.400	35.200	86.500
6	rwj	15.600	46.400	16.400	5.240	241	31.300	35.300	88.500
7	rwj	7.400	44.900	15.700	5.260	183	29.800	35	85.400
8	rwj	21.100	35.900	11.900	4.250	343	28	33.100	84.500

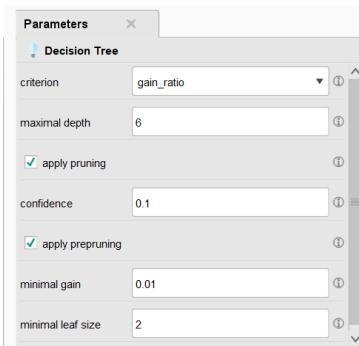
Berikut pada gambar diatas hasil pelabelan data yang terjadi, yang dalam hal ini variabel sorce telah ditetapkan sebagai label yang memiliki type data binomial.

## 8. Membangun Model

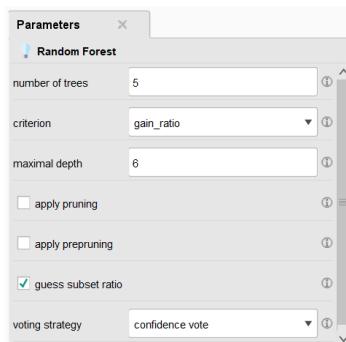
### 8.1 Menyiapkan Parameter Model



Dataset ini jika ditinjau dari setiap variabel dan studi kasusnya, model yang cocok digunakan pada data ini adalah model klasifikasi. Pada model ini kita akan membuat 2 model untuk melakukan perbandingan terhadap tingkat akurasi dengan menggunakan model decision tree dan random forest. Nilai pada maximal depth dapat berubah-ubah disesuaikan dengan pengujian yang dilakukan.

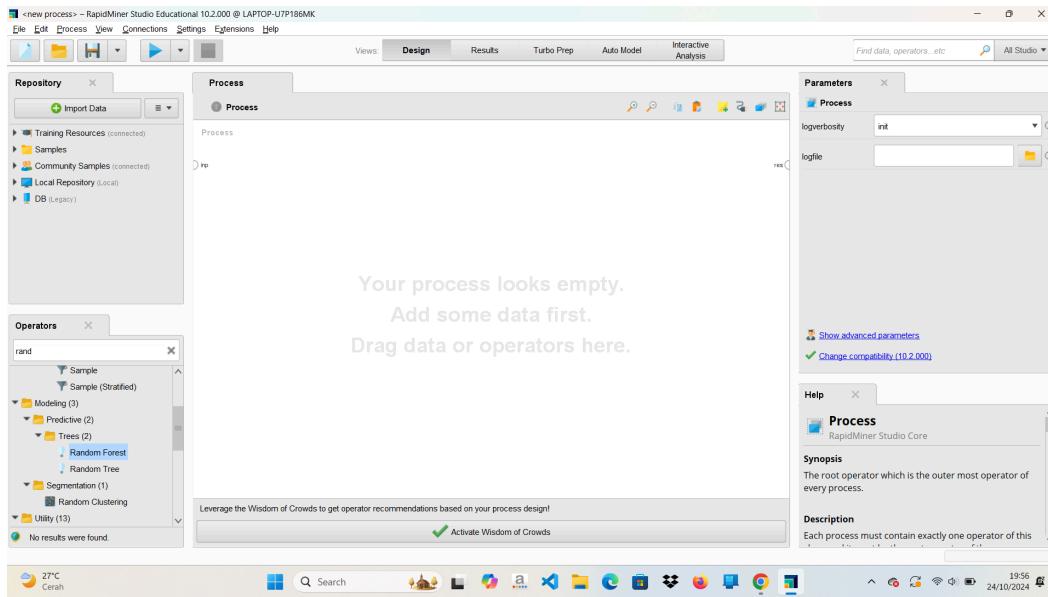


Pada parameter yang tertampil di atas untuk operator decision tree dimana diuji coba menggunakan maximal depth sebesar 6.



Dan pada parameter yang tertampil di atas untuk operator random forest dimana diuji coba menggunakan nilai yang sama dengan maximal depth sebesar 6.

### 8.2 Menggunakan Tools Pemodelan

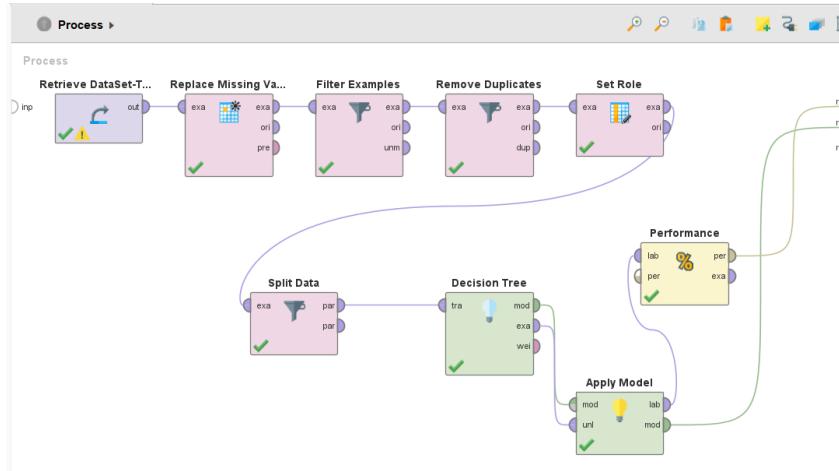


Tools pemodelan yang digunakan pada pelatihan ini untuk menyelesaiannya menggunakan tools Rapidminer Studio 10.2.000.

## 9. Mengevaluasi Hasil Pemodelan

### 9.1 Menggunakan Model pada Data

Model yang digunakan pertama untuk menyelesaikan kasus ini adalah model decision tree.



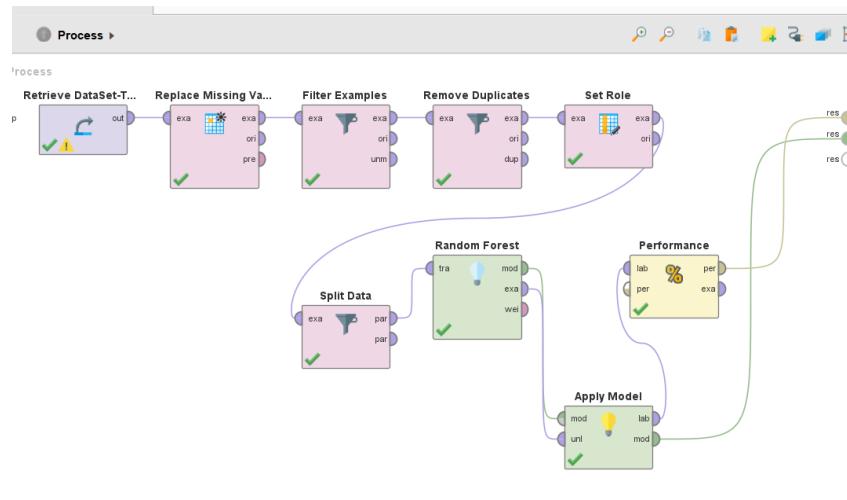
Pada proses decision tree digunakan beberapa operator untuk memperoleh hasil pemodelan diantaranya:

- **Retrieve data**, merupakan proses untuk mengambil dan menginput dataset yang akan digunakan ini pada pelatihan ini.
- **Replace missing data**, merupakan proses untuk menghilangkan missing value yang bersifat numerik yaitu pada variabel data ini terdapat missing value di ‘Jumlah Lekosit’.
- **Filter examples**, merupakan proses untuk melakukan penyaringan terhadap data variabel yang bersifat bukan nomerik dan tujuannya untuk menghapus data missing value tersebut. Pada filter

examples ini, variabel yang difilter untuk penghapusan missing value terdapat pada variabel target yaitu 'Source' yang memiliki 2 missing value.

- **Remove duplicates**, merupakan proses untuk menghapus semua variabel duplikat pada dataset ini.
- **Set role**, merupakan proses yang dilakukan untuk pelabelan data pada variabel target yang digunakan yakni pelabelan pada variabel 'Source'.
- **Split data**, merupakan tahapan untuk membagi menjadi 2 data yakni data training dan data testing. Data training yang digunakan berupa 70% dan data testing berupa 30%..
- **Decision tree**, merupakan proses model yang digunakan memperoleh prediksi berdasarkan klasifikasi dataset ini. Gain ratio atau yang dikenal dengan pemisahan data pada setiap node yang digunakan adalah maximal depth sebesar 6.
- **Apply model**, merupakan proses untuk menggunakan model yang dipilih pada operator yang digunakan.
- **Performance**, merupakan proses akhir untuk menampilkan hasil pemodelan data dari decision tree.

Model yang digunakan kedua untuk menyelesaikan kasus ini adalah model random forest.



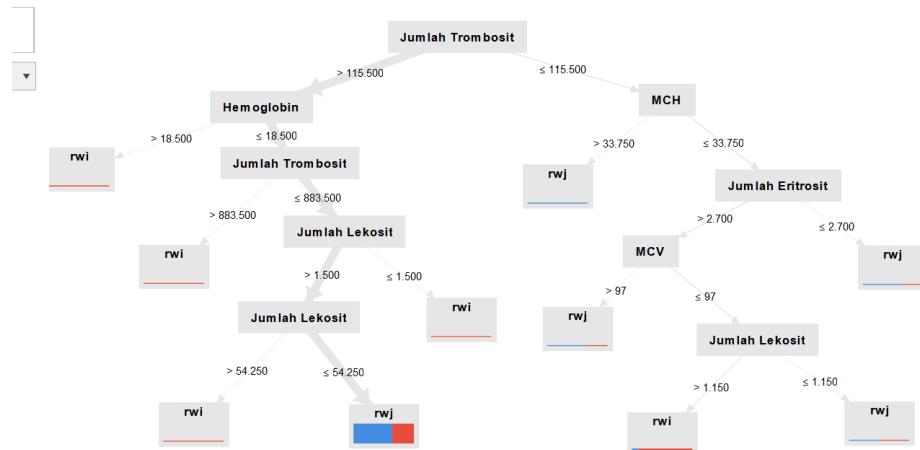
Pada proses random forest digunakan beberapa operator untuk memperoleh hasil pemodelan diantaranya:

- **Retrieve data**, merupakan proses untuk mengambil dan menginput dataset yang akan digunakan ini pada pelatihan ini.
- **Replace missing data**, merupakan proses untuk menghilangkan missing value yang bersifat numerik yaitu pada variabel data ini terdapat missing value di 'Jumlah Lekosit'.
- **Filter examples**, merupakan proses untuk melakukan penyaringan terhadap data variabel yang bersifat bukan nomerik dan tujuannya untuk menghapus data missing value tersebut. Pada filter examples ini, variabel yang difilter untuk penghapusan missing value terdapat pada variabel target yaitu 'Source' yang memiliki 2 missing value.
- **Remove duplicates**, merupakan proses untuk menghapus semua variabel duplikat pada dataset ini.

- **Set role**, merupakan proses yang dilakukan untuk pelabelan data pada variabel target yang digunakan yakni pelabelan pada variabel 'Source'.
- **Split data**, merupakan tahapan untuk membagi menjadi 2 data yakni data training dan data testing. Data training yang digunakan berupa 70% dan data testing berupa 30%..
- **Random forest**, merupakan proses model yang digunakan memperoleh prediksi berdasarkan klasifikasi dataset ini. Gain ratio atau yang dikenal dengan pemisahan data pada setiap node yang digunakan adalah maximal depth sebesar 6.
- **Apply model**, merupakan proses untuk menggunakan model yang dipilih pada operator yang digunakan.
- **Performance**, merupakan proses akhir untuk menampilkan hasil pemodelan data dari random forest.

## 9.2 Menilai Hasil Pemodelan

Hasil pemodelan yang dilakukan pada model decision tree.



Pada gambar diatas ini merupakan hasil pohon keputusan yang dibuat menggunakan model decision tree. Dapat dilihat pada gambar yang menjadi akar dari pohon ini adalah 'Jumlah trombosit'.

		Nilai sebenarnya	
		TRUE	FALSE
Nilai prediksi	TRUE	TP (True Positive) Correct result	FP (False Positive) Unexpected result
	FALSE	FN (False Negative) Missing result	TN (True Negative) Correct absence of result

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Pada gambar diatas ini merupakan rumus untuk mencari nilai prediksi yaitu akurasi dan presisi.

accuracy: 67.17%			
	true rwj	true rwi	class precision
pred. rwj	1808	1003	64.32%
pred. rwi	34	314	90.23%
class recall	98.15%	23.84%	

Pada hasil pemodelan yang diperoleh ternyata jika menggunakan model decision tree kita memperoleh nilai akurasi 67.17%. Dimana nilai akurasi ini dapat diperoleh dari:

- **Akurasi** =  $TP + TN / TP + TN + FP + FN = 1808 + 314 / 1808 + 314 + 1003 + 34 = 2122 / 3159 = 0,6717$  atau setara dengan 67.17%

Dan untuk presisi pada decision tree ini diperoleh dari:

- **Presisi false** =  $TP / TP + FP = 1808 / 1808 + 1003 = 1808 / 2811 = 0,6431$  atau setara dengan 64,32%.

- **Presisi true** =  $TN / TN + FN = 314 / 314 + 34 = 314 / 348 = 0,9022$  atau setara dengan 90,23%.

Model yang digunakan kedua untuk menyelesaikan kasus ini adalah random forest.

Pada gambar diatas ini merupakan hasil pohon keputusan yang dibuat menggunakan model random forest. Dapat dilihat pada gambar yang menjadi akar dari pohon ini adalah 'Jumlah trombosit'.

accuracy: 64.99%			
	true rwj	true rwi	class precision
pred. rwj	1814	1078	62.72%
pred. rwi	28	239	89.51%
class recall	98.48%	18.15%	

Pada hasil pemodelan yang diperoleh ternyata jika menggunakan model decision tree kita memperoleh nilai akurasi 64.99%. Dimana nilai akurasi ini dapat diperoleh dari:

- **Akurasi** =  $TP + TN / TP + TN + FP + FN = 1814 + 239 / 1814 + 239 + 1078 + 28 = 2053 / 3159 = 0,6498$  atau setara dengan 64.99%.

Dan untuk presisi pada decision tree ini diperoleh dari:

- **Presisi false** =  $TP / TP + FP = 1814 / 1814 + 1078 = 1814 / 2892 = 0,6272$  atau setara dengan 62,72%.

- **Presisi true** =  $TN / TN + FN = 239 / 239 + 28 = 239 / 267 = 0,8951$  atau setara dengan 89,51%.