A. Deskripsi Studi Kasus

Perguruan tinggi adalah tahap akhir yang bersifat opsional pada pendidikan formal. Tahap akhir ini juga merupakan salah satu cara yang bermanfaat untuk memajukan kesejahteraan umum dan mencerdaskan kehidupan bangsa. Pada studi kasus ini, minat dan kebutuhan siswa SMA/SMK akan perguruan tinggi berpengaruh terhadap keputusan untuk mengambil pendidikan ini. Berdasarkan penelitian yang dilakukan oleh Ari Putro Wicaksono yang berjudul "Faktor–Faktor yang Mempengaruhi Minat Melanjutkan Studi ke Perguruan Tinggi Siswa Kelas XII Jurusan Teknik Sepeda Motor di SMK Negeri 11 Malang Tahun Ajaran 2012/2013", terdapat dua faktor yang mempengaruhi minat seseorang untuk melanjutkan studi ke perguruan tinggi dapat berasal dari faktor internal (fisiologis, dan psikologis) ataupun *external* (sosial, dan non sosial).

Berdasarkan faktor-faktor ini disertai dengan data yang tepat, dapat dibangun sebuah sistem yang mampu memprediksi keputusan dari siswa. Dengan penerapan data mining, faktor-faktor yang mempengaruhi ketidakmampuan siswa melanjutkan ke perguruan tinggi dapat diperbaiki dan dicari solusinya. Dengan harapan meningkatkan jumlah sarjana dalam upaya mencerdaskan kehidupan bangsa.

A. Deskripsi Dataset

Dataset dirancang secara artifisial menggunakan *package* sklearn pada python menggunakan fungsi make_classification. Data yang dihasilkan umumnya bersifat numerik dan memerlukan proses klasterisasi untuk mendapatkan data *nominal*. Pada data ini didefinisikan 10 kolom fitur dan 1 kolom label. Selain itu juga dibuat 1000 baris data dengan sebaran 50% untuk label "False" dan 50% untuk label "True". Maksud dari "True" adalah siswa melanjutkan ke perguruan tinggi sedangkan "False" berarti tidak melanjutkan ke perguruan tinggi. Berikut adalah fitur-fitur yang ada pada dataset:

No	Nama	Tipe	Rentang Nilai data	Penjelasan Atribut
1	asal_sekolah	*Nominal*	SMA, SMK	Asal sekolah dari siswa
2	akreditasi_sekolah	*Ordinal*	A, B	Akreditasi sekolah dari siswa
3	jenis_kelamin	*Nominal*	Laki-Laki, Perempuan	Jenis kelamin siswa
4	minat	*Ordinal*	Sangat berminat,	Seberapa minat siswa

			Cukup berminat, Ragu-ragu, Kurang berminat, Tidak berminat	terhadap melanjutkan ke perguruan tinggi
5	asal_daerah	*Nominal*	Perkotaan, Pedesaan	Asal daerah siswa
6	umur_orang_tua	*Numeric*	40 - 65	Umur dari orang tua siswa
7	gaji	*Numeric*	1.000.000 - 10.000.000	Gaji dari orang tua siswa
8	luas_rumah	*Numeric*	20 - 120	Luas rumah orang tua siswa
9	rerata_nilai	*Numeric*	75 - 98	Rerata keseluruhan nilai siswa
10	orang_tua_sarjana	*Ordinal*	True, False	Apakah orang tua siswa lulusan sarjana
11	kuliah	*Ordinal*	True, False	Apakah siswa melanjutkan ke perguruan tinggi

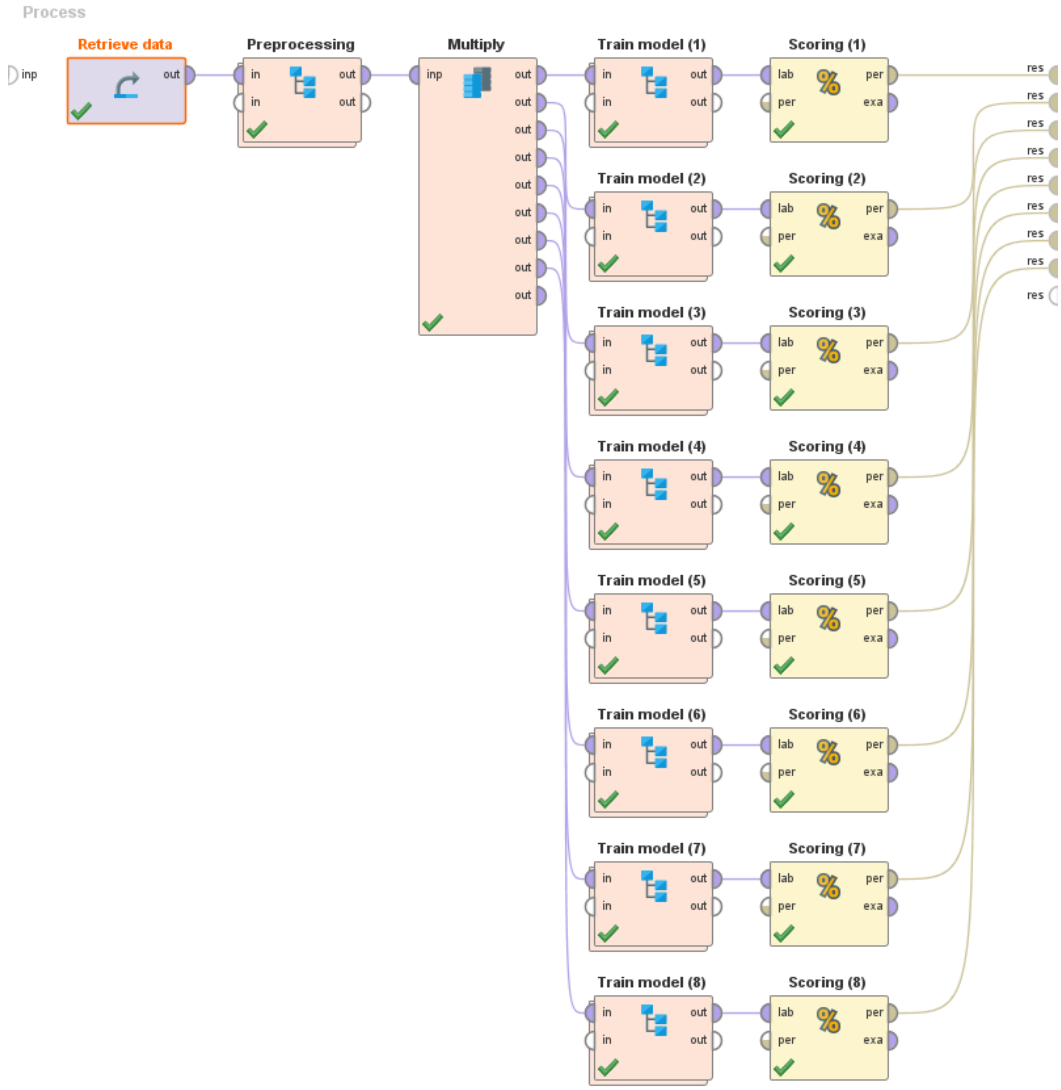
B. Metode Eksperimen

Pada penelitian ini akan terdapat beberapa skenario yang akan diuji untuk mendapatkan hasil akurasi yang berbeda. Skenario melibatkan 2 *classifier* yang berbeda dengan masing-masing 4 *hyperparameter* yang juga berbeda. Sehingga total kombinasi didapatkan 8 kombinasi skenario yang berbeda. Berikut adalah daftar delapan skenario yang akan diuji:

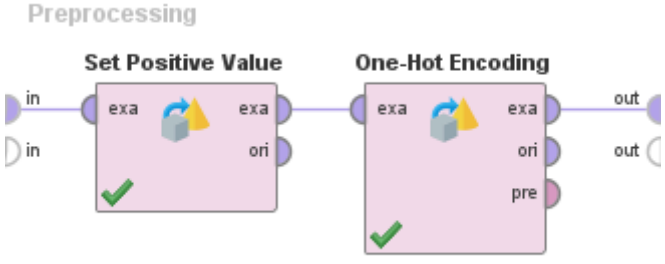
No.	Classifier	Hyperparameters
1	Decision Tree	criterion: gain_ratio, maximal depth: 10
2	Decision Tree	criterion: gain_ratio, maximal depth: 15
3	Decision Tree	criterion: information_gain, maximal depth: 10
4	Decision Tree	criterion: information_gain, maximal depth: 15
5	Support Vector Machine	kernel type: dot, C: 0
6	Support Vector Machine	kernel type: dot, C: 5.5
7	Support Vector Machine	kernel type: radial, C: 0
8	Support Vector Machine	kernel type: radial, C: 5.5

Berdasarkan skenario diatas, pengujian dilakukan secara simultan untuk mendapatkan hasil yang akurat dan presisi di setiap pengujian. Berikut adalah *pipeline* yang

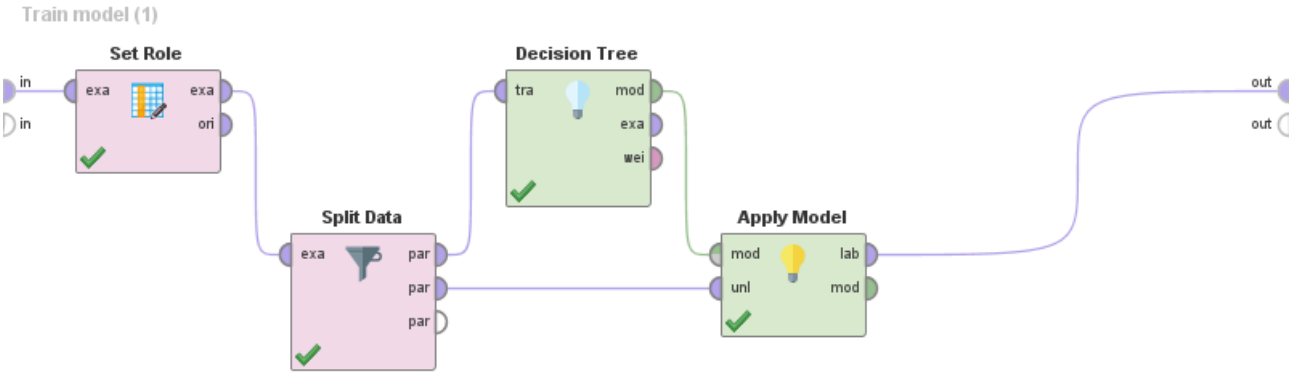
digunakan untuk menguji skenario tersebut dimana terdapat pemuatan data, *preprocessing*, *training model*, dan skoring.



Sub proses p*reprocessing* pada *pipeline* diatas melibatkan proses *Set Positive Value* untuk menentukan kelas positif dan *One-Hot Encoding* untuk data *nominal* sehingga semua data bertipe numerik.



Sedangkan subproses *train model* terdiri dari proses *Set Role* untuk menentukan kolom target label, *Split Data* untuk membagi *dataset* menjadi data train dan test dengan metode *stratified sampling dimana random seed* diatur ke 42, *classifier* dan *hyperparameter*-nya untuk melakukan pelatihan model dari data *training*, dan *Apply Model* untuk menguji model terhadap data *test*.



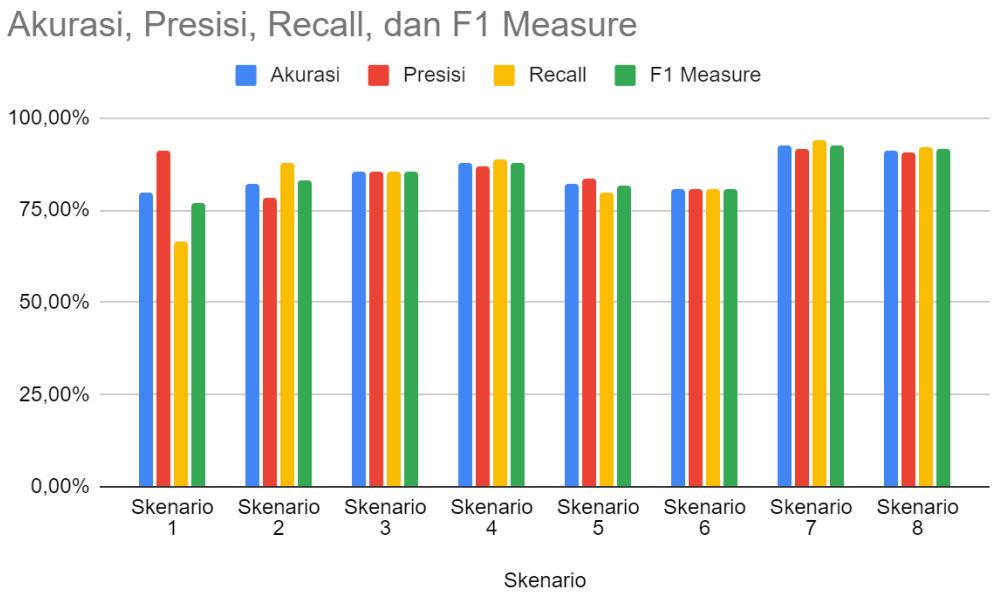
C. Hasil Eksperimen

Hasil delapan eksperimen menghasilkan delapan hasil yang berbeda. Pengukuran dilakukan menggunakan empat metrik yaitu akurasi, presisi, *recall*, dan *f-measure*. Berikut adalah tabel hasil eksperimen dari delapan skenario.

Skenario	Metrik			
	Akurasi	Presisi	Recall	F-Measure
Skenario 1	80,00%	90,91%	66,67%	76,92%
Skenario 2	82,00%	78,57%	88,00%	83,02%
Skenario 3	85,33%	85,33%	85,33%	85,33%
Skenario 4	87,67%	86,93%	88,67%	87,79%
Skenario 5	82,00%	83,33%	80,00%	81,63%
Skenario 6	80,67%	80,67%	80,67%	80,67%
Skenario 7	92,67%	91,56%	94,00%	92,76%
Skenario 8	91,33%	90,79%	92,00%	91,39%

Dapat dilihat bahwa rata-rata skenario yang menggunakan *classifier Support Vector Machine* (SVM) memiliki akurasi yang lebih tinggi dibanding *Decision Tree*. Sehingga secara umum metode menggunakan SVM memiliki performa lebih baik daripada *Decision Tree*. Selain itu, pada skenario 7 dan 8 memiliki performa yang lebih tinggi daripada skenario yang lain.

D. Kesimpulan



Berdasarkan diagram diatas dapat disimpulkan bahwa metode terbaik diperoleh pada skenario ke-7 dengan akurasi sebesar 92,67%, presisi sebesar 91,56%, recall sebesar 94,00%, dan f-measure sebesar 92,76%. *Classifier* yang digunakan adalah *Support Vector Machine* dengan *hyperparameter* jenis kernel radial dan nilai C sama dengan 0. Hal ini disebabkan oleh sifat data yang tidak linear. Sehingga *classifier* SVM dengan kernel radial yang mampu memetakan fitur ke dimensi yang lebih tinggi dengan wajar memiliki performa yang lebih baik.