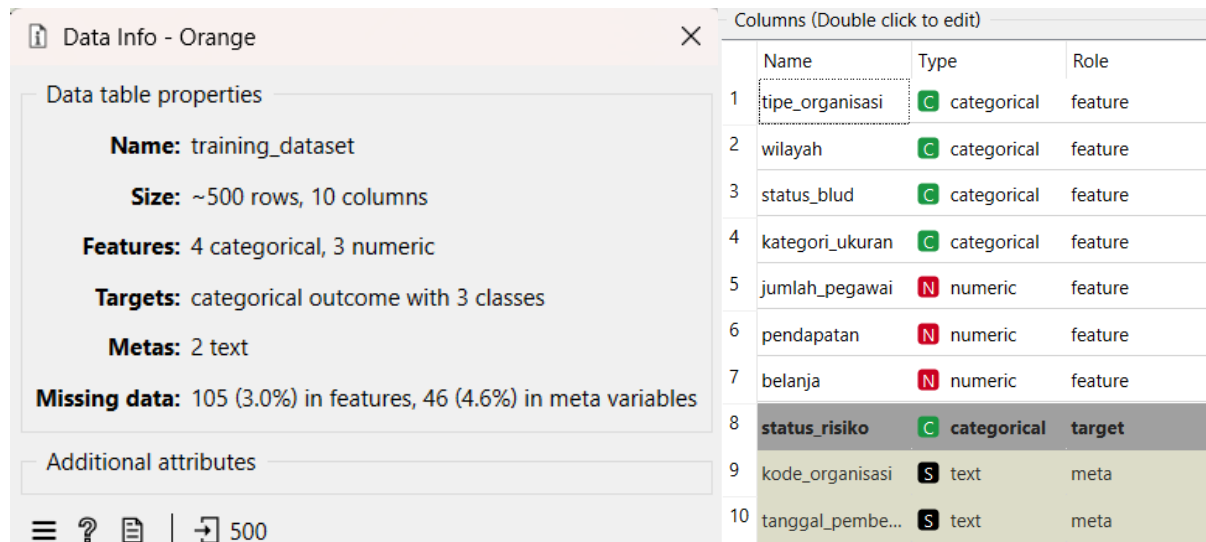


Pengerjaan ulang UTS ADKSP

Soal 1 Klasifikasi

Pertanyaan yang harus dijawab

1. Jelaskan karakteristik dataset yang digunakan (sertakan screen shoot)



- Jumlah record = 500
- Jumlah atribut/variabel/feature = 10
- Tipe kategorikal = tipe_organisasi, wilayah, status_blud, kategori_ukuran, status_risiko
- Tipe numeric = jumlah_pegawai, pendapatan, belanja
- 2 meta = kode_organisasi dan tanggal_pembe

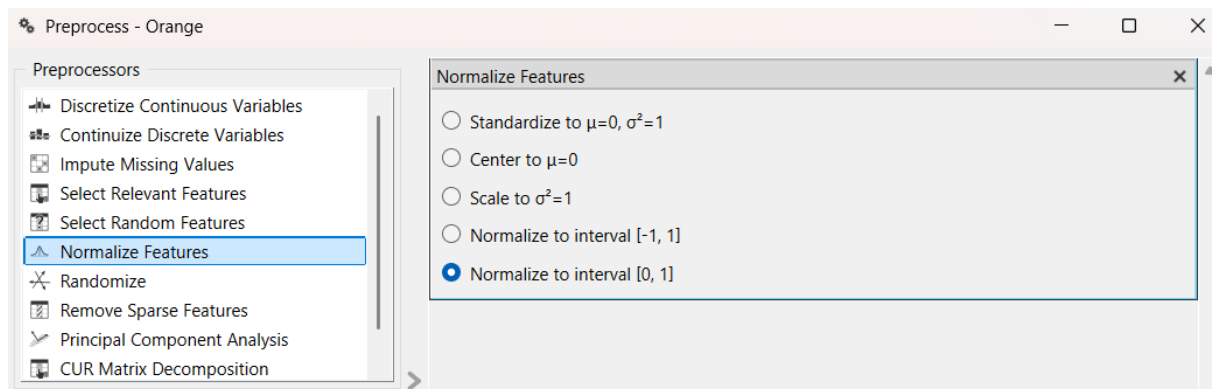
2. Apakah ditemukan missing value, class imbalance, maupun outlier? Jelaskan dan (sertakan screen shoot).



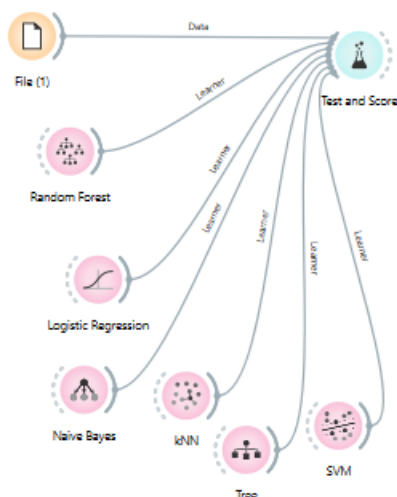
- Missing value = 105 (3%) in features, 46 (4,6%) in meta

- b) Class imbalance = dari widget Distributions, terlihat bahwa ada ketimpangan dalam data risiko rendah yang sangat banyak, diikuti risiko sedang jauh di bawah, dan risiko tinggi tidak jauh di bawah risiko sedang
- 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

3. Preprocessing apa saja yang Saudara lakukan? Mengapa preprocessing tersebut diperlukan? (sertakan screen shoot). Simpan hasil preprocessing dan beri nama file tersebut sebagai cleaned_dataset.csv.
- Preprocessing yang dilakukan adalah normalize, hal ini diperlukan untuk menyamakan skala nilai antar-atribut numerik agar variabel dengan rentang nilai yang berbeda tidak terlalu mendominasi proses analisis atau pemodelan.



4. Algoritma apa saja yang digunakan? (sertakan screen shoot).



5. Algoritma manakah yang memberikan performa terbaik? Sertakan tabel hasil Test and Score (sertakan screen shoot).

Test and Score - Orange

File Edit View Window Help

☒ Cross validation

Number of folds: 10

☒ Stratified

☐ Cross validation by feature

☐ Random sampling

Repeat train/test: 10

Training set size: 95 %

☒ Stratified

☐ Leave one out

☐ Test on train data

☐ Test on test data

Evaluation results for target (None, show average over classes)

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 SVM adalah pilihan terbaik secara keseluruhan jika mempertimbangkan kombinasi metrik utama, karena memiliki Accuracy tertinggi (71,8%), Precision tertinggi (64,3%), dan Recall tertinggi (71,8%), sementara F1-nya juga sangat tinggi (0,618).

6. Sertakan Confusion Matrix dari model terbaik beserta interpretasinya (sertakan screen shoot).

Confusion Matrix - Orange

Learners

Random Forest

Logistic Regression

Naive Bayes

kNN

Tree

SVM

Predicted

		Predicted			Σ
		Risiko Rendah	Risiko Sedang	Risiko Tinggi	
Actual	Risiko Rendah	316	3	0	319
	Risiko Sedang	79	7	0	86
	Risiko Tinggi	45	0	0	45
Σ		440	10	0	450

“Meskipun SVM menghasilkan accuracy sebesar 71,8% pada Test & Score, confusion matrix menunjukkan bahwa performa model belum merata pada setiap kelas. Model sangat baik dalam mengidentifikasi Risiko Rendah, tetapi masih lemah dalam mengidentifikasi Risiko Sedang dan Risiko Tinggi. Hal ini kemungkinan dipengaruhi oleh ketidakseimbangan jumlah data antar kelas.”

7. Hasil prediksi

Predictions - Orange

Show probabilities for: Classes known to the model

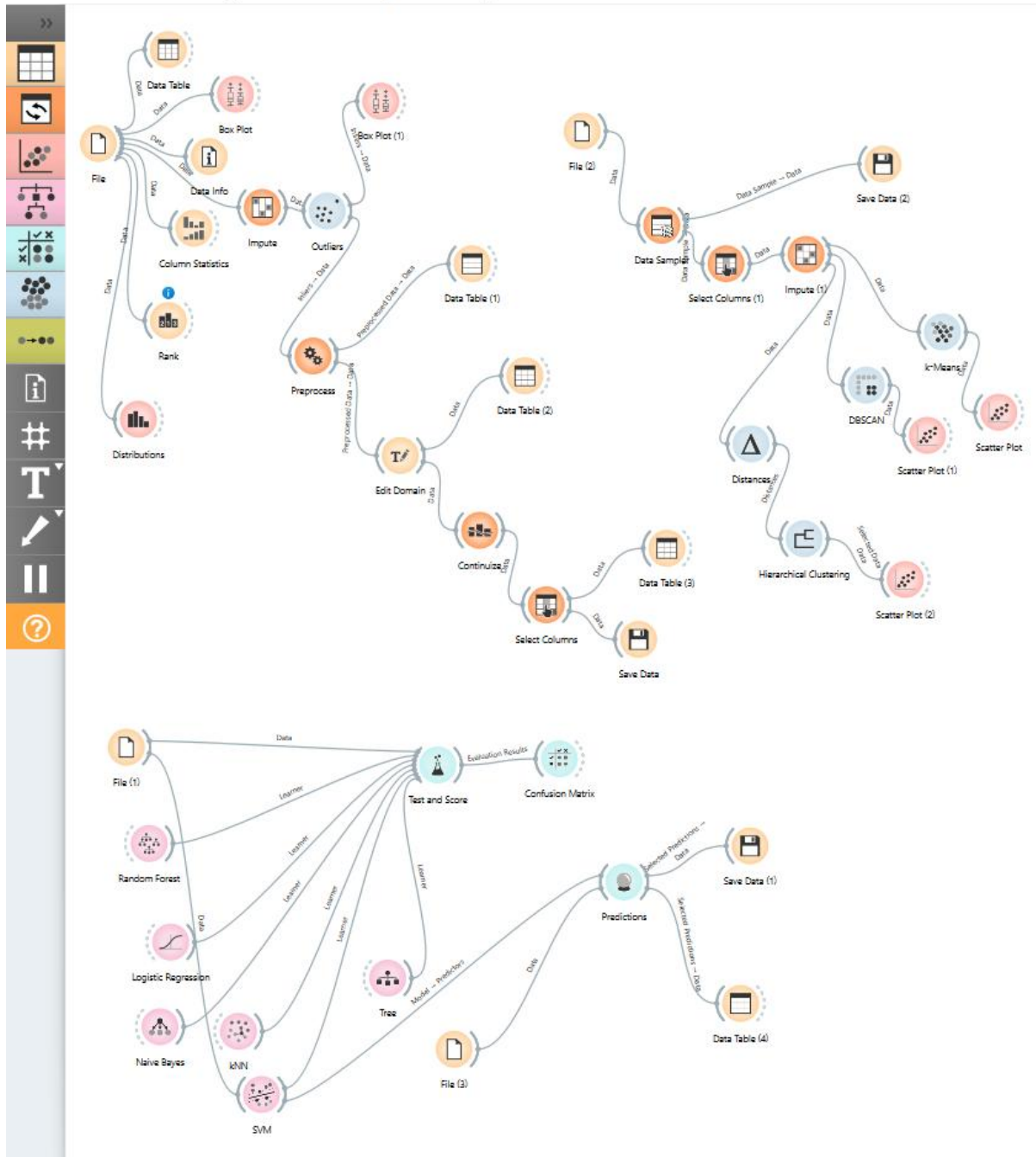
Restore Original Order

	SVM	yggal_pembentuk	kode_organisasi	tipe_organisasi	wilayah	status_blud	kategori_ukuran	jumlah_pegawai	pendapatan	belanja
1	0.75 : 0.16 : 0.09 → Risiko Rendah	2014-05-21	TEST-001	Dinas	Kota Medan	False	Kecil	142	9452028196651...	6954309808899...
2	0.75 : 0.16 : 0.09 → Risiko Rendah	2004-03-13	TEST-002	BUMD	Kota Yogyakarta	True	Menengah	53	3816379995169...	3454596008115...
3	0.75 : 0.16 : 0.09 → Risiko Rendah	2012-06-07	TEST-003	BLUD	Kota Semarang	False	Besar	699	2656056976456...	1679600844746...
4	0.75 : 0.16 : 0.09 → Risiko Rendah	2003-09-16	TEST-004	Rumah Sakit Da...	Kota Bandung	True	Besar	245	3646623954530...	2923834684735...
5	0.75 : 0.16 : 0.09 → Risiko Rendah	2006-10-22	TEST-005	BUMD	Kota Yogyakarta	True	Besar	148	5807960390530...	3583196548662...
6	0.75 : 0.16 : 0.09 → Risiko Rendah	2005-03-01	TEST-006	Rumah Sakit Da...	Kota Semarang	False	Kecil	650	1640487343152...	1416447543782...
7	0.75 : 0.16 : 0.09 → Risiko Rendah	2001-06-06	TEST-007	Badan	Kota Yogyakarta	False	Kecil	487	4180848144315...	3844916359158...
8	0.75 : 0.16 : 0.09 → Risiko Rendah	2009-08-08	TEST-008	BLUD	Kabupaten Ban...	False	Besar	67	3729614377694...	2496167372414...
9	0.75 : 0.16 : 0.09 → Risiko Rendah	2013-07-10	TEST-009	Sekolah Negeri	Kota Medan	False	Menengah	825	8336630418853...	5041103507182...
10	0.75 : 0.16 : 0.09 → Risiko Rendah	2017-09-16	TEST-010	Badan	Kota Semarang	False	Besar	670	7046063303328...	4449836727304...
11	0.75 : 0.16 : 0.09 → Risiko Rendah	2005-10-11	TEST-011	BLUD	Kabupaten Ban...	False	Sangat Besar	47	3927936627005...	2563154728510...
12	0.75 : 0.16 : 0.09 → Risiko Rendah	2000-01-21	TEST-012	Rumah Sakit Da...	Kabupaten Ban...	False	Sangat Besar	854	2886316309864...	2580009229251...
13	0.75 : 0.16 : 0.09 → Risiko Rendah	2020-05-08	TEST-013	Rumah Sakit Da...	Kabupaten Ban...	False	Kecil	151	3293986934334...	2871457496976...
14	0.75 : 0.16 : 0.09 → Risiko Rendah	2005-11-05	TEST-014	BLUD	Kabupaten Ban...	False	Kecil	247	2198811412087...	1734584635723...
15	0.75 : 0.16 : 0.09 → Risiko Rendah	2018-12-13	TEST-015	Dinas	Kabupaten Ban...	True	Sangat Besar	156	1026503954696...	971621920729...
16	0.75 : 0.16 : 0.09 → Risiko Rendah	2019-11-16	TEST-016	Kecamatan	Kota Semarang	False	Menengah	766	1111773641821...	7317951407349...
17	0.75 : 0.16 : 0.09 → Risiko Rendah	2023-10-27	TEST-017	Badan	Kabupaten Sle...	True	Menengah	695	1649405703763...	1064684226740...
18	0.75 : 0.16 : 0.09 → Risiko Rendah	2009-12-13	TEST-018	Badan	Kabupaten Ban...	True	Menengah	748	3311082614055...	2779842728172...
19	0.75 : 0.16 : 0.09 → Risiko Rendah	2022-04-07	TEST-019	Sekolah Negeri	Kota Bandung	True	Kecil	607	675093777510...	490871062887...
20	0.75 : 0.16 : 0.09 → Risiko Rendah	2015-08-12	TEST-020	BUMD	Kabupaten Ban...	False	Menengah	410	2491652035300...	1818434570015...
21	0.75 : 0.16 : 0.09 → Risiko Rendah	2011-04-07	TEST-021	Sekolah Negeri	Kabupaten Sle...	True	Besar	502	1005177291569...	8991839176108...
22	0.75 : 0.16 : 0.09 → Risiko Rendah	2001-09-26	TEST-022	Kelurahan	Kota Bandung	True	Kecil	242	3076643655455...	2704494896283...
23	0.75 : 0.16 : 0.09 → Risiko Rendah	2021-12-07	TEST-023	Dinas	Kabupaten Sle...	True	Besar	195	1797341639635...	1252135122474...
24	0.75 : 0.16 : 0.09 → Risiko Rendah	2020-02-01	TEST-024	Kelurahan	Kabupaten Gresik	True	Kecil	139	2067441520708...	1828275044681...
25	0.75 : 0.16 : 0.09 → Risiko Rendah	2008-10-07	TEST-025	BLUD	Kota Bandung	True	Sangat Besar	550	3896806476690...	3298671310720...
26	0.75 : 0.16 : 0.09 → Risiko Rendah	2002-02-22	TEST-026	Dinas	Kabupaten Sle...	True	Menengah	862	2155985448079...	1831331635324...
27	0.75 : 0.16 : 0.09 → Risiko Rendah	2005-02-14	TEST-027	Sekolah Negeri	Kabupaten Ban...	False	Sangat Besar	750	1968168062509...	1318324405623...
28	0.75 : 0.16 : 0.09 → Risiko Rendah	2015-12-09	TEST-028	Rumah Sakit Da...	Kota Bandung	False	Kecil	487	7269607281567...	5714590796732...
29	0.75 : 0.16 : 0.09 → Risiko Rendah	2008-12-03	TEST-029	Kelurahan	Kabupaten Sle...	False	Besar	717	1188041058960...	762928759328...

Kelas Hasil Prediksi	Jumlah
Risiko Rendah	30
Risiko Sedang	0
Risiko Tinggi	
Total	30

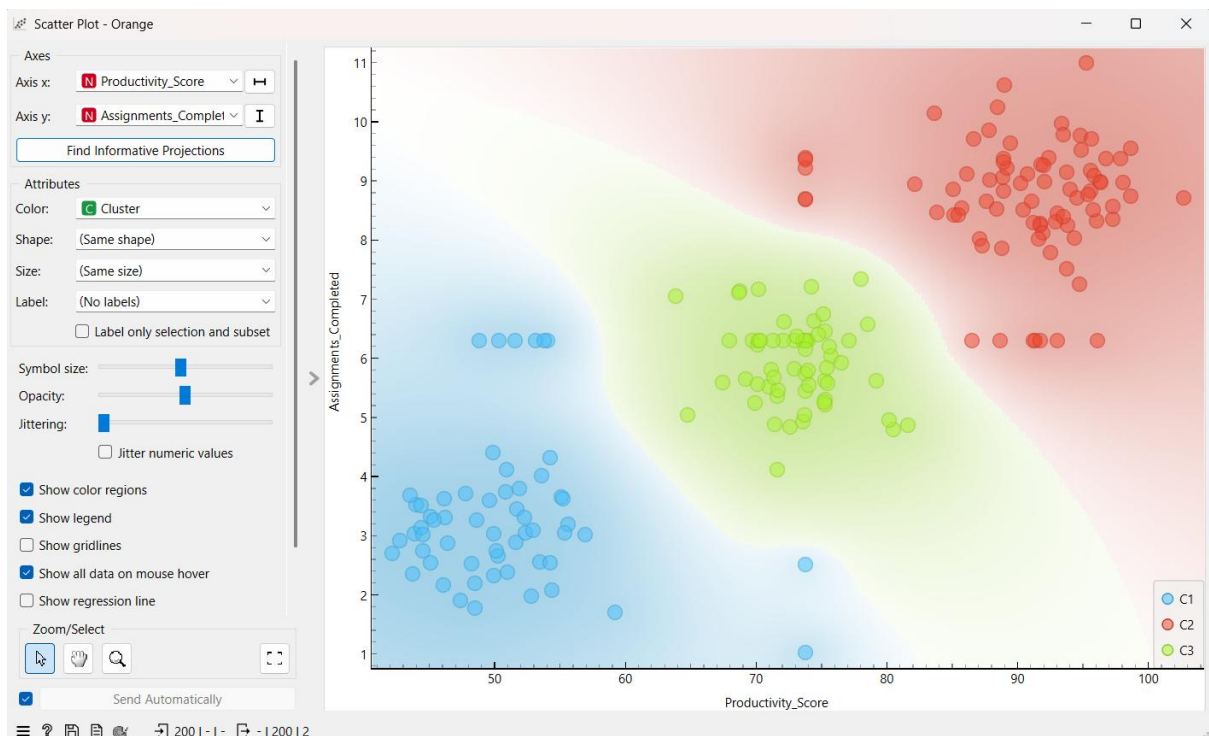
8. atribut yang paling berpengaruh adalah tipe organisasi karena rank menunjukkan demikian
9. 3 insight
 - a. Bahwa SVM menjadi model terbaik untuk melakukan prediksi. Instansi pemerintah dapat menggunakan SVM untuk memprediksi organisasi mana yang berpotensi risiko agar proses monitoring, evaluasi, dan pengambilan keputusan dapat dilakukan lebih 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. SS workflow

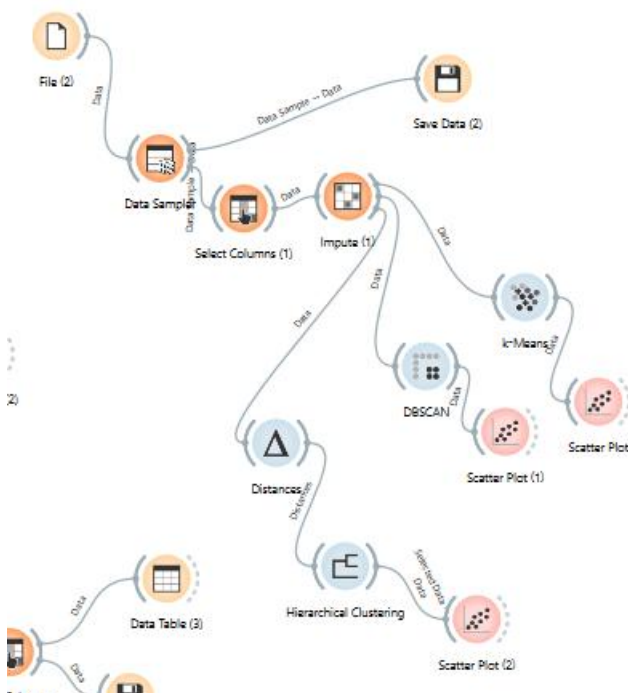


Soal 2 clustering

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



3 soal yang paling banyak dirubah:

1. Pertanyaan 5 pada soal 1 klasifikasi: mengubah dari data terbaik tree menjadi SVM
2. Pertanyaan 6 pada soal 1 klasifikasi: paragraf interpretasi confusion matrix untuk SVM
3. Pertanyaan 7 pada soal 1 klasifikasi: memperbaiki alur widget prediksi, mengubah jumlah data pada masing-masing kelas

- Apa yang salah/kurang dari jawaban sebelumnya?
 - Pemilihan model terbaik dari hasil test and score salah, prediksi tidak berhasil karena tidak ada variabel target
- Apa yang anda perbaiki?
 - Model tree diganti menjadi model SVM, sambungkan alur yang benar untuk prediksi
- Apa yang sekarang anda pahami?
 - Bahwa ketika memencet AUC, CA, F1, prec, recall, mcc, itu tidak memunculkan yang paling tinggi di atas.
 - Bahwa prediksi tidak butuh variabel target untuk data testingnya

Model	AUC	CA	F1	Prec	Recall	MCC
Model 1	0.800	0.600	0.600	0.555	0.600	0.600