`

LAPORAN

CLEANSING PREPROCESSING

NAMA ANGGOTA KELOMPOK :

1. Adie Gunawan Alwani (5200411486)
2. Alfia Candra Kusumapratiwi (5200411487)
3. Arieska Restu Harpian Dwika (5200411488)

Coding & Machine Learning E

S1 INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

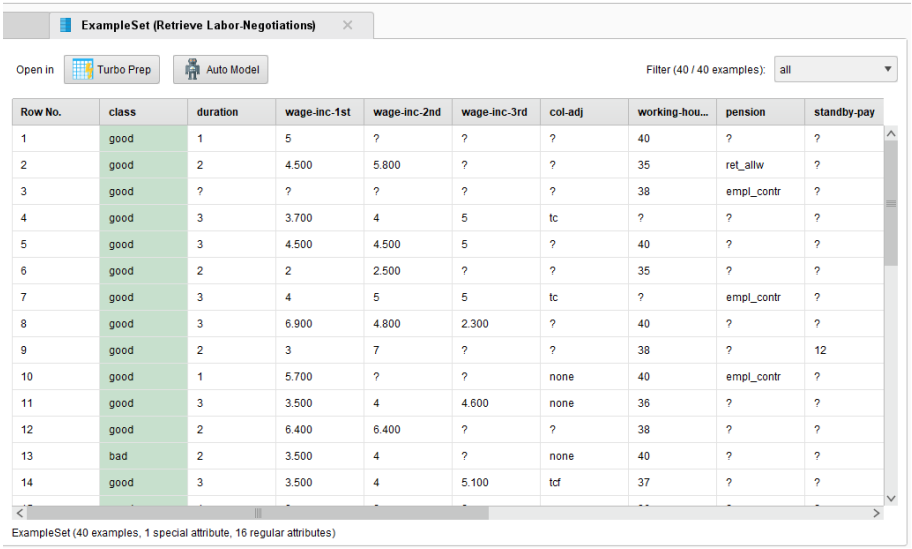
UNIVERSITAS TEKNOLOGI YOGYAKARTA

2021/2022

`

Data Original

Pada tugas ini kami menggunakan data labor negotiation yang terdapat pada sample di rapidminer. Isi data tersebut dapat dilihat pada gambar beikut:



ExampleSet (Retrieve Labor-Negotiations)

Open in Turbo Prep Auto Model Filter (40 / 40 examples): all

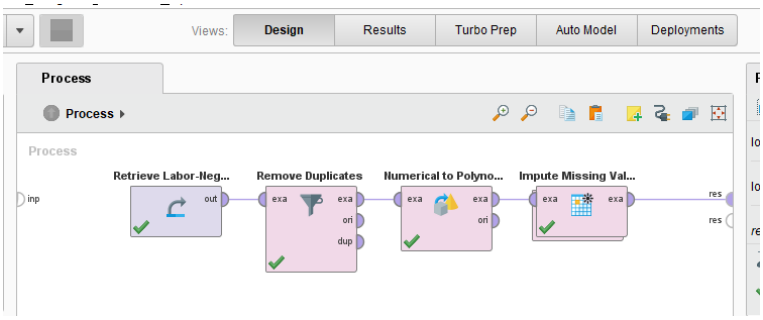
Row No.	class	duration	wage-inc-1st	wage-inc-2nd	wage-inc-3rd	col-adj	working-hou...	pension	standby-pay
1	good	1	5	?	?	?	40	?	?
2	good	2	4.500	5.800	?	?	35	ret_allw	?
3	good	?	?	?	?	?	38	empl_contr	?
4	good	3	3.700	4	5	tc	?	?	?
5	good	3	4.500	4.500	5	?	40	?	?
6	good	2	2	2.500	?	?	35	?	?
7	good	3	4	5	5	tc	?	empl_contr	?
8	good	3	6.900	4.800	2.300	?	40	?	?
9	good	2	3	7	?	?	38	?	12
10	good	1	5.700	?	?	none	40	empl_contr	?
11	good	3	3.500	4	4.600	none	36	?	?
12	good	2	6.400	6.400	?	?	38	?	?
13	bad	2	3.500	4	?	none	40	?	?
14	good	3	3.500	4	5.100	tcf	37	?	?

ExampleSet (40 examples, 1 special attribute, 16 regular attributes)

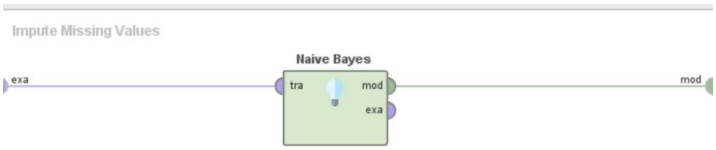
Selanjutnya kami melakukan cleansing preprocessing pada data tersebut dengan 3 metode yang berbeda. Metode-metode tersebut menggunakan operator **Impute Missing Values**. Operator ini memperkirakan nilai untuk nilai yang hilang dari atribut yang dipilih dengan menerapkan model yang dipelajari untuk nilai yang hilang. Akan tetapi, dengan menggunakan algoritma yang berbeda. Algoritma yang kami gunakan yakni, Naive Bayes, Decision Tree, dan KNN.

Metode 1 - Naive Bayes

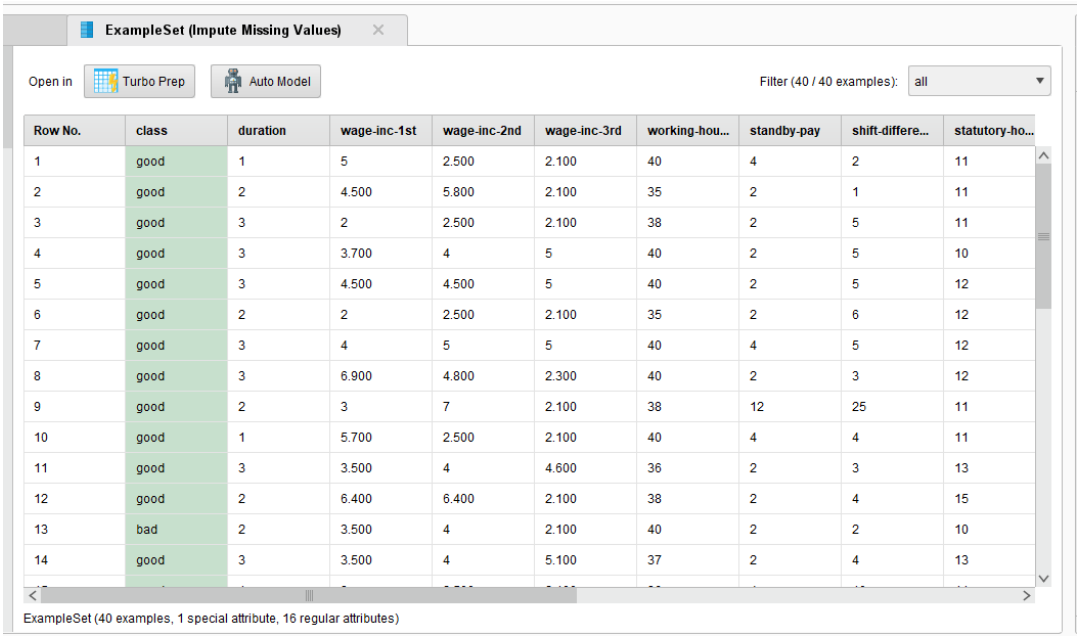
Pada metode 1, kami menggunakan algoritma Naïve Bayes. Naïve Bayes adalah pengklasifikasi bias tinggi, varians rendah, dan dapat membangun model yang baik bahkan dengan kumpulan data kecil. Kasus penggunaan yang umum melibatkan kategorisasi teks, termasuk deteksi spam, analisis sentimen, dan sistem pemberi rekomendasi. Secara komputasi tidak mahal. Kasus penggunaan umum melibatkan kategorisasi teks, termasuk deteksi spam, analisis sentimen, dan sistem pemberi rekomendasi. Proses yang kami lakukan pada rapidminer yaitu seperti gambar berikut.



`



Data tersebut diberi operator **Remove Duplicates**. Operator ini berfungsi untuk menghilangkan data yang sama. Setelah itu, dihubungkan dengan operator **Numerical to Polynominal** agar attribute yang numerical berubah menjadi polynominal. Kemudian dihubungkan dengan operator **Impute Missing Values**. Pada operator tersebut diberi operator **Naïve Bayes**. Hasil dari proses tersebut yaitu pada gambar berikut.



Row No.	class	duration	wage-inc-1st	wage-inc-2nd	wage-inc-3rd	working-hou...	standby-pay	shift-differe...	statutory-ho...
1	good	1	5	2.500	2.100	40	4	2	11
2	good	2	4.500	5.800	2.100	35	2	1	11
3	good	3	2	2.500	2.100	38	2	5	11
4	good	3	3.700	4	5	40	2	5	10
5	good	3	4.500	4.500	5	40	2	5	12
6	good	2	2	2.500	2.100	35	2	6	12
7	good	3	4	5	5	40	4	5	12
8	good	3	6.900	4.800	2.300	40	2	3	12
9	good	2	3	7	2.100	38	12	25	11
10	good	1	5.700	2.500	2.100	40	4	4	11
11	good	3	3.500	4	4.600	36	2	3	13
12	good	2	6.400	6.400	2.100	38	2	4	15
13	bad	2	3.500	4	2.100	40	2	2	10
14	good	3	3.500	4	5.100	37	2	4	13

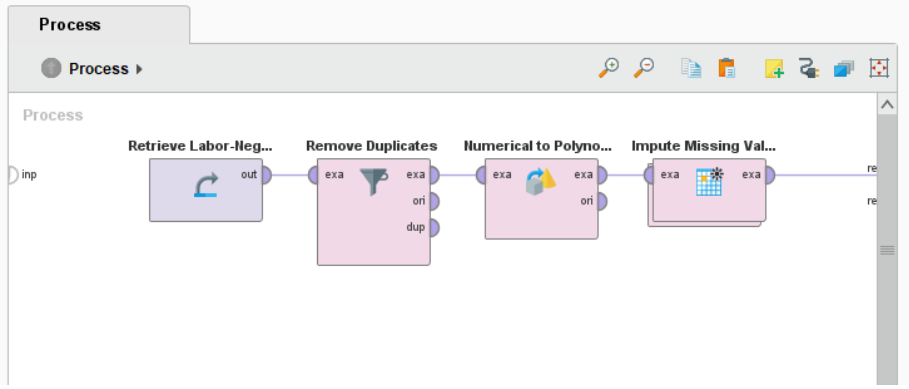
Hasil dari proses tersebut yaitu nilai yang missing diberi nilai berdasarkan algoritma Naïve Bayes. Berbeda dengan data yang sebelum diproses, data-data setelah diproses tidak terdapat data yang missing dan tidak terdapat data yang sama.

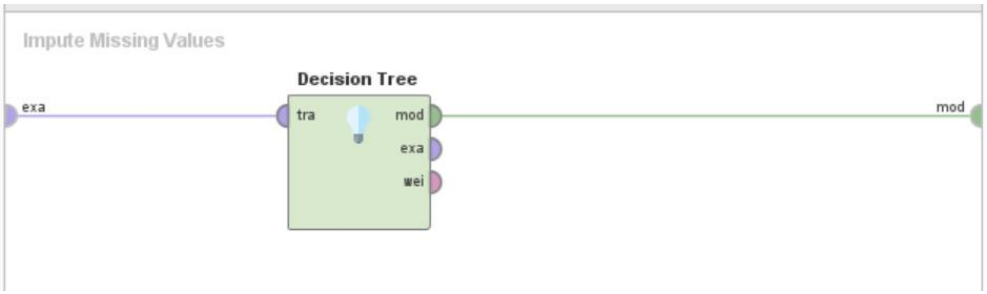
Metode 2 - Decision Tree

Pada metode 2, kami menggunakan algoritma Decision Tree. Decision Tree adalah kumpulan simpul seperti pohon yang dimaksudkan untuk membuat keputusan tentang afiliasi nilai ke kelas atau perkiraan nilai target numerik. Setiap node mewakili aturan pemisahan untuk satu Atribut tertentu.

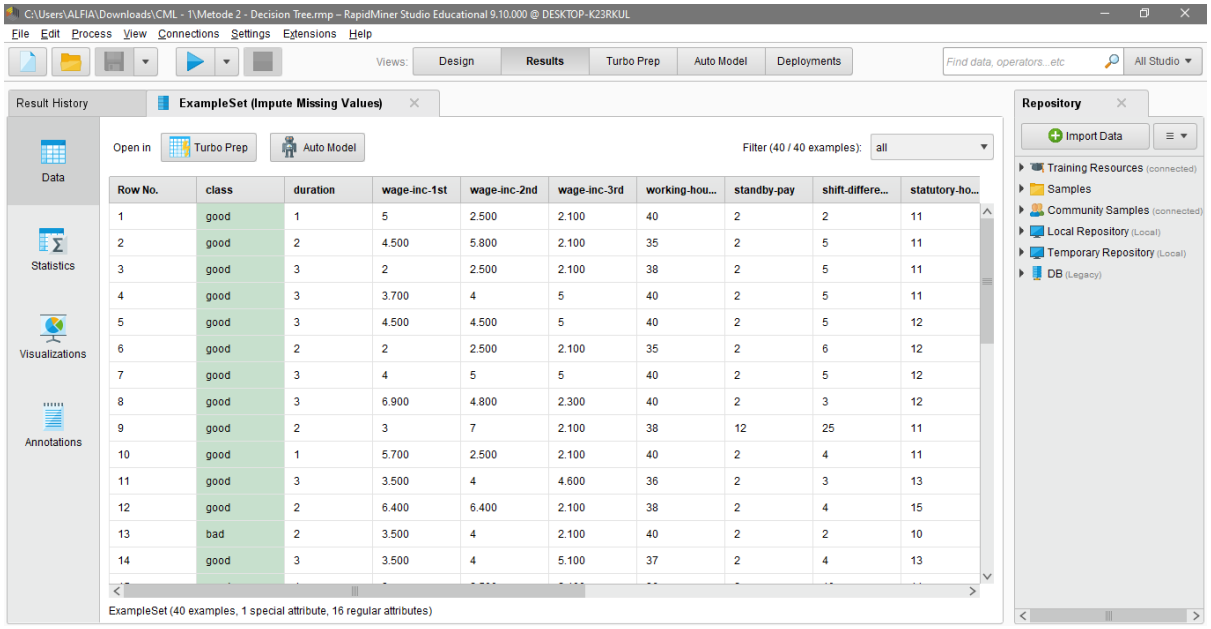
Untuk klasifikasi aturan ini memisahkan nilai-nilai milik kelas yang berbeda, untuk regresi itu memisahkan mereka untuk mengurangi kesalahan secara optimal untuk kriteria parameter yang dipilih. Proses yang kami lakukan pada rapidminer yaitu seperti gambar berikut.

`





Pada proses tersebut, kami menggunakan operator yang sama, tapi pada operator **Impute Missing Values** diberi operator **Decision Tree**. Hasil dari proses tersebut yaitu seperti pada gambar berikut.



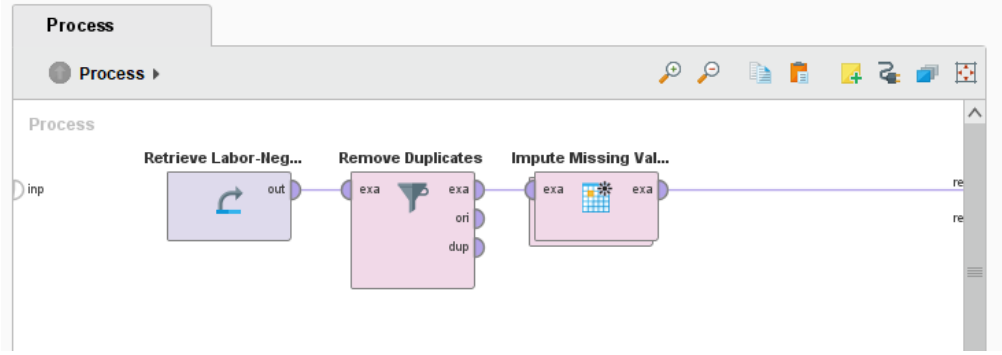
Row No.	class	duration	wage-inc-1st	wage-inc-2nd	wage-inc-3rd	working-hou...	standby-pay	shift-differe...	statutory-ho...
1	good	1	5	2.500	2.100	40	2	2	11
2	good	2	4.500	5.800	2.100	35	2	5	11
3	good	3	2	2.500	2.100	38	2	5	11
4	good	3	3.700	4	5	40	2	5	11
5	good	3	4.500	4.500	5	40	2	5	12
6	good	2	2	2.500	2.100	35	2	6	12
7	good	3	4	5	5	40	2	5	12
8	good	3	6.900	4.800	2.300	40	2	3	12
9	good	2	3	7	2.100	38	12	25	11
10	good	1	5.700	2.500	2.100	40	2	4	11
11	good	3	3.500	4	4.600	36	2	3	13
12	good	2	6.400	6.400	2.100	38	2	4	15
13	bad	2	3.500	4	2.100	40	2	2	10
14	good	3	3.500	4	5.100	37	2	4	13

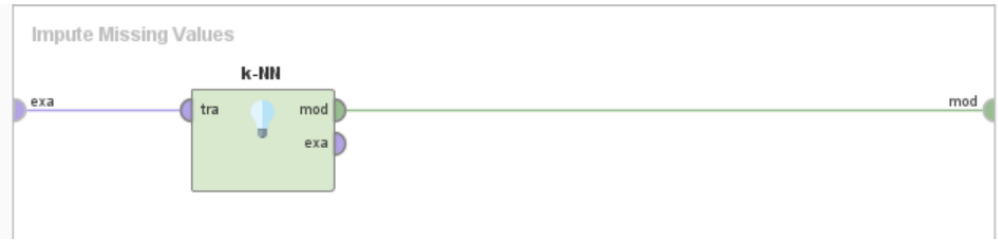
`

Hasil dari proses tersebut yaitu nilai yang missing diberi nilai berdasarkan algoritma Decision Tree. Berbeda dengan data yang sebelum diproses, data-data setelah diproses tidak terdapat data yang missing dan tidak terdapat data yang sama.

Metode 3 - KNN

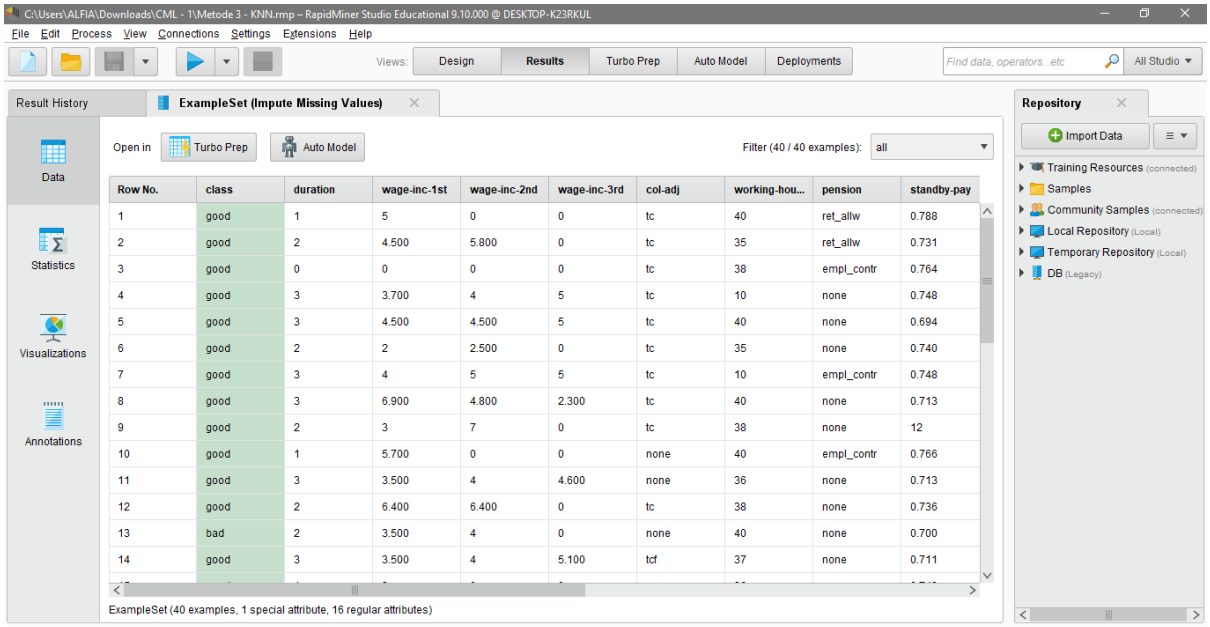
Pada metode 2, kami menggunakan algoritma KNN. Algoritma k-Nearest Neighbor didasarkan pada perbandingan Contoh yang tidak diketahui dengan Contoh pelatihan k yang merupakan tetangga terdekat dari Contoh yang tidak diketahui. Proses yang kami lakukan pada rapidminer yaitu seperti gambar berikut.





Pada proses tersebut, kami menggunakan operator **Remove Duplicate** dan **Impute Missing Values**. Pada operator **Impute Missing Values** diberi operator **KNN**. Hasil dari proses tersebut yaitu seperti pada gambar berikut.

`



Hasil dari proses tersebut yaitu nilai yang missing diberi nilai berdasarkan KNN. Berbeda dengan data yang sebelum diproses, data-data setelah diproses tidak terdapat data yang missing dan tidak terdapat data yang sama.