

CRISP-DM: FINAL PHASE

Nur Fajar – NPM.187006102

nurfajar.tech@gmail.com

CRISP-DM(1)

BUSINESS UNDERSTANDING

Program pengelolaan kesehatan sehingga dapat menghindari pelanggan dari kemungkinan terkena serangan jantung.

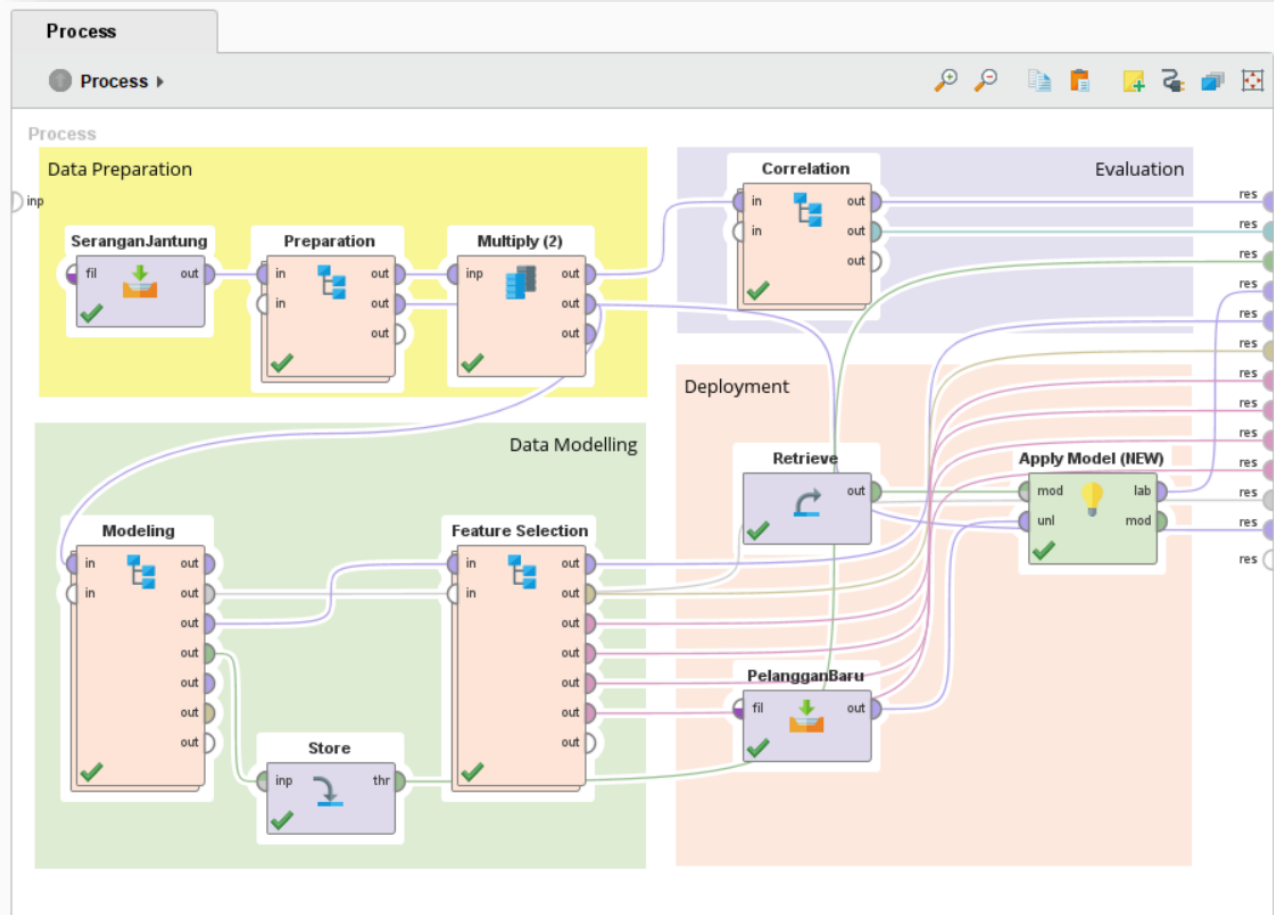
Untuk bisa sukses melakukan program ini, Budi harus:

menemukan pola pelanggan asuransi dengan profile seperti apa yang kemungkinan terkena serangan jantung.

DATA UNDERSTANDING

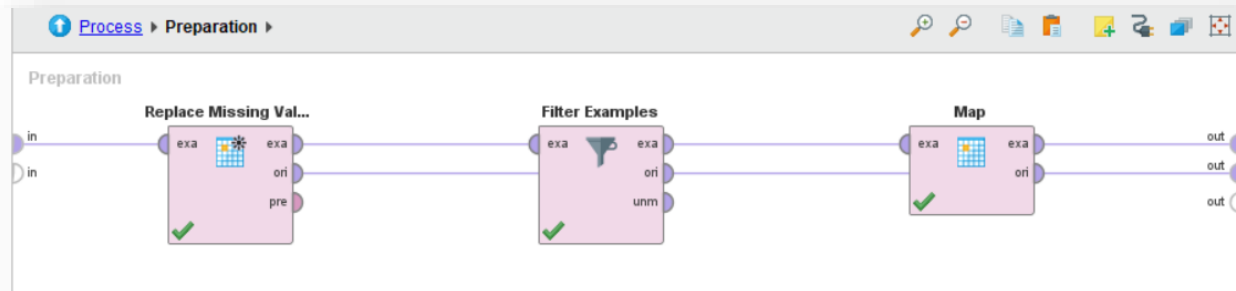
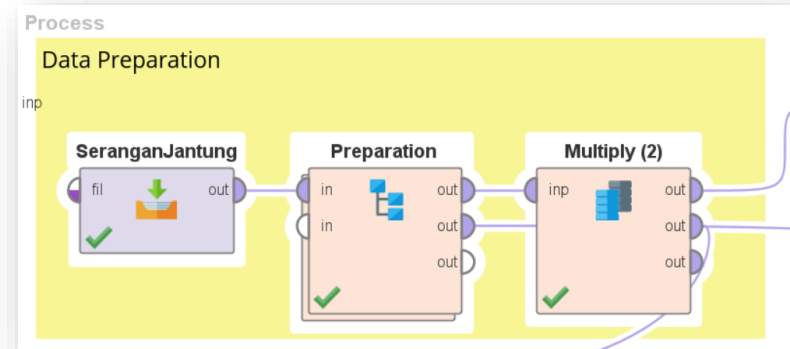
ATRIBUT	KETERANGAN			
Umur	Maksimal 116 tahun			
Status Pernikahan	0 = Belum Menikah	1 = Sudah Menikah	2 = Ditinggal Cerai	3 = Ditinggal Meninggal
Jenis Kelamin	0 = Perempuan		1 = Laki-laki	
Kategori Berat Badan	0 = Normal	1 = Kelebihan Berat Badan		2 = Obesitas
Kolesterol	0 – 400			
Pelatihan Pengelolaan Stress	0 = Tidak Mengikuti		1 = Mengikuti	
Tingkat Stress	0 – 100 %			
Serangan Jantung	0 = Tidak		1 = Ya	

Tampilan Process Rapidminer



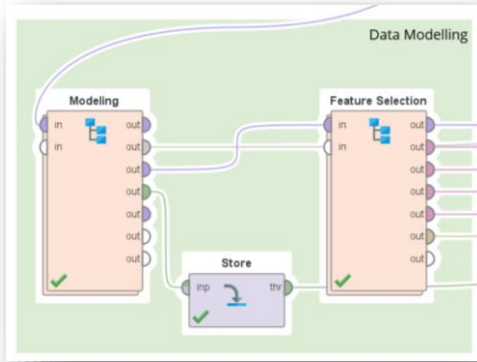
DATA PREPARATION

- Mengganti Missing Value dengan Nilai Rerata.
- Menghilangkan Data yang berada di luar jangkauan data.
- Mengubah Value Yes menjadi 1 dan No menjadi 0 pada Atribut Serangan Jantung.



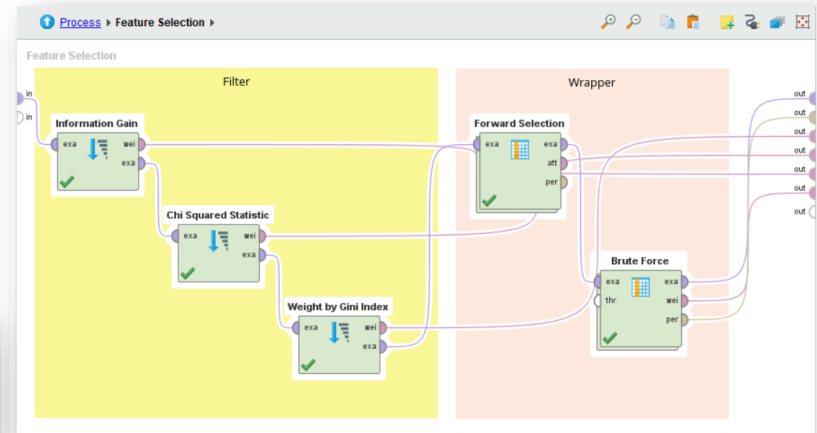
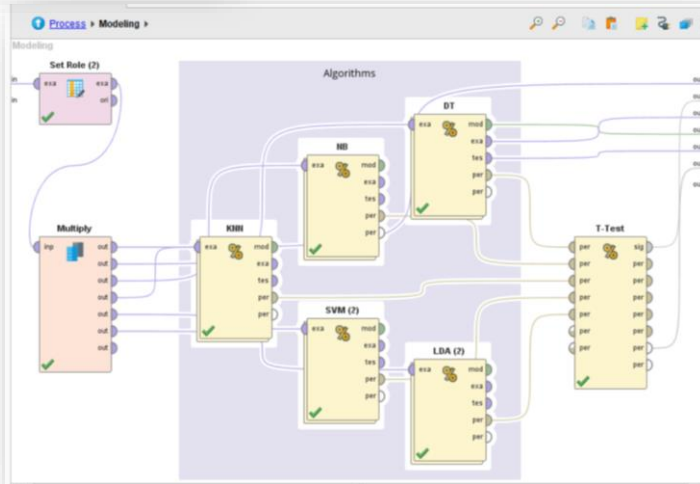
CRISP-DM (4)

MODELLING



Membandingkan Algoritma:

1. Decision Tree
2. Naïve Bayes
3. KNN
4. SVM
5. LDA






Menggunakan Feature Selection dengan 2 Kategori:

1. Filter dengan Information gain, Chi Squared, dan Gini Index.
2. Wrapper dengan Forward Selection dan Brute Force.

Urutan Algoritma Berdasar t-Test:

- | | | |
|----|---------------|----------|
| 0. | Decision Tree | : 97.0 % |
| 1. | Naïve Bayes | : 88.6 % |
| 2. | KNN | : 79.7 % |
| 3. | SVM | : 87.9 % |
| 4. | LDA | : 89.3 % |

Didapatkan, **Algoritma Decision Tree** merupakan algoritma terbaik dengan akurasi tertinggi sebesar 97.0%.

 T-test significance	Pairwise t-Test Probabilities for random values with the same result: <pre> ----- 0.034 0.002 0.055 0.030 ----- 0.125 0.895 0.872 ----- ----- 0.196 0.085 ----- ----- 0.781 ----- ----- ----- </pre>
 Description	Values smaller than alpha=0.050 indicate a probably significant difference between the mean values! List of performance values: <pre> 0: 0.970 +/- 0.051 1: 0.886 +/- 0.105 2: 0.797 +/- 0.139 3: 0.879 +/- 0.132 4: 0.893 +/- 0.091 </pre>
 Annotations	

Information Gain

attribute	weight ↓
Kategori Berat Badan	0.357
Status Pernikahan	0.251
Tingkat Stress	0.201
Pelatihan Pengelolaan Stress	0.173
Umur	0.158
Jenis Kelamin	0.075
Kolesterol	0.024

Chi Squared

attribute	weight ↓
Kategori Berat Badan	68.061
Status Pernikahan	55.822
Tingkat Stress	52.094
Pelatihan Pengelolaan Stress	30.104
Umur	27.622
Jenis Kelamin	13.322
Kolesterol	6.005

Forward Selection

attribute	weight ↓
Status Pernikahan	1
Kategori Berat Badan	1
Umur	0
Jenis Kelamin	0
Kolesterol	0
Pelatihan Pengelolaan Stress	0
Tingkat Stress	0

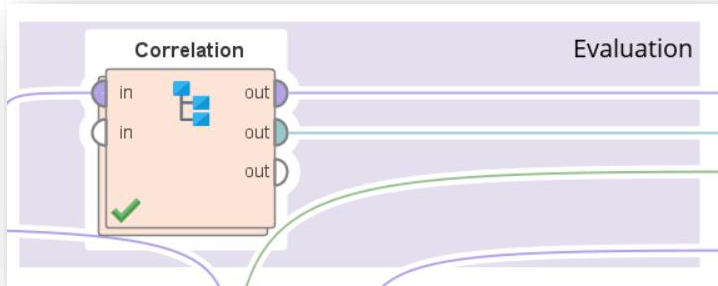
Atribut **Kategori Berat Badan** dan **Status Pernikahan** selalu menjadi 2 atribut teratas sehingga Atribut ini sangat berpengaruh pada Kemungkinan Seseorang mendapatkan Serangan Jantung.

Setelah dilakukan
Feature Selection,
Akurasi Algoritma
Decision Tree **Meningkat**
menjadi **98.11%**

Akurasi Algoritma Decision Tree setelah Feature Selection

Criterion	<input checked="" type="radio"/> Table View <input type="radio"/> Plot View		
accuracy	accuracy: 98.11%		
precision			
recall			
AUC (optimistic)			
AUC			
AUC (pessimistic)			
	true 1	true 0	class precision
pred. 1	26	1	96.30%
pred. 0	0	26	100.00%
class recall	100.00%	96.30%	

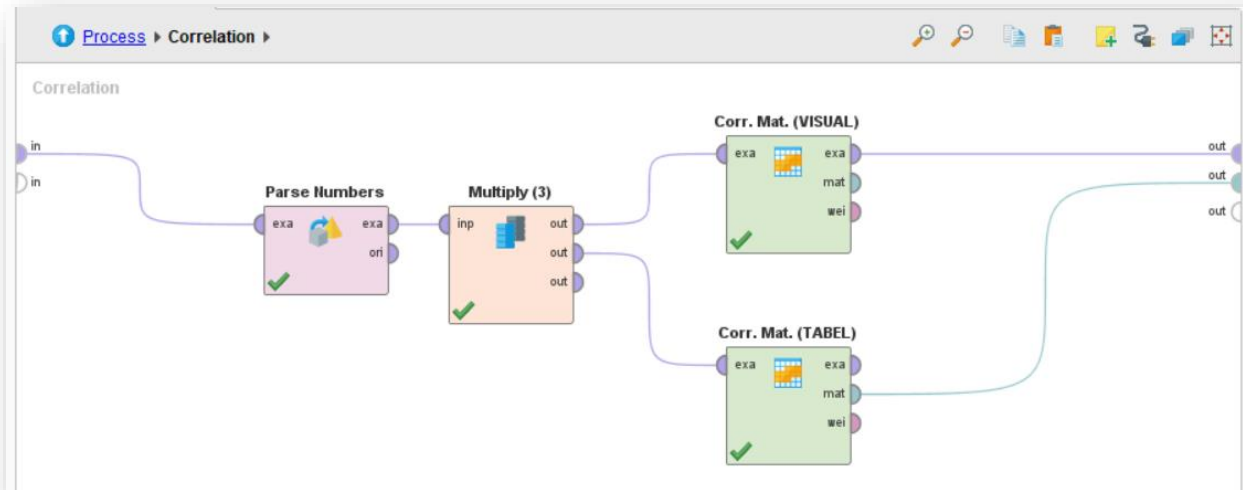
CRISP-DM (5) EVALUATION



Menggunakan 2 Buah Operator Correlation Matrix dikarenakan ketika menggunakan hanya 1 Operator Correlation Matrix **tidak dapat memunculkan Example data dan Matrix secara bersamaan.**



Operator Parse Numbers digunakan untuk **mengubah type atribut Serangan Jantung menjadi bertipe numeric** agar bisa diproses pada Operator Correlation Matrix.



CRISP-DM (5)

EVALUATION

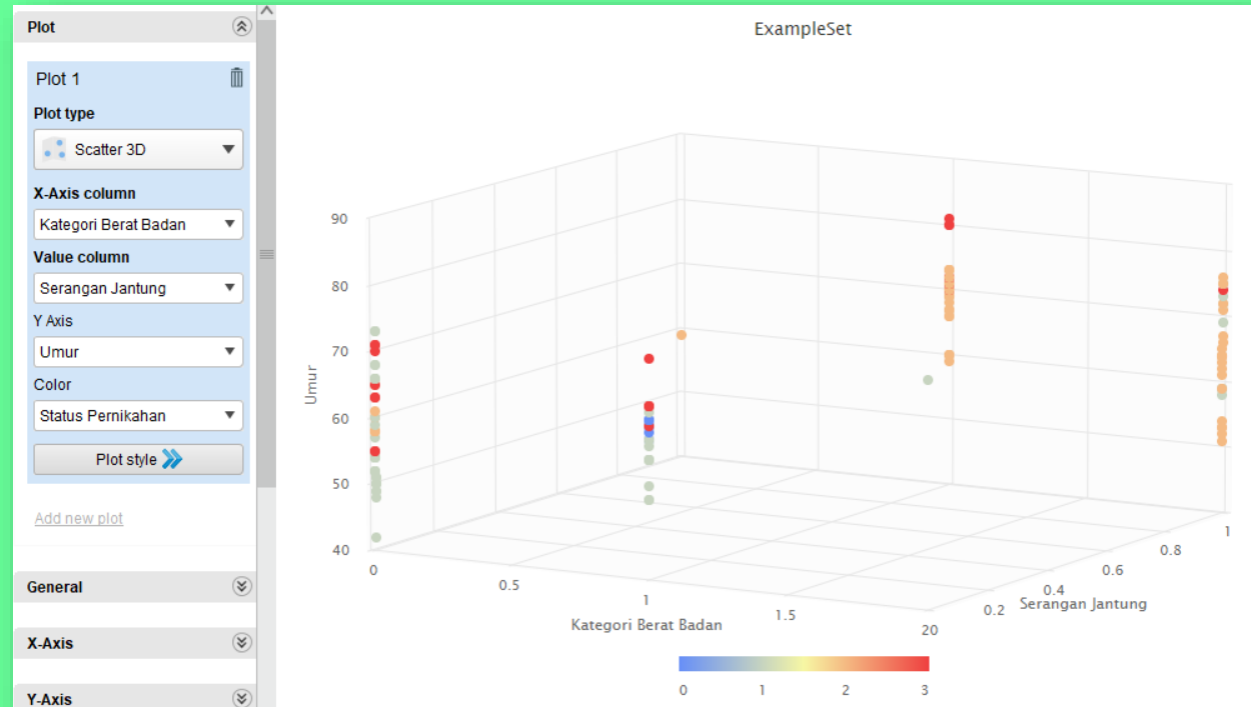
Attributes	Serangan ...	Umur	Status Pernik...	Jenis Kelamin	Kategori Berat Badan	Kolesterol	Pelatihan Pe...	Tingkat Stress
Serangan Jantung	1	0.448	0.388	0.318	0.715	0.111	-0.478	0.477
Umur	0.448	1	0.394	0.011	0.314	0.170	-0.343	0.586
Status Pernikahan	0.388	0.394	1	-0.004	0.076	0.072	-0.326	0.238
Jenis Kelamin	0.318	0.011	-0.004	1	0.432	-0.017	-0.261	0.066
Kategori Berat Badan	0.715	0.314	0.076	0.432	1	-0.005	-0.371	0.469
Kolesterol	0.111	0.170	0.072	-0.017	-0.005	1	-0.113	0.070
Pelatihan Pengelolaan Stress	-0.478	-0.343	-0.326	-0.261	-0.371	-0.113	1	-0.377
Tingkat Stress	0.477	0.586	0.238	0.066	0.469	0.070	-0.377	1

Pada Correlation Matrix di atas diperlihatkan

Serangan Jantung Memiliki **Korelasi Positif Kuat dengan Kategori Berat Badan** yaitu dengan nilai 0.715, dan hanya memiliki **korelasi negative dengan Pelatihan Pengelolaan Stress** yaitu dengan nilai -0.478.

CRISP-DM (5) EVALUATION

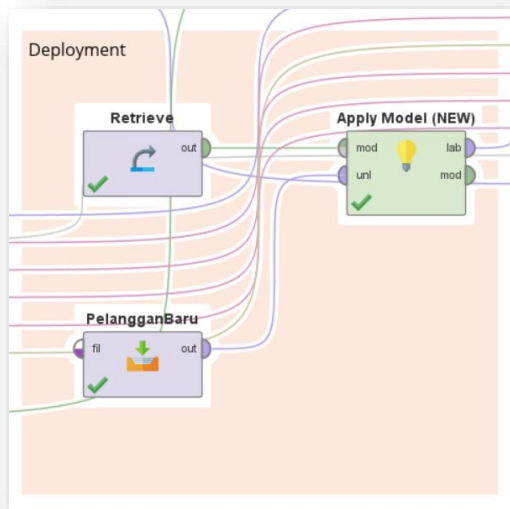
Visualisasi dengan menggunakan Scatter 3D menunjukan pelanggan yang Memiliki Serangan Jantung, kebanyakan Memiliki **Status Pernikahan pada level 2 (Ditinggal Cerai)** dan memiliki kategori berat badan pada level 2 (Obesitas).



CRISP-DM (6)

DEPLOYMENT

Deployment mencoba untuk menggunakan model (terbaik) sebelumnya lalu diterapkan pada dataset baru berjumlah 10 data untuk menerapkan prediksi yang dihasilkan dari tahap model sebelumnya.



Dari Hasil deployment didapatkan tingkat confident berikut:

Data Nomor 2, 4, 5, 9 diprediksi 100% Dia berpotensi untuk memiliki Serangan Jantung.

Data Nomor 1, 3, 8, 10 diprediksi 100% Dia tidak berpotensi untuk memiliki Serangan Jantung.

Kebanyakan Data yang diprediksi memiliki Serangan Jantung adalah memiliki Status Pernikahan (2) ditinggal cerai.

Row No.	prediction(S...	confidence(1) ↓	confidence(0)	Umur	Status Pe...	Jenis K...	Kategori B...	Kolesterol	Pelatihan ...	Tingkat Stre...
2	1	1	0	55	2	1	2	163	0	40
4	1	1	0	58	1	1	2	206	0	70
5	1	1	0	62	2	1	1	148	1	50
9	1	1	0	67	2	1	1	172	0	60
6	0	0.024	0.976	70	1	0	0	172	0	60
7	0	0.024	0.976	52	1	0	0	171	1	35
1	0	0	1	61	0	1	1	139	1	50
3	0	0	1	53	1	1	1	172	0	55
8	0	0	1	50	1	1	1	172	0	55
10	0	0	1	62	1	1	1	166	1	50

KESIMPULAN

1. Menurut hasil t-Test, Algoritma **Decision Tree** memiliki **Akurasi Paling Tinggi** yaitu: 97.0% dan meningkat menjadi 98.11% setelah dilakukan feature selection.
2. **Korelasi positif terkuat** yang dimiliki Serangan Jantung adalah dengan Kategori Berat Badan dengan nilai korelasi sebesar **0.715**.
3. Pelanggan Asuransi yang memiliki status pernikahan **ditinggal cerai** dan memiliki kategori berat badan **obesitas** akan **sangat memungkinkan untuk mendapatkan serangan jantung**.

TERIMA KASIH.

Nur Fajar – NPM.187006102

nurfajar.tech@gmail.com