

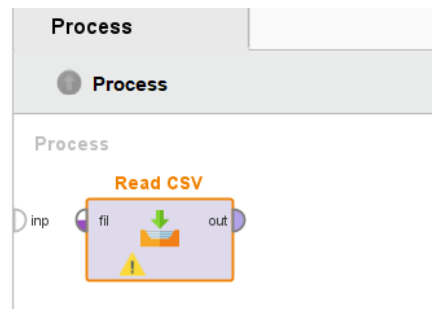
Nama : Fenny Anggraini

Tema pelatihan : ADS – ASSOCIATE DATA SCIENTIST

## PROSES Pengerjaan

### DATA UNDERSTANDING

1. Melakukan pembacaan pada data dengan menggunakan operator “Read CSV” untuk membaca data.



2. Hasil dari operator “Read CSV” dimana menampilkan output dari dataset yang ada seperti ini, dimana terdiri dari 16 atribut dan sebanyak 7665 records data.

Altair AI Studio Trial 2024.0.3 @ DESKTOP-7GDV6U8

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Interactive Analysis

Find data, operators, etc. All Studio

Result History ExampleSet (Read CSV)

Open in Turbo Prep Auto Model Interactive Analysis

Filter (7,665 / 7,665 examples): all

Row No.	Date	Weekday	Holiday	Month	Season	Temperature	Humidity	Windy	Outlook	Crowdedness	EmailCampa...	Maintenance...	ID
1	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Cleaning Ame...	A
2	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Cleaning Ame...	B
3	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Cleaning Ame...	C
4	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Cleaning Ame...	D
5	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Cleaning Ame...	E
6	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Cleaning Ame...	F
7	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Cleaning Ame...	G
8	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	[]	A
9	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	[]	B
10	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	[]	C
11	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	[]	D
12	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	[]	E
13	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	[]	F
14	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	[]	G
15	Jan 3, 2021	6	0	Jan	Winter	2.300	58	0	overcast	0.700	Overcast skie...	[]	A
16	Jan 3, 2021	6	0	Jan	Winter	2.300	58	0	overcast	0.700	Overcast skie...	[]	B
17	Jan 3, 2021	6	0	Jan	Winter	2.300	58	0	overcast	0.700	Overcast skie...	[]	C
18	Jan 3, 2021	6	0	Jan	Winter	2.300	58	0	overcast	0.700	Overcast skie...	[]	D

ExampleSet (7,665 examples, 0 special attributes, 16 regular attributes)

### 3. Memeriksa data / menelaah data

- Missing value, dari seluruh atribut yang ada terdapat missing value pada atribut “review” sebanyak 6313, namun secara khusus dapat dilihat bahwa atribut ini tidak akan berpengaruh pada data yang artinya kemungkinan tidak akan termasuk kedalam atribut yang akan dipilih karena tidak terlalu berpengaruh. Dan juga terdapat pada atribut “Maintenance task” sebanyak 1820, atribut ini akan di isi nilai yang hilangnya karena akan berpengaruh pada data.

Altair AI Studio Trial 2024.0.3 @ DESKTOP-7GDV6U8

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Interactive Analysis

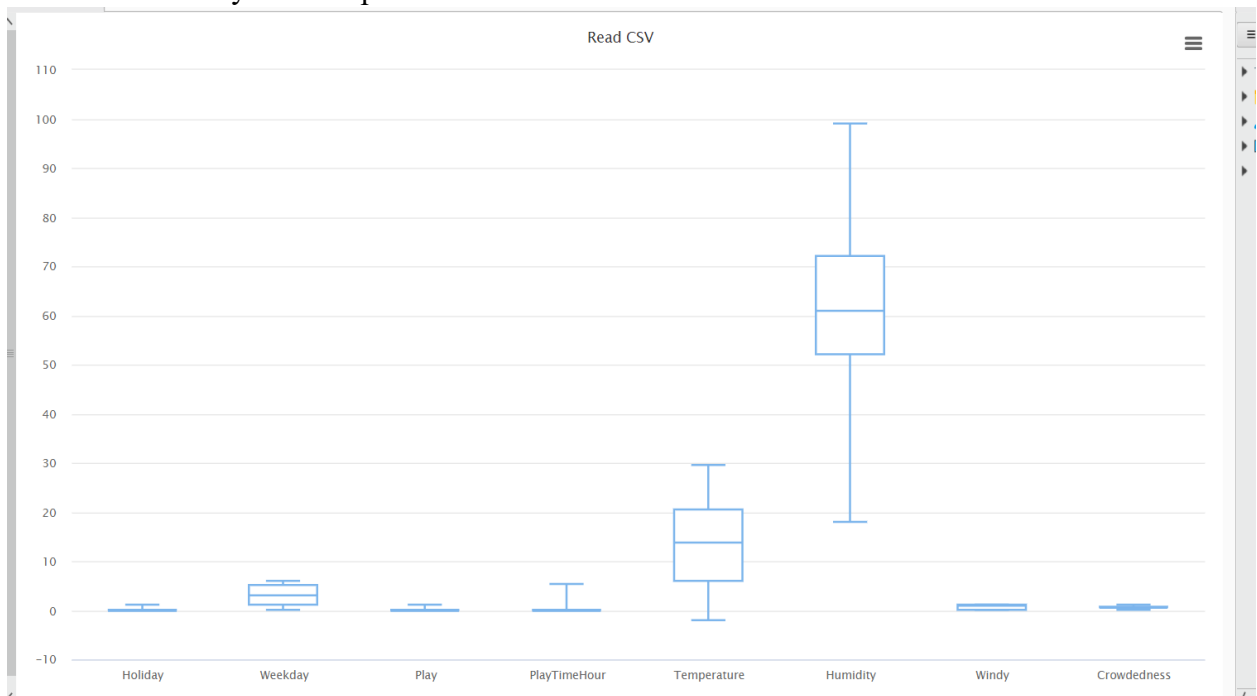
Find data, operators, etc. All Studio

Result History ExampleSet (Read CSV)

Name	Type	Missing	Statistics	Filter (16 / 16 attributes):
Windy	Integer	0	0 1 0.508	
Outlook	Nominal	0	Least snow (378) Most sunny (3241) Values sunny (3241), overcast (2660), ... [2 more]	
Crowdedness	Real	0	Min 0 Max 1 Average 0.621	
EmailCampaign	Nominal	0	Least Your Sun [...] kies! (7) Most 100% Hum [...] ated. (7) Values 100% Hum [...] hydrated. (7), 22.22°C [...] of golf. (7), ... [1093 more]	
MaintenanceTask	Nominal	1820	Least [Wateri [...] nce] (7) Most [Cleani [...] s] (546) Values [Cleaning Amenities] (546), [Sprinkler Checks] (455), ... [191 more]	
ID	Nominal	0	Least G (1095) Most A (1095) Values A (1095), B (1095), ... [5 more]	
Play	Integer	0	Min 0 Max 1 Average 0.183	
PlayTimeHour	Real	0	Min 0 Max 5.200 Average 0.642	
Review	Nominal	6313	Least You coul [...] ease! (1) Most Ah, the [...] ture! (2) Values Ah, the [...] d nature! (2), Oh, what [...] fing joy! (2), ... [1348 more]	

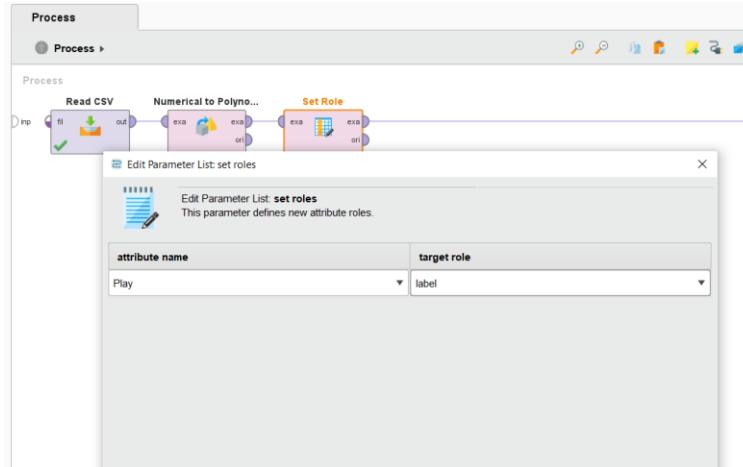
Showing attributes 1 - 16 Examples: 7,665 Special Attributes: 0 Regular Attributes: 16

- Outliers, mengidentifikasi terdapat data yang pencilan atau outlier tidak pada data, terlihat bahwa tidak adanya outlier pada data.



## DATA PREPROCESSING

1. Memberikan label pada data menggunakan operator “Set role” dimana label disini dari keseluruhan atribut yang dapat dijadikan label yaitu atribut “Play” dimana Menunjukkan apakah golf dimainkan pada hari tersebut (1 untuk ya, 0 untuk tidak) – Ini adalah **label atau target**.

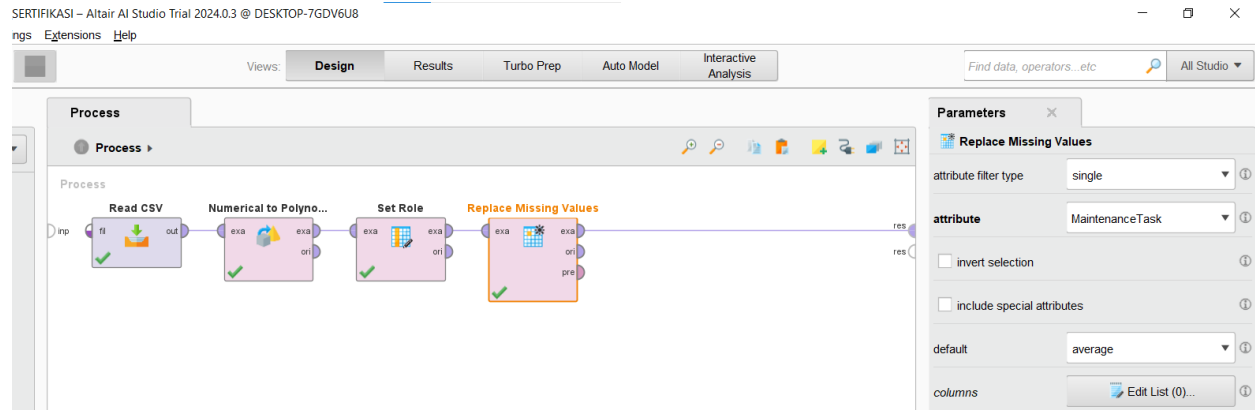


Ketika label sudah ditentukan maka output yang didapatkan seperti ini, kolom dengan warna hijau sendiri merupakan label/target.

The screenshot shows the 'Results' view in RapidMiner for the 'ExampleSet (Set Role)'. The table contains 18 rows of data. The 'Play' column is highlighted in green, indicating it is the target variable. The table has 15 regular attributes and 1 special attribute.

Row No.	Play	Date	Weekday	Holiday	Month	Season	Temperature	Humidity	Windy	Outlook	Crowdedness	EmailCampa...	Mainte...
1	1	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Clean
2	0	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Clean
3	0	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Clean
4	1	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Clean
5	1	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Clean
6	0	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Clean
7	0	Jan 1, 2021	4	1	Jan	Winter	3.300	49	1	sunny	0.730	Happy New Ye...	[Clean
8	1	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	?
9	0	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	?
10	1	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	?
11	0	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	?
12	0	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	?
13	0	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	?
14	0	Jan 2, 2021	5	0	Jan	Winter	2.900	53	0	sunny	0.740	Beautiful sunn...	?
15	0	Jan 3, 2021	6	0	Jan	Winter	2.300	58	0	overcast	0.700	Overcast skie...	?
16	0	Jan 3, 2021	6	0	Jan	Winter	2.300	58	0	overcast	0.700	Overcast skie...	?
17	1	Jan 3, 2021	6	0	Jan	Winter	2.300	58	0	overcast	0.700	Overcast skie...	?
18	0	Jan 3, 2021	6	0	Jan	Winter	2.300	58	0	overcast	0.700	Overcast skie...	?

2. Melakukan “Replace missing value” dimana mengisi nilai yang hilang menggunakan operator tersebut di rapidminer, dengan parameter dimana hanya mengisi nilai yang hilang pada atribut “Maintenance task” dengan menggunakan metode average (nilai rata-rata).



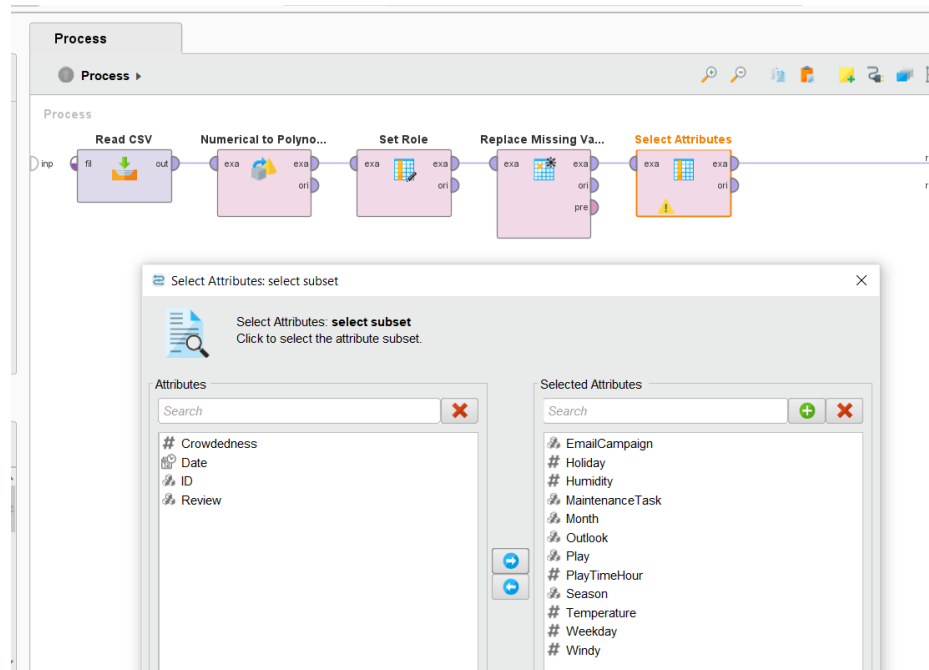
Output yang dihasilkan akan seperti ini untuk atribut tersebut sudah tidak lagi terdapat missing value

Altair AI Studio - Results view for ExampleSet (Replace Missing Values)

Name	Type	Missing	Statistics
Play	Nominal	0	Least 1 (1399), Most 0 (6266), Values 0 (6266), 1 (1399)
MaintenanceTask	Polynomial	0	Least [Wateri [...] nce] (7), Most [Cleani [...] ] (2366), Values [Cleaning Amenities] (2366), [Sprinkler Checks] (455), ...[191 more]
Date	Date	0	Earliest date Jan 1, 2021, Latest date Dec 31, 2023, Duration 1094 days
Weekday	Integer	0	Min 0, Max 6, Average 3.005
Holiday	Integer	0	Min 0, Max 1, Average 0.045
Month	Polynomial	0	Least Feb (588), Most Aug (651), Values Aug (651), Dec (651), ...[10 more]
Season	Polynomial	0	Least Winter (1869), Most Summer (1974), Values Summer (1974), Spring (1953), ...[2 more]
Temperature	Real	0	Min -2, Max 29.500, Average 13.436
Humidity	Integer	0	Min 18, Max 99, Average 61.525

Showing attributes 1 - 16. Examples: 7,665. Special Attributes: 1. Regular Attributes: 15.

- Pemilihan fitur yang berpengaruh, disini menggunakan operator “Select attribute” dimana memilih fitur yang akan berpengaruh seperti berikut ini



### Atribut yang Berpengaruh untuk Pemodelan:

Untuk tahap pemodelan, terutama jika tujuan Anda adalah untuk memprediksi apakah golf akan dimainkan (target variabel Play), atribut-atribut berikut bisa sangat berpengaruh:

- **Weekday:** Pengaruh hari dalam seminggu terhadap keputusan untuk bermain golf.
- **Holiday:** Apakah hari libur dapat mempengaruhi keputusan bermain golf.
- **Month:** Musim dan bulan dapat berpengaruh pada kebiasaan bermain golf.
- **Season:** Musim yang dapat mempengaruhi cuaca dan suasana bermain golf.
- **Temperature:** Suhu dapat mempengaruhi kenyamanan saat bermain golf.
- **Humidity:** Kelembaban juga dapat mempengaruhi kenyamanan dan keputusan bermain.
- **Windy:** Apakah hari berangin dapat mempengaruhi keputusan bermain golf.
- **Outlook:** Kondisi cuaca dapat sangat mempengaruhi keputusan bermain golf.
- **Crowdedness:** Tingkat keramaian lapangan golf dapat mempengaruhi keputusan.
- **Maintenance Task:** Pemeliharaan lapangan mungkin mempengaruhi ketersediaan lapangan.
- **PlayTimeHour:** Durasi bermain golf yang mungkin berkaitan dengan keputusan untuk bermain.

Berikut ini hasil dataset setelah dilakukan pemilihan atribut

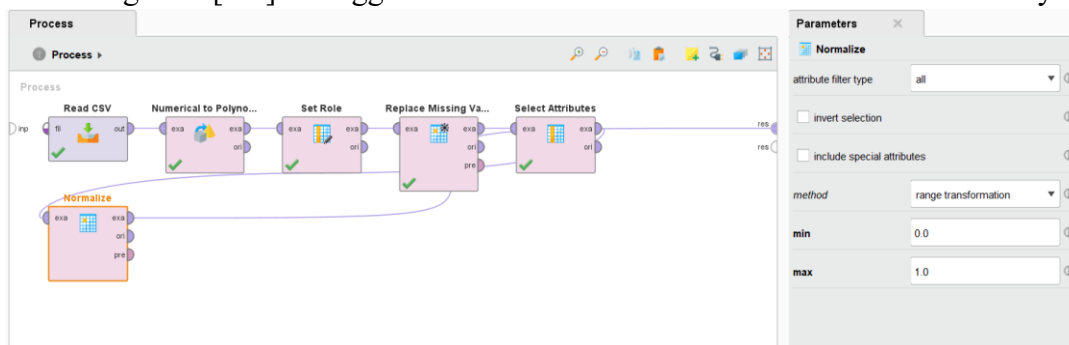
ExampleSet (Select Attributes)

Open in Turbo Prep Auto Model Interactive Analysis

Filter (7,665 / 7,665 examples): all

Row No.	Play	Maintenance...	Weekday	Holiday	Month	Season	Temperature	Humidity	Windy	Outlook	EmailCampa...	PlayTimeHour
1	1	[Cleaning Ame...	4	1	Jan	Winter	3.300	49	1	sunny	Happy New Ye...	3.100
2	0	[Cleaning Ame...	4	1	Jan	Winter	3.300	49	1	sunny	Happy New Ye...	0
3	0	[Cleaning Ame...	4	1	Jan	Winter	3.300	49	1	sunny	Happy New Ye...	0
4	1	[Cleaning Ame...	4	1	Jan	Winter	3.300	49	1	sunny	Happy New Ye...	3.600
5	1	[Cleaning Ame...	4	1	Jan	Winter	3.300	49	1	sunny	Happy New Ye...	3.400
6	0	[Cleaning Ame...	4	1	Jan	Winter	3.300	49	1	sunny	Happy New Ye...	0
7	0	[Cleaning Ame...	4	1	Jan	Winter	3.300	49	1	sunny	Happy New Ye...	0

4. Normalisasi data menggunakan operator “Normalize” agar data memiliki rentang nilai yang sama disini menggunakan teknik range transformation yang merupakan Min-Max scaler data memiliki rentang nilai [0-1] sehingga akan lebih mudah untuk diolah oleh mesin nantinya.



Output dataset yang didapatkan sebagai berikut, dimana setiap atribut memiliki rentang nilai yang sama [0-1]

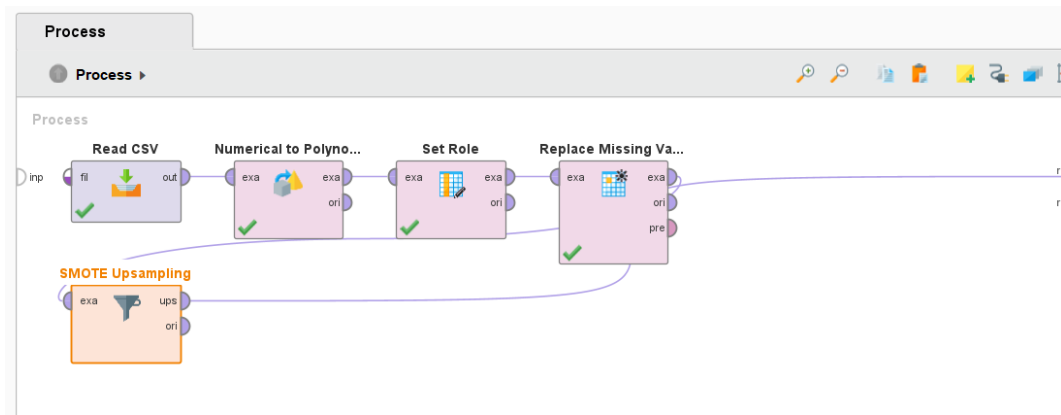
ExampleSet (Normalize)

Open in Turbo Prep Auto Model Interactive Analysis

Filter (7,665 / 7,665 examples): all

Row No.	Play	Weekday	Holiday	Temperature	Humidity	Windy	PlayTimeHour	Maintenance...	Month	Season	Outlook	EmailCampa...
1	1	0.667	1	0.168	0.383	1	0.596	[Cleaning Ame...	Jan	Winter	sunny	Happy New Ye...
2	0	0.667	1	0.168	0.383	1	0	[Cleaning Ame...	Jan	Winter	sunny	Happy New Ye...
3	0	0.667	1	0.168	0.383	1	0	[Cleaning Ame...	Jan	Winter	sunny	Happy New Ye...
4	1	0.667	1	0.168	0.383	1	0.692	[Cleaning Ame...	Jan	Winter	sunny	Happy New Ye...
5	1	0.667	1	0.168	0.383	1	0.654	[Cleaning Ame...	Jan	Winter	sunny	Happy New Ye...
6	0	0.667	1	0.168	0.383	1	0	[Cleaning Ame...	Jan	Winter	sunny	Happy New Ye...
7	0	0.667	1	0.168	0.383	1	0	[Cleaning Ame...	Jan	Winter	sunny	Happy New Ye...
8	1	0.833	0	0.156	0.432	0	0.635	[Cleaning Ame...	Jan	Winter	sunny	Beautiful sunn...
9	0	0.833	0	0.156	0.432	0	0	[Cleaning Ame...	Jan	Winter	sunny	Beautiful sunn...
10	1	0.833	0	0.156	0.432	0	0.885	[Cleaning Ame...	Jan	Winter	sunny	Beautiful sunn...
11	0	0.833	0	0.156	0.432	0	0	[Cleaning Ame...	Jan	Winter	sunny	Beautiful sunn...
12	0	0.833	0	0.156	0.432	0	0	[Cleaning Ame...	Jan	Winter	sunny	Beautiful sunn...
13	0	0.833	0	0.156	0.432	0	0	[Cleaning Ame...	Jan	Winter	sunny	Beautiful sunn...
14	0	0.833	0	0.156	0.432	0	0	[Cleaning Ame...	Jan	Winter	sunny	Beautiful sunn...
15	0	1	0	0.137	0.494	0	0	[Cleaning Ame...	Jan	Winter	overcast	Overcast ske...
16	0	1	0	0.137	0.494	0	0	[Cleaning Ame...	Jan	Winter	overcast	Overcast ske...
17	1	1	0	0.137	0.494	0	0.885	[Cleaning Ame...	Jan	Winter	overcast	Overcast ske...

5. Handling imbalance data pada target/label dengan menggunakan metode SMOTE UPSAMPLING, dapat terlihat antara target kedua kelas tidak seimbang yaitu lebih banyak target di kelas 0 daripada 1 maka dilakukan penyeimbangan data

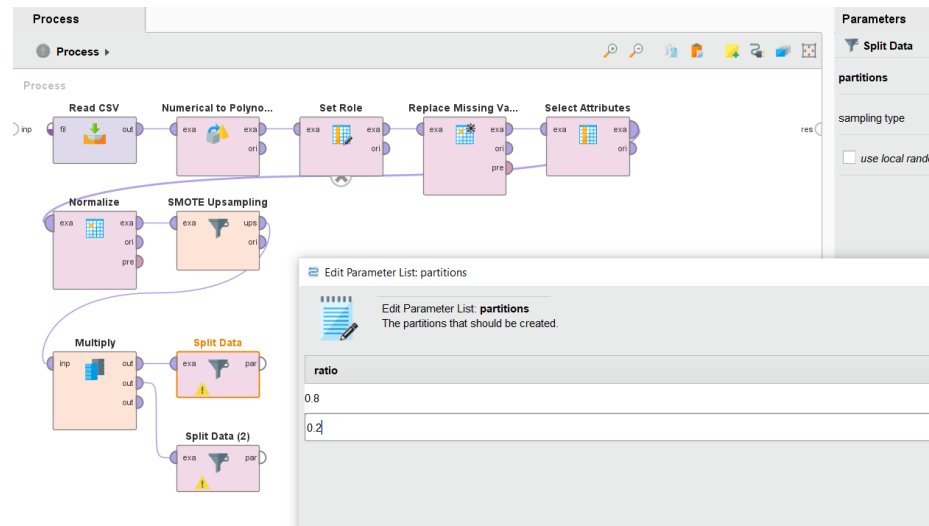


Maka akan didapatkanlah hasil seperti berikut ini, kedua kelas sudah seimbang.

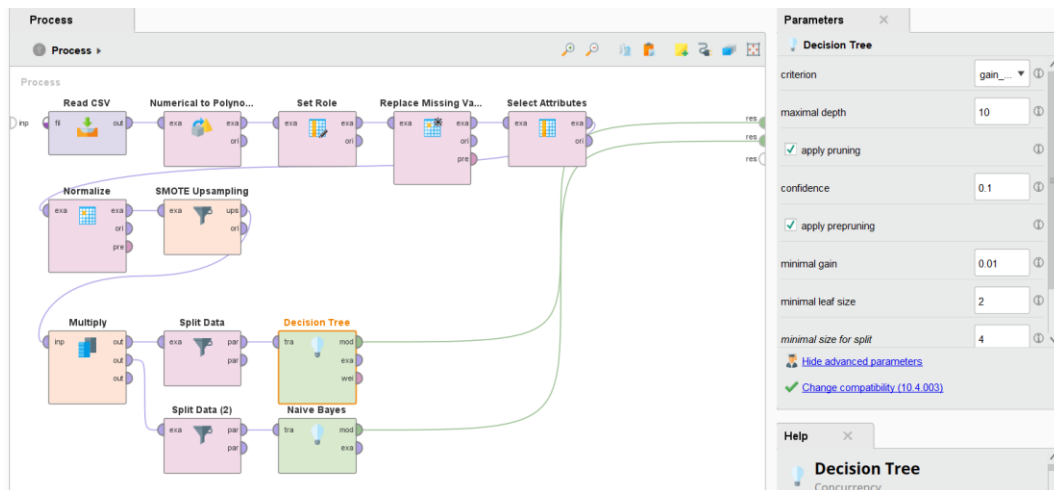
ExampleSet (SMOTE Upsampling)					
Name	Type	Missing	Statistics		
Filter (16 / 16 attributes): <input type="text" value="Search for Attributes"/>					
Label <b>Play</b>	Nominal	0	Least 1 (6266)	Most 0 (6266)	Values 0 (6266), 1 (6266)
<b>Maintenance Task</b>	Polynomial	0	Least [Wateri [...] ent] (7)	Most [Cleani [...] ] (4451)	Values [Cleaning Amenities] (4451), [Sprinkler Checks] (655), ...[191 more]

## DATA MODELLING

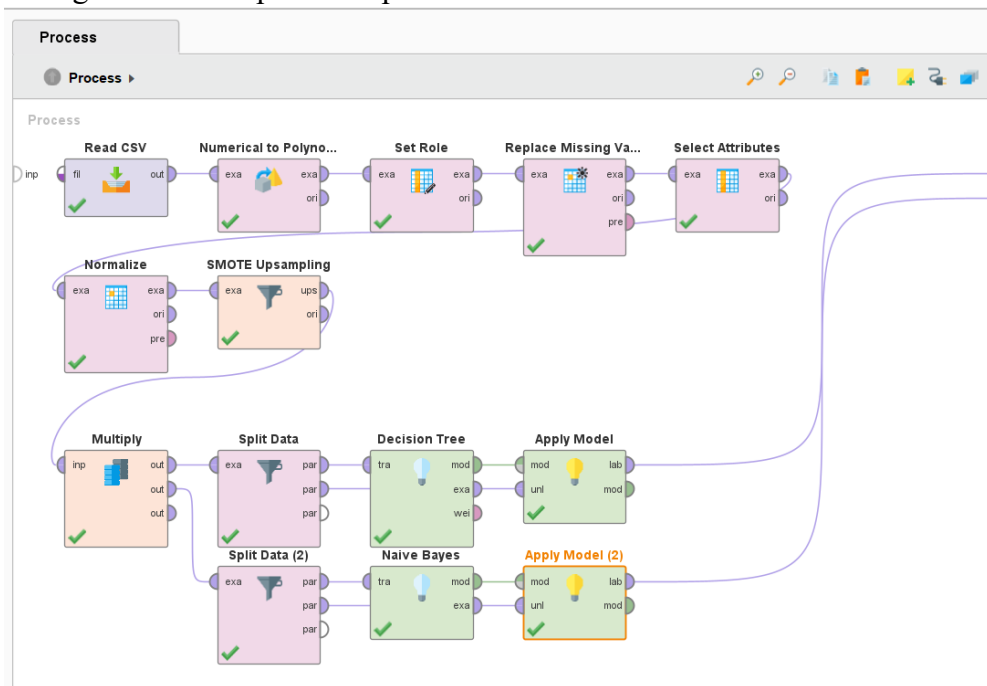
1. Melakukan pembagian data menjadi data training dan data testing menggunakan operator “splitting data” dimana disini dibagi menjadi rasio 80% untuk data training dan 20% untuk data testing.



2. Melakukan pemodelan dengan menggunakan dua algoritma yaitu “Decission tree” dan “Naive bayes” dengan parameter deafulat yang telah ditentukan.

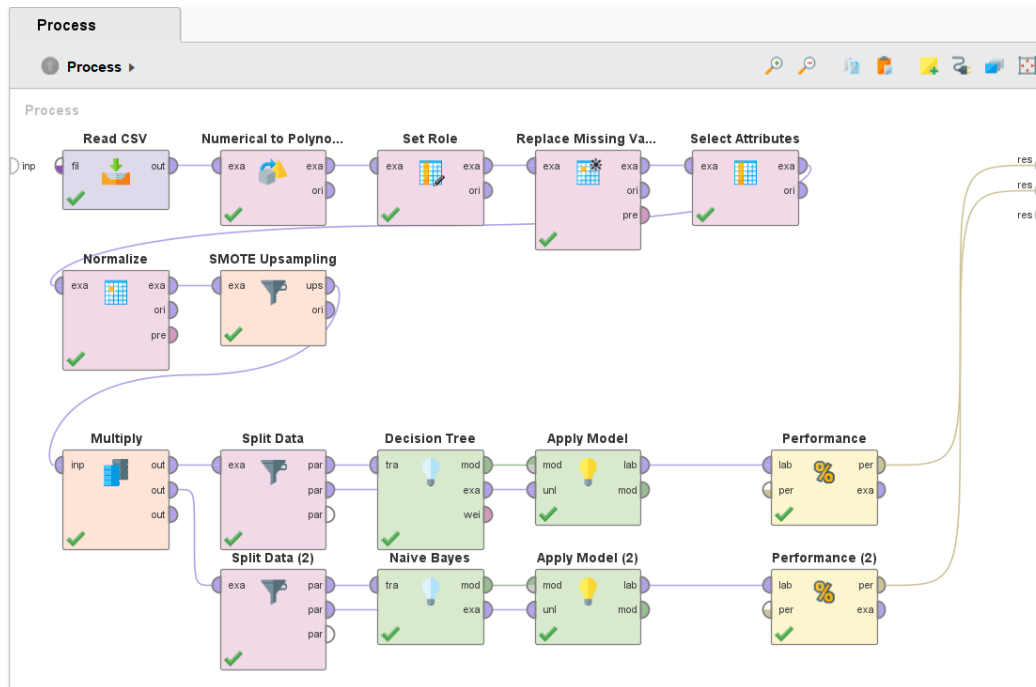


- Proses peneraapan model dengan menggunakan operator “Apply model” dimana algoritma yang telah dicoba diterapkan kedalam pemodelan sehingga menghasilkan prediksi antara data sebenarnya dengan data hasil prediksi apakah sudah sesuai atau belum



- Model evaluation, dimana disini menilai atau mengukur kinerja dari performa suatu model dengan menggunakan metriks evaluasi berupa accuracy.





## PERFORMA MODEL

1. Decision tree, memiliki performa sangat baik dimana nilai akurasi yang dihasilkan mencapai 100% ini artinya model dapat memberikan prediksi yang tepat

Result History

PerformanceVector (Performance (2))

PerformanceVector (Performance)

Criterion: accuracy

Table View Plot View

accuracy: 100.00%

	true 1	true 0	class precision
pred. 1	1253	0	100.00%
pred. 0	0	1253	100.00%
class recall	100.00%	100.00%	

2. Naïve bayes, sama halnya dengan algoritma sebelumnya juga memberikan nilai akurasi yang sangat baik 100%, artinya model sangat handal dalam melakukan prediksi.

History

PerformanceVector (Performance (2))

PerformanceVector (Performance)

Criterion: accuracy

Table View Plot View

accuracy: 100.00%

	true 1	true 0	class precision
pred. 1	1253	0	100.00%
pred. 0	0	1253	100.00%
class recall	100.00%	100.00%	

Kedua model memberikan gambaran pada confusion matrix :

- Ada **1.253 sampel dari kelas '1'** yang diprediksi dengan benar sebagai '1' (True Positives).
- Ada **1.253 sampel dari kelas '0'** yang diprediksi dengan benar sebagai '0' (True Negatives).
- Tidak ada kesalahan prediksi, karena tidak ada **False Positives** (kelas '1' diprediksi sebagai '0') dan tidak ada **False Negatives** (kelas '0' diprediksi sebagai '1').

Interpretasi

Kedua model ini sangat baik dalam klasifikasi, dengan akurasi sempurna pada data uji. Namun, hasil akurasi 100% bisa juga menimbulkan kekhawatiran tentang potensi overfitting, terutama jika data uji dan data pelatihan terlalu mirip atau jika model terlalu kompleks dan tidak generalisasi dengan baik pada data yang belum pernah dilihat sebelumnya.