Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 606−613
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2160

Komparasi Algoritma Klasifikasi Data Mining Menggunakan Optimize Selection untuk Peminatan Program Studi

Khaerul Anam, Bani Nurhakim*, Christina Juliane

Program Studi Sistem Informasi, STMIK LIKMI, Bandung, Indonesia
Email: [1]jodiust9@gmail.com, [2,*]baninurhakim@gmail.com, [3]christina.juliane@likmi.ac.id
Email Penulis Korespondensi: baninurhakim@gmail.com
Submitted:**21/08/2022**; Accepted:**07/09/2022**; Published**: 30/09/2022**

Abstrak− Pemilihan Program Studi menjadi tantangan tersendiri bagi calon mahasiswa. STMIK IKMI Cirebon menjadi provider KIP Kuliah dengan tiga pilihan program studi bagi calon mahasiswa. Permasalahan penelitian adalah belum tersedianya model peminataan siswa terhadap program studi sehingga perlu dilakukan sebuah peminatan program studi dengan menerapkan algoritma pada model klasifikasi. Algortima yang digunakan sebagai komparasi yaitu algoritma Decision Tree (C4.5), Naive Bayes, k-Nearest Neighbour dan Support Vector Machine. Model klasifikasi menerapkan operator Optimize Selection dengan mencari atribut dominan dalam pengaruhnya pada peminatan program studi mahasiswa. Terakhir, model perbandingan akan dilakukan uji parametrik t-test dalam rangka menguji signifikansi algoritma. Hasil dari pengujian akurasi algoritma didapatkan algoritma SVM memiliki akurasi terbaik dengan nilai 80,76%. Sedangkan algoritma dengan akurasi paling rendah adalah Naive Bayes dengan nilai 74,64%. Sedangkan dua algoritma lainnya memiliki tingkat akurasi berurutan yaitu 80,47% untuk Decision Tree dan 76,09% untuk k-NN. Hasil dari penelitian ini digunakan untuk klasifikasi minat program studi calon mahasiswa STMIK IKMI Cirebon yang bermanfaat untuk memprediksi minat terhadap program studi berdasarkan latar belakang calon mahasiswa.

Kata Kunci: Komparasi; Data Mining; Algoritma Klasifikasi; Optimize Selection; Peminatan Program Studi

Abstract−The selection of a study program is a unique opportunity for a student. STMIK IKMI Cirebon is now a KIP Kuliah provider, offering three study program. The research problem is the unavailability of a model of student interest in the study program, so it is necessary to carry out an interest in the study program by applying an algorithm to the classification model. The algorithm used as a comparison is the Decision Tree algorithm (C4.5), Naive Bayes, k-Nearest Neighbor and Support Vector Machine. The classification model applies the Optimize Selection operator by looking for the dominant attribute in its influence on the specialization of the student study program. Finally, the comparison model will be tested by parametric t-test in order to test the significance of the algorithm. The results of the algorithm accuracy test obtained that the SVM algorithm has the best accuracy with a value of 80.76%. While the algorithm with the lowest accuracy is Naive Bayes with a value of 74.64%. While the other two algorithms have a sequential accuracy rate of 80.47% for Decision Tree and 76.09% for k-NN. The results of this study are used to classify study preference for students in STMIK IKMI Cirebon which is useful for predicting study interest based on the background of students**.**

Keywords: Comparison; Data Mining; Classification Algorithm; Optimize Selection; Study Program Specialization

1. PENDAHULUAN

Salah satu tujuan selanjutnya bagi siswa yang telah lulus dari Sekolah Menengah Atas (SMA) dan sederajat adalah menempuh studi di perguruan tinggi. Pemilihan program studi yang akan diampu menjadi tantangan tersendiri bagi calon mahasiswa. Pemilihan ini dipengaruhi berbagai faktor seperti tingkat ekonomi, geografis, prestasi akademik, kompetitor dan lain sebagainya [1]. Pemilihan program studi ini sangat penting, karena dapat menimbulkan efek negatif jika salah dalam proses pengambilan keputusan terhadap program studi yang dipilih seperti ketidakoptimalan prestasi, penguasaan materi yang kurang, serta ketidakpuasan terhadap hasil studi mahasiswa [2].

STMIK IKMI Cirebon ditahun 2020 hingga saat ini menjadi salah satu kampus di daerah Kota Cirebon yang menjadi *provider* KIP Kuliah. Pengarahan bagi calon mahasiswa dalam peminatan program studi yang tepat adalah hal yang penting untuk dilakukan [3]. Namun disisi lain dapat pula terjadi dari masing-masing mahasiswa tidak tahu pilihan terhadap minat dan kemampuannya seperti kesulitan dalam menentukan program studi. Penentuan program studi yang dipilih dapat terjadi karena keinginan dari mahasiswa itu sendiri atau dapat pula terjadi karena ikut-ikutan teman. Untuk itu diperlukan preferensi dalam membantu mahasiswa untuk menghindari adanya dampak negatif yang tidak diinginkan serta untuk mencapai keberhasilan program KIP Kuliah itu sendiri.

Metode *data mining* dapat menjadi instrumen pembantu dalam melakukan prediksi dan klasifikasi dalam upaya penyelesaian kasus di lembaga pendidikan [3]. Klasifikasi merupakan model analisis data yang dapat memprediksi label kelas sampel. Banyak metode klasifikasi telah diusulkan dalam bidang pembelajaran mesin, sistem pakar dan statistik. *Software* repositori telah dikembangkan oleh penelitian pada tahun 1990 untuk memperoleh pemahaman terhadap data yang lebih mendalam. Pemanfaatan dengan metode *Decision Tree* (C4.5), *Naïve Bayes*, *Decision Stump*, *K-Nearest Neighbor* (k-NN), *Random Forest* yang membandingkan algoritma C4.5, *Random Forest*, dan *SimpleCART* dan menghasilkan pemeriksaan bahwa algoritma C4.5 bekerja lebih baik serta sangat layak untuk sistem. Selain itu, dapat juga meminimalkan untuk atribut dan numerik. Dari dua puluh metode klasifikasi, *Bayes Net*, *Random Forest*, *Naïve Bayes*, Regresi Logistik, dan Klasifikasi melalui Regresi adalah metode klasifikasi terbaik. Untuk dataset atribut numerik, Klasifikasi *NBTree*, Regresi, dan multiclass *Classifier* adalah metode pengelompokan terbaik. Untuk atribut langsung dari dataset *NB-Tree*, Pengelompokan melalui strategi Regresi dan metode *Bayes Net* sangat ideal. Dari kelima prinsip tersebut di atas metode klasifikasi dilihat dari teknik *PART* dan *Decision Tree* adalah yang terbaik.

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 606−613
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2160

Namun bagaimanapun juga, beberapa metode telah menunjukkan hasil terbaik di mana prosedur yang tepat dipilih untuk data yang tepat. Tidak ada pengklasifikasi khusus yang berkinerja terbaik untuk semua dataset [4].

Dalam rangka memilih algoritma klasifikasi dengan akurasi terbaik maka perlu dilakukan perbandingan antar algoritma. Studi mengenai perbandingan algoritma klasifikasi telah banyak dilakukan. Salah satunya oleh I. Hidayanti dkk [5] untuk menetukan konsentrasi Jurusan mahasiswa di Universitas Bina Darma Palembang. Menggunakan data alumni mahasiswa Teknik Informatika peneliti membandingkan Algoritma C4.5 dan *naive bayes* yang menghasilkan kinerja akurasi C4.5 sebesar 48,06 % dan *naive bayes* sebesar 42,79%. Hasil akurasi ini masih sangat rendah dan memerlukan perbaikan dalam penentuan label.

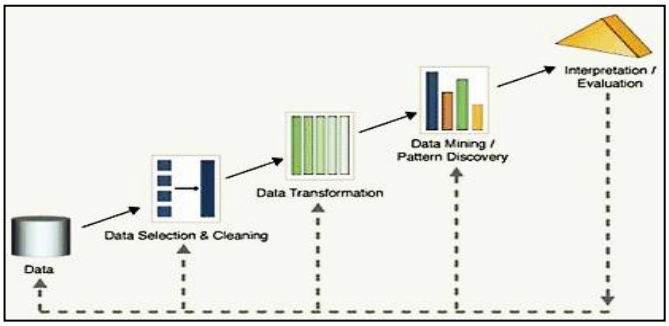
Selanjutnya dilakukan juga oleh O. Arifin dkk [6], yakni perbandingan performa dari metode *Support Vector Machine* dan *naive bayes* untuk klasifikasi minat jalur SMA. Perbandingan dengan menggunakan kernel anova dan parameter C sebesar 5.0 menghasilkan algoritma *SVM* yang relatif lebih unggul dibandingkan dengan *naive bayes*. Algoritma *SVM* menghasilkan akurasi sebesar 97,01 % sedangkan naive bayes sebesar 90,86 %.

Adapun penelitian yang akan dilakukan adalah dengan menggunakan perbandingan tingkat akurasi terhadap beberapa algoritma klasifikasi seperti *Decision Tree* (C4.5), *Naive Bayes*, *k-Nearest Neighbour* dan *support vector machine* pada peminatan mahasiswa terhadap program studi Teknik Informatika (S1), Manajemen Informatika (D3), dan Komputerisasi Akuntansi (D3) di STMIK IKMI Cirebon. Indikator yang digunakan untuk mengukur kinerja algoritma adalah menggunakan tabel *confusion matrix*. Selain itu, pada model klasifikasi diterapkan *operator optimize selection* dengan tujuan mencari atribut yang memiliki bobot tinggi dalam pengaruhnya pada peminatan program studi mahasiswa. Hasil dari penelitian ini digunakan untuk klasifikasi minat program studi calon mahasiswa STMIK IKMI Cirebon yang bermanfaat untuk memprediksi minat terhadap program studi berdasarkan latar belakang calon mahasiswa.

2. METODOLOGI PENELITIAN

2.1. *Knowledge Discovery in Database* (KDD)

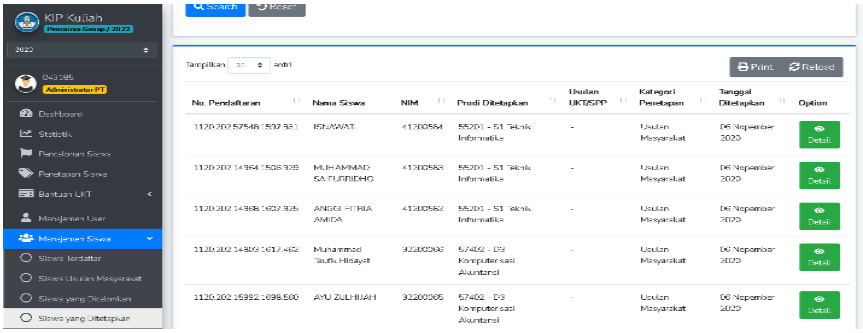
Data mining merupakan bagian dari *Knowledge Discovery Data* (KDD) yang merupakan proses ekstraksi informasi yang berguna, tidak diketahui sebelumnya dan tersembunyi dari data [7]. Tahapan Data Mining dalam metode KDD seperti digambarkan dalam gambar gambar 1.



Gambar 1. Tahapan KDD [7]

2.2 *Data*

Dataset dalam penelitian ini merupakan data sekunder yang didapatkan dari Sistem Informasi Pendaftaran Mahasiswa Baru STMIK IKMI Cirebon dan Sistem Informasi KIP Kuliah seperti pada gambar 2, yakni data penerimaan mahasiswa baru STMIK IKMI Cirebon dari tahun 2018 sampai tahun 2022. *Dataset* ini memiliki 43 atribut dengan total *record* 2051 data. Untuk pengujian algoritma *dataset* akan dibagi menjadi dua yakni data latih *(training)* diambil dari data pendaftaran dari tahun 2018 - 2021 dengan jumlah 1708 data dan data uji *(testing)* diambil dari data pendaftaran dari tahun 2022 berjumlah 343 data.



Gambar 2. Sistem Informasi KIP Kuliah

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 606−613
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2160

2.3 *Pre-Processing*

Pre-Processing merupakan proses yang mengontrol data mentah yang dapat dieksekusi lebih lanjut di dalam kumpulan data. Kekurangan, ketidakkuatan, dan kekeliruan data harus dikelola. Data secara teratur terfragmentasi, membingungkan, dua kali lipat dan gempar di dalam kenyataan. Ini mungkin tidak mengandung kualitas kualitas yang memadai atau hanya data secara total. Dengan adanya permasalahan pada data ini, dan variasi dalam nama atribut dapat juga memiliki masalah, yang dapat mempengaruhi sirkulasi umum terhadap data. Karena permasalahan data ini dapat menyebabkan prediksi yang salah, kami menggunakan cara *pre-processing* ini untuk menangani pertanyaan sebelum berhasil dan terjadi. Karena pendekatan ini sangat bagus, ini berarti bahwa indeks data benar-benar bersih selama *pre-processing*, dengan indeks data uji tidak memiliki nilai yang hilang. Tidak ada yang tidak dapat dipahami, redudansi atau interupsi. Selain itu, semua atribut yang ada memiliki nilai numerik *(floating)* dan tidak bersifat mutlak [8].

Dalam proses ini dilakukan pembersihan dataset dengan menghilangkan atribut-atribut yang kurang relevan serta memiliki banyak missing value. Hanya dipilih 10 atribut dengan interpretasi yang sesuai dan tidak memiliki missing value seperti pada tabel 2.

Tabel 1. Atribut dan Interpretasinya

Atribut	Interpretasi	Missing Value
NIK	Individu Mahasiswa (id)	0
Asal Sekolah	Latar Pendidikan	0
Kab/Kota	Asal Daerah	0
Tanggal Lahir	Usia	0
Jenis Kelamin	*Gender*	0
Pekerjaan Ayah	Latar Sosial	0
Penghasilan Ayah	Latar Ekonomi	0
Pekerjaan Ibu	Latar Sosial	0
Penghasilan Ibu	Latar Ekonomi	0
Program Studi	Minat Studi (label)	0

2.4 *Data Transformation*

Transformasi data merupakan proses pengubahan data dan penggabungan data ke dalam format tertentu. Hal ini karena proses data mining membutuhkan format data khusus sebelum diaplikasikan. Pada proses ini, dilakukan transformasi pada atribut "asal sekolah" dan "tanggal lahir" agar lebih memudahkan dalam pembacan pola *(pattern)* yang dihasilkan dari proses *data mining*.

Data asal sekolah mahasiswa dikategorikan menjadi 6 (enam) jenis sekolah yaitu, SMAN yang mewakili SMA Negeri, SMAS mewakili SMA Swasta, SMKN mewakili SMK Negeri, SMKS mewakili SMK swasta, MAN mewakili MA Negeri, MAS mewakili MA Swasta, dan PKBM mewakili Paket C. Sedangkan untuk atribut tanggal lahir ditransformasikan menjadi usia/umur dari mahasiswa pada saat melakukan pendaftaran kuliah. Tabel transformasi seperti pada tabel 2.

Tabel 2. Data Trassformation

Atribut	Transformasi	Isi data
Asal Sekolah	Jenis Sekolah	SMAN, SMAS, SMKN, SMKS, MAN, MAS, PKBM
Tanggal Lahir	Umur/Usia	Usia Mahasiswa

2.5 Decision Tree (C4.5)

Ketepatan dan kekuatan yang tinggi dari klasifikasi yaitu sebuah keunggulan dari algoritma *decision tree*. Kerugiannya yaitu bahwa pohon keputusan secara efektif dipengaruhi oleh contoh dan dalam pohon keputusan sub pilihan keputusan dapat diulang beberapa kali. Masalah *over-stitching* dapat diselesaikan dengan inovasi Teknik *cutting* dan *k-fold cross validation*. Dengan melihat cabang yang banyak dapat dihilangkan terlebih dahulu melalui pemangkasan, instalasi secara berlebihan juga dapat dihentikan. Isu berikutnya menyangkut tahap *pra-filter*, yang dapat dipakai untuk membuang sebagian dari kemampuan yang tidak relevan dari tahap *pra-pemrosesan* data. Ini memungkinkan terjadi pohon keputusan menjadi lebih sederhana dan menjaga jarak strategis dari masalah pohon keputusan yang salah. Pohon keputusan merupakan salah satu model klasifikasi yang paling banyak dikenal dalam penulisan [8].

2.6 *Naive Bayes*

Naive Bayes merupakan *classifier* untuk prediksi anggota kelas, sebagai ketentuan suatu tuple pada kelas yang dibuat. Klasifikasi *Naive Bayes* memberikan asumsi terhadap pengaruh atribut yang dimiliki pada kelas yang diberikan tidak dapat bergantung terhadap atribut nilai lainnya. Pendapat tersebut merupakan suatu kondisional dari kelas yang dibuat, agar terbentuk secara sederhana dengan perhitungan yang dilibatkan [9].

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 606−613
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2160

$$P(H|X) = \left(\frac{P(X|H)P(H)}{P(H)}\right) \quad (1)$$

2.7 *K-Nearest Neighbor*

Cara kerja Metode *K-NN* adalah dengan menghitung jarak terdekat antara data untuk selanjutnya dievaluasi dengan K tetangga *(neighbor)* terdekat dalam data latih *(data training)*. Sementara itu data training ditampilkan ke ruang yang berdimensi banyak dimana masing- masing dimensi menjelaskan fitur dari data. Adapun tahap-tahap untuk menghitung *K-NN* adalah sebagai berikut [10] :

a. Ambil nilai K secara acak
b. Menghitung jarak antar data dengan persamaan (2)

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (2)$$

Keterangan : X = sampel data Y = data uji D = Jarak

c. Mengurutkan hasil pengukuran jarak dan menetapkan data neighbor terdekat menurut jarak terdekat terhadap -K.
d. Menggunakan dominasi dari kelas tetangga terdekat sebagai angka prediksi data baru

Adapun untuk menghitung nilai prediksi *k-NN* digunakan persamaan (3)

$$Y = \frac{1}{K}\sum_{i=1}^{k} yi \quad (3)$$

Dimana: Y = Perkiraan K= jumlah tetanga terdekat Yi = output tetangga terdekat

2.8 *Support Vector Machine (SVM)*

SVM merupakan metode yang dapat menyelesaikan masalah secara linear maupun non-linear. Dalam memecahkan masalah non-linier di *SVM*, konsep kernel digunakan di ruang kerja berdimensi tinggi dengan mencari *hyperplane* untuk memaksimalkan margin antar kelas data. Dalam mengklasifikasikan data menggunakan metode *SVM* digunakan fungsi kernel K (xi, xd). Kernel yang digunakan dalam penelitian ini adalah sebagai berikut [11]:

$$\mathrm{K}(\mathrm{x}_i, \mathrm{x}_d) = \left(x_j^T X_j + \mathrm{C}\right)^{\mathrm{d}}, \gamma > 0 \quad (4)$$

Pengolahan data latih menggunakan algoritma pembelajaran sekuensial karena merupakan algoritma yang sederhana yang tidak memakan banyak waktu, dengan tahapan perhitungan sebagai berikut:

a. Tentukan parameter yang dibutuhkan oleh *SVM*
b. Hitung matriks *Hessian* yang diperoleh dari perkalian antara polinomial dan kernel y, yang merupakan vektor 1 dan -1. Persamaan matriks *Hessian* adalah:

$$d_{ij} = y_i y_i (K(x_i, y_i) + \lambda^2 \quad (5)$$

c. Hitung iterasi i ke j dengan rumus sebagai berikut :
   1. $E_i = \sum_j^i a_j \, D_{ij}$
   2. $\delta\alpha_i = \min\left(\max(\gamma(1 - E_i), a_i), C - a_i\right)$
   3. $\alpha_i = \alpha_i + \delta\alpha_i$
d. Lakukan 3 langkah diatas sampai didapatkan batas maksimum iterasi.
e. Proses pembelajaran sekuensial akan mendapatkan nilai dari *support vector* (SV), dimana SV = αi> SV *threshold*. Setelah itu perlu dilakukan perhitungan nilai bias b seperti pada persamaan 6.

$$b = -\frac{1}{2}\left(\sum_{i=0}^{N} \alpha_i \gamma_i K(x_i, x^-) + \sum_{i=0}^{N} \alpha_i \gamma_i K(x_i, x^+)\right) \quad (6)$$

Untuk mengetahui hasil klasifikasi komentar pada kelas sentimen tertentu, fungsi f(x) proses perhitungan dilakukan. Jika hasil dari fungsi ini negatif, maka polaritas sentimen dari komentar tersebut diidentifikasi sebagai sentimen negatif. Jika nilai fungsinya positif, maka polaritas sentimen dari komentar tersebut diidentifikasi sebagai sentimen positif. Fungsi f(x) diperoleh dari persamaan berikut:

$$f(x) = \sum_{i=0}^{N} \alpha_i \gamma_i K(x_i, x) + b \quad (7)$$

2.9 *Confusion Matrix*

Confusion matrix sebagai indikator analisis performa *classifier* dalam mengidentifikasi tupel dari kelas yang berbeda. Pada literatur berbeda *confusion matrix* juga dikenal dengan istilah *True positive* atau *tuple positif* dengan label benar, sedangkan *True negative* adalah *tuple negatif* dengan label benar. Ada juga *false positive* yang merupakan *tuple negatif* dengan label salah, dan *false negative* adalah *tuple positif* dengan label salah [12]. Tabel *confusion matrix* terlihat seperti pada tabel 3 [13] :

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 606−613
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2160

Tabel 3. Confusion Matrix

Klasifikasi benar	Klasifikasi	
	Positif (+)	Negatif (-)
Positif Benar	True Positif (TP)	False Negatif (FN)
Negatif Benar	False Positif (FP)	True Negatif (TN)

$$Akurasi = \frac{(TP+TN)}{TP+TN+FP+FN}\ x\ 100\% \quad (8)$$

$$Presisi = \frac{(TP+TN)}{TP+FP}\ x\ 100\% \quad (9)$$

$$Recall = \frac{TP}{TP+FN}\ x\ 100\% \quad (10)$$

2.10 *Optimize Selection*

Operator rapidminer ini digunakan untuk memilih atribut yang paling relevan dari *ExampleSet* yang diberikan. Terdapat dua algoritma pemilihan fitur yaitu, *forward selection* dan *backward elimination* yang dapat digunakan untuk pemilihan atribut paling berpengaruh dari dataset [14].

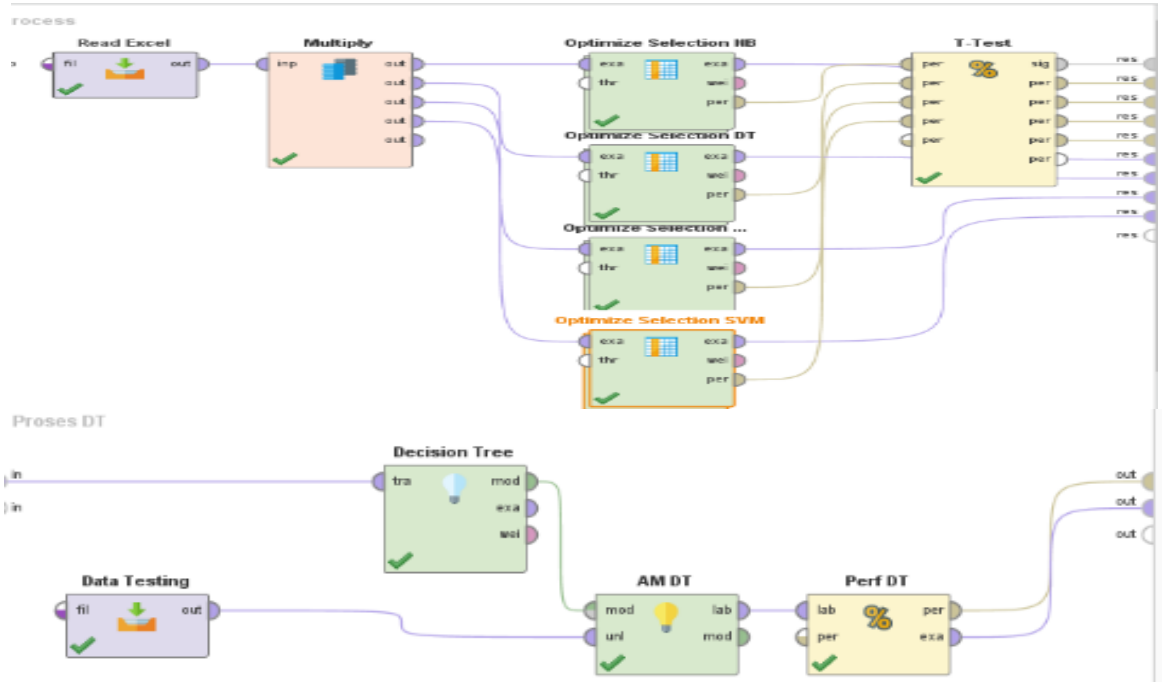
2.11 *T-Test*

T-test adalah membandingkan hubungan antara dua variabel yaitu variabel respon dan variabel *predictor*. *T-test* sample bepasangan *(paired-sample t-test)* digunakan untuk mengukur dari perbandingan selisih dua rata-rata dari dua sampel yang berpasangan dengan asumsi bahwa data yang ada itu terdistribusi normal [4]. Uji-t berpasangan adalah pengujian hipotesis nol bahwa perbedaan antara dua tanggapan yang diukur pada unit statistik yang sama memiliki nilai rata-rata nol. Sebagai contoh untuk mengukur ukuran tumor pasien kanker sebelum dan sesudah perawatan. Jika pengobatan efektif, diharapkan ukuran tumor untuk banyak pasien menjadi lebih kecil setelah pengobatan. Ini sering disebut sebagai uji-t berpasangan atau pengukuran berulang [15].

3. HASIL DAN PEMBAHASAN

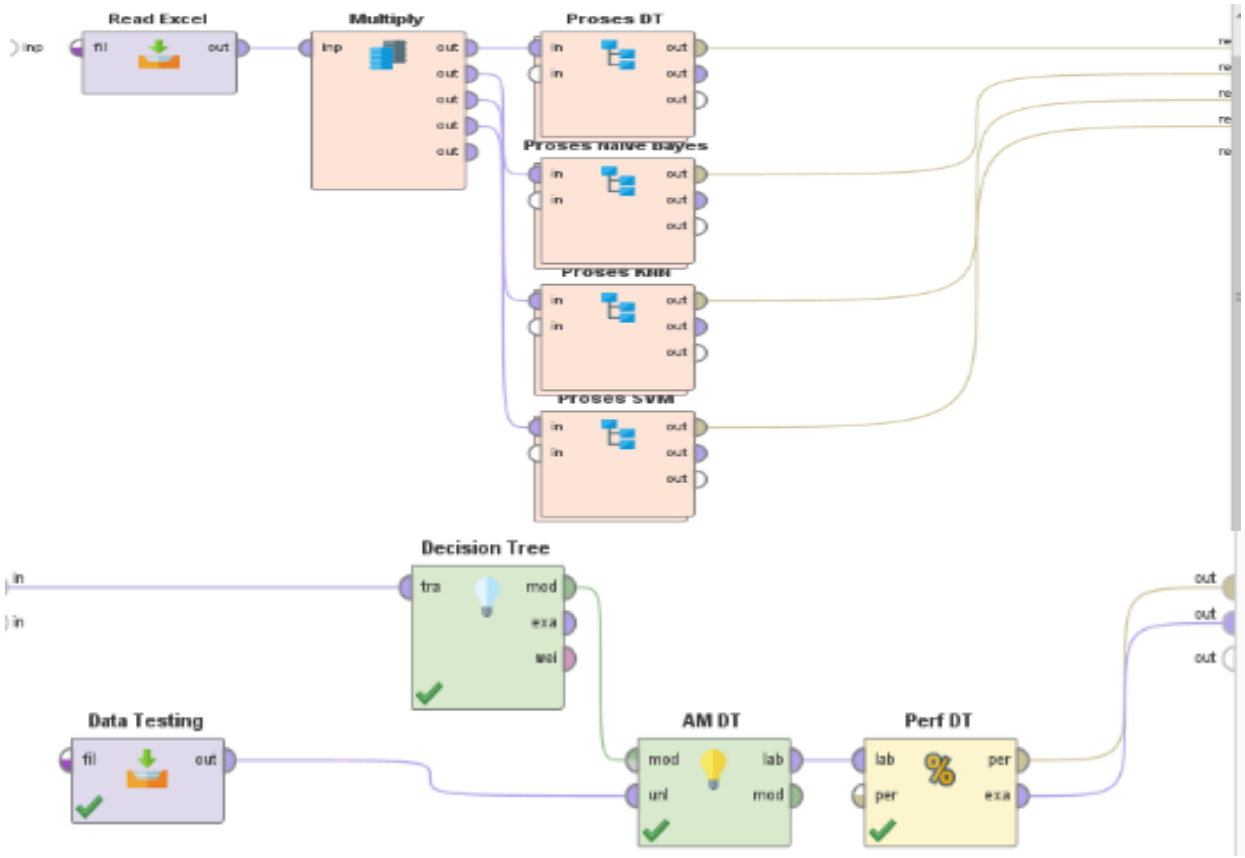
3.1 *Model Data Mining*

Pemodelan *data mining* komparasi dilakukan pada aplikasi rapidminer version 9.10. Dilakukan dua pemodelan komparasi dimana satu model menggunakan operator *Optimize Selection* seperti pada gambar 3 dan model lainnya tidak menggunakan operator *Optimize Selection* seperti gambar 4. Dataset sebagai *data training* dibaca melalui operator read excel untuk kemudian dibagikan ke model-model *subproses* algoritma menggunakan *operator multiply*. Kemudian didalam *subproses data training* akan disambungkan ke operator algoritma sedangkan *data testing* akan disambungkan ke operator *apply model*. Terakhir diterapkan operator *performance* untuk mengukur kinerja dari model klasifikasi dan operator *t-test* untuk mengukur signifikansi antar algoritma.



Gambar 3. Model Komparasi Menggunakan *Optimize Selection*

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 606−613
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2160

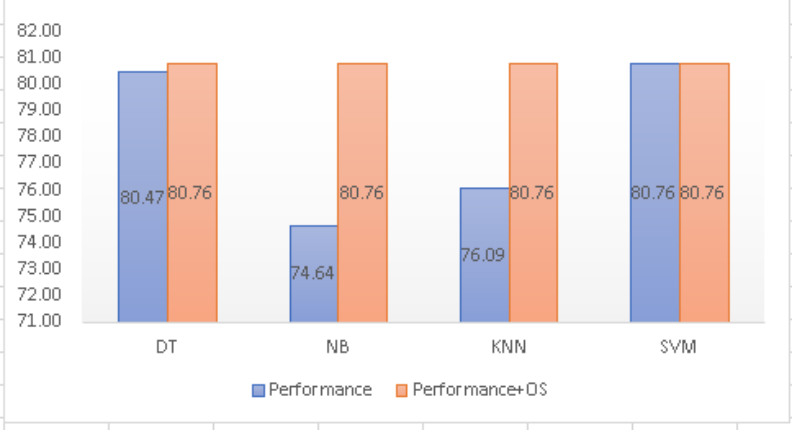


Gambar 4. Model Komparasi Tanpa *Optimize Selection*

Dari hasil eksperimen pada kedua model didapatkan hasil akurasi algoritma *Decision Tree* adalah 80,47% pada model tanpa *Optimize Selection* dan meningkat 80,76% pada model menggunakan *Optimize Selection*. Pada algoritma *Naive Bayes* juga yang mengalami peningkatan performa yang signifikan dimana tanpa *Optimize Selection* memiliki performa 74,64% dan dengan *Optimize Selection* menjadi 80,76% sama halnya pada algoritma k-NN perfroma juga meningkatkan dari 76,09% menjadi 80,76%. Adapun pada *SVM* akurasi tidak mengalami peningkatan dengan performa 80,76% pada semua model seperti pada gambar 5. Pada hasil pengujian bobot atribut didapatkan atribut dominan dalam proses klasifikasi berdasarkan empat algoritma yang diujikan. Pada algoritma *Decision Tree*, *Naive Bayes*, dan *SVM* adalah atribut "Jenis Sekolah". Sedangkan pada algoritma k-NN yang menjadi atribut dominan adalah "Umur" seperti pada tabel 4.

Tabel 4. Akurasi Algoritma dengan dan tanpa Optimze Selection

Algoritma	Performance (%)	Performance +OS (%)	Atribut Dominan
Decision Tree	80.47	80.76	Jenis Sekolah
Naive Bayes	74.64	80.76	Jenis Sekolah
k-NN	76.09	80.76	Umur
SVM	**80.76**	80.76	Jenis Sekolah



Gambar 5. Grafik Komparasi Algoritma dengan dan tanpa Optimze Selection

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 606−613
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2160

Dari hasil perbandingan pengujian akurasi algoritma didapatkan algoritma *SVM* yang memiliki akurasi terbaik dalam mengklasifikasikan minat studi pada penerimaan mahasiswa baru di STMIK IKMI Cirebon dengan nilai 80,76%. Sedangkan algoritma yang memiliki akurasi paling rendah adalah *Naive Bayes* dengan nilai 74,64%. Sedangkan dua algoritma lainnya memiliki tingkat akurasi berurutan yaitu 80,47% untuk *Decision Tree* dan 76,09% untuk k-NN. Hasil akurasi untuk algoritma *Decision Tree*, *Naive Bayes*, dan k-NN mengalami peningkatan dengan menerapkan Operator *Optimize Selection* menjadi 80,76%. Berbeda dengan *SVM* yang memiliki niai akurasi yang tetap yakni 80,76% pada setiap model pengujian.

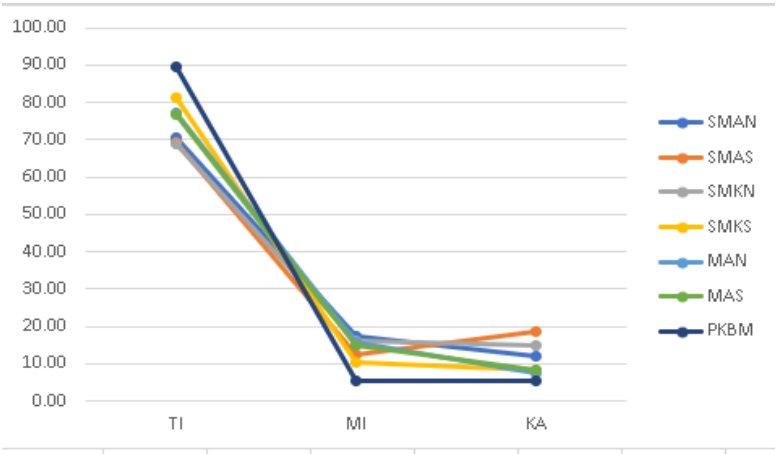
Pada hasil pengujian operator *t-test*, didapat hasil bahwasannya tidak ada nilai perbandingan yang lebih kecil dari nilai alpha 0.050. Hal ini diartikan bahwasannya keempat algoritma yang dikomparasi memiliki kinerja yang sama-sama baik dan tidak memiliki nilai signifikan satu sama lain seperti ditunjukan pada gambar 6.

A	B	C	D	E
	0.746 +/- ?	0.737 +/- ?	0.737 +/- ?	0.737 +/- ?
0.746 +/- ?		1.000	1.000	1.000
0.737 +/- ?			1.000	1.000
0.737 +/- ?				1.000
0.737 +/- ?				

Gambar 6. **Hasil uji beda *t-test***

3.2 *Analisa Atribut Dominan*

Atribut “Jenis Sekolah” sebagai atribut dominan dapat digunakan untuk membuat grafik probabilitas jenis sekolah asal mahasiswa terhadap minat studi seperti pada gambar 7. Terlihat bahwa dari setiap jenis sekolah mahasiswa memiliki probabilitas tinggi terhadap program studi Teknik informatika dengan probabilitas tertinggi dari jenis sekolah PKBM dengan angka 89,47%. Adapun pada program studi Manajemen Informatika probabilitas tertinggi dari jenis sekolah SMAN dengan angka 17,16% dan pada program studi Komputerisasi Akuntansi probabilitas tertinggi dari jenis sekolah SMAS dengan angka 18,46%.



Gambar 7. Grafik Probabilitas Jenis Sekolah Terhadap Minat Studi

Hasil pola minat studi mahasiswa pada penelitian ini sekaligus menjadi gap dengan penelitian sebelumnya [5][6][16], yakni pada penerapan operator *Optimize Selection*. Hasil pencarian atribut dominan ini digunakan untuk membuat pola minat studi mahasiswa baru di STMIK IKMI Cirebon. Pola yang dihasilkan ini dapat menjadi instrumen pembantu dalam melakukan kebijakan-kebijakan seperti promosi dan konseling bagi mahasiswa baru yang akan melanjutkan studi di STMIK IKMI Cirebon pada masa mendatang.

4. KESIMPULAN

Hasil eksperimen komparasi algoritma klasifikasi menggunakan *optimize selection* terhadap *dataset* minat Program Studi mahasiswa di STMIK IKMI Cirebon menghasilkan nilai akurasi dari algoritma *Decision Tree* sebesar 80,47%, *Naive Bayes* 74,64%, *k-NN* 76,09% dan *SVM* 80,76%. Nilai akurasi ini meningkat ketika diterapkan operator *Optimize Selection* terutama pada algoritma *Naive Bayes* yang meningkat dari 74,64% menjadi 80,76% dan k-NN dari 76,09% menjadi 80,76%. Pada hasil pengujian bobot atribut didapatkan atribut dominan dalam proses klasifikasi *Decision Tree*, *Naive Bayes*, dan *SVM* adalah atribut “Jenis Sekolah”. Sedangkan pada algoritma k-NN yang menjadi atribut dominan adalah atribut “Umur”. Atribut dominan ini digunakan untuk membuat grafik probabilitas jenis sekolah asal

Building of Informatics, Technology and Science (BITS)
Volume 4, No 2, September 2022 Page: 606−613
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v4i2.2160

mahasiswa terhadap minat studi yang dapat menjadi instrumen pembantu dalam melakukan kebijakan-kebijakan seperti promosi dan konseling bagi mahasiswa baru yang akan melanjutkan studi di STMIK IKMI Cirebon pada masa mendatang.

REFERENCES

[1] A. R. Pratama, Rio Rizki Aryanto, and Lizda Iswari, "Studi Komparasi Model Klasifikasi Berbasis Pembelajaran Mesin untuk Sistem Rekomendasi Program Studi," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 5, pp. 853–862, 2021, doi: 10.29207/resti.v5i5.3392.

[2] U. Enri, "PENERAPAN ALGORITMA C4.5 DALAM PEMILIHAN PROGRAM STUDI FAKULTAS ILMU KOMPUTER (Studi Kasus Sekolah Menengah Atas Negeri 1 Tambun Utara)," *J. Rekayasa Inf.*, vol. 7, no. 1, pp. 1–7, 2018.

[3] T. B. Sasongko, "Komparasi dan Analisis Kinerja Model Algoritma SVM dan PSO-SVM (Studi Kasus Klasifikasi Jalur Minat SMA)," *J. Tek. Inform. dan Sist. Inf.*, vol. 2, no. 2, pp. 244–253, 2016, doi: 10.28932/jutisi.v2i2.476.

[4] R. Annisa, "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung," *J. Tek. Inform. Kaputama*, vol. 3, no. 1, pp. 22–28, 2019, [Online]. Available: https://jurnal.kaputama.ac.id/index.php/JTIK/article/view/141/156

[5] I. Hidayanti, T. B. Kurniawan, and A. Afriyudi, "Perbandingan Dan Analisis Metode Klasifikasi Untuk Menentukan Konsentrasi Jurusan," *J. Ilm. Inform. Glob.*, vol. 11, no. 1, pp. 16–21, 2020, doi: 10.36982/jig.v11i1.1067.

[6] O. Arifin and T. B. Sasongko, "Analisa perbandingan tingkat performansi metode support vector machine dan naïve bayes classifier," *Semin. Nas. Teknol. Inf. dan Multimed. 2018*, vol. 6, no. 1, pp. 67–72, 2018.

[7] R. T. Prasetio and E. Ripandi, "Optimasi Klasifikasi Jenis Hutan Menggunakan Deep Learning Berbasis Optimize Selection," *J. Inform.*, vol. 6, no. 1, pp. 100–106, 2019, doi: 10.31311/ji.v6i1.5176.

[8] Y. A. Wijaya, N. Suarna, Iin, R. Hamonangan, and R. Nining, "Comparison of machine learning algorithm for Santander dataset," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012032, 2021, doi: 10.1088/1757-899x/1088/1/012032.

[9] A. M. Hilda, I. Rahmadi, A. R. Dzikrillah, and D. Mugisidi, "UHAMKA DENGAN MENGGUNAKAN ALGORITMA CLASSIFIER," no. 6, pp. 1–6.

[10] M. E. Lasulika, "Komparasi Naïve Bayes, Support Vector Machine Dan K-Nearest Neighbor Untuk Mengetahui Akurasi Tertinggi Pada Prediksi Kelancaran Pembayaran Tv Kabel," *Ilk. J. Ilm.*, vol. 11, no. 1, pp. 11–16, 2019, doi: 10.33096/ilkom.v11i1.408.11-16.

[11] G. A. Pradnyana, I. G. M. Darmawiguna, D. K. S. Suditresna Jaya, and A. Sasmita, "Performance analysis of support vector machines with polynomial kernel for sentiment polarity identification: A case study in lecturer's performance questionnaire," *J. Phys. Conf. Ser.*, vol. 1810, no. 1, 2021, doi: 10.1088/1742-6596/1810/1/012033.

[12] S. A. Deni Gunawan, Dwiza Riana, Dian Ardiansyah, Fajar Akbar, "Komparasi Algoritma Support Vector Machine Dan Naïve Bayes Dengan Algoritma Genetika Pada Analisis Sentimen Calon Gubernur," vol. V, no. 1, pp. 135–138, 2019, doi: 10.31294/jtk.v4i2.

[13] H. Apriyani and K. Kurniati, "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Dalam Klasifikasi Penyakit Diabetes Melitus," *J. Inf. Technol. Ampera*, vol. 1, no. 3, pp. 133–143, 2020, doi: 10.51519/journalita.volume1.isssue3.year2020.page133-143.

[14] "Optimize Selection - RapidMiner Documentation." https://docs.rapidminer.com/latest/studio/operators/modeling/optimization/feature_selection/optimize_selection.html (accessed Jul. 22, 2022).

[15] "T-Test - RapidMiner Documentation." https://docs.rapidminer.com/latest/studio/operators/validation/performance/significance_tests/t_test.html (accessed Jul. 23, 2022).

[16] O. Nurdiawan, D. A. Kurnia, D. Solihudin, T. Hartati, and T. Suprapti, "Comparison of the K-Nearest Neighbor algorithm and the decision tree on moisture classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1088, no. 1, p. 012031, 2021, doi: 10.1088/1757-899x/1088/1/012031.