

Pembuatan Predictive Model Decision Tree untuk Segmentasi Pasar Pelanggan The Body Shop

Dokumentasi Project

COMP6140 - Data Mining

Tema: The Body Shop



Jacelyn Angraini - 2201789896

Kenny ongko – 2201798686

LF01

Daftar Isi

1. Pendahuluan.....	1
1.1 Tujuan.....	1
1.2 Manfaat.....	1
1.2 Dasar Teori.....	2
1.2.1 The Body Shop	2
1.2.2 Knowledge Discovery of Data (KDD)	3
1.2.3 Market Segmentation	4
1.2.4 Decision Tree	5
2. Metodologi.....	5
2.1 Data Collection.....	5
2.2 Data Cleaning	10
2.3 Data Integration.....	14
2.4 Data Selection	15
2.5 Data Transformation	17
2.6 Data Mining.....	20
2.7 Evaluation and Presentation	28
3. Kesimpulan.....	29
4. Referensi	30

1. Pendahuluan

Pada project ini, kami melakukan segmentasi pasar (*market segmentation*) terhadap data pelanggan The Body Shop untuk menentukan ciri-ciri dari pelanggan yang menyukai kategori produk tertentu yang ditawarkan oleh The Body Shop. Untuk mendapatkan karakteristik pelanggan tersebut, maka kami perlu melakukan data mining dengan menggunakan metode klasifikasi. Klasifikasi memerlukan penggunaan algoritma *machine learning* yang mempelajari cara menetapkan label kelas ke contoh dari domain masalah. Ada beberapa algoritma populer yang digunakan untuk melakukan klasifikasi yaitu seperti *k-Nearest Neighbours* (KNN), *Decision Trees*, *Support Vector Machine* (SVM), *Naïve Bayes*, dan lainnya. Pada project ini, kami memutuskan untuk menggunakan *Decision Tree*.

1.1 Tujuan

Tujuan kami membuat project ini yaitu sebagai berikut:

- Untuk mencari tahu kategori produk yang paling disukai oleh pelanggan The Body Shop.
- Untuk mencari tahu karakteristik demografi dari pelanggan yang menyukai kategori produk tertentu dari The Body Shop

1.2 Manfaat

Manfaat yang didapat dari project kami yaitu sebagai berikut:

- Dengan mengetahui segmentasi dari pelanggan The Body Shop, informasi ini dapat memudahkan pihak marketing untuk memberikan iklan yang lebih personal atau lebih sesuai dengan selera setiap pelanggan secara individu. Dengan begitu, marketing yang dilakukan akan lebih efektif sehingga dapat meningkatkan loyalitas pelanggan dan juga menambah pendapatan perusahaan.
- Informasi yang didapat juga akan membantu pihak marketing dalam memberikan iklan personal kepada pelanggan baru berdasarkan karakteristiknya. Sehingga, meskipun pelanggan tersebut belum sering berbelanja, pihak marketing mempunyai gambaran umum terhadap preferensi dari pelanggan tersebut berdasarkan ciri-ciri demografinya.

1.2 Dasar Teori

1.2.1 The Body Shop

The Body Shop adalah perusahaan kosmetik dan kecantikan global yang mendapat inspirasi dari alam dan menghasilkan produk – produk yang bersandar pada nilai nilai etika. Pertama kali The Body Shop didirikan pada tahun 1976 oleh Dame Anita Roddick di Inggris, Saat pertama kali membuka toko, Anita hanya mampu menjual 25 jenis Produk kecantikan yang dibuat dengan tangan (hand-made), namun berkembang dengan sangat pesat hingga terdapat cabang di seluruh dunia. The Body Shop terus berkembang dari tahun ke tahun hingga pada 2006 The Body Shop di beli L’Oreal dan kini The Body Shop memiliki rangkaian produk sebanyak 1,200 macam meliputi produk kosmetik dan makeup di 2,500 toko yang tersebar di 61 negara di dunia.

The Body Shop Indonesia pertama kali membuka tokonya di Pondok Indah Mall pada tanggal 12 Desember 1992 dan sampai saat ini terus memperbanyak gerainya di wilayah Indonesia. The Body Shop dalam menjalankan usahanya yang di wujudkan melalui kepedulian dan tanggung jawab terhadap perubahan social dan lingkungan. Nilai – nilai (values) The Body Shop ini akhirnya dipandang sebagai value added yang sangat signifikan dalam meningkatkan gaya hidup konsumennya.

Produk The Body Shop di bagi menjadi beberapa category yaitu Wellbeing, Make-up, Bath and Body, Skin Care, Men’s, Home Fragrance, Fragrance, Hair, Accessories, dan Gifts. Produk – produk The Body Shop ini umumnya ditujukan untuk perempuan sehingga sebagian besar konsumen The Body Shop adalah perempuan, namun ada juga rangkaian produk yang ditujukan untuk konsumen pria sehingga target konsumen tidak hanya terbatas pada kaum wanita saja. The Body Shop merupakan salah satu perusahaan kosmetik paling berpengaruh di dunia karena selalu berpegang teguh pada filosofi serta misi mereka yang salah satunya adalah berusaha untuk melakukan perubahan sosial yang lebih baik.

1.2.2 Knowledge Discovery of Data (KDD)

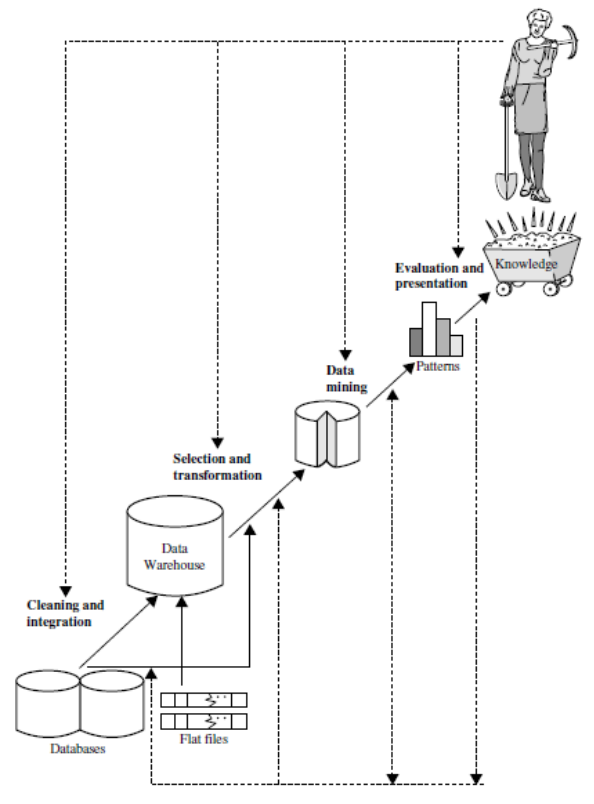


Figure 1.4 Data mining as a step in the process of knowledge discovery.

Dalam proses KDD (knowledge discovery from data) ada beberapa langkah yang harus dilakukan yaitu:

1. Cleaning and integration

Data cleaning digunakan untuk menghilangkan data yang inkonsisten dan salah dari data set. Data integration dilakukan untuk menggabungkan data yang berasal dari sumber yang berbeda ke dalam satu tempat.

2. Selection and transformation

Data selection adalah tahapan dimana data yang relevan pada analysis yang akan dilakukan diambil dari database. Data transformation adalah tahapan dimana data diubah kedalam bentuk yang sesuai dengan melakukan operasi summary atau aggregate.

3.Data mining

Data mining adalah tahapan dimana algoritma digunakan untuk mengekstrak pattern yang bermakna dan berguna dari data set.

4.Evaluation and presentation

Pattern evaluation digunakan untuk mengidentifikasi pola yang mewakili pengetahuan berdasarkan ukuran yang diberikan. Knowledge presentation didefinisikan sebagai teknik yang menggunakan alat visualisasi untuk merepresentasikan hasil data mining.

1.2.3 Market Segmentation

Strategic planning bergantung pada kemampuan untuk mengumpulkan informasi mengenai keinginan pelanggan dan dapat memproses informasi itu menjadi prediksi persyaratan di masa depan. Dalam situasi di mana keinginan pelanggan besar, diperlukan metode untuk mengurangi jumlah skenario yang mungkin terjadi. Pengurangan skenario umumnya dicapai dalam bisnis melalui proses yang dikenal sebagai segmentasi pasar. Penelitian awal telah menunjukkan bahwa memperkirakan kelompok pelanggan (atau produk) lebih akurat daripada perkiraan individu agregat karena efek positif dari *error smoothing* dan *error cancelation*.

Segmentasi Pasar adalah proses mempartisi pasar ke dalam kelompok pelanggan dan prospek dengan kebutuhan dan / atau karakteristik yang sama yang cenderung menunjukkan perilaku pembelian serupa. Segmentasi pasar adalah tugas penting dalam pemasaran. Hal ini memungkinkan staf pemasaran untuk mengetahui metode pemasaran apa yang dapat mereka gunakan untuk kelompok tertentu di pasar. Mereka dapat mencampur dan mencocokkan kombinasi yang berbeda dari harga, promosi, dan tempat produk. Mereka juga menggunakan segmentasi pasar untuk mengetahui pelanggan mana yang dapat mempertahankan loyalitasnya atau pelanggan yang kemungkinan akan lebih bersedia membeli produk mereka.

Segmentasi pasar dapat membantu perusahaan untuk menentukan dan lebih memahami audiens target dan pelanggan idealnya. Segmentasi pasar menawarkan opsi iklan yang lebih tepat sasaran dan untuk menyesuaikan konten mereka untuk grup audiens yang berbeda. Segmentasi pasar memungkinkan perusahaan untuk menargetkan konten ke orang yang tepat dengan cara yang benar, daripada menargetkan seluruh audiens dengan pesan umum. Ketika pesan yang dikirim tidak dioptimalkan untuk audiens, yang terjadi adalah perusahaan tersebut akan berakhir dengan banyak biaya iklan yang terbuang sia-sia. Segmentasi pasar membantu meningkatkan kemungkinan orang berinteraksi dengan iklan atau konten yang ditawarkan, menghasilkan kampanye yang lebih efisien dan peningkatan laba atas investasi (ROI).

1.2.4 Decision Tree

Decision tree adalah salah satu cara data mining dalam memprediksi masa depan dengan membangun klasifikasi atau regresi model dalam bentuk struktur pohon (*tree*). Hasil akhir dari proses tersebut adalah pohon dengan node *decision* dan node *leaf*. Sebuah node *v* (misalnya, Cuaca/ Outlook) memiliki dua atau lebih cabang misalnya, Panas, Berawan dan Hujan.

2. Metodologi

2.1 Data Collection

Untuk melaksanakan project ini, kami membuat data dummy pelanggan yang totalnya berjumlah 260 baris. Kami membagi data yang kami buat menjadi 2 data set dimana 80,77% (210 data) dari data digunakan sebagai data training dan 19,23% (50 data) sisanya digunakan sebagai data testing. Berikut atribut yang terdapat di data dummy yang kami buat:

- CustomerID
- CustomerName
- CustomerGender
- CustomerDOB
- CustomerAddress
- CustomerNumber

- CustomerEmail
- MaritalStatus
- WebsiteActivity
- PaymentMethod
- MajorProductCategory

Atribut “MajorProductCategory” bertindak sebagai label atau hasil yang ingin kita dapat yaitu kategori favorit dari pelanggan dengan karakteristik demografi tertentu. Oleh karena atribut ini merupakan label, maka pada data set testing tidak terdapat atribut “MajorProductCategory”. Ada 5 kemungkinan nilai dari label ini yaitu:

- Shower Gel
- EDT
- Body Mist
- Body Lotion
- Shampoo

Kami memilih untuk menggunakan 5 nilai ini karena merupakan kategori favorit dari pelanggan The Body Shop. The Body Shop sendiri mempunyai banyak kategori produk yang ditawarkan yaitu sekitar 98 buah menurut website resminya yaitu www.thebodyshop.co.id. Menurut kami jumlah tersebut terlalu besar dan tidak efektif untuk menggunakan semua kategori produk yang ada untuk skala project kami, sehingga kami memutuskan untuk memilih 5 kategori produk terbaik yang ada di The Body Shop.

Data set Training dan Testing (260 data)

AutoSave Off TBS_All - Excel KENNY ONGKO

File Home Insert Page Layout Formulas Data Review View Help Search

Clipboard Font Alignment Number Styles Cells Editing Ideas

POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format. Don't show again Save As...

CustomerID;CustomerName;CustomerGender;CustomerDOB;CustomerAddress;CustomerNumber;CustomerEmail;MaritalStatus;WebsiteActivity;PaymentMethod;MajorProductCategory










	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	CustomerID	CustomerName	CustomerGender	CustomerDOB	CustomerAddress	CustomerNumber	CustomerEmail	MaritalStatus	WebsiteActivity	PaymentMethod	MajorProductCategory																		
2	C0001	Bengkulu	Ja	Teng	0817-555-863	bagussuv.arno@gmail.com	Single	3	Internet	Payment	Body Lotion																		
3	C0002	D	Pagar Al	KalBar	0839-555-382	damanw.acana@gmail.com	Single	4	Credit	Shampoo																			
4	C0003	Li	Administrasi	Jakarta Utara	0838-555-603	lmarjagawidodo@gmail.com	Single	1	Bank	Transfer	Shampoo																		
5	C0004	H	Mojokerto	Ja	Tim	0878-555-460	harjobalaminmansur@gmail.com	Single	2	Credit	Shampoo																		
6	C0005	P	Kupang	TKI	0853-555-222	praba_haihm@gmail.com	Single	2	Bank	Transfer	Shampoo																		
7	C0006	A	Bima 711	Aceh	0878-555-301	artajalans.kom@gmail.com	Single	2	Internet	Payment	Shampoo																		
8	C0007	Ei	Balkpap	KalTim	0850-555-628	entengmaridisaraghm.ti@gmail.com	Single	4	Credit	Shower Gel																			
9	C0008	D	Kediri 20	KalTim	0816-555-873	di.hendisuv.arno@gmail.com	Single	1	Bank	Transfer	Shower Gel																		
10	C0009	B	Madun 1	SulTeng	0838-555-323	bagiagunanto@gmail.com	Single	1	Virtual Account	Shampoo																			
11	C0010	D	Bandar L	Aceh	0856-555-633	dadputra@gmail.com	Single	1	Credit	Shampoo																			
12	C0011	Pu	Binjai 32	SulTira	0818-555-333	puvazuli.aman@gmail.com	Single	2	Bank	Transfer	Shampoo																		
13	C0012	Tu	Bandar 7	Ja	Bar	0838-555-105	tugmanistompul@gmail.com	Single	5	Virtual Account	EDT																		
14	C0013	Li	Baru 233	PapBar	0836-555-920	lamanarpati@gmail.com	Married	5	Bank	Transfer	Shampoo																		
15	C0014	D	Administ	SunSel	0870-555-917	dalmanutama@gmail.com	Married	2	Internet	Payment	Shampoo																		
16	C0015	K	Sawahlu	Maluku	0850-555-732	kusumav.atarpanov@gmail.com	Married	2	Internet	Payment	Shampoo																		
17	C0016	D	Bandar L	SulTira	0853-555-211	danugaranhutaapea@gmail.com	Married	3	Credit	EDT																			
18	C0017	D	Palemba	KalSel	0815-555-836	damarpragofligantoro@gmail.com	Married	4	Bank	Transfer	Shower Gel																		
19	C0018	Ar	Madun 6	Brau	0-837-555-605	artantoluhursantoso@gmail.com	Married	5	Internet	Payment	Shampoo																		
20	C0019	H	Paranasi	Banten	0818-555-778	himav.anklupono@gmail.com	Married	5	Credit	Shower Gel																			
21	C0020	A	Baru 531	Papua	0830-555-801	lumanis.amos@gmail.com	Married	4	Bank	Transfer	Shampoo																		
22	C0021	M	Magelan	SunBar	0836-555-416	malikajellasturyanis.kom@gmail.com	Married	5	Virtual Account	EDT																			
23	C0022	P	Lubuklin	KepR	0855-555-156	putiparisuatinis.pd@gmail.com	Married	3	Credit	Body Mist																			
24	C0023	S	Pasuraa	PapBar	0837-555-876	sakurasusanti@gmail.com	Married	3	Bank	Transfer	EDT																		
25	C0024	Z	Palangki	DIY	0836-555-163	zaenabyuniar@gmail.com	Married	1	Bank	Transfer	EDT																		
26	C0025	Ei	Temane	SunSel	0838-555-541	elishahayiah@gmail.com	Married	1	Virtual Account	EDT																			
27	C0026	D	Tanjung	Ja	Bar	0838-555-293	onkurv.anda@gmail.com	Married	3	Bank	Transfer	Shower Gel																	
28	C0027	V	Blitar 95	PapBar	0839-555-916	victoriamaldanasyiah@gmail.com	Married	5	Internet	Payment	Shower Gel																		

TBS All

Ready Type here to search

8:24 PM 12/13/2020

Meta Data dari Data set Training dan Testing:

Name 	 Type	Missing
 CustomerAddress	Polynominal	0
 CustomerDOB	Date	2
 CustomerEmail	Polynominal	0
 CustomerGender	Polynominal	5
 CustomerName	Polynominal	0
 CustomerNumber	Polynominal	0
 MajorProductCategory	Polynominal	0

Data set Training (210 data)

AutoSaveOff

TBS_Training - Excel

KENNY ONGKO

FileHomeInsertPage LayoutFormulasDataReviewViewHelp

Search

ShareComments

Paste

Clipboard

Calibri11

Font

Wrap Text

Alignment

General

Number

Conditional Formatting

Styles

Format as Table

Cell Styles

Insert

Cells

Delete

Cells

Format

Cells

Sort & Filter

Editing

Find & Select

Editing

Ideas

Ideas

POSSIBLE DATA LOSS

Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

Don't show again

Save As...

A1

CustomerID;CustomerName ;CustomerGender;CustomerDOB;CustomerAddress;CustomerNumber;CustomerEmail;MaritalStatus;WebsiteActivity;PaymentMethod;MajorPro

ductCategory

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	CustomerID	CustomerName	CustomerGender	CustomerDOB	CustomerAddress	CustomerNumber	CustomerEmail	MaritalStatus	WebsiteActivity	PaymentMethod	MajorProductCategory																		
2	C0001B. Bengkulu	Ja	0817-555-863	bagussuv.arno@gmail.com	Single	3	Internet	Payment	Body Lotion																				
3	C0002D. Pagar Al	Kal	0839-555-382	damanw.acana@gmail.com	Single	4	Credit	Shampoo																					
4	C0003L. Administrasi	Jakarta	0838-555-603	lmajagawidodo@gmail.com	Single	1	Bank	Transfer	Shampoo																				
5	C0004H. Mojokert	Ja	0878-555-460	haryobakamartiansur@gmail.com	Single	2	Credit	Shampoo																					
6	C0005P. Kupang	DKI	0859-555-222	praba_hajajasa_halm@gmail.com	Single	2	Bank	Transfer	Shampoo																				
7	C0006A. Bima	711	Aceh	0878-555-301	artajallanz.kom@gmail.com	Single	2	Internet	Payment	Shampoo																			
8	C0007E. Balikpapan	Kal	0856-555-628	erengmarladizragihm.t@gmail.com	Single	4	Credit	Shower Gel																					
9	C0008D. Kediri	20	Kal	0816-555-673	dhendrisu.arno@gmail.com	Single	1	Bank	Transfer	Shower Gel																			
10	C0009B. Madun	1	Sul	0838-555-323	bagayaganto@gmail.com	Single	1	Virtual Account	Shampoo																				
11	C0010D. Bandar	L	Aceh	0856-555-639	dadputra@gmail.com	Single	1	Credit	Shampoo																				
12	C0011P. Binjai	32	Sul	0818-555-393	puvazul.karnan@gmail.com	Single	2	Bank	Transfer	Shampoo																			
13	C0012T. Banjar	7	Ja	0838-555-105	rugimanustompul@gmail.com	Single	5	Virtual Account	EDT																				
14	C0013L. Batu	23	Pap	0836-555-920	lamamapari@gmail.com	Married	5	Bank	Transfer	Shampoo																			
15	C0014D. Administ	Sun	0878-555-317	dalmamutama@gmail.com	Married	2	Internet	Payment	Shampoo																				
16	C0015K. Sawahlu	Maluku	0858-555-782	luzumaw.antapanov@gmail.com	Married	2	Internet	Payment	Shampoo																				
17	C0016D. Bandar	L	Sul	0853-555-211	daugaranhutapea@gmail.com	Married	3	Credit	EDT																				
18	C0017D. Palembang	Kal	0815-555-696	damanprayogilgantoro@gmail.com	Married	4	Bank	Transfer	Shower Gel																				
19	C0018Ar	Madun	6	Riau	0-897-555-605	atanoluhurasantoso@gmail.com	Married	5	Internet	Payment	Shampoo																		
20	C0019H. Parame	Banten	0878-555-778	lilmaw.antiapono@gmail.com	Married	5	Credit	Shower Gel																					
21	C0020A. Batu	531	Papu	0838-555-601	umarizamas@gmail.com	Married	4	Bank	Transfer	Shampoo																			
22	C0021M. Magelan	Sun	0896-555-418	malikagelkasuryatmis.kom@gmail.com	Married	5	Virtual Account	EDT																					
23	C0022P. Lubuklin	Kep	0855-555-158	putiparisuatinis.pd@gmail.com	Married	3	Credit	Body Mist																					
24	C0023S. Pasurus	Pap	0897-555-878	rakurasusanti@gmail.com	Married	3	Bank	Transfer	EDT																				
25	C0024Z. Palang	DKI	0846-555-153	caenabiyunsa@gmail.com	Married	1	Bank	Transfer	EDT																				
26	C0025E. Ternate	Sun	0838-555-554	elichaharyah@gmail.com	Married	1	Virtual Account	EDT																					
27	C0026O. Tanjung	Ja	0838-555-293	onikusw.andari@gmail.com	Married	3	Bank	Transfer	Shower Gel																				
28	C0027V. Bitar	95	Pap	0839-555-916	victoriaimaidanasyah@gmail.com	Married	5	Internet	Payment	Shower Gel																			

TBS_Training

Ready

Data set Testing (50 data)

FileHomeInsertPage LayoutFormulasDataReviewViewHelp

Search

ShareComments

Paste

Clipboard

Font

Alignment

Number

Styles

Cells

Editing

Ideas

Calibri11A⁺

BBIU

Wrap Text

General

Conditional FormattingTableCell Styles

InsertDeleteFormat

Sort & Find & Filter > Select >

Ideas

POSSIBLE DATA LOSS

Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

Don't show again

Save As...

A1

CustomerID;CustomerName ;CustomerGender;CustomerDOB;CustomerAddress;CustomerNumber;CustomerEmail;MaritalStatus;WebsiteActivity;PaymentMethod

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	CustomerID	CustomerName	CustomerGender	CustomerDOB	CustomerAddress	CustomerNumber	CustomerEmail	MaritalStatus	WebsiteActivity	PaymentMethod																			
2	C021	Palopo	5	Kal	0815-555-7970	293	iraiviv@gmail.com	Single	5	Credit																			
3	C021	Bogor	76	Riau	(+62) 835 7970 294	baliduan@gmail.com	Single	2	Bank	Transfer																			
4	C021	Mojokert	Maluku	(+62) 835 7970 295	kuncama@gmail.com	Single	4	Credit																					
5	C021	Manado	DKI	(+62) 835 7970 296	galari@gmail.com	Single	2	Bank	Transfer																				
6	C021	Magelan	Maluku	(+62) 835 7970 297	raulaspd@gmail.com	Single	5	Internet	Payment																				
7	C021	Administ	Pap	0815-555-220	2626	salasara@gmail.com	Single	4	Credit																				
8	C021	Serang	6	Sul	Bar-1934	zulakso@gmail.com	Single	1	Virtual Account																				
9	C021	Ambon	1	Sun	Bar-1931	jasmaing@gmail.com	Single	2	Credit																				
10	C021	Chokseu	Ja	Teng-1574	endib@gmail.com	Single	5	Virtual Account																					
11	C022	Lingsia	1	Lampung	-1672	nurul@gmail.com	Single	3	Credit																				
12	C022	Bandar A	Sul	-2179	anastasi@gmail.com	Single	5	Bank	Transfer																				
13	C022	Prabum	Riau	-1982	kamalno@gmail.com	Single	5	Bank	Transfer																				
14	C022	Dumai	8	Sul	Bar-1660	baligs@gmail.com	Married	2	Bank	Transfer																			
15	C022	Bandar L	Bengkulu	-1988	lami@gmail.com	Married	2	Internet	Payment																				
16	C022	Mojokert	Sul	Bar-1754	gnak@gmail.com	Married	4	Internet	Payment																				
17	C022	Binjai	57	Sun	Bar-2147	devanil@gmail.com	Married	4	Credit																				
18	C022	Brung	4	Jambi	-2241	engno@gmail.com	Married	3	Credit																				
19	C022	Kendari	1	NTB	-1937	endahut@gmail.com	Married	1	Bank	Transfer																			
20	C022	Medan	3	Lampung	-1509	sabiono@gmail.com	Married	4	Virtual Account																				
21	C022	Binjai	32	Sul	Bar-2176	vadmon@gmail.com	Married	5	Credit																				
22	C022	Bandar	7	Ja	Bar-1900	teksur@gmail.com	Married	5	Internet	Payment																			
23	C022	Batu	23	Pap	Bar-1710	kamah@gmail.com	Married	1	Credit																				
24	C022	Administ	Sun	Bar-1912	maimupd@gmail.com	Married	3	Credit																					
25	C022	Sawahlu	Maluku	(+62) 816 2200 2827	kamari@gmail.com	Married	4	Bank	Transfer																				
26	C022	Bandar L	Sul	Bar-1710	2200 2828	melindar@gmail.com	Married	3	Virtual Account																				
27	C022	Palemba	Kal	Bar-1420	816 2200 2829	cecemal@gmail.com	Married	3	Credit																				
28	C022	Madun	6	Riau	(+62) 816 2200 2830	dariadan@gmail.com	Married	4	Virtual Account																				

TBS_Testing

70%

Type here to search

12/13/20208:23 PM

2.2 Data Cleaning

Sebelum data dapat diproses, data harus dibersihkan terlebih dahulu untuk memastikan informasi yang dihasilkan akurat. Pada data set dummy yang kita buat terdapat beberapa data “kotor” yang dapat mempengaruhi akurasi model yang dibuat, oleh karena itu kita harus mengatasi data-data tersebut terlebih dahulu. Berikut merupakan tahapan yang kami lakukan untuk mengatasi data-data kotor atau tidak valid yang terdapat di data set kami yaitu:

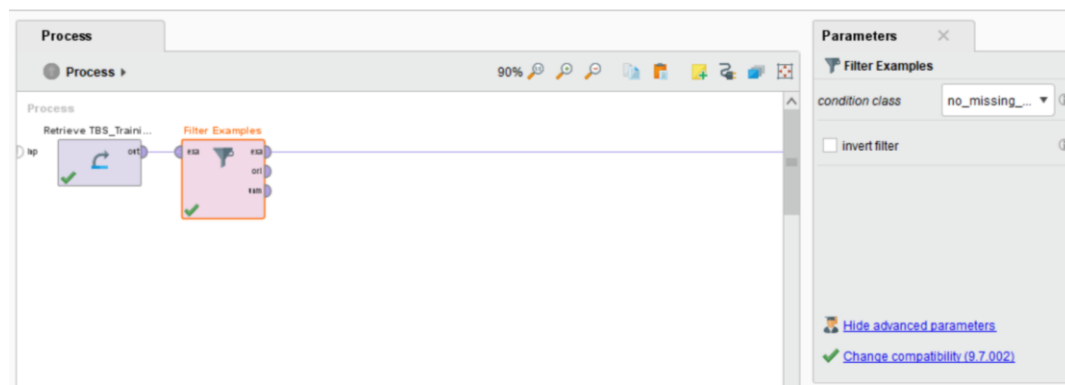
- Menghapus data yang tidak lengkap

Gambar di bawah ini menunjukkan adanya record data yang tidak lengkap. Oleh karena itu, kita perlu melakukan filtering agar data tersebut tidak termasuk dalam analisa kita.

Open in Turbo Prep Auto Model Filter (7 / 260 examples): missing_attributes ▾

Row No.	idCust...	Custom...	Custom...	Custom...	Custom...	Custom...	Custom...	Marital...	Websit...	Paymen...	MajorPr...
1	C0102	Warsa S...	?	Jan 5, 19...	Ds. Suka...	0811 38...	warsara...	Single	5	Bank Tra...	Shower ...
2	C0104	Makuta B...	?	Aug 1, 2...	Jr. Dipati...	0811 38...	makutm...	Single	2	Bank Tra...	Body Mist
3	C0105	Lurhur L...	?	Feb 12, ...	Kpg. Ban...	0816 77...	lurhomm...	Married	4	Internet ...	Shower ...
4	C0107	Diana P...	?	Jun 9, 19...	Jr. Samp...	0816 77...	dianti@g...	Married	5	Bank Tra...	Shower ...
5	C0108	Daru Ka...	Male	?	Psr. Mad...	0816 77...	darukom...	Married	4	Credit	Shower ...
6	C0109	Sarah H...	?	Oct 7, 19...	Ds. Sam...	0816 77...	sarapd...	Married	2	Credit	Shower ...
7	C0110	Ajiono G...	Male	?	Jln. Cut ...	0816 77...	ajionota...	Married	1	Bank Tra...	Shampoo

Untuk melakukan filtering maka dapat digunakan operator “filter examples” di RapidMiner dengan parameter yaitu “no_missing_attributes”.



Berikut adalah perbandingan data sebelum dan sesudah dilakukan filtering:

- Sebelum filtering

Row No.	CustomerNo...	CustomerNa...	CustomerGe...	CustomerD...	CustomerAd...	CustomerNu...	CustomerE...	MaritalStatus
99	C0099	Victoria Iriana...	Female	Mar 2, 1984	Jln. Sudiarto ...	0838-555-588	victoriairianar...	Married
100	C0100	Raisa Laksita	Female	Jun 3, 1976	Dk. Nanas N...	0856-555-091	raisalaksita...	Married
101	C0101	Dina Nurdyanti	Female	Dec 23, 2002	Ds. Thamrin ...	0816-555-565	dinanurdyant...	Married
102	C0102	Warsa Saputra	?	Jan 5, 1995	Ds. Sukajadi ...	0811 3843 1...	warsara@gm...	Single
103	C0103	Warji Prayoga	Male	Jan 3, 1920	Gg. Kyai Ged...	0811 3843 1...	warjioga@g...	Single
104	C0104	Makuta Bahu...	?	Aug 1, 2002	Jr. Dipatiukur ...	0811 3843 1...	makutmo@g...	Single
105	C0105	Lurhur Lulut ...	?	Feb 12, 1989	Kpg. Banal N...	0816 7756 2...	lurhomm@g...	Married
106	C0106	Ifa Palastri	Female	Mar 9, 1919	Ki. Suniaraja ...	0816 7756 2...	ifapaltri@gm...	Married
107	C0107	Diana Pudjia...	?	Jun 9, 1989	Jr. Sampang...	0816 7756 2...	dianti@gmail...	Married
108	C0108	Daru Kanda ...	Male	?	Psr. Madiun ...	0816 7756 2...	darukom@g...	Married
109	C0109	Sarah Hasan...	?	Oct 7, 1985	Ds. Samanh...	0816 7756 2...	sarapd@gm...	Married
110	C0110	Ajiono Galih ...	Male	?	Jln. Cut Nyak ...	0816 7756 2...	ajionota@gm...	Married
111	C0111	Ridwan Irawa...	Male	Jan 2, 1974	Psr. Sadang ...	0878-555-804	ridwanirwan...	Single

ExampleSet (210 examples, 0 special attributes, 11 regular attributes)

o Sesudah filtering

ExampleSet (Filter Examples)								
Open in		Turbo Prep		Auto Model		Filter (203 / 203 examples): all		
Row No.	CustomerNo...	CustomerNa...	CustomerGe...	CustomerD...	CustomerAd...	CustomerNu...	CustomerE...	MaritalStatus
96	C0096	Kartika Malik...	Female	Aug 8, 1993	Jln. Sam Rat...	0838-555-498	kartikamalika...	Married
97	C0097	Zahra Ami Us...	Female	Aug 15, 1970	Jr. Laswi No. ...	0813-555-163	zahraamiusa...	Married
98	C0098	Ghaliyati Pudj...	Female	Aug 23, 1975	Ki. Siliwangi ...	0819-555-693	ghaliyatipudji...	Married
99	C0099	Victoria Iriana...	Female	Mar 2, 1984	Jln. Sudiarto ...	0838-555-588	victoriairianar...	Married
100	C0100	Raisa Laksita	Female	Jun 3, 1976	Dk. Nanas N...	0856-555-091	raisalaksita...	Married
101	C0101	Dina Nurdyanti	Female	Dec 23, 2002	Ds. Thamrin ...	0816-555-565	dinanurdyant...	Married
102	C0103	Warji Prayoga	Male	Jan 3, 1920	Gg. Kyai Ged...	0811 3843 1...	warjioga@g...	Single
103	C0106	Ifa Palastri	Female	Mar 9, 1919	Ki. Suniaraja ...	0816 7756 2...	ifapaltri@gm...	Married
104	C0111	Ridwan Irawa...	Male	Jan 2, 1974	Psr. Sadang ...	0878-555-804	ridwanirwan...	Single
105	C0112	Cakrabirawa ...	Male	May 3, 1974	Psr. K.H. Mas...	0896-555-174	cakrabirawa...	Single
106	C0113	Halima Pad...	Male	Apr 11, 1975	Psr. Cikapay...	0838-555-396	halima09@g...	Single
107	C0114	Rahayu Wula...	Male	Feb 11, 1974	Jr. Zamrud N...	0838-555-173	Rahayuwulan...	Single
108	C0115	Cinta Gina Fa...	Male	Dec 4, 1975	Dk. Dago No....	0838-555-319	cintaGF@gm...	Single

ExampleSet (203 examples, 0 special attributes, 11 regular attributes)

- Menghapus data yang berulang

Pada gambar data set training dibawah ini terdapat data pelanggan yang berulang yaitu pada baris 57 dan 58.

57	C0057	Kamila A...	Female	Apr 8, 19...	Jr. Kalim...	0898-55...	kamilaa...	Married	2	Bank Tra...	EDT
58	C0058	Kamila A...	Female	Apr 8, 19...	Jr. Kalim...	0898-55...	kamilaa...	Married	2	Bank Tra...	EDT

Untuk mengatasi hal ini kita dapat menggunakan operator “remove duplicates” dan menggunakan parameter sebagai berikut:

The image displays two screenshots from the RapidMiner Studio interface. The top screenshot shows a process flow with three operators: 'Retrieve TBS_Traini...', 'Filter Examples', and 'Remove Duplicates'. The 'Remove Duplicates' operator is highlighted, and its parameters are shown on the right. The parameters are: 'attribute filter type' set to 'subset', 'attributes' set to 'Select Attri...', 'invert selection' unchecked, 'include special attributes' checked, and 'treat missing values as duplicates' unchecked. The bottom screenshot shows the 'Select Attributes: attributes' dialog box. It has a search bar and a list of attributes on the left, including 'CustomerAddress', 'CustomerEmail', 'CustomerNumber', 'MajorProductCategory', 'MaritalStatus', 'PaymentMethod', 'WebsiteActivity', and 'CustomerID'. On the right, there is a 'Selected Attributes' list containing 'CustomerDOB', 'CustomerGender', and 'CustomerName'. The 'Apply' button is highlighted.

Setelah dijalankan, maka akan menghasilkan data set yang tidak memiliki data berulang.

ExampleSet (Remove Duplicates)

Open in Turbo Prep Auto Model

Filter (202 / 202 examples): all

Row No.	CustomerNo...	CustomerNa...	CustomerGe...	CustomerD...	CustomerAd...	CustomerNu...	CustomerE...	MaritalStatus
49	C0049	Nrima Prako...	Male	Feb 27, 1997	Kpg. Baik No...	0838-555-283	nrimaprakos...	Married
50	C0050	Tirta Waskita ...	Male	May 24, 1971	Ki. Bambon ...	0838-555-923	tirtawaskitas.i...	Married
51	C0051	Viman Pradip...	Male	Feb 10, 1987	Kpg. Baja No...	0838-555-942	vimanpradipt...	Married
52	C0052	Emin Saputra	Male	Dec 7, 1979	Ds. Moch. Ya...	0899-555-620	eminsaputra...	Married
53	C0053	Ajiono Prano...	Male	May 8, 2000	Dk. Agus Sali...	0896-555-373	ajionoprano...	Married
54	C0054	Kamal Saeful...	Male	Jul 2, 1970	Psr. Basuki N...	0819-555-870	kamalsaefull...	Married
55	C0055	Maman Hardi...	Male	Jan 2, 1982	Psr. Gedeba...	0858-555-387	mamanhardi...	Married
56	C0056	Fitria Yolanda	Female	Aug 12, 1972	Ds. Achmad ...	0878-555-672	fitriayolanda...	Married
57	C0057	Kamila Agne...	Female	Apr 8, 1991	Jr. Kalimalan...	0898-555-264	kamilaagnes...	Married
58	C0059	Rahmi Azale...	Female	Nov 10, 1993	Dk. Dipenogo...	0838-555-857	rahmiazaleau...	Married
59	C0060	Carla Yulianti	Female	May 12, 1988	Gg. Tangkub...	0878-555-828	carlayulianti@...	Married
60	C0061	Nurul Ratih N...	Female	Aug 11, 1990	Psr. Cikutra B...	0838-555-691	nurulrathinas...	Married
61	C0062	Elma Permata	Female	Apr 1, 2003	Jln. Tentara P...	0838-555-872	elmapermata...	Married

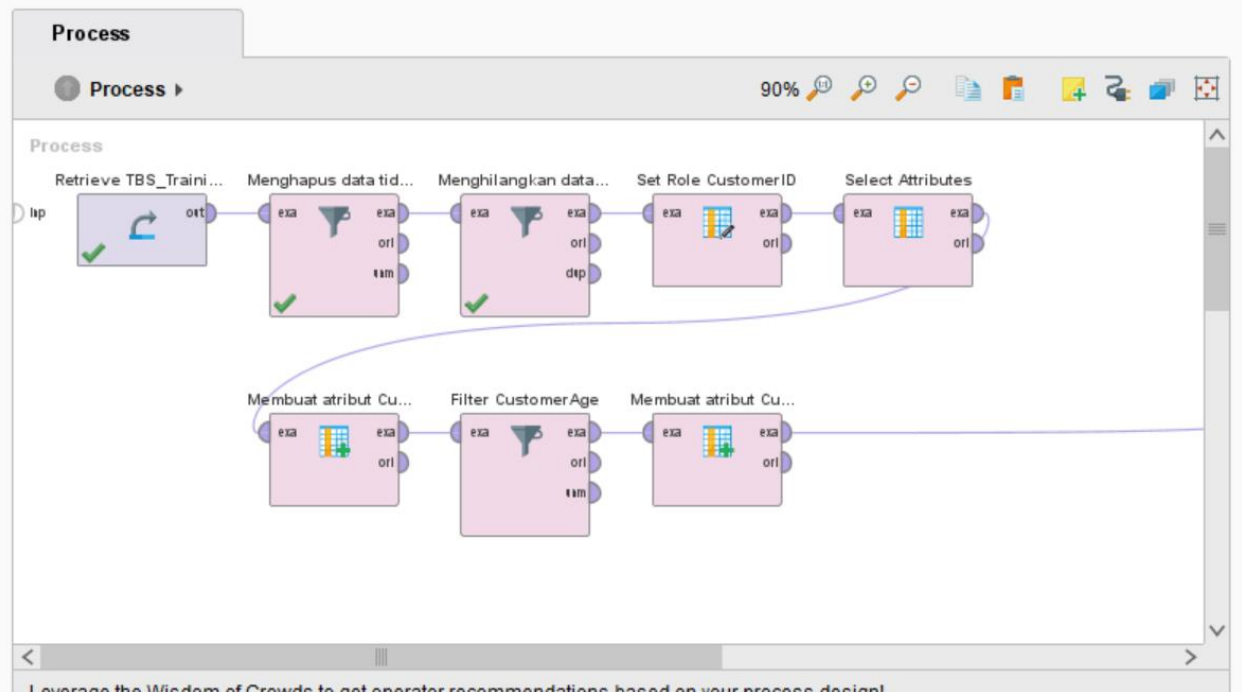
ExampleSet (202 examples, 0 special attributes, 11 regular attributes)

- Menghapus data yang tidak akurat

Kami juga perlu menghapus data-data yang tidak akurat seperti umur yang terkesan tidak masuk akal. Oleh karena itu, kami membatasi umur yang ada yaitu tidak lebih dari 75 tahun.

Row No.	CustomerNo...	CustomerNa...	CustomerGe...	CustomerD...	MaritalStatus	PaymentMet...	MajorProdu...	CustomerAge
96	C0097	Zahra Ami Us...	Female	Aug 15, 1970	Married	Bank Transfer	Body Lotion	50
97	C0098	Ghaliyati Pudj...	Female	Aug 23, 1975	Married	Credit	Shower Gel	45
98	C0099	Victoria Iriana...	Female	Mar 2, 1984	Married	Bank Transfer	Shower Gel	36
99	C0100	Raisa Laksita	Female	Jun 3, 1976	Married	Internet Paym...	Shower Gel	44
100	C0101	Dina Nurdianti	Female	Dec 23, 2002	Married	Credit	Body Mist	17
101	C0103	Warji Prayoga	Male	Jan 3, 1920	Single	Virtual Account	Body Lotion	100
102	C0106	Ifa Palastri	Female	Mar 9, 1919	Married	Credit	Shower Gel	101
103	C0111	Ridwan Irawa...	Male	Jan 2, 1974	Single	Internet Paym...	Body Lotion	46
104	C0112	Cakrabirawa ...	Male	May 3, 1974	Single	Credit	Body Lotion	46
105	C0113	Halima Pad...	Male	Apr 11, 1975	Single	Virtual Account	Body Lotion	45
106	C0114	Rahayu Wula	Male	Feb 11 1974	Single	Bank Transfer	Body Lotion	46

Untuk mengatasi hal ini, kita dapat menggunakan operator “generate attribute” untuk menghasilkan atribut yang menghitung umur pelanggan. Kemudian, kita menggunakan operator “filter example” untuk melakukan filtering sehingga data yang termasuk hanya data pelanggan yang berumur kurang dari atau sama dengan 75 tahun.



attribute name	function expressions
CustomerAge	<code>floor(date_diff(CustomerDOB,date_now()))/(1000*60*60*24*365.25))</code>

Create Filters: filters

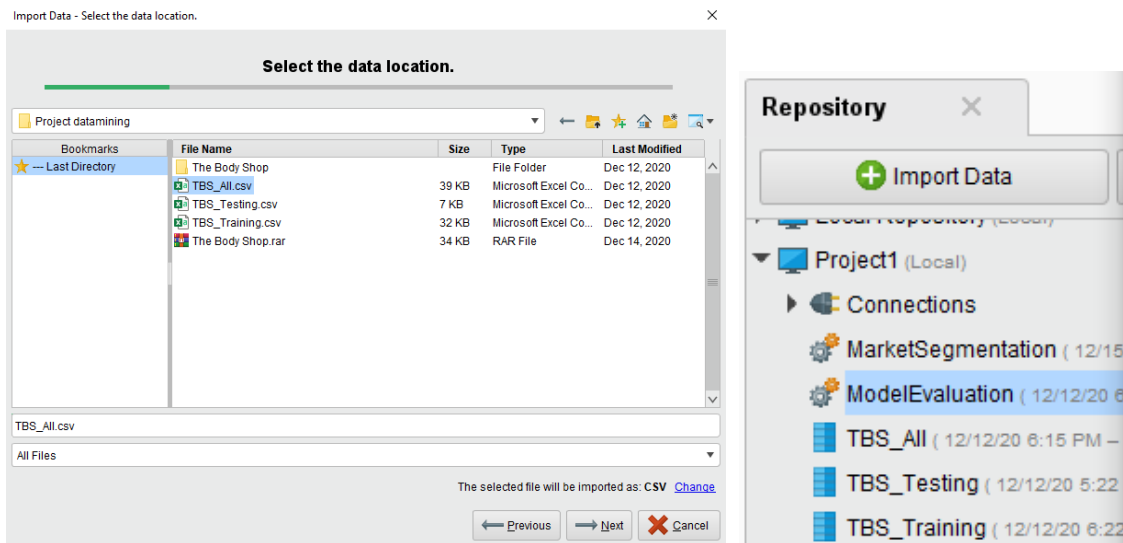
Create Filters: **filters**
Defines the list of filters to apply.

CustomerAge ≤ 75

2.3 Data Integration

Proses data integration adalah tahapan dimana kami harus menggabungkan dua atau lebih data dari beberapa sumber database yang berbeda kedalam suatu penyimpanan data warehouse, tujuannya agar tidak adanya duplikasi data dan mensesederhanakan proses menganalisa pengambilan keputusan. Berhubung data set kami hanya terdiri dari data csv, maka kami melakukan import data di RapidMiner dan memasukan data set Training, Testing, dan

gabungan dari keduanya (All) untuk keperluan evaluasi model. Kemudian, data-data tersebut kami masukan ke dalam sebuah repository local.



2.4 Data Selection

Pada tahapan ini, data yang relevan untuk dianalisa akan diambil dari data set. Menurut kami, data yang relevan untuk membuat predictive model yang memprediksi produk favorit pelanggan yaitu CustomerID, CustomerGender, CustomerDOB, MaritalStatus, dan MajorProductCategory (bagi data set training). Atribut lain seperti CustomerName, CustomerAddress, CustomerPhone, dan CustomerEmail tidak perlu digunakan karena nilai yang ada pasti berbeda-beda diantara setiap pelanggan sehingga tidak membantu proses klasifikasi. Selain itu, attribute WebsiteActivity dan PaymentMethod juga tidak kami masukan karena menurut kami atribut tersebut kurang mempengaruhi preferensi produk favorit pelanggan.

Row No.	CustomerNo...	CustomerName...	CustomerGender...	CustomerDOB...	MaritalStatus	PaymentMethod...	MajorProduct...	CustomerAge
1	C0001	Bagus Suwar...	Male	Jan 26, 1974	Single	Internet Paym...	Body Lotion	46
2	C0002	Darman Wac...	Male	Sep 6, 1983	Single	Credit	Shampoo	37
3	C0003	Limar Jaga...	Male	Jan 30, 1977	Single	Bank Transfer	Shampoo	43
4	C0004	Harjo Balam...	Male	May 13, 1982	Single	Credit	Shampoo	38
5	C0005	Praba_Harja...	Male	Jan 2, 1975	Single	Bank Transfer	Shampoo	45
6	C0006	Arta Jailani S...	Male	Oct 24, 1987	Single	Internet Paym...	Shampoo	33
7	C0007	Enteng Maria...	Male	Nov 20, 1999	Single	Credit	Shower Gel	21
8	C0008	Dr.Hendri Su...	Male	Sep 17, 1992	Single	Bank Transfer	Shower Gel	28
9	C0009	Bagiya Gunarto	Male	Aug 9, 1982	Single	Virtual Account	Shampoo	38
10	C0010	Dadi Putra	Male	Dec 4, 1978	Single	Credit	Shampoo	42
11	C0011	Purwa Zulkar...	Male	Nov 11, 1990	Single	Bank Transfer	Shampoo	30
12	C0012	Tugiman Sito...	Male	May 8, 2004	Single	Virtual Account	EDT	16
13	C0013	Lamar Narpati	Male	Feb 6, 1989	Married	Bank Transfer	Shampoo	31
14	C0014	Daliman Uta...	Male	Jun 1, 1980	Married	Internet Paym...	Shampoo	40
15	C0015	Kusuma Wart...	Male	Jul 7, 1984	Married	Internet Paym...	Shampoo	36

Dari data di atas kita dapat melihat bahwa data yang ada sudah bersih namun tidak setiap atribut data yang ada pada table tersebut kita perlukan untuk melakukan analisa, maka dari itu kita perlu melakukan tahap data selection sehingga menghasilkan data set seperti gambar di bawah ini.

- Data set Training setelah melalui data selection:

Row No.	CustomerNo...	CustomerGender...	CustomerDOB...	MaritalStatus	MajorProduct...
1	C0001	Male	Jan 26, 1974	Single	Body Lotion
2	C0002	Male	Sep 6, 1983	Single	Shampoo
3	C0003	Male	Jan 30, 1977	Single	Shampoo
4	C0004	Male	May 13, 1982	Single	Shampoo
5	C0005	Male	Jan 2, 1975	Single	Shampoo
6	C0006	Male	Oct 24, 1987	Single	Shampoo
7	C0007	Male	Nov 20, 1999	Single	Shower Gel
8	C0008	Male	Sep 17, 1992	Single	Shower Gel
9	C0009	Male	Aug 9, 1982	Single	Shampoo
10	C0010	Male	Dec 4, 1978	Single	Shampoo
11	C0011	Male	Nov 11, 1990	Single	Shampoo
12	C0012	Male	May 8, 2004	Single	EDT
13	C0013	Male	Feb 6, 1989	Married	Shampoo
14	C0014	Male	Jun 1, 1980	Married	Shampoo


ExampleSet (202 examples, 1 special attribute, 4 regular attributes)

2.5 Data Transformation

Setelah dilakukan data selection, maka data set kami sekarang terdiri dari atribut-atribut yang relevan terhadap analisa yang akan kami lakukan. Pada tahap data transformation, kami perlu mengubah nilai dari beberapa atribut yang ada ke dalam bentuk yang lebih sesuai untuk di analisa. Selain itu, kami juga melakukan filtering terhadap data set untuk menghasilkan data set dengan kriteria tertentu. Transformasi data yang kami lakukan yaitu:

- Membuat sebuah kolom baru yaitu CustomerAge yang didapat dari CustomerDOB.
- Mengganti nilai dari CustomerAge menjadi range dari umur yang ada di data set sehingga beragam nilai yang ada dikelompokkan dan dimasukkan kedalam atribut baru yaitu CustomerAgeGroup. Berikut range umur yang terdapat di data set:
 - o ≤ 20
 - o 21-30
 - o 31-45
 - o 46-50
 - o 50+

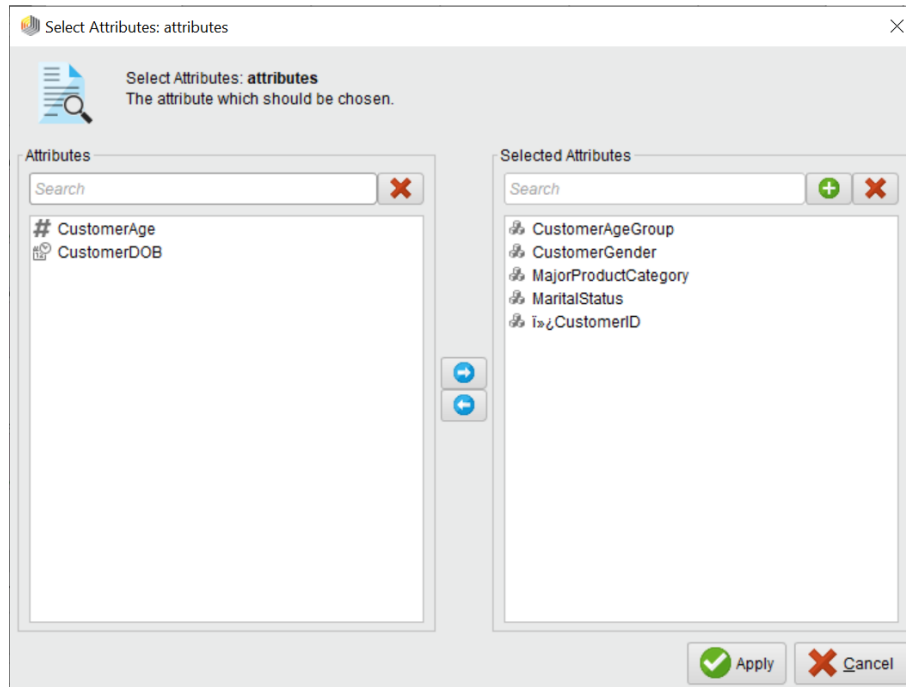
Berikut adalah perhitungan yang digunakan untuk mengelompokkan usia pelanggan dengan operator “Generate Attributes”:



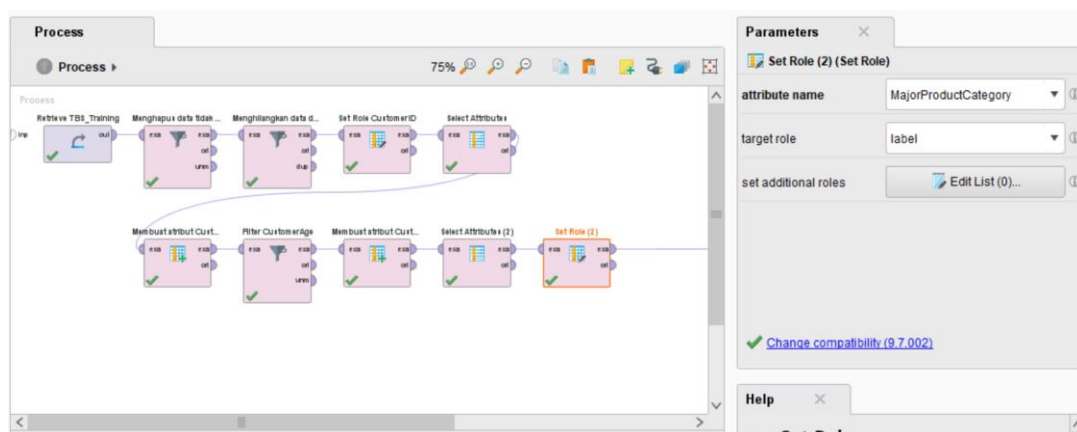
```
1 if(CustomerAge<=20, "[<=20]",
2   if(CustomerAge>=21&&CustomerAge<=30, "[21-30]",
3     if(CustomerAge>=31&&CustomerAge<=45, "[31-45]",
4       if(CustomerAge>=46&&CustomerAge<=50, "[46-50]", "[50+]"))))
```

Info: Expression is syntactically correct.

- Memilih kolom atribut yang akan digunakan dengan menggunakan operator “Select Attributes”. Kolom atribut yang dipilih yaitu CustomerID, CustomerAgeGroup, CustomerGender, MajorProductCategory (untuk data training), dan MaritalStatus.

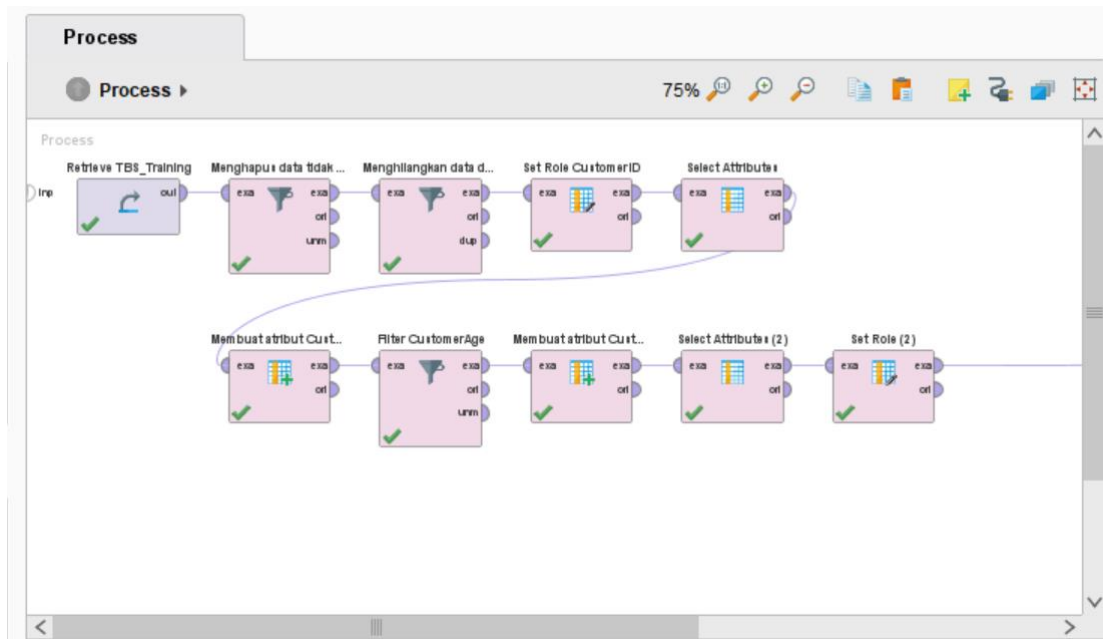


- Menetapkan atribut “MajorProductCategory” sebagai label.



Berikut adalah penampakan data setelah melalui proses data transformation.

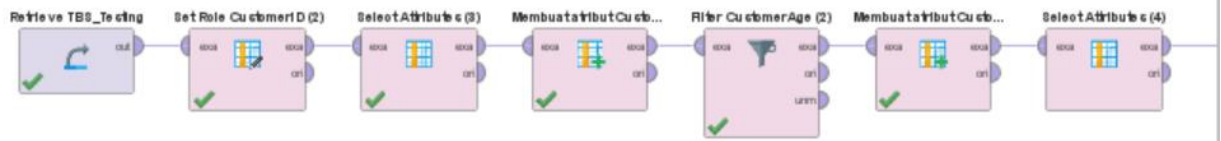
- Data Training



Row No.	Id Customer...	MajorProdu...	CustomerGe...	MaritalStatus	CustomerAg...
1	C0001	Body Lotion	Male	Single	[46-50]
2	C0002	Shampoo	Male	Single	[31-45]
3	C0003	Shampoo	Male	Single	[31-45]
4	C0004	Shampoo	Male	Single	[31-45]
5	C0005	Shampoo	Male	Single	[31-45]
6	C0006	Shampoo	Male	Single	[31-45]
7	C0007	Shower Gel	Male	Single	[21-30]
8	C0008	Shower Gel	Male	Single	[21-30]
9	C0009	Shampoo	Male	Single	[31-45]
10	C0010	Shampoo	Male	Single	[31-45]
11	C0011	Shampoo	Male	Single	[21-30]
12	C0012	EDT	Male	Single	[<=20]
13	C0013	Shampoo	Male	Married	[31-45]
14	C0014	Shampoo	Male	Married	[31-45]

ExampleSet (200 examples, 2 special attributes, 3 regular attributes)

- Data Testing



Row No.	CustomerID	CustomerGe...	MaritalStatus	CustomerAg...
1	C0211	Female	Single	[46-50]
2	C0212	Male	Single	[46-50]
3	C0213	Male	Single	[46-50]
4	C0214	Male	Single	[46-50]
5	C0215	Male	Single	[46-50]
6	C0216	Male	Single	[31-45]
7	C0217	Female	Single	[31-45]
8	C0218	Male	Single	[31-45]
9	C0219	Male	Single	[31-45]
10	C0220	Female	Single	[31-45]
11	C0221	Female	Single	[31-45]
12	C0222	Male	Single	[31-45]
13	C0223	Male	Married	[31-45]
14	C0224	Female	Married	[31-45]

ExampleSet (50 examples, 1 special attribute, 3 regular attributes)

2.6 Data Mining

- Pembuatan Predictive Model

Kami memutuskan untuk membuat sebuah predictive model dalam bentuk decision tree. Decision tree ini akan membagi data set pelanggan yang ada dan mencari atribut yang paling menentukan produk favorit dari pelanggan dengan karakteristik tertentu. Kami menggunakan parameter sebagai berikut:

Parameter	Value
criterion	information_gain
maximal depth	10
apply pruning	<input checked="" type="checkbox"/>
confidence	0.1
apply prepruning	<input checked="" type="checkbox"/>
minimal gain	0.01
minimal leaf size	2
minimal size for split	4
number of prepruning alternatives	3

Kriteria yang kami gunakan dalam pembuatan model ini adalah information gain. Information gain adalah sebuah metrik yang digunakan untuk mengukur kualitas dari pembagian data. Untuk menghitung information gain, maka kita terlebih dahulu harus menghitung information entropy yaitu berapa besar varians yang ada dalam data tersebut atau keberagaman dari suatu data. Entropy merupakan sebuah ukuran ketidakpastian yang terkait dengan variable. Semakin besar entropy maka semakin besar ketidakpastian yang ada. Rumus dari perhitungan entropy adalah sebagai berikut:

$$H(Y) = -\sum_{i=1}^m p_i \log(p_i), \text{ where } P(Y = y_i)$$

Untuk menghitung information gain atau informasi tambahan yang di dapat maka kita perlu mencari informasi yang diperlukan untuk melakukan klasifikasi dengan rumus sebagai berikut:

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times Info(D_j)$$

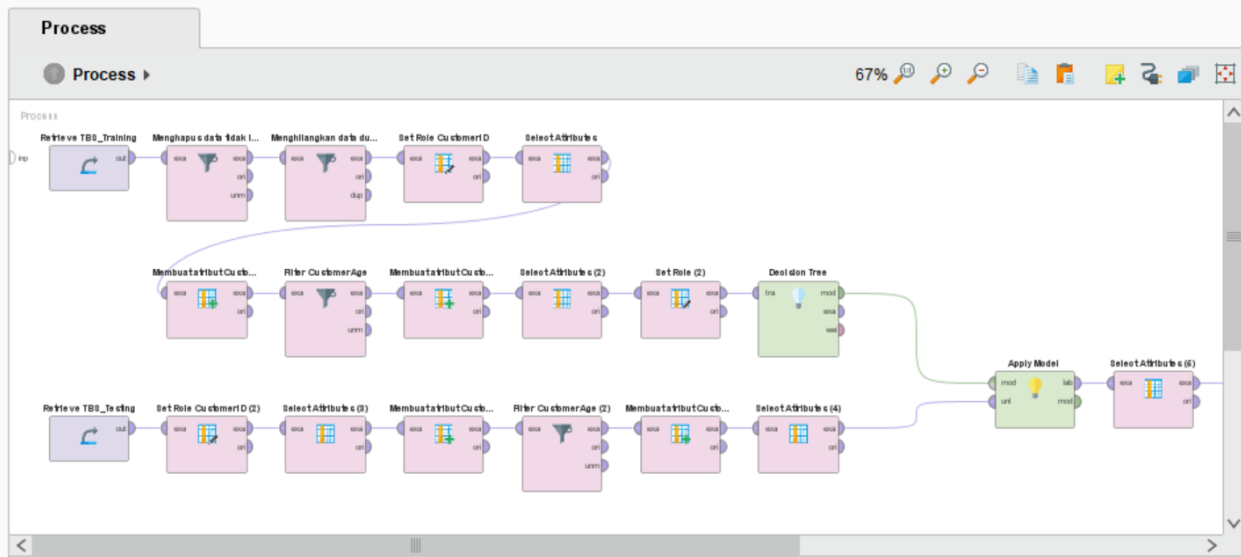
Setelah mendapat entropy dan informasi yang dibutuhkan untuk melakukan klasifikasi, barulah kita dapat menghitung informasi yang kita dapat dengan rumus sebagai berikut:

$$Gain(A) = Info(D) - Info_A(D)$$

Parameter lain yang kami gunakan dalam pembuatan model yaitu pruning dan prepruning. Pruning artinya beberapa cabang dapat digantikan dengan leaf tergantung pada confidence parameter yang ditetapkan yaitu 0.1. Kami juga menggunakan parameter prepruning yang menentukan kriteria berhenti pembagian selain maximal depth yaitu minimal gain, minimal leaf size, minimal size for split, dan number of prepruning alternatives.

- Hasil Model Decision Tree

Setelah kami berhasil membuat model dari decision tree, kami menggunakan operator “Apply Model” untuk mengaplikasikan model tersebut ke data set testing yang sudah kami buat. Setelah diaplikasikan ke model, nilai label prediksi yang dihasilkan akan muncul.



Berikut adalah hasil dari prediksi yang dilakukan model decision tree yang kami buat berdasarkan data set training:

Row No.	CustomerID	prediction(M...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	CustomerGe...
1	C0211	Body Lotion	0.850	0	0	0	0	0.150	Female
2	C0212	Body Lotion	0.850	0	0	0	0	0.150	Male
3	C0213	Body Lotion	0.850	0	0	0	0	0.150	Male
4	C0214	Body Lotion	0.850	0	0	0	0	0.150	Male
5	C0215	Body Lotion	0.850	0	0	0	0	0.150	Male
6	C0216	Shampoo	0.087	0.489	0.391	0.033	0	0	Male
7	C0217	Shower Gel	0.111	0.044	0.778	0.067	0	0	Female
8	C0218	Shampoo	0.087	0.489	0.391	0.033	0	0	Male
9	C0219	Shampoo	0.087	0.489	0.391	0.033	0	0	Male
10	C0220	Shower Gel	0.111	0.044	0.778	0.067	0	0	Female
11	C0221	Shower Gel	0.111	0.044	0.778	0.067	0	0	Female
12	C0222	Shampoo	0.087	0.489	0.391	0.033	0	0	Male
13	C0223	Shampoo	0.087	0.489	0.391	0.033	0	0	Male

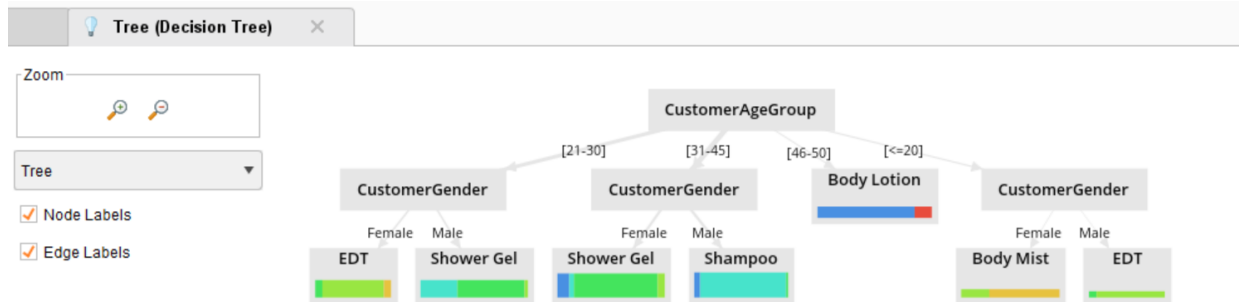
ExampleSet (50 examples, 8 special attributes, 3 regular attributes)

Agar lebih mudah untuk melihat informasi yang dihasilkan, maka kami menggunakan operator “Select Attribute” untuk melakukan selection terhadap kolom atribut yang kami inginkan.

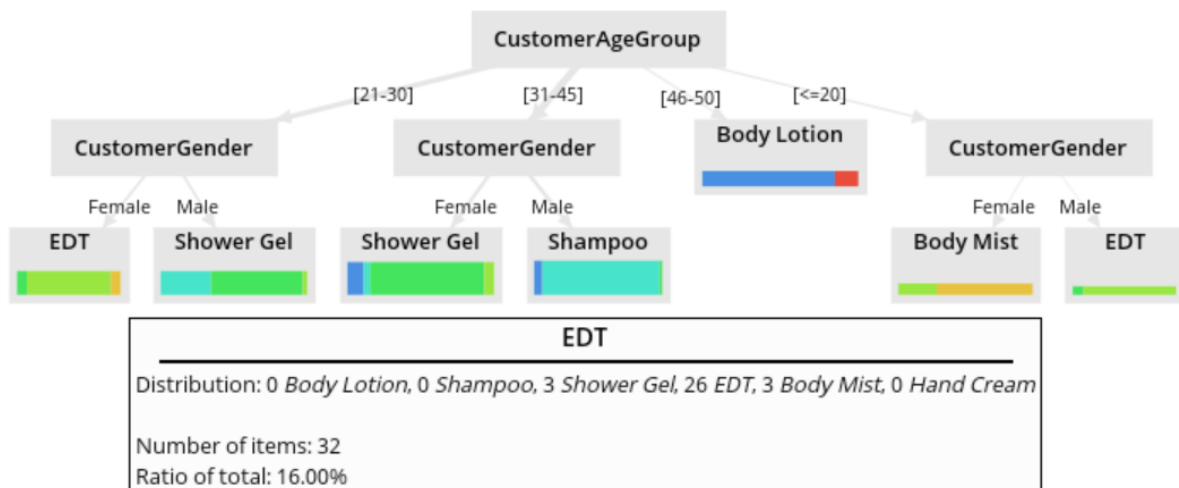
Row No.	CustomerID	prediction(MajorProductCategory)	CustomerGender	CustomerAgeGroup
1	C0211	Body Lotion	Female	[46-50]
2	C0212	Body Lotion	Male	[46-50]
3	C0213	Body Lotion	Male	[46-50]
4	C0214	Body Lotion	Male	[46-50]
5	C0215	Body Lotion	Male	[46-50]
6	C0216	Shampoo	Male	[31-45]
7	C0217	Shower Gel	Female	[31-45]
8	C0218	Shampoo	Male	[31-45]
9	C0219	Shampoo	Male	[31-45]
10	C0220	Shower Gel	Female	[31-45]
11	C0221	Shower Gel	Female	[31-45]
12	C0222	Shampoo	Male	[31-45]
13	C0223	Shampoo	Male	[31-45]
14	C0224	Shower Gel	Female	[31-45]

ExampleSet (50 examples, 2 special attributes, 2 regular attributes)

Berikut adalah hasil akhir predictive model decision tree yang telah kami buat:



Bar yang ada di bawah setiap label menunjukkan distribusi dari label tersebut di data set.



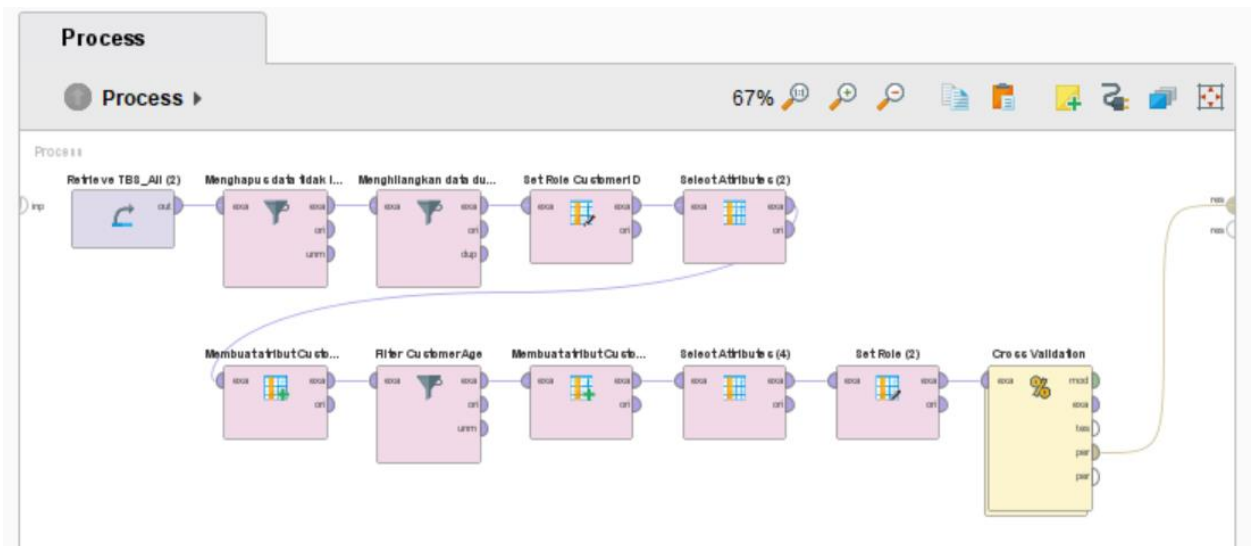
Gambar diatas menunjukkan distribusi label EDT pada CustomerAgeGroup “[21-30]” dengan nilai CustomerGender yaitu “Female”. Jumlah distribusi label EDT yaitu 26 buah dari 32 buah label pada karakteristik pelanggan tersebut, sedangkan “Ratio of total: 16.00%” merupakan rasio jumlah label tersebut dibandingkan dengan jumlah keseluruhan label yang ada di data set.

$$\text{Ratio of total} = \frac{32}{200} \times 100\% = 16\%$$

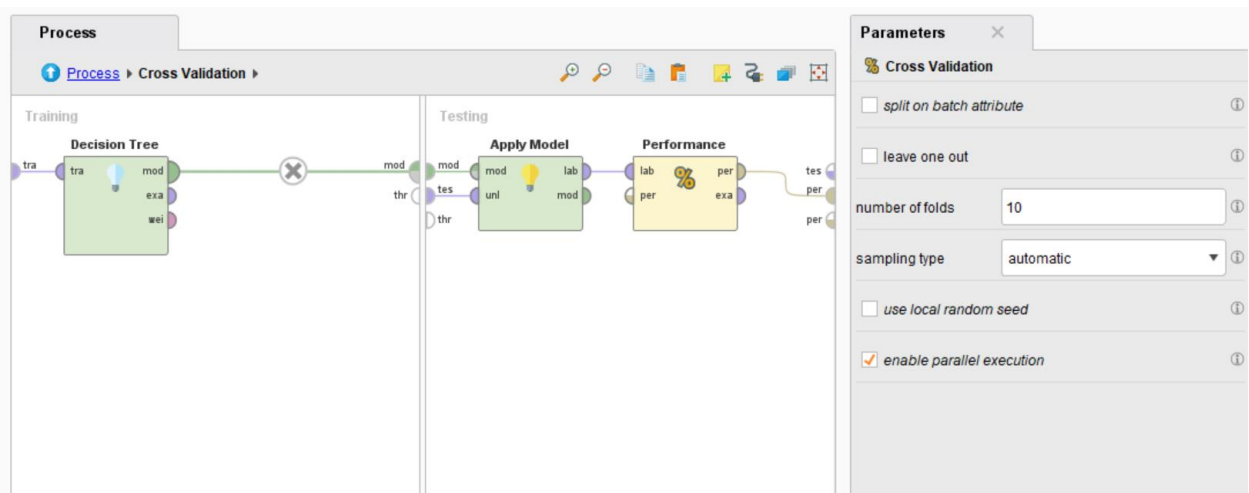
- Evaluasi Model

Setelah membuat predictive model, langkah yang kami lakukan selanjutnya yaitu menghitung performance dari model yang kami buat. Untuk melakukan evaluasi model ini, kami menggunakan metode cross-validation. K-fold cross-validation adalah prosedur

resampling yang digunakan untuk mengevaluasi performa dari model machine learning yang mempunyai sampel data terbatas. Kami memilih menggunakan metode ini karena cukup mudah untuk diimplementasikan namun memberikan hasil yang mengandung nilai bias yang lebih rendah disbanding dengan metode lain. Berikut adalah proses yang kami lakukan di RapidMiner:



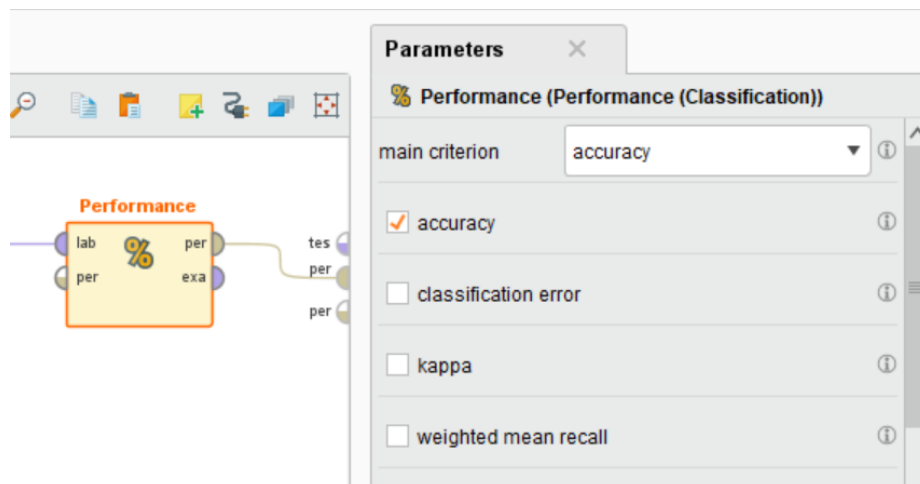
Untuk melaksanakan cross validation di RapidMiner, kami menggunakan operator “Cross Validation”. Data set yang kami gunakan untuk evaluasi adalah data set gabungan dari training dan testing. Oleh karena data training kami mengandung data-data kotor, maka kami harus melakukan data cleaning, selection, dan transformation terlebih dahulu sebelum diaplikasikan cross validation.



Di dalam operator cross validation data set yang kami sambungkan akan dibagi menjadi 2 bagian yaitu training dan testing. Untuk data training, kami menggunakan decision tree untuk menghasilkan predictive model dengan parameter yang sama dengan pembuatan model sebelumnya pada data set training. Untuk data testing perlu digunakan operator “Apply Model” untuk mengaplikasikan predictive model tersebut ke data testing dan kemudian diukur performance nya menggunakan operator “Performance (classification)” karena model yang kita buat termasuk dalam algoritma klasifikasi.

Parameter yang terdapat di sebelah kanan adalah parameter yang diterapkan pada cross validation tersebut. Disini kita menentukan jumlah k adalah 10, artinya data kami akan terbagi menjadi 10 group data. Sampling type adalah metode yang harus kami gunakan untuk membuat subset dari data yang kami sediakan, disini sampling type kami tetapkan ke automatic yang artinya sampling type yang kami pakai adalah stratified sampling. Stratified sampling adalah metode sampling yang membuat subset secara random dan memastikan bahwa distribusi label di setiap subset jumlahnya sama, sehingga menghindari satu subset terdiri sepenuhnya dari 1 label karena hal tersebut dapat menyebabkan kesalahan dalam menghitung performance dari sebuah model.

Untuk mengukur performance dari model kami, kami menggunakan kriteria “Accuracy” yang akan menghasilkan akurasi dari predictive model yang telah kami buat. Saat dijalankan, maka akan dihasilkan sebuah confusion matrix dan juga accuracy dari model kami yaitu 85.60%.



- Perhitungan Confusion Matrix

accuracy: 85.60% +/- 6.59% (micro average: 85.60%)

	true Body Lotion	true Shampoo	true Shower Gel	true EDT	true Body Mist	true Hand Cre...	class precision
pred. Body Lot...	25	0	0	0	0	2	92.59%
pred. Shampoo	4	56	1	0	0	0	91.80%
pred. Shower ...	5	6	78	3	0	1	83.87%
pred. EDT	0	0	4	43	4	1	82.69%
pred. Body Mist	0	0	0	5	12	0	70.59%
pred. Hand Cr...	0	0	0	0	0	0	0.00%
class recall	73.53%	90.32%	93.98%	84.31%	75.00%	0.00%	

○ Precision

$$Precision = \frac{Tp}{Tp + Fp}$$

$$Precision\ Body\ Lotion = \frac{25}{25 + 2} = 0.9259 \times 100\% = 92.59\%$$

$$Precision\ Shower\ Gel = \frac{78}{78 + 15} = 0.8387 \times 100\% = 83.87\%$$

$$Precision\ Shampoo = \frac{56}{56 + 5} = 0.9180 \times 100\% = 91.80\%$$

$$Precision\ EDT = \frac{43}{43 + 9} = 0.8269 \times 100\% = 82.69\%$$

$$Precision\ Body\ Mist = \frac{12}{12 + 5} = 0.7059 \times 100\% = 70.59\%$$

$$Precision\ HandCream = \frac{0}{0 + 2} = 0 \times 100\% = 0\%$$

○ Recall

$$Recall = \frac{Tp}{Tp + Fn}$$

$$Recall\ Body\ lotion = \frac{25}{25 + 9} = 0.7353 \times 100\% = 73.53\%$$

$$\text{Recall Shampoo} = \frac{56}{56 + 6} = 0.9032 \times 100\% = 90.32\%$$

$$\text{Recall Shower Gel} = \frac{78}{78 + 5} = 0.9398 \times 100\% = 93.98\%$$

$$\text{Recall EDT} = \frac{43}{43 + 8} = 0.8431 \times 100\% = 84.31\%$$

$$\text{Recall Body Mist} = \frac{12}{12 + 4} = 0.75 \times 100\% = 75.00\%$$

$$\text{Recall Hand Cream} = \frac{0}{0 + 0} = 0 \times 100\% = 0\%$$

- Accuracy

$$\text{Accuracy} = \frac{Tp + Tn}{All}$$

$$\text{Accuracy} = \frac{214}{250} = 0.856 \times 100\% = 85.60\%$$

2.7 Evaluation and Presentation

Dari predictive model yang telah kami buat dengan menggunakan decision tree dengan metric information gain, didapat hasil bahwa atribut yang memberikan information gain terbesar adalah CustomerAgeGroup sehingga atribut tersebut menjadi parent node atau root. Setelah itu, untuk masing-masing cabang akan dicari atribut yang menghasilkan information gain terbanyak dan membagi data.

Berikut hasil yang kita dapat:

- Untuk CustomerAgeGroup dengan rentang usia 21-30 tahun, atribut yang menjadi pembagi adalah CustomerGender.
 - Jika CustomerGender adalah “Female” maka kemungkinan besar kategori produk favoritnya adalah “EDT”.
 - Jika CustomerGender adalah “Male” maka kemungkinan besar kategori produk favoritnya adalah “Shower Gel”.

- Untuk CustomerAgeGroup dengan rentang usia 31-45 tahun, atribut yang menjadi pembagi adalah CustomerGender.
 - o Jika CustomerGender adalah “Female” maka kemungkinan besar kategori produk favoritnya adalah “Shower Gel”.
 - o Jika CustomerGender adalah “Male” maka kemungkinan besar kategori produk favoritnya adalah “Shampoo”.
- Untuk CustomerAgeGroup dengan rentang usia 46-50, semua data yang ada menghasilkan 1 label yaitu “Body Lotion”. Hal ini berarti kemungkinan besar preferensi dari pelanggan yang berusia 46-50 adalah “Body Lotion”.
- Untuk CustomerAgeGroup dengan rentang usia kurang dari atau sama dengan 20 tahun, atribut yang menjadi pembagi adalah CustomerGender.
 - o Jika CustomerGender adalah “Female” maka kemungkinan besar kategori produk favoritnya adalah “Body Mist”.
 - o Jika CustomerGender adalah “Male” maka kemungkinan besar kategori produk favoritnya adalah “EDT”.

3. Kesimpulan

Project yang kami buat memanfaatkan data pelanggan dari perusahaan The Body Shop untuk mencari tahu kategori produk dari pelanggan. Hal ini dapat dicapai dengan melakukan segmentasi pasar dengan menghasilkan predictive model yang menggunakan algoritma decision tree. Decision tree adalah sebuah algoritma yang digunakan untuk melakukan klasifikasi data di *machine learning*. Dari model yang telah kami buat, kami mendapat kesimpulan bahwa pembagi utama data set pelanggan di The Body Shop adalah CustomerAgeGroup atau usia pelanggan. Setelah itu, pembagi lainnya adalah CustomerGender atau jenis kelamin dari pelanggan.

Hasil dari predictive model ini dapat memberi prediksi kategori produk dari pelanggan dengan akurasi sebesar 85.60% berdasarkan metode resampling menggunakan cross-validation. Adanya data ini dapat membantu perusahaan The Body Shop untuk melakukan *targeted marketing* kepada pelanggan nya dan juga menambahkan ketertarikan pelanggan baru karena kita dapat memberi rekomendasi produk yang sesuai dengan preferensinya. Hal ini

tentu akan berdampak baik bagi perusahaan karena kemungkinan besar loyalitas pelanggan lama dan baru akan bertambah sehingga menambah pendapatan perusahaan.

4. Referensi

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

<https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222>

https://www.researchgate.net/publication/316312766_Market_segmentation_through_data_mining_A_method_to_extract_behaviors_from_a_noisy_data_set