PERBANDINGAN ALGORITMA C4.5 DAN K-NN DALAM IDENTIFIKASI MAHASISWA BERPOTENSI DROP OUT

Yeyen Dwi Atma[1)], Arif Setyanto[2)]

[1), 2)] *Magister Teknik Informatika Universitas Amikom Yogyakarta*
[1), 2)] *Jln. Ring Road Utara, Condongcatur, Depok, Kec. Depok, Kabupaten Sleman, DI. Yogyakarta 55281*
Email : yeyenduwy@gmail.com[1)], arief_s@amikom.ac.id[2)]

Abstrak

Dalam dunia pendidikan, data mining dikenal dengan istilah Educational Data Mining (EDM). EDM merupakan teknikuntuk menggali data pada ranah pendidikan,penggunaan EDM ditujukan untuk lebih memahami data mahasiswa. Dalam dunia pendidikan EDM membantu pendidik menganalisis hasil kinerja mahasiswa, mendeteksi mahasiswa yang memerlukan dukungan agar tidak mengalami kegagalan. Kegagalan akademik mahasiswa dalam menempuh pendidikan merupakan masalah yang penting untuk dilakukan analisa dan prediksi dalam meminimalisir banyaknya kasus mahasiswa putus studi atau drop out.Prediksi diberlakukan untuk memeberikan peringatan dini gunamencegah kegagalan mahasiswa. Algoritma data mining untuk proses klasifikasi dan prediksiyang banyak digunakan diataranyaadalah C4.5 dan K-NN. Dalam penelitian ini C4.5 dan K-NN memanfaatkan fitur seleksi Forward selection dengan memanfaatkan karakteristik data itu sendiri. Dalam penelitian ini K-NNberbasi Forward selection lebih akurat dalam mengklasifikasikan status mahasiswa dengan hasil akurasi 99.46% *dan termasuk dalam kategori "excellent classification".*

Kata kunci:*data mining, C4.5, K-NN, drop out, Forward selection.*

1. PENDAHULUAN

1.1 Latar Belakang Masalah

Pada perguruan tinggi, mahasiswa merupakan parameter penting dalam pergerakan bisnis proses dan evaluasi penyelenggaraan program studi. Menjaga hal yang berkaitan dengan prestasi, kompetensi, dan presensi mahasiswa seharusnya mendapatkan perhatian yang serius dalam evaluasi kinerja mahasiswa. Selain itu bentuk kinerja yang baik sebagai mahasiswa adalah dengan lulus tepat waktu dan tidak terancam sangsi drop out atau mengundurkan diri. Dalam peraturan Menteri Ristek Dikti nomor 44 tahun 2015 menyebutkan standar proses pembelajaran maksimum dalam masa studi adalah 7 tahun, IPK diatas 2.0, minimum sks adalah 144 untuk program Sarjana. Meskipun masa studi maksimum yang ditentukan cukup panjang 7 tahun dan kriteria lain IPK > 2.0, minimum sks adalah 144 sudah di uraikan, perguruan tinggi sebaiknya memiliki batasan tersendiri untuk menjaga kualitas dan kinerja mahasiswa.

STMIK Balikpapan merupakan salah satu perguruan tinggi swasta yang mengacu pada bidang teknologi informasi dan bergerak demi meningkatkan kualitas dari sumberdaya manusia. Salah satu yang dijaga demi meningkatkan kualitas SDM yaitu dengan menjaga kualitas mahasiswaSTMIK Balikpapan. Untuk menjaga kualitas mahasiswa perlu adanya proses monitoring secara berkala dalam mendapatkan pendukung keputusan seperti mahasiswa yang masih layak lanjut studi ataupun harus *drop out* karena berbagai macam faktor. Tingkat pengambilan keputusan yang akurat tersebut tidak hanya berpengaruh untuk mahasiswa itu sendiri, tapi dapat berpengaruh juga pada rasio dan poin akreditasi perguruan tinggi tersebut.

Salah satu bidang ilmu yang berkembang dalam pemanfaatan data pada teknologi informasi dalam bidang pendidikan saat ini adalah data mining. Data mining memiliki banyak manfaat dalam berbagai masalah pengolahan data, data yang diolah umumnya akan memiliki informasi penting yang dapat digali dan dianalisa[1]. Menurut [2] data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui. Salah satu penggunaan data mining dalam proses pengetahuan terhadap data adalah teknik klasifikasi. Menurut [3]teknik klasifikasi merupakan kegiatan dalam mengekstraksi data dan kemudian memprediksi label kategori untuk masing-masing data.

Menurut [2]klasifikasi pada dasarnya adalah melakukan penggolongan terhadap data yang memiliki target variable kategori. Dengan semakin berkembangnya teknologi, klasifikasi bisa dilakukan dengan berbagai algoritma sesuai dengan kebutuhan. Menurut [8]beberapa algoritma klasifikasi data mining terbaik diantaranya adalah algoritma model *C4.5* dan *K-Nearest Neighbor.* Dalam penelitian terdahulu terkait teknik klasifikasi yang menggunakan metode *C4.5* dan *K-Nearest Neighbor* diantaranya: Penelitian [14]dalam menganalisis perbandingan algoritma *classification* untuk *authentication* uang kertas menghasilkan akurasi

pendeteksian keaslian uang sangat tinggi. Persentase akurasi yang sangat tinggi adalah menggunakan metode *Decision Tree C4.5* dengan nilai akurasi sebesar 98,5 %, sedangkan *Neural Network* sebesar 95%, dan *Navies Bayes* sebesar 85%.

Penelitian yang dilakukan [11]dengan judul implementasi pengolahan citra dan klasifikasi K-*Nearest Neighbour* untuk membangun aplikasi pembeda daging sapi dan daging babi berbasis web, penelitian tersebut berhasil melakukan klasifikasi perbedaan daging babi dan daging sapi berdasarkan extraksi ciri tekstur dengan akurasi 88,75%.

Berdasarkan dari beberapa penelitian sebelumnya dapat dilihat bahwa proses klasifikasi data sangat membantu dalam proses pengerjaan komputerisasi sehingga mempermudah untuk menarik data jika diperlukan suatu saat untukanalisa mahasiswa yang masuk dalam dua kelas kategori*drop out*dan tidak *drop out* setiap tahunnya. Namun pada penelitian sebelumnya masih sedikit yang membahas manakah metode yang paling akurat dalam prediksi mahasiswa *drop out* menggunakan algoritma *C4.5* dan *K-Nearest Neighbour*dengan penambahanfitur*Forward selection*. Penelitian ini bertujuan untuk melakukan perbandingan dalam klasifikasi mahasiswa *drop out* menggunakan algortima *C4.5* dan algoritma *K-Nearest Neighbour*dengan fitur*Forward selection* serta mengevaluasi performa algoritma *C4.5* dan algoritma *K-Nearest Neighbour* dalam menghasilkan tingkat keakuratan klasifikasi mahasiswa.Dari hasil penelitian ini diharapkan dapat digunakan sebagai rekomendasi dan masukan bagi peneliti selanjutnya, masyarakat, dan institusi pendidikan terkait sehingga diketahui manakah metode klasifikasi yang akurat.

2.TINJAUANPUSTAKA

2.1 Tinjauan Studi

Penelitian terkait tentang klasifikasi yang menggunakan algoritma C4.5, K-NN, dan *Forward selection*:

Penelitian [15]melakukan klasifikasi kelulusan mahasiswa menggunakan algoritma Naïve Bayes dengan fitur *Forward Selection*, Dalam penelitian yang dilakukan Naive Bayes memanfaatkan fungsi seleksi fitur dari *Forward selection* untuk pemilihan atribut data dengan karakteristik data itu sendiri, dan meningkatkan ketepatan klasifikasi Naïve Bayes. *Forward selection* berbasis Naive Bayes lebih akurat dan efektif dalam mengklasifikasikan status kelulusan mahasiswa dengan hasil akurasi 97,14% dan termasuk dalam kategori "excellent classification". Dengan memperoleh atribut yang berpengaruh yaitu: *status pekerjaan* dan *IPK semester 4*.

Penelitian [11] membandingkan akurasi prediksi kategori Indeks Prestasi (IP) semester pertama mahasiswa Fakultas Teknologi Informasi (FTI) Universitas Kristen Duta Wacana (UKDW) menggunakan algoritma C4.5 dan CART. Penelitian ini juga mengeksplorasi berbagai parameter seperti kategorisasi atribut numerik, keseimbangan data, jumlah kategori IP, dan ketersediaan atribut yang berbeda karena perbedaan ketersediaan data antara jalur prestasi dan jalur non-prestasi. Hasil penelitian ini algoritma C4.5 dan CART memiliki akurasi yang sama untuk memprediksi kategori IP mahasiswa baru pada jalur prestasi (data non numerik), yaitu sebesar 86,86%.

2.2 Landasan Teori

Menurut[6]data mining adalah istilah yang digunakan untuk menemukan pengetahuan yang tersembunyi di dalam database. Data mining merupakan proses semi otomatik yang menggunakan teknik stratistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi pengetahuan potensial dan berguna yang bermanfaat yang tersimpan dalam database besar.Menurut [4]data mining didefinisikan sebagai sebuah proses untuk menemukan hubungan, pola dan tren baru yang bermakna dengan menyaring data yang sangat besar, yang tersimpan dalam penyimpanan, menggunakan teknik pengenalan pola seperti teknik Statistik dan Matematika.

Dalam data mining terdapat tahapan-tahapan yang menjadi pedoman untuk mencari pengetahuan terhadap data dengan istilah *Knowledge Discovery In Databases (KDD). KKD* merupakan proses untuk mengetahui pola dalam kumpulan data dengan jumlah besar. Secara umum tahapan-tahapan proses *KKD* terdiri dari [4]:

1. *Data Cleaning.* Proses menghilangkan noise dari data yang tidak konsisten.
2. *Data Integration.* Penggabungan Data dari berbagai database ke dalam satu database baru.
3. *Data Selection.* Proses pemilihan data yang relevan yang didapat dari database.
4. *Data Tranformation.* Data diubah ke dalam format yang sesuai untuk diproses dalam Data Mining.
5. *Data Mining.* Suatu metode yang diterapkan untuk menemukan pengetahuan berharga yang tersembunyi dari data.
6. *Pattern Evaluation.* Mengidentifikasi pola-pola menarik untuk dipresentasikan ke dalam knowlwdge based.
7. *Knowledge Presentation.* Visualisasi dan penyajian pengetahuan mengenai teknik yang digunakan untuk memperoleh pengetahuan yang diperoleh oleh user.

2.2.1.1 C4.5

Algoritma *C4.5* adalah kelompok dari algoritma *Decision Tree*. Algoritma ini mempunyai dua input yaitu training samples dan testing samples. Training samples adalah data contoh yang akan digunakan untuk membangun sebuah tree yang telah diuji kebenarannya sebelumnya. Sedangkan samples merupakan field-field data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data. Algoritma *C4.5* ini juga merupakan salah satu metode untuk membuat decision tree berdasarkan training data yang telah disediakan. Algoritma *C4.5* dibuat oleh Ross Quinlan yang adalah pengembangan dari *ID3* yang juga dibuat oleh Quinlan. Beberapa pengembangan yang dilakukan pada *C4.5* adalah sebagai berikut: dapat mengatasi missing value, dapat mengatasi continue data, dan pruning.

Terdapat beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma *C4.5*, yaitu [2]:

1. Menyiapkan data training. Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon. Akar akan diambil dari atribut yang terpilih, dengan cara menghitung nilai gain dari masing-masing atribut, nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung gain dari atribut, hitung dahulu nilai entropy menggunakan persamaan 1 sebagai berikut:

   $$Entrophy\ (s) = \sum_{i=1}^{n} - pi * \log_2 pi \quad \ldots(1)$$

   Keterangan:
   S : himpunan kasus
   A : atribut
   n : jumlah partisi S
   pi : proporsi dari Si terhadap S

3. Menentukan nilai gain dengan metode *informasi gain*:

   $$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{Si}{S} * Entropy(Si) \ldots(2)$$

   Keterangan:
   S : himpunan kasus
   A : atribut
   n : jumlah partisi atribut A
   |Si| : jumlah kasus pada partisi ke-i
   |S| : jumlah kasus dalam S

4. Ulangi langkah ke-2 hingga semua kasus terpartisi.
5. Proses partisi pohon keputusan akan berhenti saat:
   a. Semua kasus dalam node N mendapat kelas yang sama.
   b. Tidak ada atribut di dalam kasus yang dipartisi lagi.
   c. Tidak ada kasus di dalam cabang yang kosong.

2.2.1.2 K-Nearest Neighbour

Algoritma *K-Nearest Neighbor* merupakan metode klasifikasi yang mengelompokkan data baru berdasarkan jarak data baru itu kedalam beberapa data tetangga (negihbord) terdekat. Teknik *K-Nearest Neighbor* dengan melakukan langkah-langkah yaitu, input : data training, label data training, k, data testing [5].

Proses yang dilakukan dalam K-NN untuk mendapatkan clas kategori adalah dengan menghitung kemiripan antara data baru dengan tiap-tiap data yang sebelumnya telah dikategotrikan. Dengan kata lain algoritma ini bekerja berdasarkan dari data baru terhadap data yang sudah ada atau latih.

Data latih tersebut kemudian diurutkan mulai dari data yang memiliki nilai kemiripan paling besar dengan data yang akan dikategorikan. Data dipilih sebanyak *k* data dengan nilai kemiripan terbesar, kemudian memprediksikan data yang baru ke dalam kategori data tersebut.Adapun rumus untuk melakukan perhitungan kedekatan antara dua kasus dapat dilihat pada persamaan 3 sebagai berikut;

$$\text{Similarity } (T,S) = x = \frac{\sum_{i=1}^{n} f(Ti,Si) * wi}{\text{wi}} \quad \ldots(3)$$

Keterangan:
T : kasus baru
S : kasus yang ada dalam penyimpanan
n : jumlah atribut dalam setiap kasus
i : atribut individu antara 1 s.d. n
f : fungsi *similarity* atribut *i* antara kasus T dan kasus S
w : bobot yang diberikan pada atribut ke-*i*

kedekatan biasanya berada pada nilai antara 0 s.d. 1. Nilai 0 artinya kedua kasus mutlak tidak mirip, sebaliknya untuk nilai 1 kasus mirip dengan mutlak.

2.2.1.2 Seleksi Fitur *Forward selection*

Metode *Forward selection* atau metode seleksi maju adalah algoritma pencarian paling sederhana. *Forward selection* didasarkan pada model Regresi Linear. *Forward selection* adalah salah satu teknik untuk mereduksi dimensi dataset dengan menghilangkan atribut-atribut yang tidak relevan atau redudan [17]. Tujuan dari seleksi fitur ini adalah untuk mengurangi tingkat kompleksitas dari sebuah algoritma klasifikasi, meningkatkan akurasi dari algoritma klasifikasi tersebut, dan mampu mengetahui fitur-fitur yang paling berpengaruh terhadap tingkat akurasi.

Metode *forward selection* adalah pemodelan dimulai dari nol peubah (empty model), kemudian satu persatu peubah dimasukan sampai kriteria tertentu dipenuhi[18]. Langkah-langkah metode *forward selection* adalah sebagai berikut:

a. Membuat model dengan meregresikan variabel respon Y dengan setiap variabel prediktor. Kemudian dipilih model yang mempunyai nilai R^2 tertinggi. Misal model tersebut adalah yang memuat prediktor X_a, yaitu:
$\hat{Y} = b_0 + b_a X_a$ …(4)
b. Meregresikan variabel respon Y, dengan prediktor X_a, ditambah dengan setiap prediktor selain X_a dan prediktor lain. Kemudian dipilih model yang nilai R_2 nya tertinggi, misal mengandung tambahan prediktor X_b, yaitu model
$\hat{Y} = b_0 - b_a X_a + b_x X_b$ …(5)
Prediktor terpilih X_b berarti mempunyai $F_{sequensial}$ tertinggi. Formula $F_{sequensial}$ untuk X_b adalah
$F_{seq} = R(\beta_b \beta_0, \beta_a) MSE\ db$ …(6)
Nilai Fsequensial untuk X_b juga dapat diperoleh dengan cara mengkuadratkan nilai statistik uji T prediktor $X_{b.}$
c. Proses diulang sampai didapatkan $F_{sequensial} > F_{in}$. Nilai $F_{in} = F(1,v,\alpha_{in})$, sehingga model terbaik yang dipilih adalah model yang tidak mempunyai prediktor dengan $F_{sequensial} < F_{in}$.

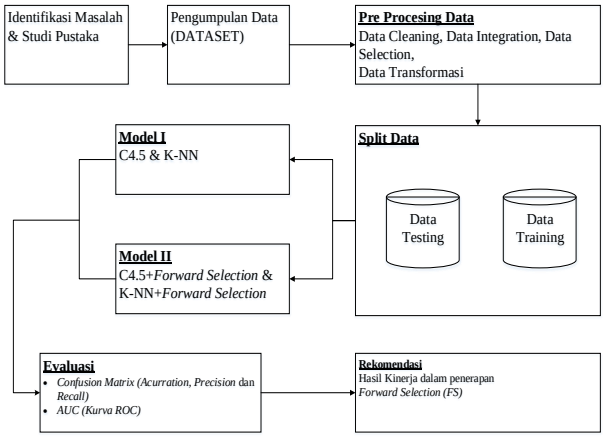
2.2.1.3*RapidMiner*

RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. RapidMiner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. RapidMiner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan menggunakan Bahasa pemrograman java sehingga dapat bekerja di semua sistem operasi[19].

3.METODE PENELITIAN

3.1 Alur Penelitian

Dalam penelitian ini metode yang dilakukan mengacu kepada teknik observasi yang meliputi identifikasi masalah dan studi pustaka dalam melakukan penelitian ini, kemudian dilanjutkan kedalam tahapan pengumpulan data, *pre processing data*, yang selanjutnya dilakukan proses mining untuk mengetahui hasil akhir dari penelitian ini. Pada penelitian ini model klasifikasi dimulai dari dataset yang akan dibagi menjadi data trining dan testing menggunakan *split data*yang selanjutnya dilakukan mining dengan algoritma klasifikasi sehingga dihasilkan model klasifikasi dan memunculkan parameter evaluasi. Model yang ada dalam penelitian ini adalah perbandingan optimasi C4.5, dan K-Nearest Neighbor, C45 (FS), dan K-NN (FS) yang dijabarkan pada gambar 1 berikut:



Gambar 1. Alur Penelitian

1.2 Pengujian Model

Pada tahapan ini akan dijelaskan tentang teknik pengujian yang digunakan dalam penelitian.Tahapan yang dilakukan untuk mengkalsifikasi status mahasiswa dengan menggunakan dua metode yaitu *C4.5, K-NN*dan *C4.5, K-NN* dengan *Forward selection*. Proses eksperimen dan pengujian menggunakan dataset mahasiswa STMIK Balikpapan yang memiliki 2 class mahasiswa dan 15 atribut dengan jumlah 3034 record. Prosesexperiment dan pengujian yang dilakukan mengikuti proses klasifikasi menggunakan RapidMiner.

1.3 Evaluasi Hasil

Pada tahap ini akan dibahas mengenai hasil evaluasi dan experimen yang telah dikerjakan. Phase evaluasi ini akan dilakukan perbandingan kuantitatif dengan mempertimbangkan nilai komparasi *confusion matrix*(*Accuracy*, *Precision* dan *Recall*).

4. HASIL DAN PEMBAHASAN

4.1 Dataset Mahasiswa

Langkah pertama proses klasifikasi diawali dengan penentuan dataset yang disimpan dalam format excel (*.xls) seperti pada gambar 2 berikut:

JURUSAN	JK	WAKTU KULIAH	STATUS PEKERJAAN	IPS1	IPS2	IPS3	IPS4	IPS5	IPS6	IPS7	IPS8	SKS	STATUS
TEKNIK INFOR	L	MALAM	TIDAK BEKERJA	2.94	2.82	3	3.21	3.04	3.26	3.00	2.11	141	Tidak Drop Out
TEKNIK INFOR	L	MALAM	TIDAK BEKERJA	2.2	2.21	2.41	2.05	2.55	2.5	2.50	3.00	153	Tidak Drop Out
TEKNIK INFOR	L	MALAM	TIDAK BEKERJA	2.2	2.26	2.63	2.36	1.74	2.88	1.50	3.22	138	Tidak Drop Out
TEKNIK INFOR	L	PAGI	TIDAK BEKERJA	1.89	0	0	0	0	0	0	0	22	Drop Out
TEKNIK INFOR	L	MALAM	TIDAK BEKERJA	2.2	2.26	2.63	2.13	2.41	3.39	2.52	1.55	157	Tidak Drop Out
TEKNIK INFOR	L	MALAM	BEKERJA	2.11	2.26	2.26	0	0	0	0	0	64	Drop Out
TEKNIK INFOR	L	MALAM	BEKERJA	2.26	2.13	2.37	0	0	0	0	0	64	Drop Out
TEKNIK INFOR	L	MALAM	TIDAK BEKERJA	3.11	3.26	3.19	3.17	3.64	3.61	3.27	0	154	Tidak Drop Out
TEKNIK INFOR	L	MALAM	BEKERJA	2.05	2.26	2.74	2.26	0	0	0	0	85	Drop Out
TEKNIK INFOR	P	MALAM	BEKERJA	2.63	2.74	2.63	0	0	0	0	0	64	Drop Out
TEKNIK INFOR	L	MALAM	BEKERJA	2.37	2.26	2	0	0	0	0	0	64	Drop Out
TEKNIK INFOR	P	MALAM	TIDAK BEKERJA	2.11	2.11	2	2.88	2.35	2.83	4.00	2.33	144	Tidak Drop Out
TEKNIK INFOR	L	MALAM	TIDAK BEKERJA	2.21	2.26	2.63	1.91	1.89	2.58	0	0	123	Tidak Drop Out
TEKNIK INFOR	L	MALAM	TIDAK BEKERJA	3.05	3.32	3.71	3.21	3.28	3.13	3.18	0	154	Tidak Drop Out
TEKNIK INFOR	L	MALAM	BEKERJA	3.42	2.63	2.29	1.89	0	0	0	0	85	Drop Out
TEKNIK INFOR	L	MALAM	TIDAK BEKERJA	2.32	2.26	2.63	2.26	2.81	2.65	2.04	2.16	152	Tidak Drop Out
TEKNIK INFOR	L	MALAM	TIDAK BEKERJA	2.3	2.26	2.63	2.26	2.21	2.89	2.09	3.25	148	Tidak Drop Out
TEKNIK INFOR	L	PAGI	TIDAK BEKERJA	1.53	0	0	0	0	0	0	0	22	Drop Out
TEKNIK INFOR	L	MALAM	BEKERJA	2.3	2.26	2.26	2.26	0	0	0	0	85	Tidak Drop Out
TEKNIK INFOR	P	PAGI	TIDAK BEKERJA	2.32	2.37	0	0	0	0	0	0	45	Drop Out
TEKNIK INFOR	P	MALAM	TIDAK BEKERJA	2.47	2.68	2.84	2.76	3.29	3.05	4.00	3.00	144	Tidak Drop Out
TEKNIK INFOR	L	MALAM	TIDAK BEKERJA	2.26	2.26	2.28	2.26	0.81	2.31	0.83	2.38	146	Tidak Drop Out
TEKNIK INFOR	P	MALAM	TIDAK BEKERJA	2.47	2.58	2.32	2.5	2.89	2.85	2.45	2.44	143	Tidak Drop Out

Gambar 2. Potongan Data Status Mahasiswa

Ketentuan dan tipe data attribut dan label pada data indikator dataset mahasiswa drop out dalam penelitian ini dapat dijabarkan seperti pada table 1 berikut:

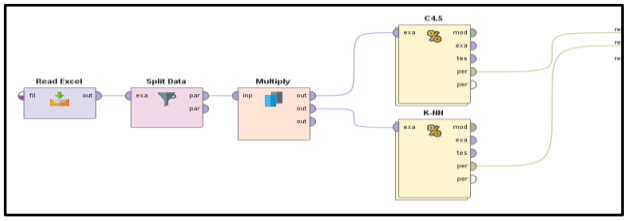
Tabel 1. Pembagian Variabel dan Kelas Data

Variabel	Nama Field	Kelas Data	Jenis Kelas Data
X1	NIM	Nomor induk Mahasiswa	Polinominal
X2	Nama	Nama Mahasiswa	Polinominal
X3	Jurusan	S1 ; D3	Binominal
X4	Jenis Kelamin	Pria ; Wanita	Binominal
X5	Waktu Kuliah	Pagi ; Malam	Binominal
X6	Status Pekerjaan	Bekerja ; Tidak Bekerja	Binominal
X7	IPS 1	0 s/d 4,00	Numeric
X8	IPS 3	0 s/d 4,00	Numeric
X9	IPS 4	0 s/d 4,00	Numeric
X10	IPS 5	0 s/d 4,00	Numeric
X11	IPS 6	0 s/d 4,00	Numeric
X12	IPS 7	0 s/d 4,00	Numeric
X13	IPS 8	0 s/d 4,00	Numeric
X14	SKS	0 s/d 4,00	Numeric
X15	Status	Class : Tidak Drop Out ;Drop Out	Binominal

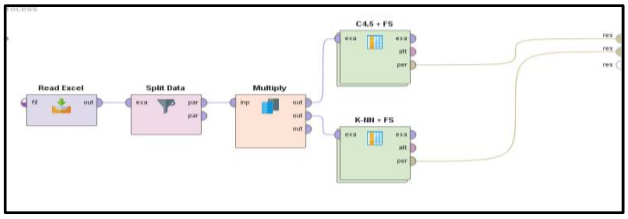
4.1.1 Implementasi dengan RapidMiner

Pada penelitian ini dilakukan pengujian keakuratan klasifikasi mahasiswa drop out dengan algoritma *C4.5, K-NN* dan *C4.5, K-NN* dengan kombinasi *Forward selection*sebagai fitur seleksi. Pemodelan yang dilakukan dalam penelitian ini untuk membagi data testing dan training menggunakan model Split data dengan 80% data training dan 20% data testing dan menambahkan fitur koneksi multiply sebagai percabangan untuk integrasi kedalam dataset mahasiswa drop out dengan model sebagai berikut:

1. Klasifikasi dataset menggunakan algoritma *C4.5, K-NN.*
2. Klasifikasi dataset menggunakan algoritma *C4.5, K-NN*dengan kombinasi seleksi fitur*Forward selection* (*C4.5+FS)* dan*(K-NN+FS).*



Gambar 3. Proses Algoritma *C4.5, K-NN*di RapidMiner



Gambar 4. Proses Algoritma *C4.5, K-NN* Dengan*Forward selection* di RapidMiner

4.1.2 Analisa Hasil

Pengujian dilakukan untuk mengetahui tingkat akurasi dari algoritma C4.5 dan K-NN dengan tambahan metode fitur *Forward selection* yang dilakukan pada dataset mahasiswa sebanyak 3034. Dalam hasil percobaan menujukkan bahwa penggunaan *Forward selection* meningkatkan akurasi algoritma C4.5 mencapai 96% sedangkan pada algoritma K-NN mencapai 99.46%.

Tabel 1. Hasil Perbandingan Algoritma C4.5 dan K-NN

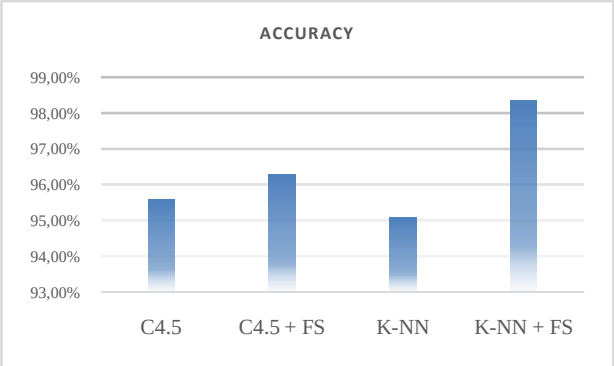
Model	Accuracy	Precision	Recall
C4.5	95.96%	99.08%	84.64%
C4.5 + FS	96.66%	99.81%	85.25%
K-NN	95.07%	95.35%	84.44%
K-NN + FS	98.34%	100%	93.37%

Penambahan seleksi fitur *forward selection*menghasilkan tingkat akurasi yang lebih baik daripada algoritma tanpa tambahan fitur *forward selection*yang hanya mecapai tingkat akurasi sebesar 96.66% pada algoritma C4.5 dan pada K-Nearest Neighbour sebesar 99.46%.

ISSN : 2580-1503

accuracy: 95.96% +/- 1.13% (mikro: 95.95%)	C4.5		
	true Tidak Drop Out	true Drop Out	class precision
pred. Tidak Drop Out	1849	95	95.11%
pred. Drop Out	5	523	99.05%
class recall	99.73%	84.63%	
accuracy: 96.28%	C4.5 + Forward Selection		
	true Tidak Drop Out	true Drop Out	class precision
pred. Tidak Drop Out	1853	91	95.32%
pred. Drop Out	1	527	99.81%
class recall	99.95%	85.28%	
accuracy: 95.07% +/- 1.36% (mikro: 95.06%)	K-NN		
	true Tidak Drop Out	true Drop Out	class precision
pred. Tidak Drop Out	1828	96	95.01%
pred. Drop Out	26	522	95.26%
class recall	98.60%	84.47%	
accuracy: 98.34%	K-NN Forward Selection		
	true Tidak Drop Out	true Drop Out	class precision
pred. Tidak Drop Out	1854	41	97.84%
pred. Drop Out	0	577	100.00%
class recall	100.00%	93.37%	

Gambar 5. Hasil Uji Algoritma C4.5 dan K-NN



Gambar 5. Grafik Hasil Uji Komparasi Algoritma C4.5 dan K-NN.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil implementasi algoritma C4.5 dan K-NN pada kasus prediksi mahasiswa drop out dapat diambil beberapa kesimpulan sebagai berikut:

1. Dengan menggunakan dataset yang sama, penggunaan fitur *Forward selection*pada algoritma C4.5 dan K-NN pada kasus prediksi mahasiswa drop out dengan atribut : nim, nama, jurusan, jenis kelamin, waktu kuliah, status pekerjaan, indeks prestasi semester, jumlah satuan kredit semester mampu meningkatkan hasil akurasi.
2. Hasil dari pengujian algoritma klasifikasi untuk kasus prediksi mahasiswa drop out untuk algoritma C4.5 tanpa penambahan fitur seleksi *Forward selection* didapatkan akurasi sebesar 95.96%, kemudian setelah ditambahkan fitur seleksi *Forward selection* meningkat menjadi 96.66%. Sedangkan pada algoritma K-Nearest Neighbour tanpa penambahan fitur seleksi *Forward selection* diperoleh nilai akurasi sebesar 95.07% setelah ditambahkan fitur seleksi *Forward selection* meningkat menjadi 98.34%.
3. Dari hasil pengujian, kinerja antara algoritma K-Nearest Neighbour ditambah fitur seleksi *Forward selection*lebih ungggul bila dibandingkan dengan algoritma C4.5 ditambah fitur seleksi *Forward selection* pada kasus prediksi mahasiswa drop out.

5.2 Saran

Berdasarkan hasil pengujian dan kesimpulan yang telah jelaskan sebelumnya, maka beberapa saran dalam penelitian ini yang diusulkan adalah sebagai berikut:

1. Melakukan pengujian dan perbandingan pada algoritma lain untuk mendapatkan pengetahuan komparasi yang lebih luas.
2. Menggunakan metode seleksi lain seperti *Backward Elimination*, *Brute force*, *Evolutionary*dan lain sebagainya.
3. Metode *Forward selection* berbasis K-Nearet Neighbour terbukti akurat dalam klasifikasi mahasiswa drop out dari dataset, tetapi dalam penelitian ini terdapat pertimbangan penggunaan prosedur ini tidak selalu mengarahkan ke model pemilihan atribut yang terbaik. *Forward selection* berbasis *K-Nearest Neighbour* hanya mempertimbangkan sebuah subset atribut kecil dari semua model-model yang mungkin, sehingga resiko melewatkan atau kehilangan model terbaik akan bertambah, seiring dengan penambahan jumlah variabel bebas.

6. Daftar Pustaka

[1] Kantardzic, M., 2003, Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons.

[2] Kusrini dan Luthfi, E. T., 2009. Algoritma Data Mining. Yogyakarta : Penerbit Andi.

[3] Kamber, M., & Han, J. 2006, Data Mining, Concepts and Techniques Second Edition, Morgan Kaufmann Publishers, San Francisco.

[4] Larose, Daniel T. 2005. Discovering Knowledge in Data: An Introduction to Marketing, Sales, Customer Relationship Management. Second Edition. Wiley MIT Press.

[5] Santoso, B. (2007). Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis (1 ed.). Yogyakarta: Graha Ilmu.

[6] Turban, Efraim., Jay E. Aronson, dan Ting Peng Liang. 2005. Decision Support System and Intelligent Systems (Sistem Pendukung Keputusan dan Sistem Cerdas). Edisi 7 Jilid 1, Andi Offset. Yogyakarta.

[7] Jiawei Han, Data Mining Concept And Technique, 2nd ed., Asma Stephan, Ed. Champaign, United States of America: Multiscience Press, 2007.

[8] U. Fayyad , Smyth P, 1996, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", in Communications of the ACM, Vol. 39, No. 11. (pp. 27-34)

[9] Wu, X., & Kumar, V, 2009, The Top Ten Algorithms in Data Mining, CRC Press, Boca Raton

[10] Achmad Rifai, Rizki Aulianita, 2018, Komparasi Algoritma Klasifikasi C4.5 dan Naïve Bayes Berbasis Particle Swarm Optimization Untuk Penentuan Resiko Kredit, Journal Speed – Sentra Penelitian Engineering dan Edukasi, ISSN : 1979-9330, Volume 10 No 2 – 2018.

[11] Dea Alverina, A. R. Chrismanto, and R. G. Santosa, 2018, Perbandingan Akurasi Algoritma C4.5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa, Jurnal Teknologi dan Sistem Komputer, vol. 6, no. 2, Apr. 2018. doi: 10.14710/jtsiskom.6.2.2018.76-83, ISSN:2338-0403.

[12] Budianita, E., Jasril, J., & Handayani, L. (2015). Implementasi Pengolahan Citra dan Klasifikasi K-Nearest Neighbour Untuk Membangun Aplikasi Pembeda Daging Sapi dan Babi Berbasis Web. Jurnal Sains dan Teknologi Industri, 12(2), 242-247.

[13] Geeta Kashyap, Ekta Chauhan, 2016, Parametric Comparisons of Classification Techniques in Data Mining Applications, International Journal of Engineering Development and Research, Vol 4, Issue 2, ISSN: 2321-9939

[14] Khairul Sani, Wing Wahyu Winarno, Silmi Fauziati, 2016, Analisis Perbandingan Algoritma Classification Untuk Authentication Uang Kertas (Studi Kasus: Banknote Authentication), Jurnal Informatika Vol. 10, No. 1

[15] Mohammad Fajariaditya Nugroho, Setyoningsih Wibowo, 2017, Fitur Seleksi *Forward selection* Untuk Menetukan Atribut Yang Berpengaruh Pada Klasifikasi Kelulusan Mahasiswa Fakultas Ilmu Komputer UNAKI Semarang Menggunakan Algoritma Naive Bayes, Jurnal Informatika Vol. 3, N. 1: ISSN: 2460-4801/2447-6645

[16] Sari Dewi, 2016, Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan, Jurnal Techno Nusa Mandiri, ISSN 1978-2136, Vol. 60 XIII, No. 1

[17] Wahyuni, S. (2018). IMPLEMENTASI RAPIDMINER DALAM MENGANALISA DATA MAHASISWA DROP OUT. Jurnal Abdi Ilmu, 10(2), 1899-1902.

[18] Draper, N., Smith, H., 1992, Analisis Regresi Terapan Edisi Kedua, PT. Gramedia Pustaka Utama, Jakarta.

[19] RapidMiner, 16 September 2018, RapidMiner Documentation, http://docs.rapidminer.com/