UNIVERSITAS BUDI LUHUR
FAKULTAS TEKNOLOGI INFORMASI
UJIAN TENGAH SEMESTER GENAP TAHUN AJARAN 2022/2023

Mata Kuliah	:	Penambangan Data dan BI	Kelompok	:	XM
Hari/Tanggal	:	Sabtu, 20 Mei 2023	Waktu	:	1 Minggu
Dosen	:	RUSDAH	Sifat Ujian	:	BUKA BUKU

NIM	:	2211600818
NAMA	:	Sirin Mazaya Rochmah Shahab

Tahap 1. Business Understanding (Nilai 5)

Tahap pertama ini merupakan tahapan yang penting. Tujuan utama dari tahap ini adalah memahami bagaimana proses bisnis, permasalahan yang dihadapi, tujuan penambangan data, dan segala hal yang mendukung pemahaman terhadap apa yang terjadi di instansi.

- **Permasalahan**: Kanker payudara dilaporkan sebagai penyebab kematian kelima di dunia berdasarkan jenis kanker pada tahun 2015 (WHO, 2015). Kanker merupakan penyakit yang sulit proses penyembuhannya. Selain itu memerlukan biaya yang cukup besar untuk pengobatan dan perawatannya. Seseorang yang telah mengidap kanker payudara dapat disembuhkan dengan berbagai pengobatan untuk menghambat pertumbuhan sel kanker, seperti operasi, kemoterapi, radio terapi dan terapi hormonal. Meski demikian, pasien yang telah menjalani pengobatan tidak dapat sembuh sepenuhnya. Kekambuhan kanker payudara merupakan penyebab utama kematian pada pasien kanker payudara. Seseorang yang kambuh memerlukan penanganan yang lebih serius. Sehingga diperlukan model yang dapat memprediksi apakah seorang pasien berpotensi akan kambuh atau tidak.
- **Tujuan**: untuk memprediksi kekambuhan pasien kanker payudara.

Anda diminta untuk menentukan tujuan dari proyek data sains dari penjelasan di atas.

Tujuannya ialah untuk mengetahui faktor-faktor apa saja yang dapat mempengaruhi kambuhnya kanker payudara pada pasien.

Ada banyak atribut yang digunakan untuk memetakan kondisi pasien, dari banyaknya atribut tersebut dapat dilakukan pemodelan untuk memprediksi apakah kanker payudara pada pasien akan kambuh lagi atau tidak.

Tahap 2: Data Understanding (Nilai 10)

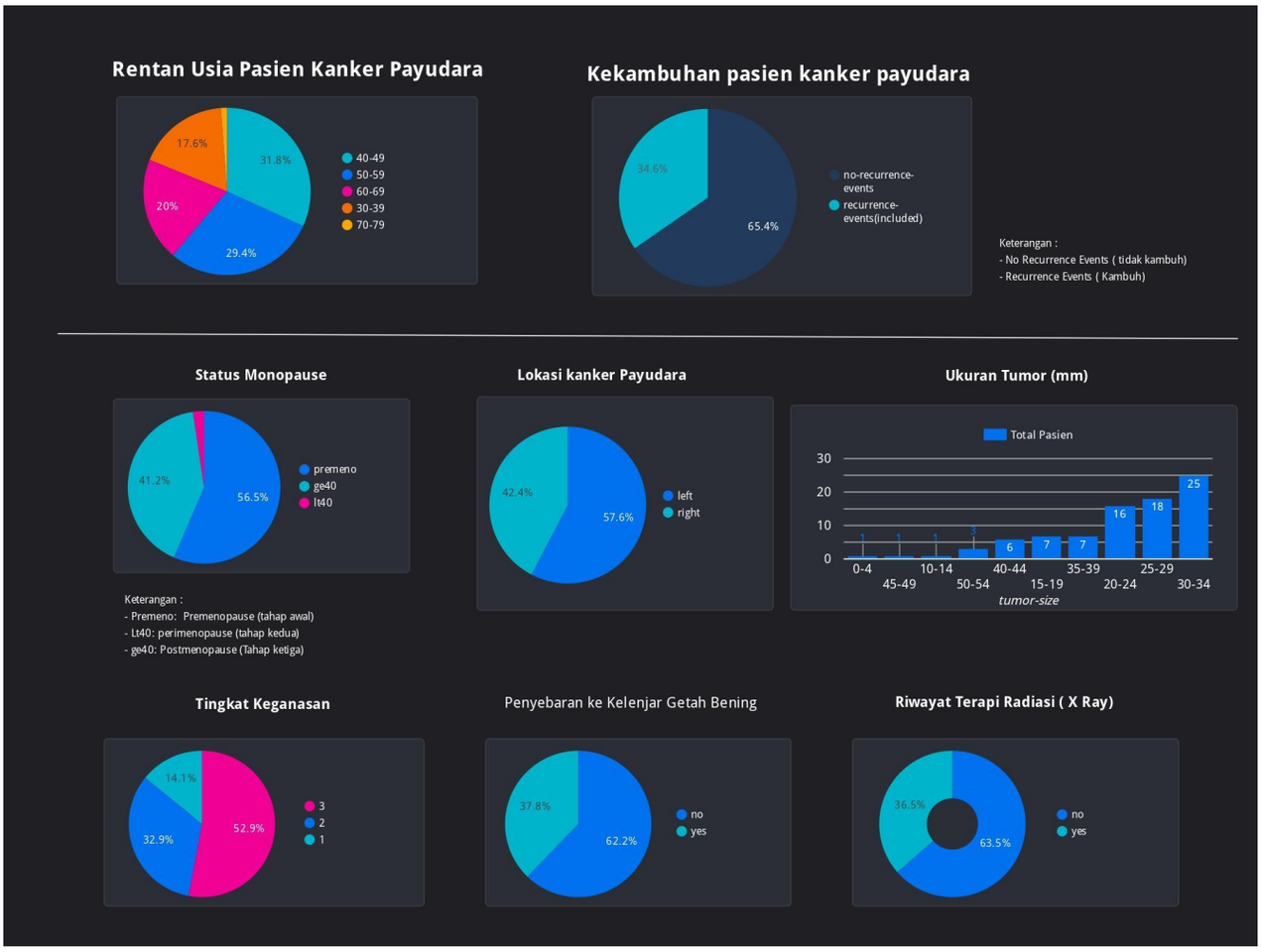
- Tahap kedua ini berfokus pada data yang akan diolah dan dianalisis menggunakan teknik data mining. Data harus dipahami secara utuh agar metode dan hasil penambangan data menjadi lebih optimal. Dataset "**Breast Cancer.csv**" terdiri dari 286 data dengan 10 atribut. Berikut ini penjelasan dari masing-masing atribut:

Atribut	Deskripsi	Isi Atribut
Age	Usia pasien pada saat diagnosis	10-19, 20-29, 30-39, 40-49, 50-59, 60-69,

		70-79, 80-89, 90-99
Menopause	status menopause (masa berakhirnya siklus menstruasi) pasien pada saat diagnosis	premeno : Premenopause (tahap awal) Lt40: perimenopause (tahap kedua) ge40: Postmenopause (Tahap ketiga)
Tumor size	ukuran tumor (dalam mm)	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
Inv-nodes	Kelenjar getah bening aksila (ketiak) Rentang nilai antara 0 - 39 yang menunjukkan kanker payudara pada saat pemeriksaan histologis	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
Node caps	Penyebaran ke kelenjar getah bening	Yes, No
Deg-Malig (Degree of malignancy)	Tingkat keganasan	1, 2, 3
Breast	Lokasi kanker payudara	Left, Right
Breast-Quad (Breast quadrant)	Kuadran lokasi Payudara	Left_up, Left_low, Right_up, Right_low, central
Irradiat (Irradiation)	Riwayat terapi menggunakan radiasi (X-Ray)	Yes, No
Class	Atribut kelas	no-recurrence-events (tidak kambuh), recurrence-events (kambuh)

- Lakukan eksplorasi data, laporkan fakta-fakta yang Anda anggap menarik dalam data tersebut. Misalnya dengan memvisualisasikan data tersebut dalam bentuk grafik dan beri penjelasan.

Presentasi Kambuhnya kanker payudara pada pasien



Penjelasan Visualisasi Data Kanker Payudara

Dari data didapatkan bahwa sebesar 33,6% pasien paling banyak terkena kanker payudara rentan umur 50-59 tahun dan paling banyak disaat memasuki status monopause tahap awal.

Dari banyaknya data sebesar 65,4% pasien tidak mengalami kambuhnya kanker payudara dan 34,6 % mengalami kambuh kanker payudara

Dari 34,6 % pasien yang mengalami kambuh kanker payudara paling banyak lokasi kanker di payudara sebelah kiri dan ukuran tumor berada diukuran 30-34 mm.

Pada pasien yang mengalami kambunnya kanker payudara memiliki tingkat keganasan kanker pada level 3 dan juga tidak melakukan terapi radiasi (X ray)

Link Visualisasi : [Laporan Data Kanker Payudara (google.com)]

Tahap 3: Data Preparation (Nilai 10)

Tahapan ini bertujuan untuk mempersiapkan data agar siap diolah. Seperti yang sudah dipelajari, tahap ini dapat berupa data cleaning, data reduction, data transformation, data discretization dan data integration. Lakukan **load** dataset ke Rapidminer, dan lakukan hal-hal berikut ini:

1. Periksa apakah ada data yang harus diperbaiki, dihilangkan, outlier, dll? Jika ada, apa yang Anda lakukan? Jelaskan! Penjelasan dapat disertai gambar
2. Tampilkan cuplikan data setelah dilakukan proses data preparation.

1. **Menghilangkan data duplikat**

Before (sebelum dihilangkan jumlah data sebanyak **286 example**)

ExampleSet (Retrieve breast-cancer) | ExampleSet (Remove Duplicates) | ExampleSet (Replace Missing Values)

Open in Turbo Prep | Auto Model | Filter (286 / 286 examples): all

Row No.	class	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast
1	no-recurrence-...	30-39	premeno	30-34	0-2	no	3	left	left_low
2	no-recurrence-...	40-49	premeno	20-24	0-2	no	2	right	right_u
3	no-recurrence-...	40-49	premeno	20-24	0-2	no	2	left	left_low
4	no-recurrence-...	60-69	ge40	15-19	0-2	no	2	right	left_up
5	no-recurrence-...	40-49	premeno	0-4	0-2	no	2	right	right_lo
6	no-recurrence-...	60-69	ge40	15-19	0-2	no	2	left	left_low
7	no-recurrence-...	50-59	premeno	25-29	0-2	no	2	left	left_low
8	no-recurrence-...	60-69	ge40	20-24	0-2	no	1	left	left_low
9	no-recurrence-...	40-49	premeno	50-54	0-2	no	2	left	left_low
10	no-recurrence-...	40-49	premeno	20-24	0-2	no	2	right	left_up
11	no-recurrence-...	40-49	premeno	0-4	0-2	no	3	left	central
12	no-recurrence-...	50-59	ge40	25-29	0-2	no	2	left	left_low
13	no-recurrence-...	60-69	lt40	10-14	0-2	no	1	left	right_u
14	no-recurrence-...	50-59	ge40	25-29	0-2	no	3	left	right_u
15	no-recurrence-...	40-49	premeno	30-34	0-2	no	3	left	left_up
16	no-recurrence-...	60-69	lt40	30-34	0-2	no	1	left	left_low

ExampleSet (286 examples, 1 special attribute, 9 regular attributes)

After Remove Duplicat (total data sebanyak **266 example**), sehingga terdapat 20 data duplikat

Open in Turbo Prep Auto Model Filter (266 / 266 examples): all

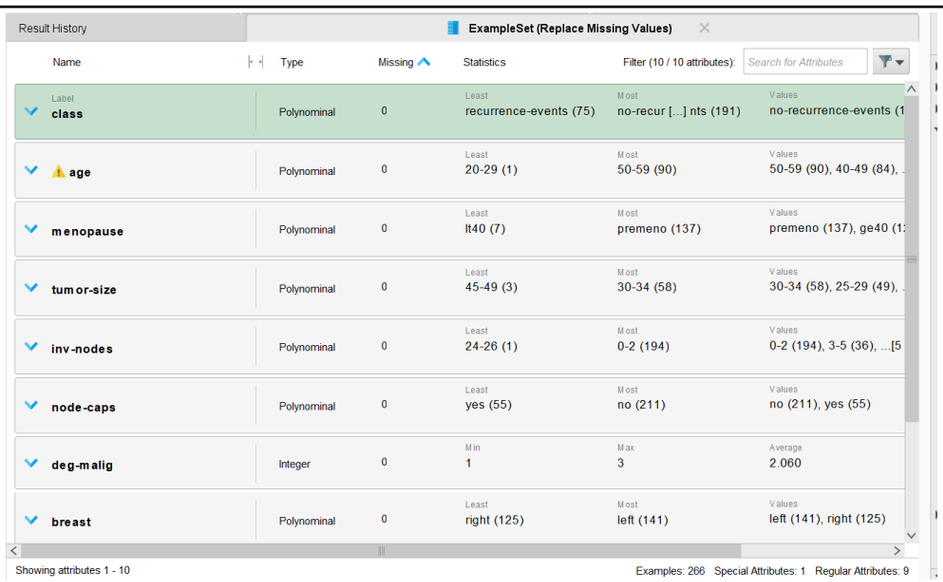
Row No.	class	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad
1	no-recurrence-...	30-39	premeno	30-34	0-2	no	3	left	left_low
2	no-recurrence-...	40-49	premeno	20-24	0-2	no	2	right	right_up
3	no-recurrence-...	40-49	premeno	20-24	0-2	no	2	left	left_low
4	no-recurrence-...	60-69	ge40	15-19	0-2	no	2	right	left_up
5	no-recurrence-...	40-49	premeno	0-4	0-2	no	2	right	right_low
6	no-recurrence-...	60-69	ge40	15-19	0-2	no	2	left	left_low
7	no-recurrence-...	50-59	premeno	25-29	0-2	no	2	left	left_low
8	no-recurrence-...	60-69	ge40	20-24	0-2	no	1	left	left_low
9	no-recurrence-...	40-49	premeno	50-54	0-2	no	2	left	left_low
10	no-recurrence-...	40-49	premeno	20-24	0-2	no	2	right	left_up
11	no-recurrence-...	40-49	premeno	0-4	0-2	no	3	left	central
12	no-recurrence-...	50-59	ge40	25-29	0-2	no	2	left	left_low
13	no-recurrence-...	60-69	lt40	10-14	0-2	no	1	left	right_up
14	no-recurrence-...	50-59	ge40	25-29	0-2	no	3	left	right_up
15	no-recurrence-...	40-49	premeno	30-34	0-2	no	3	left	left_up
16	no-recurrence-...	60-69	lt40	30-34	0-2	no	1	left	left_low

2. Menghilangkan data Missing Value

Before (terdapat beberapa data yang missing)

Name	Type	Missing	Statistics	Filter (10 / 10 attributes): Search for Attributes	
node-caps	Polynominal	8	Least: yes (55)	Most: no (203)	Values: no (203), yes (55)
breast-quad	Polynominal	1	Least: central (19)	Most: left_low (101)	Values: left_low (101), left_up
Label: class	Polynominal	0	Least: recurrence-events (75)	Most: no-recur [...] nts (191)	Values: no-recurrence-events
age	Polynominal	0	Least: 20-29 (1)	Most: 50-59 (90)	Values: 50-59 (90), 40-49 (84)
menopause	Polynominal	0	Least: lt40 (7)	Most: premeno (137)	Values: premeno (137), ge40
tumor-size	Polynominal	0	Least: 45-49 (3)	Most: 30-34 (58)	Values: 30-34 (58), 25-29 (49)
inv-nodes	Polynominal	0	Least: 24-26 (1)	Most: 0-2 (194)	Values: 0-2 (194), 3-5 (36), ...
deg-malig	Integer	0	Min: 1	Max: 3	Average: 2.060

After Remove Missing value



Tahap 4: Modeling (Bobot = 15%)

Tahap pemodelan bertujuan untuk menerapkan model / metode data mining yang tepat untuk menganalisis data. Anda diminta membandingkan **minimal** 3 (tiga) metode klasifikasi. Jelaskan secara singkat kelebihan dan kekurangan ketiga metode tersebut, serta apakah metode tersebut cocok diterapkan pada dataset “Breast Cancer”?

1. **Model Decision tree**
   - Kelebihan Decision tree
     - Mudah diimplementasikan
     - Hipotesis yang dihasilkan mudah dipahami
     - Efisien
   - Kekurangan Decision tree
     - overfitting
     - Pohon keputusan dapat tumbuh menjadi sangat kompleks pada data yang rumit.
     - Kurang efektif dalam memprediksi hasil dari variabel kontinu

2. **Model Naive Bayes**
   - Kelebihan Naive Bayes
     - Bisa dipakai untuk data kuantitatif maupun kualitatif
     - Tidak perlu melakukan data training yang banyak
     - Mudah dipahami
   - Kekurangan Naive Bayes
     - Apabila probabilitas kondisionalnya bernilai nol, maka probabilitas prediksi juga akan bernilai nol
     - Asumsi bahwa masing-masing variabel independen membuat berkurangnya akurasi, karena biasanya ada korelasi antara variabel yang satu dengan variabel yang lain

3. **Model Random Forest**
    - Kelebihan Random Forest
        - Kuat terhadap data outlier (pencilan data).
        - Risiko overfitting lebih rendah.
        - Berjalan secara efisien pada kumpulan data yang besar.
    - Kekurangan Random Forest
        - Random Forest cenderung bias saat berhadapan dengan variabel kategorikal.
        - Waktu komputasi pada dataset berskala besar relatif lamba

Ketiga Model tersebut merupakan model teknik algoritma supervised.

Model tersebut cocok untuk untuk data Breast cancer yang sudah diketahui kelasnya, karena tujuan awalnya ialah untuk memprediksi faktor kambuh atau tidaknya kanker dengan mempelajari pola dengan mengklasifikasikan data.

Tahap 5: Evaluation (Nilai 25)

Untuk mengetahui performa dari masing-masing algoritma / metode data mining, dilakukan pengujian atau evaluasi performa. Anda diminta untuk mengevaluasi performa dari masing-masing metode data mining yang Anda gunakan pada tahap 4, dengan ketentuan sbb:

- Lakukan pengujian dengan menggunakan **10-fold Cross Validation.**
- Ukuran performa yang digunakan: **akurasi** (accuracy), **presisi** (*weighted mean precision*), dan **recall** (*weighted mean recall*)
- Screenshot hasil performa dari setiap metode yang diuji (tabel confusion matrix)

- Decision Tree

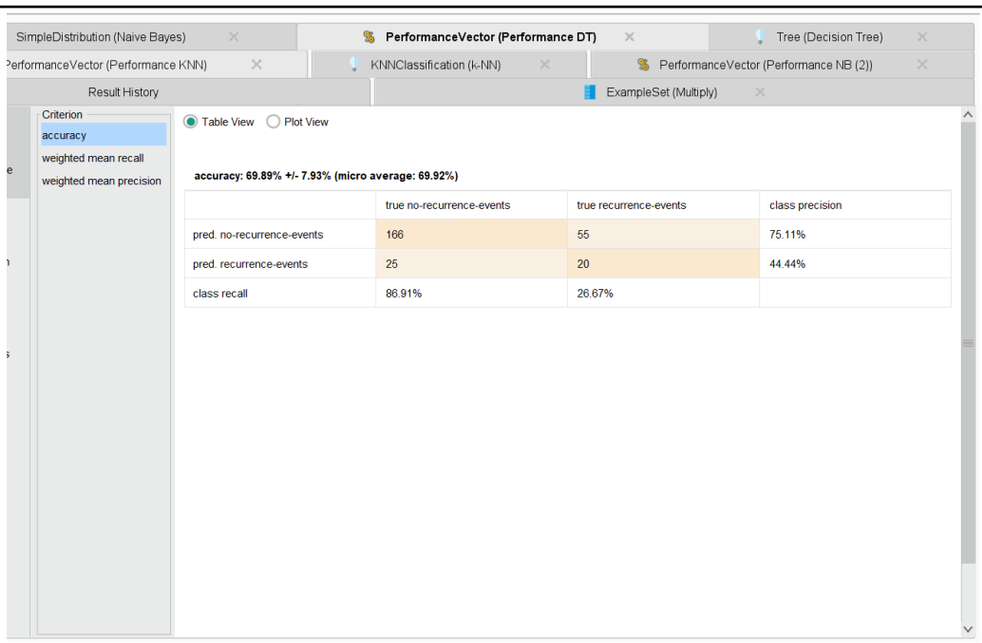


Table View Plot View

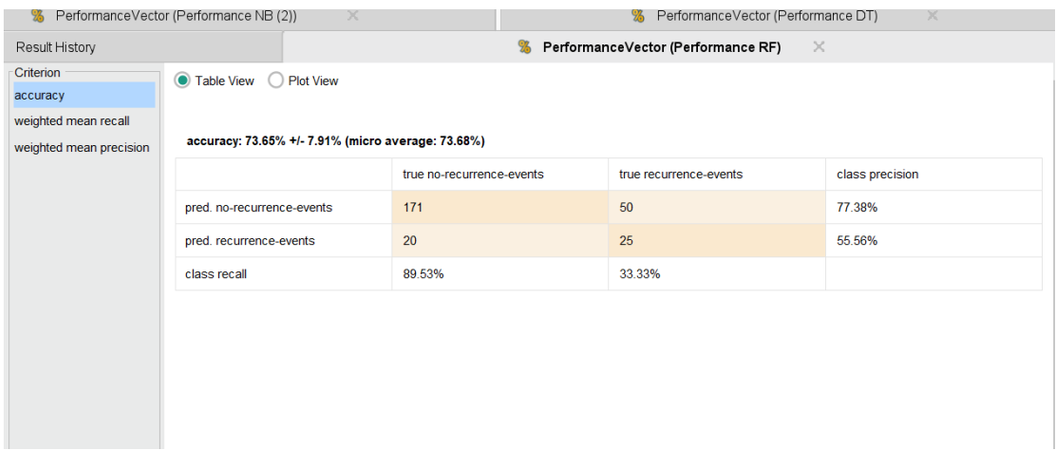
weighted_mean_recall: 56.59% +/- 10.08% (micro average: 56.79%), weights: 1, 1

	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	166	55	75.11%
pred. recurrence-events	25	20	44.44%
class recall	86.91%	26.67%	

weighted_mean_precision: 60.03% +/- 17.44% (micro average: 59.78%), weights: 1, 1

	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	166	55	75.11%
pred. recurrence-events	25	20	44.44%
class recall	86.91%	26.67%	

- Random Forest



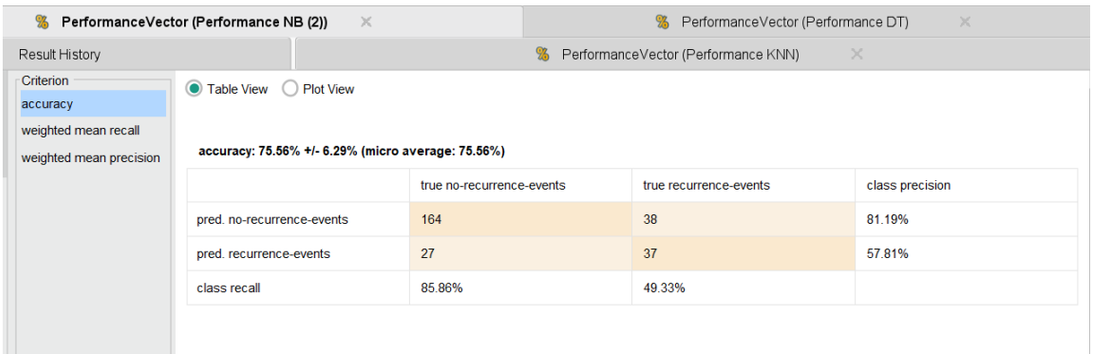
weighted_mean_recall: 61.28% +/- 11.29% (micro average: 61.43%), weights: 1, 1

	true no-recurrence-events	true recurrence-events	class precision
ıred. no-recurrence-events	171	50	77.38%
ıred. recurrence-events	20	25	55.56%
:lass recall	89.53%	33.33%	

weighted_mean_precision: 65.19% +/- 15.04% (micro average: 66.47%), weights: 1, 1

	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	171	50	77.38%
pred. recurrence-events	20	25	55.56%
class recall	89.53%	33.33%	

- Naive Bayes



weighted_mean_recall: 67.65% +/- 9.30% (micro average: 67.60%), weights: 1, 1

	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	164	38	81.19%
pred. recurrence-events	27	37	57.81%
class recall	85.86%	49.33%	

weighted_mean_precision: 69.94% +/- 7.99% (micro average: 69.50%), weights: 1, 1

	true no-recurrence-events	true recurrence-events	class precision
pred. no-recurrence-events	164	38	81.19%
pred. recurrence-events	27	37	57.81%
class recall	85.86%	49.33%	

Tahap 6: Hasil dan Pembahasan (Nilai 25)

Tahapan ini merupakan bagian penting dari project data mining. Pada tahapan ini, Anda diminta menganalisis hasil dari ujicoba data mining dan menginterpretasikan (menjelaskan) hasil ujicoba tersebut.

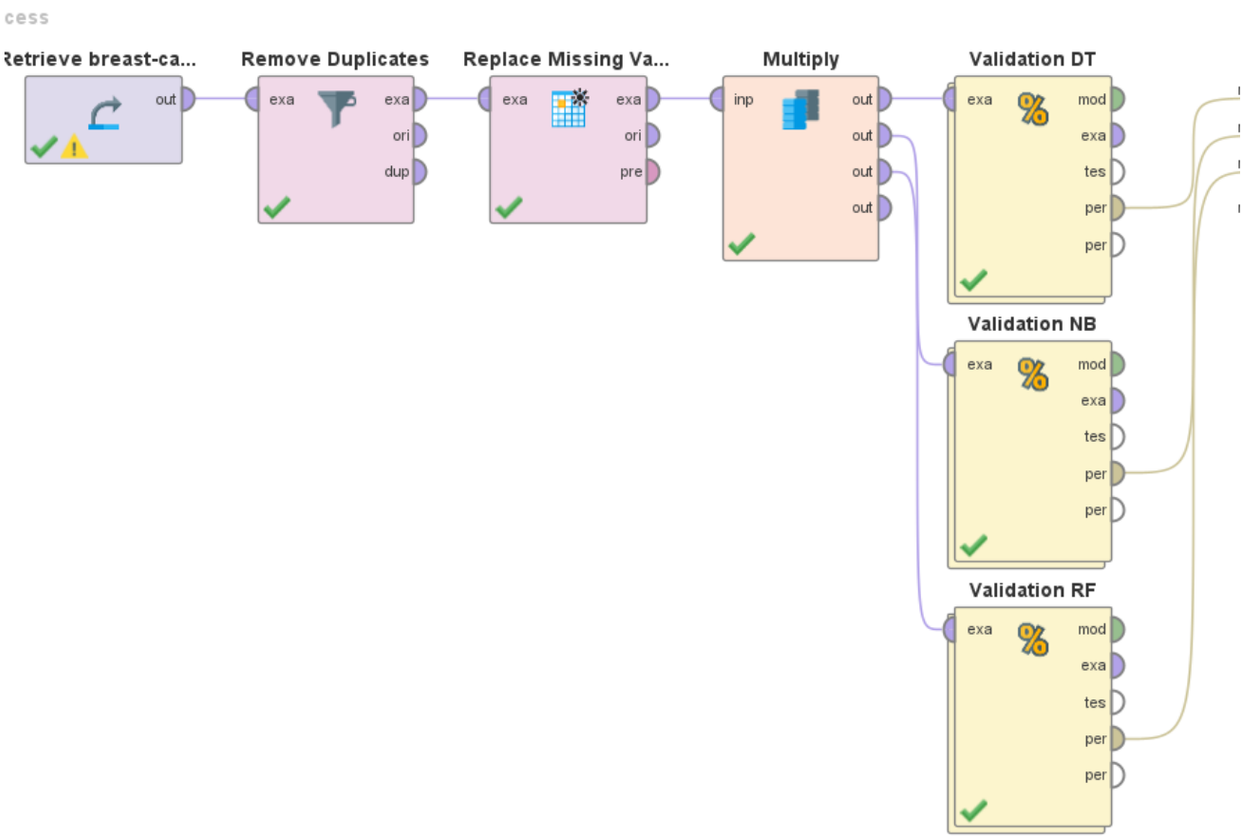
Tujuan dari UTS ini adalah membandingkan performa metode data mining. Lengkapi tabel perbandingan performa metode data mining sebagai berikut:

Metode	Akurasi	Presisi	Recall
Decision Tree	69.96%	58.86%	56.75%
Naive Bayes	75.56%	69.94%	67.65%
Random Forest	73.65%	65.19%	61.28%

Berdasarkan hasil perbandingan di atas, interpretasikan dan jelaskan hasil tersebut! Menurut Anda, algoritma / metode apa yang memiliki performa terbaik? Nyatakan kesimpulan Anda!

Dari hasil perbandingan performa antara model Decision tree, Naive Bayes dan Random Forest, diketahui Naive Bayes memiliki nilai akurasi, presisi dan recall lebih besar ketimbang kedua model lainnya, maka dari hasil performance tersebut model **naive bayes yang memiliki performa terbaik**

Screenshot / sajikan model ujicoba Anda di Rapidminer!



~ Selamat Mengerjakan ~