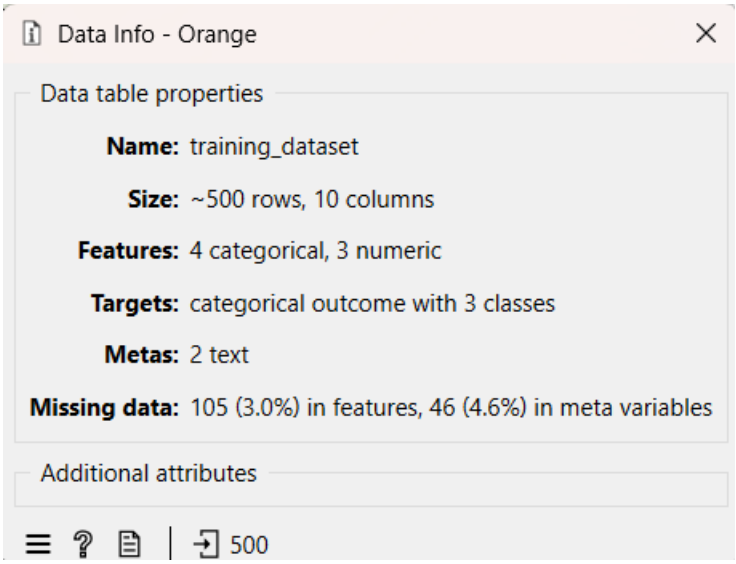
Soal 1 klasifikasi

A. Data Understanding

Identifikasi

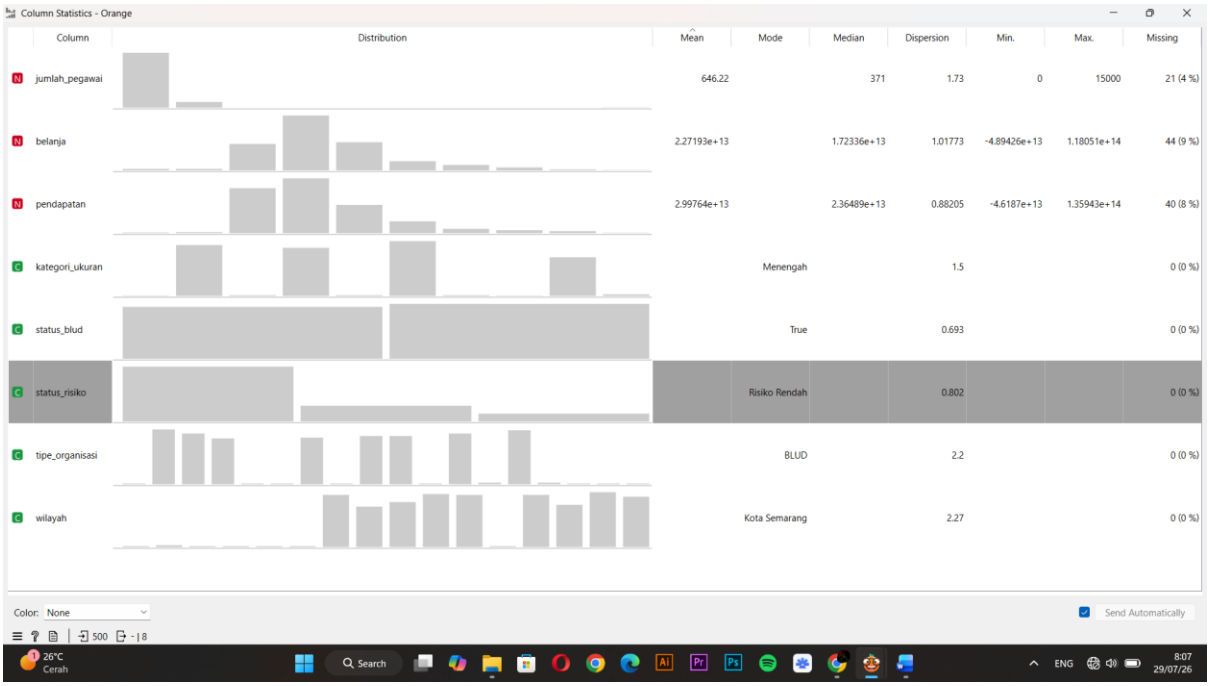


a) Jmmlah record = 500
b) Jumlah atribut/variabel/feature = 10
c) Tipe masing-masing atribut, d) atribut feature, e) atribut target (class) =

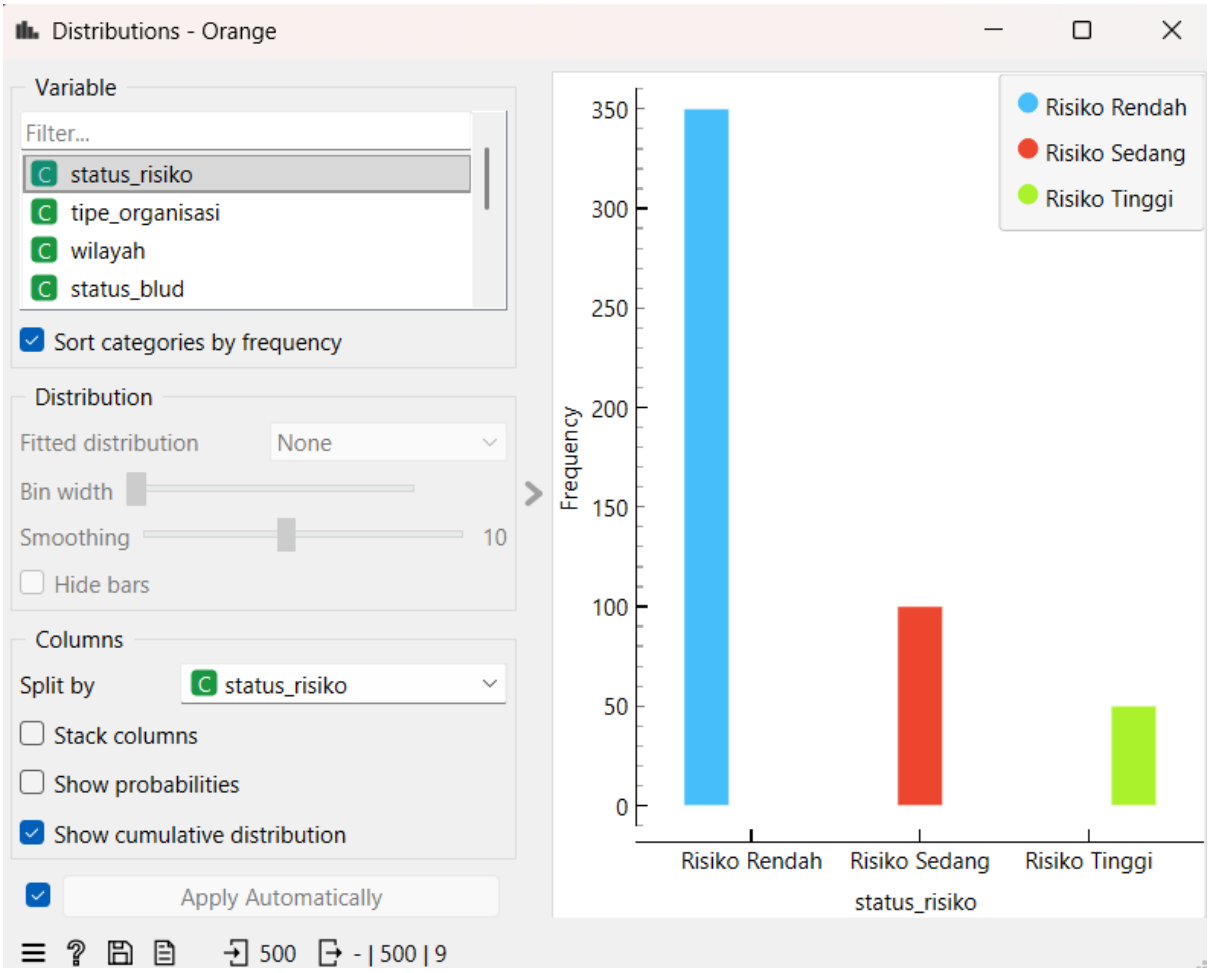
Columns (Double click to edit)

	Name	Type	Role
1	tipe_organisasi	C categorical	feature
2	wilayah	C categorical	feature
3	status_blud	C categorical	feature
4	kategori_ukuran	C categorical	feature
5	jumlah_pegawai	N numeric	feature
6	pendapatan	N numeric	feature
7	belanja	N numeric	feature
8	**status_risiko**	**C categorical**	**target**
9	kode_organisasi	S text	meta
10	tanggal_pembe...	S text	meta

Statistik deskriptif



Distribusi kelas target



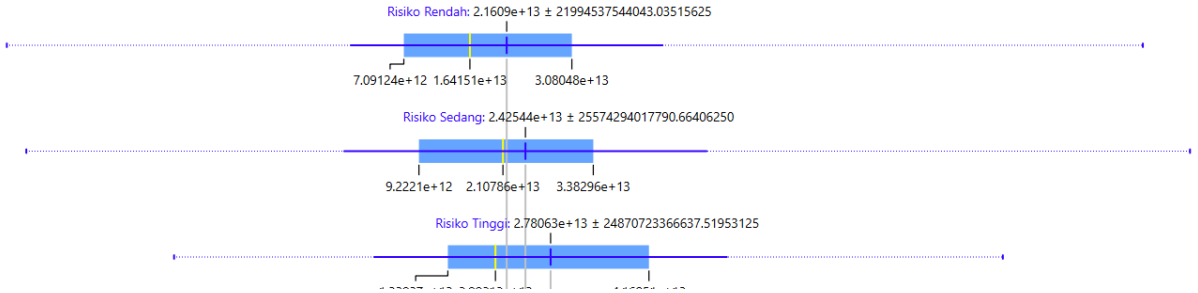
Identifikasi

a) Missing value = ada, 105 in features dan 46 in meta

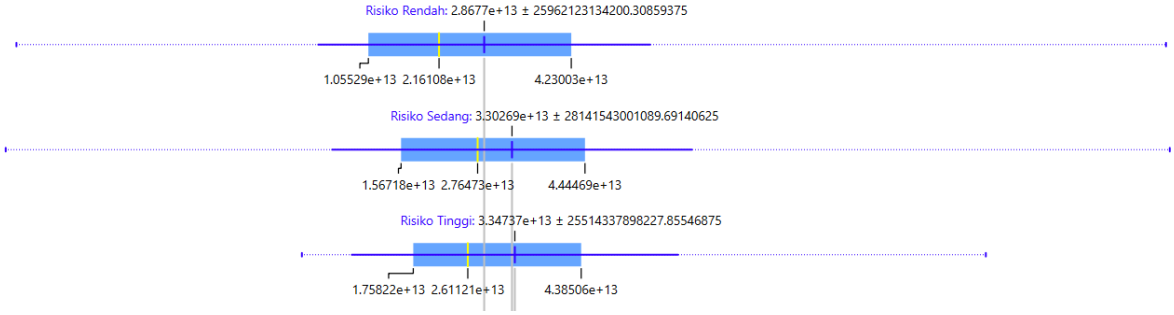
b) Class imbalance = ada, risiko rendah paling banyak, jauh dibanding risiko sedang dan risiko tinggi

c) Outlier = ada, terlihat di box plot ada titik yang sangat jauh dari sebaran normal jumlah pegawai untuk risiko rendah, begitu juga dengan feature numerik lain, terdapat titik yang melebar dari sebaran normal

belanja

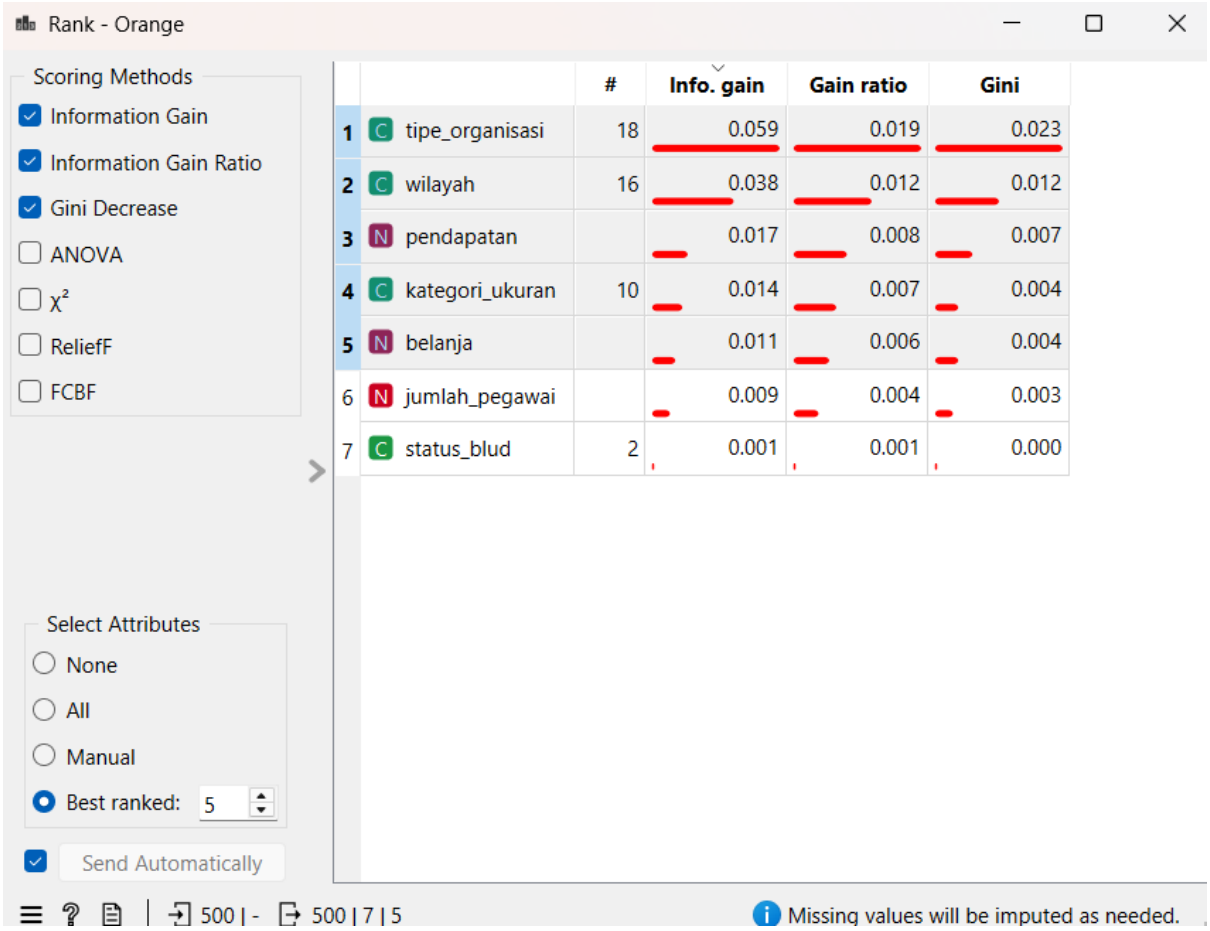


Pendapatan



Jumlah_pegawai

d) Atribut tidak relevan = ada, terlihat di rank atribut status_blud memiliki impact terhadap target relatif terlalu sedikit.



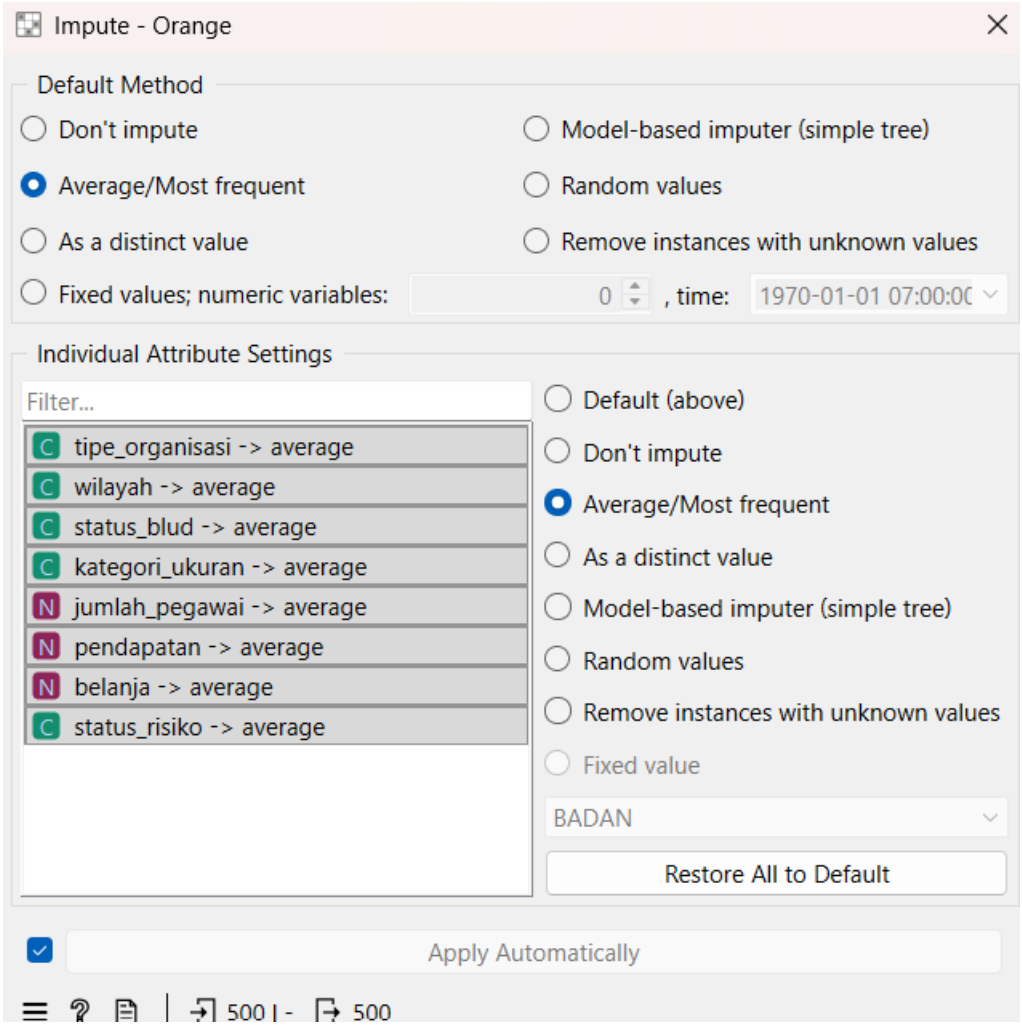
		#	Info. gain	Gain ratio	Gini
1	tipe_organisasi	18	0.059	0.019	0.023
2	wilayah	16	0.038	0.012	0.012
3	pendapatan		0.017	0.008	0.007
4	kategori_ukuran	10	0.014	0.007	0.004
5	belanja		0.011	0.006	0.004
6	jumlah_pegawai		0.009	0.004	0.003
7	status_blud	2	0.001	0.001	0.000

Identifikasi atribut yang relevan

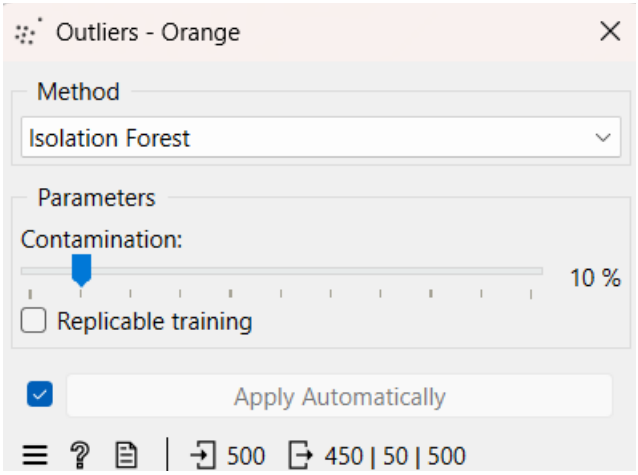
Berdasarkan rank, atribut yang relevan adalah tipe_organisasi, wilayah, pendapatan, kategori_ukuran, belanja dan jumlah_pegawai. Status_blud memiliki impact terhadap target relatif terlalu sedikit.

B. Data preparation

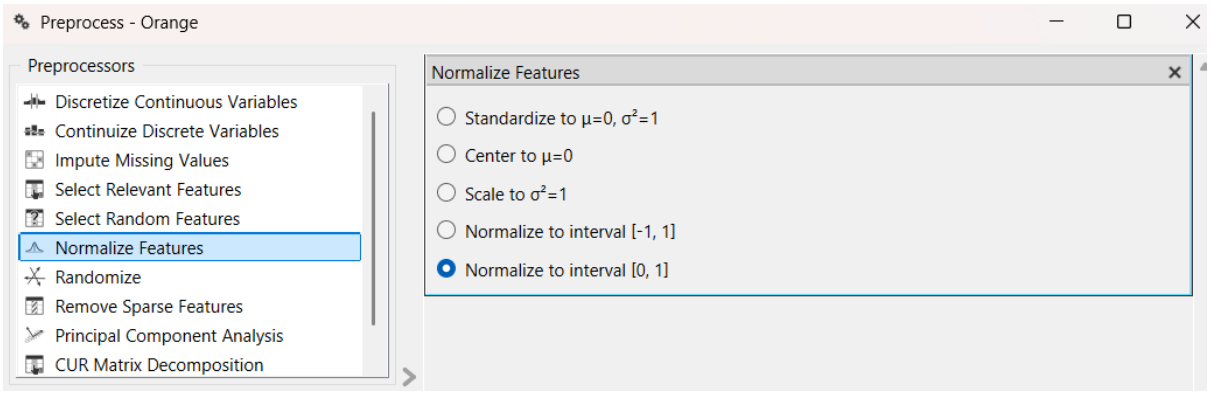
1. Menangani missing value



2. Menangani outlier

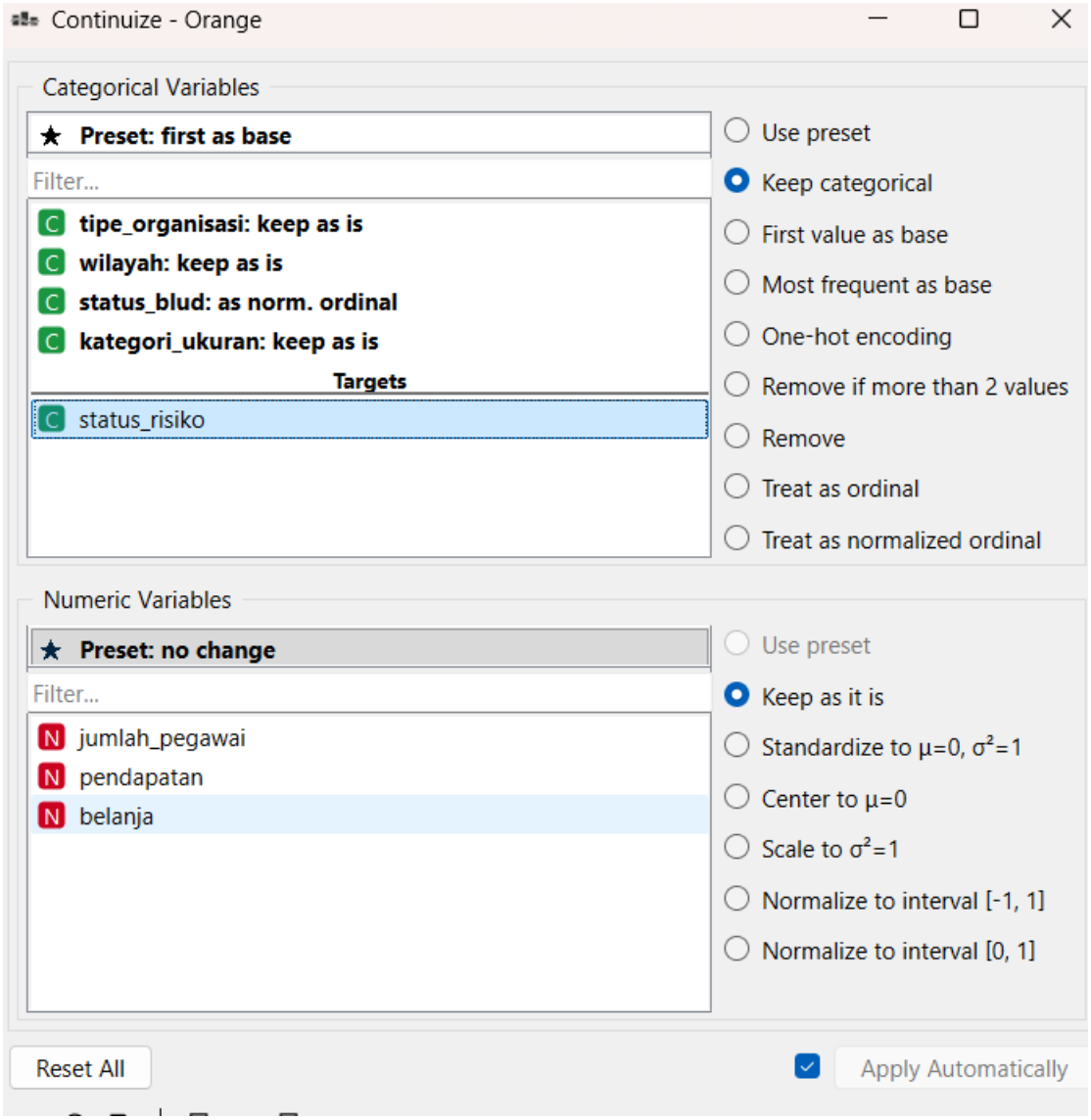


3. Normalisasi/scaling

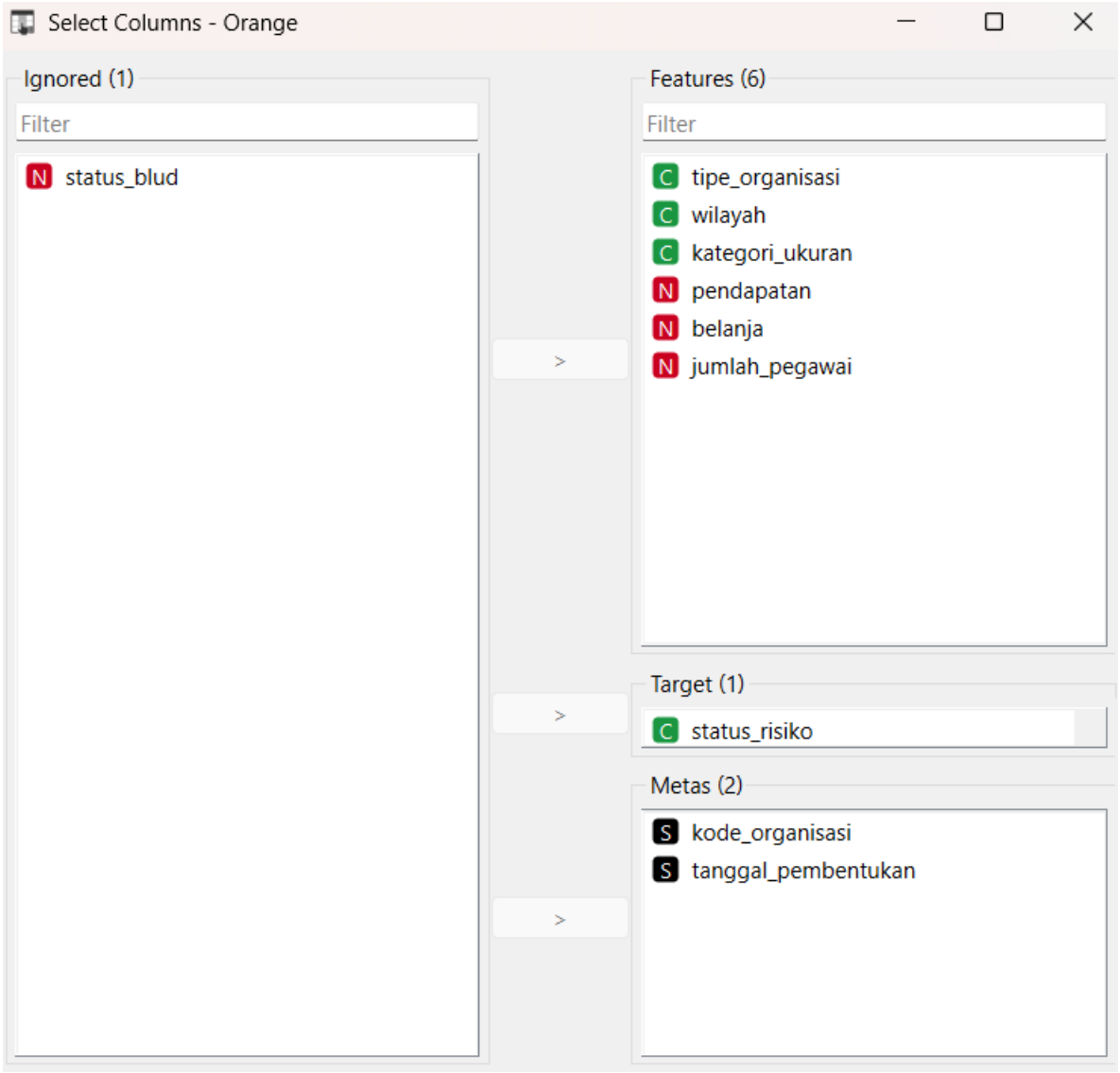


4. Encoding

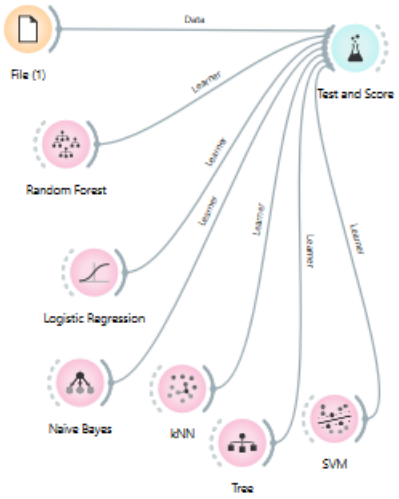
   Atribut status_blud yang berisi True dan False diubah menjadi 1 dan 0



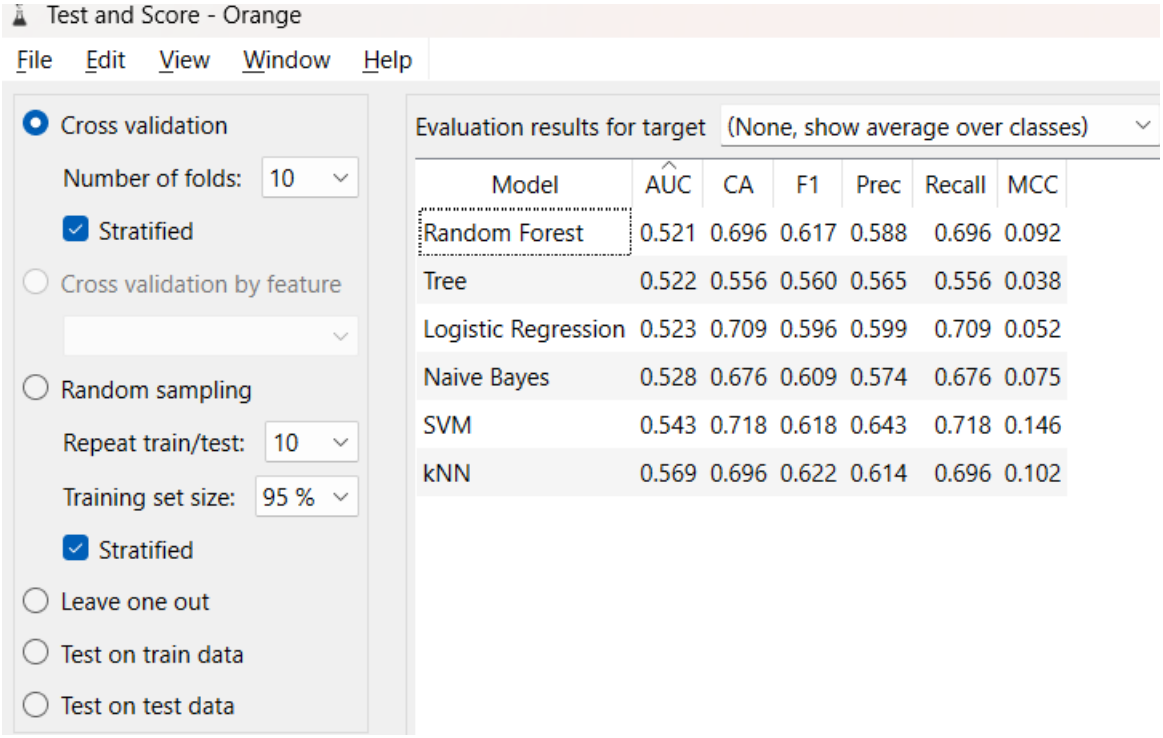
5. Memilih atribut yang digunakan sebagai feature dan class



C. Pembuatan Model



D. Evaluasi model



Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.521	0.696	0.617	0.588	0.696	0.092
Tree	0.522	0.556	0.560	0.565	0.556	0.038
Logistic Regression	0.523	0.709	0.596	0.599	0.709	0.052
Naive Bayes	0.528	0.676	0.609	0.574	0.676	0.075
SVM	0.543	0.718	0.618	0.643	0.718	0.146
kNN	0.569	0.696	0.622	0.614	0.696	0.102

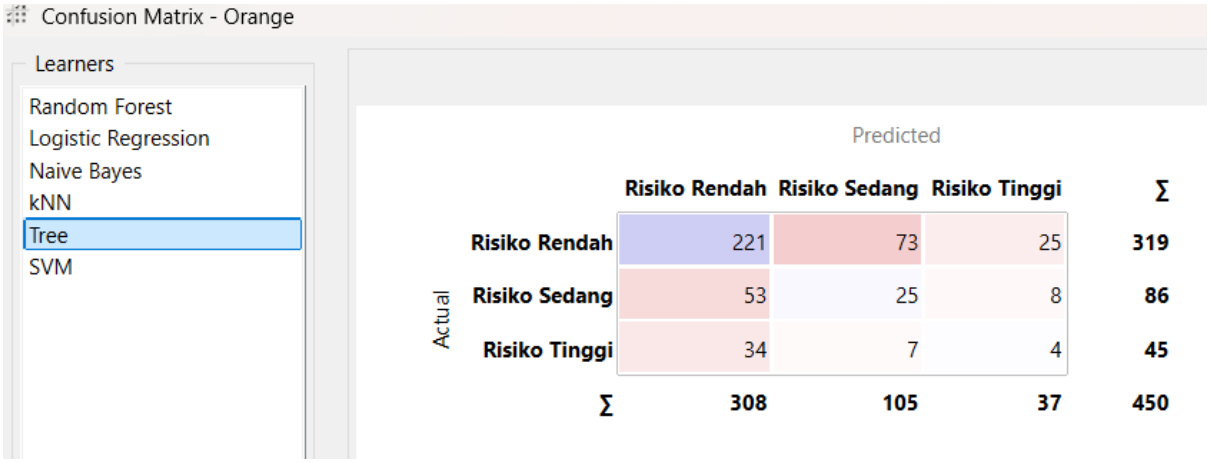
Berdasarkan evaluasi dengan test and score dan confusion matrix, ditentukan model terbaik adalah Tree karena nilai CA, F1 score, Precision, recall, dan MCC selalu paling atas (berarti Tree menunjukkan keakuratan, kemampuan diskriminasi antar target yang baik, dan keseimbangan antara precision dan recall), hanya AUC yang peringkat 2 di bawah random forest. Selain itu, dengan confusion matrix, Tree menunjukkan hasil yang paling mendekati actual.

E. Prediksi

Jml data pada masing masing kelas hasil prediksi = 6

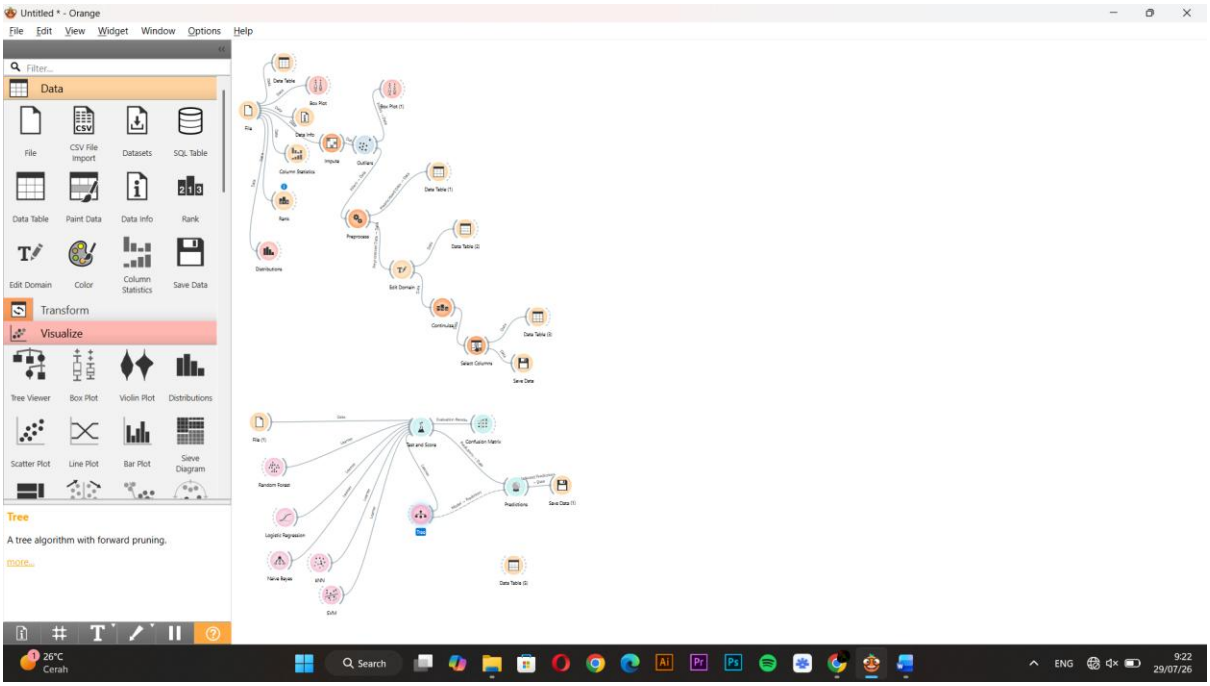
Pertanyaan

1. Sudah terjawab
2. Sudah terjawab
3. Preprocessing yang dilakukan adalah normalize, hal ini diperlukan untuk membuat konsistensi data dan mengurangi bias.
4. Sudah terjawab
5. Sudah terjawab
6. Interpretasi confusion matrix



Tree menunjukkan prediksi target yang mendekati data aktualnya

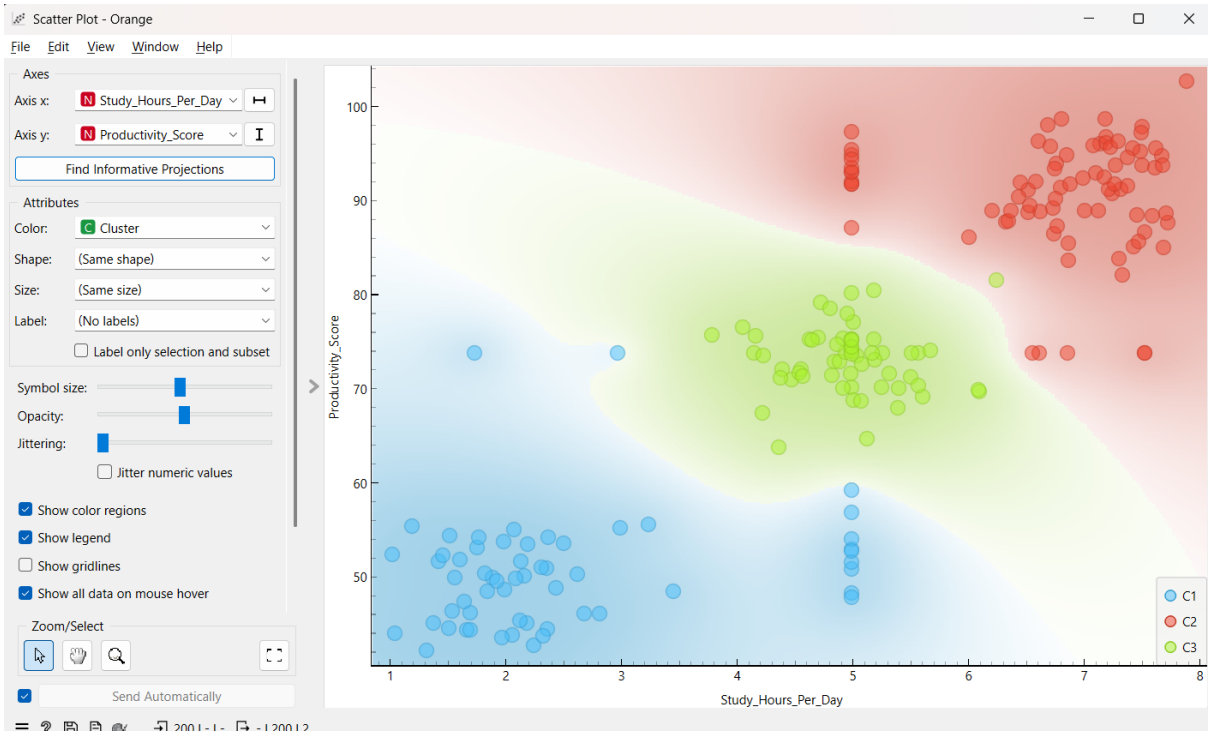
7. Hasil prediksi, jumlah data pada masing masing kelas adalah 6
8. atribut yang paling berpengaruh adalah tipe organisasi karena rank menunjukkan demikian
9. 3 insight
    a. Bahwa Tree menjadi model terbaik untuk melakukan prediksi. Instansi pemerintah dapat menggunakan Tree untuk memprediksi organisasi mana yang berpotensi risiko agar proses monitoring, evaluasi, dan pengambilan keputusan dapat dilakukan lebihi cepat, konsisten, dan berbasis data.
    b. Bahwa banyak organisasi yang risikonya rendah, tetapi masih ada beberapa yang sedang hingga tinggi. Hal ini berarti masih perlu diadakan evaluasi mendalam terkait pelaksanaan pengelolaan keuangan sektor publik di beberapa organisasi.
    c. Bahwa instansi pemerintah yang mau meningkatkan efektivitas pengelolaan keuangan perlu memperhatikan tipe_organisasi (yang berarti isi kegiatan organisasi itu seperti apa), dan wilayahnya (setiap wilayah akan berbeda keuangannya, kendala di wilayah tersebut, dsb)
10. Workflow orange



Keterangan: file data testing tidak terdapat atribut status_risiko sehingga tidak bisa digunakan

Soal 2 klasterisasi

Algoritma yang memberikan hasil clustering terbaik adalah k means (dengan jumlah cluster ditentukan adalah 3) karena dapat memisahkan data dengan baik (3 kelompok data saling mengumpul dan terpisah), hanya ada sedikit data yang masuk ke cluster data lain. Hal ini berarti kemampuan k means untuk mendiskriminasi data cukup baik



Workflow