IMPLEMENTASI DATA MINING UNTUK KLASIFIKASI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR

Inna Alvi Nikmatun
Departemen Ilmu Komputer / Informatika
Universitas Diponegoro
Email: innaalvin18@gmail.com

Indra Waspada
Departemen Ilmu Komputer / Informatika
Universitas Diponegoro
Email: indrawaspada@undip.ac.id

ABSTRAK

Akreditasi merupakan salah satu bentuk penilaian mutu dan kelayakan program studi di perguruan tinggi. Ketepatan waktu mahasiswa dalam menyelesaikan studi dan proporsi mahasiswa yang menyelesaikan studi dalam batas masa studi termasuk dalam elemen penilaian akreditasi. Hal tersebut menunjukkan diperlukan pemantauan terhadap masa studi mahasiswa. Rata-rata masa studi mahasiswa di Departemen Informatika Universitas Diponegoro masih di atas 4 tahun sehingga perlu dilakukan evaluasi dengan membangun aplikasi pengklasifikasian masa studi mahasiswa. Dengan mempertimbangkan keseimbangan data maka pengklasifikasian masa studi mahasiswa menggunakan kelas target masa studi <5 tahun dan >=5 tahun. Pada penelitian ini menggunakan data riwayat mahasiswa tahun angkatan 2007 sampai dengan 2011 yang telah lulus dengan jumlah data sebanyak 377 orang dengan 72 atribut nilai mata kuliah dan 1 kelas target berupa masa studi. Penelitian ini dilakukan dengan mengikuti tahap pengerjaan *data mining* yang mengacu pada proses *knowledge discovery in database* (KDD). Pengklasifikasian dilakukan dengan menggunakan algoritma *K-Nearest Neighbor.* Aplikasi *data mining* berhasil dibangun dengan hasil percobaan menunjukkan bahwa hasil klasifikasi masa studi terbaik diperoleh dengan memilih atribut dari semua mata kuliah pilihan dengan nilai akurasi 75.95%.

Kata kunci: mahasiswa; masa studi; KDD; *k-nearest neighbor.*

ABSTRACT

Accreditation is one form of assessment of the quality and feasibility of study programs in higher education. Timeliness of students in completing studies and the proportion of students completing studies within the study period are included in the accreditation assessment element. This shows that it is necessary to monitor the student's study period. The average study period of students in the Informatics Department of Diponegoro University is still over 4 years so it needs to be evaluated by building an application to classify student study periods. By considering the balance of data, the classifications of study periods of students use the target class of the study period <5 years and> = 5 years. In this study using historical data of students from 2007 to 2011 who have graduated with a total data of 377 people with 72 attributes of course values and 1 target class in the form of study period. This research was conducted by following the stages of data mining work that refers to the process of knowledge discovery in database (KDD). Classification is done using the K-Nearest Neighbor algorithm. Data mining applications were successfully built with experimental results showing that the best study period classification results were obtained by selecting attributes from all elective courses with an accuracy value of 75.95%.

Keywords: *students; study period; KDD; k-nearest neighbor.*

1. PENDAHULUAN

Program studi melaksanakan fungsi Tridarma Perguruan Tinggi yaitu pendidikan, penelitian dan pengabdian kepada masyarakat, serta mengelola iptek selaras yang sesuai dengan bidang studi tersebut. Untuk itu program studi harus mampu mengatur diri sendiri dalam meningkatkan dan menjamin mutu program studi. Mahasiswa merupakan aset bagi aplikasi pendidikan perguruan tinggi untuk itu perlu diperhatikan kelulusan mahasiswanya. Penilaian akreditasi ketepatan waktu menyelesaikan studi, proporsi

mahasiswa yang menyelesaikan studi dalam batas masa studi termasuk dalam elemen penilaian akreditasi pada program studi [1]. Pengaruh masa studi yang tepat waktu terhadap penilaian akreditasi perguruan tinggi maupun program studi sangat besar sehingga perlu dilakukan mekanisme evaluasi untuk optimasi ketepatan masa studi mahasiswa.

Data mahasiswa dan data kelulusan mahasiswa dapat menghasilkan informasi yang berlimpah, informasi yang tersembunyi dapat diketahui dengan cara melakukan pengelolaan pada data tersebut. Berdasarkan data yang diperoleh dari Departemen Ilmu Komputer/ Informatika Universitas Diponegoro, pada mahasiswa tahun angkatan 2007 sampai dengan 2011 dengan jumlah kelulusan 377 orang diperoleh informasi rata-rata masa studi mahasiswa masih di atas 4 tahun. Oleh karena itu perlu dilakukan analisis faktor-faktor yang mendukung tepat waktu dan juga faktor sebaliknya yang menyebabkan tidak tepat waktu. Penelitian terdahulu yang berkaitan dengan prediksi ketepatan kelulusan mahasiswa mendapatkan hasil akurasi dan 84% menggunakan algoritma K-NN [2] dan 82.08% dengan menggunakan algoritma naïve bayes [3].

Algoritma K-NN adalah salah satu algoritma yang sederhana untuk memecahkan masalah klasifikasi, algoritma K-NN sering menghasilkan hasil yang kompetitif dan signifikan [4], dengan keunggulan tersebut maka penelitian ini membangun aplikasi klasifikasi masa studi mahasiswa menggunakan algoritma *k-nearest neighbor* dengan atribut yang digunakan adalah nilai mata kuliah mahasiswa.

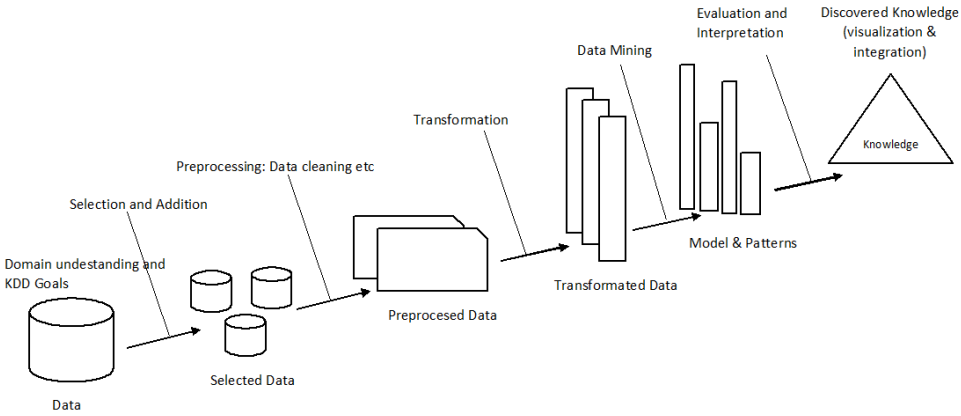
2. TINJAUAN PUSTAKA

Pada tinjauan pustaka menyajikan hasil kajian pustaka yang berhubungan dengan teori-teori dalam pembuatan penelitan. Pada bagian ini dijelaskan konsep *data mining*, klasifikasi, *k-nearest neighbor, k-fold cross validation, confusion matrix,* dan korelasi *pearson.*

2.1 Data Mining

Data mining merupakan proses penggalian informasi dan berguna dari set data besar yang melibatkan konsep interdisipliner yang relatif baru yang melibatkan analisis data dan penemuan pengetahuan dari database dan menggunakan pendekatan multi-sisi yang mencakup analisis statistik, visualisasi data, penemuan pengetahuan, pengenalan pola dan manajemen basis data [5].

Data Mining mempunyai beberapa model proses yang digunakan untuk mengarahkan pelaksanaan *data mining*, model proses yang biasa digunakan adalah *Knowledge Discovery Databases* (KDD), CRISP-DM dan SEMMA [6]. Pada penelitian ini memakai model proses *Knowledge Discovery Databases* yang mempunyai 9 langkah yang dapat dilihat pada Gambar 1.



Gambar 1. Tahapan KDD [6]

Berikut adalah penjelasan dari setiap tahapan pada KDD:

a. *Domain Understanding and KDD Goals*.
   Tujuan ditentukan dari sudut pandang *user* dan digunakan untuk mengembangkan dan pemahaman tentang domain aplikasi dan pengetahuan sebelumnya.
b. *Selection and Additions*
   Tahap kedua berfokus pada penentuan data target dan subset dari data sampel atau variabel.
c. *Preprocessing: Data Cleaning etc*

Pembersihan dan *preprocessing* data merupakan operasi dasar untuk menyelesaikan data yang konsisten tanpa *noisy*.

d. *Transformation*
Transformasi data dari satu bentuk ke bentuk lainnya sehingga data diimplementasikan dengan mudah.
e. *Data Mining (Chosing the Suitable Data Mining Task)*
Memillih metode *data mining* yang sesuai berdasarkan tujuan tertentu yang telah didefinisikan pada tahap pertama, contoh dari metode *data mining* adalah *classification*, *regression*, *clustering* dan *summarization*.
f. *Data Mining (Chosing the Suitable Data Mining Algorithm)*
Memilih algoritma yang tepat untuk pencarian pola-pola data, algoritma yang dipilih berdasarkan kecocokan kriteria dengan metode *data mining*.
g. *Data Mining (Imploying Data Mining Algorithm)*
Pada tahap ini algoritma yang telah dipilih diimplementasikan.
h. *Evaluation and Interpretation*
Tahap ini berfokus pada interpretasi dan evaluasi yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan hipotesa yang ada sebelumnya.
i. *Discovered Knowledge*
Penggunaan pengetahuan yang ditemukan dari proses KDD, dimana memutuskan apa yang akan dilakukan dengan pengetahuan dihasilkan.

2.2 Klasifikasi

Klasifikasi adalah sebuah proses untuk menemukan model yang menggambarkan dan membedakan kelas dari konsep data. Model ini diturunkan berdasarkan analisis satu set data pelatihan, model ini digunakan untuk memprediksi label kelas objek yang label kelasnya belum diketahui [7].

2.3 K-Nearest Neighbor

K-Nearest Neighbor (K-NN) merupakan salah satu algoritma yang digunakan dalam masalah pengklasifikasian. Prinsip kerja K-NN ialah mencari jarak terdekat antar data yang akan dievaluasi dengan tetangga terdekat dalam data pelatihan [8]. Algoritma *K-Nearest Neighbor* (K-NN) adalah salah satu algoritma paling sederhana untuk memecahkan masalah klasifikasi dan sering menghasilkan hasil yang kompetitif dan signifikan [4]. Untuk menghitung jarak menggunakan jarak *Euclidean.* Rumus jarak *Euclidean* didefinisikan dalam Persamaan (1):

$$d_i = \sqrt{\sum_{i=1}^{p}(X_{2i} - X_{1i})^2} \quad (1)$$

Keterangan:
X_1 = data latih
X_2 = data uji
i = variabel data
d = jarak
p = dimensi data

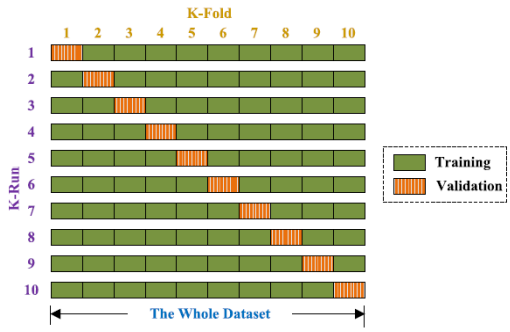
2.4 K-fold Cross Validation

K-fold cross validation merupakan sebuah metode yang digunakan untuk mengetahui rata-rata keberhasilan pengklasifikasian dengan melakukan pembagian dataset secara acak menjadi k himpunan bagian [9].

Kesalahan generalisasi pada *K-fold cross validation,* dan ditetapkan K menjadi 10 dengan dua alasan:

a. untuk menyeimbangkan antara biaya komputasi dan estimasi yang diandalkan
b. untuk perbandingan yang adil pada data latih dan data uji

Untuk 10-*fold cross validation*, dataset dibagi menjadi 10 lipatan yang saling terpisah dengan ukuran yang hampir sama. Dalam setiap *run,* 9 subset digunakan untuk pelatihan dan sisanya untuk validasi [10]. Diagram 10-*fold cross validation* dapat dilihat pada Gambar 2.



Gambar 2. Diagram 10-*fold Cross Validation* [10]

2.5 Confusion Matrix

Confusion Matrix merupakan salah satu cara untuk menganalisis kinerja model klasifikasi [7]. *Confusion Matrix* dapat dilihat pada Gambar 3.

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	*P'*	*N'*	*P + N*

Gambar 3. *Confusion Matrix* [7]

Akurasi dihitung dengan membandingkan jumlah data yang benar terklasifikasi dengan jumlah data keseluruhan. Cara perhitungannya dapat dilihat pada persamaan (2).

$$Akurasi = \frac{TP+TN}{P+N} \quad (2)$$

Precision didefinisikan sebagai ukuran ketepatan. Jika data diprediksi positif, seberapa seringkah data prediksi itu benar. Nilai *precision* dapat dilihat pada persamaan (3).

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Sedangkan *recall* adalah ukuran kelengkapan. Dari jumlah data sebenarnya yang bernilai positif, sebanyak apakah data yang diprediksi positif. Nilai *recall* dapat dilihat pada persamaan (4).

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Keterangan:

TP = nilai *true positives*
TN = nilai *true negatives*
FP = nilai *false positives*
FN = nilai *false negatives*
P + N = jumlah data

2.6 Korelasi Pearson

Korelasi *pearson* merupakan matrik statistik yang mengukur kekuatan dan hubungan linear antara dua variabel acak. Korelasi *pearson* telah diterapkan di berbagai indeks dalam statistik, seperti analisis data, klasifikasi, *clustering*, *decision making*, analisis keuangan, penelitian biologi dan lain-lain [11]. Rumus korelasi *pearson* dapat dilihat pada persamaan (5). Interval koefisien korelasi dapat dilihat pada Tabel 1.

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i^2\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i^2\right)^2}} \quad (5)$$

Keterangan:

r = Koefisien korelasi *product moment*
$\sum x$ = Jumlah total variabel x
$\sum y$ = Jumlah total variabel y
$\sum XY$ = Jumlah antara skor x dan y
N = Jumlah subjek/sampel

Tabel 1. Interval koefisien korelasi [12]

Interval Koefisien Korelasi	*Tingkat Hubungan*
0,00 – 0,199	Sangat Rendah
0,20 – 0,399	Rendah
0,40 – 0,599	Sedang
0,60 – 0,799	Kuat
0,80 – 1,00	Sangat Kuat

3. METODE PENELITIAN

Metode penelitian yang digunakan pada penelitian ini menggunakan model proses *data mining* yaitu *Knowledge Discovery Databases* (KDD), yang terdiri dari tahapan memahami domain dan tujuan, penggunaan data set dan penentuan target, pembersihan data dan pemrosesan awal data, pemilihan metode *data mining*, pemilihan algoritma *data mining*, implementasi algoritma *data mining*, evaluasi dan interpretasi

3.1 Memahami Domain Dan Tujuan

Berdasarkan data nilai mata kuliah mahasiswa Departemen Ilmu Komputer/ Informatika angkatan 2007 sampai dengan 2011 total jumlah kelulusan mahasiswa angkatan tersebut sejumlah 377 mahasiswa. Pada penelitian ini menggunakan data mahasiswa Departemen Ilmu Komputer/ Informatika angkatan 2007 sampai dengan 2011 dengan data mahasiswa yang cukup lengkap. Penelitian ini akan menggunakan *k-nearest neghbor* untuk mengklasifikasikan masa studi pada mahasiswa Departemen Ilmu Komputer/ Informatika yang kemudian hasil klasifikasi dapat digunakan sebagai salah satu alat evaluasi bagi departemen.

3.2 Penggunaan Dataset Dan Penentuan Target

Dataset yang digunakan dalam proses KDD adalah data nilai mata kuliah yang didapat dari SIA Undip yang meliputi nilai mata kuliah pada angkatan 2007 sampai dengan 2011 Pemilihan data nilai mahasiswa pada tahun 2007 sampai dengan 2011 dikarenakan pada tahun tersebut mempunyai kelengkapan data. Data mentah yang didapat sebelum dilakukan *preprocessing* berjumlah 27.842 dimana terdapat beberapa atribut yaitu ta, jalur, nim, kmk, mk, nilai bobot, nilai asli, aktif dan smt untuk contoh data awal sebelum dilakukan *preprocessing* dapat dilihat pada Tabel 3.1. Data yang akan digunakan pada penelitian ini adalah data nilai mata kuliah mahasiswa dengan atribut kode mata kuliah atau kmk, untuk penentuan kelas target didasarkan pada persentase data dengan lama studi <=4 tahun sebesar 6.4% sedangkah <5 tahun sebesar 25% dari persentase tersebut kemudian ditentukan kelas target pada penelitian ini adalah <5 tahun dan >=5 tahun agar kelas target tidak terlampau jauh dimana didapat jumlah kelas <5 sebanyak 94 dan >=5 sebanyak 283. Untuk mendapat data yang dapat digunakan pada proses penelitian ini dilakukan *preprocessing* data dengan cara penyederhanaan bentuk data yang akan dibahas pada sub-bab selanjutnya dimana hasil dari penyederhanaan menggunakan metode pivot yang akan digunakan pada penelitian ini.

3.3 Pembersihan Dan Pemrosesan Awal Data

Dari data tersebut dilakukan proses pembersihan data dan *preprocessing* untuk menangani *missing value* dan *noise* dan data yang tidak konsisten agar menghasilkan data yang berkualitas.

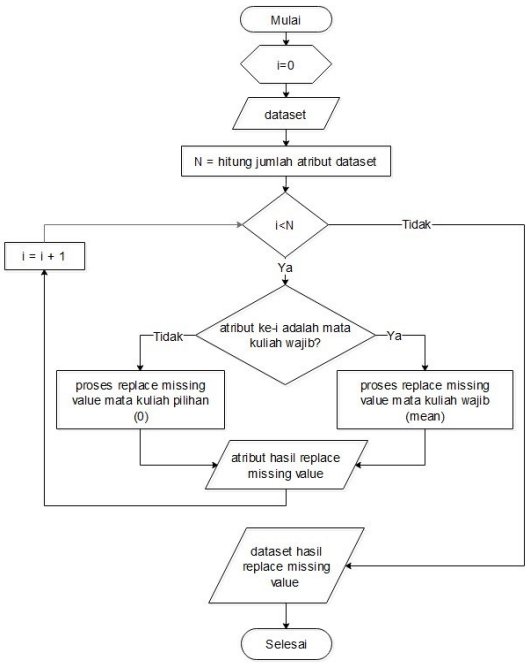
3.3.1 Pivot Data

Data yang berjumlah 27.842 data yang kemudian dipivot menjadi 377 data yang nantinya akan dipakai untuk penelitian ini dengan cara memfilter mahasiswa dengan keterangan LULUS dan tidak memakai mahasiswa dengan keterangan NON_AKTIF dan AKTIF dan menggunakan nilai 2.0 sampai dengan 4.0 dan menggunakan nilai terbaik dari nilai mata kuliah perbaikan sesuai dengan peraturan akademik Universitas Diponegoro [13] pada Pasal 31 ayat (4) butir d yang menyebutkan bahwa "Mahasiswa dinyatakan lulus mata kuliah, apabila mendapat nilai minimal C" dan Pasal 17 ayat 6 "Setiap Mahasiswa Program sarjana wajib lulus semua mata kuliah dan sejumlah mata kuliah pilihan yang tercakup dalam kurikulum program studi", dan Pasal 31 ayat (4) butir f menyebutkan bahwa "Mahasiswa yang mendapat nilai D,C,B dapat melakukan perbaikan pada semester *regular* atau remidi pada semester berjalan, dan nilai yang dipakai adalah nilai yang terbaik". Pada penelitian ini menggunakan 2 kelas yaitu <5 dan >=5 yang didapat dari semester mahasiswa, dimana semester kurang dari 10 menjadi kelas <5 dan lebih besar dari sama dengan 10 menjadi kelas >=5. Berikut adalah langkah-langkah dilakukannya pivot:

a. Menggabungkan semua data nilai mata kuliah mahasiswa dari angkatan 2007 sampai dengan 2011.
b. Menghapus data yang bernilai 0 pada kolom nilai bobot.
c. Menghapus data yang bernilai -1 pada kolom nilai bobot.
d. Menghapus data yang bernilai 1 pada kolom nilai bobot
e. Menggunakan data mahasiswa yang berstatus LULUS
f. Mengonversi seluruh data ke kurikulum 2007
g. Menghapus kolom jalur, mk, nilaiasli, aktif
h. Menghilangkan duplikasi data dan diambil yang paling besar (dari duplikasi data)
i. Membentuk pivot berdasarkan NIM (Nilai Induk Mahasiswa)
j. Menentukan Kelas
k. Menghapus kolom Nim, ta, smt

3.3.2 Replace Missing Value

Pada data yang diperoleh terdapat beberapa data yang masih kosong atau tidak ada nilainya. *Missing value* merupakan informasi yang tidak tersedia untuk sebuah kasus, *missing value* dapat terjadi karena penolakan dari responden untuk menjawab pertanyaan yang diajukan, kesalahan saat pengumpulan data misalnya pertanyaan terlewat sehingga tidak memperolah jawaban, kesalahan saat entri data, informasi yang tidak diberikan, sulit dicari atau memang informasi tersebut tidak ada [14]. Penanganan *missing value* dapat dilakukan dengan berbagai cara. Salah satu penanganan *missing value* adalah dengan mengisi nilai yang kosong dengan metode imputasi *mean*, *mean* merupakan salah satu metode yang paling umum digunakan, metode *mean* dilakukan dengan mengisi data yang *missing* dalam suatu variabel dengan nilai *mean* dari semua nilai yang diketahui dari variabel tersebut [15]. Pada penelitian ini menggunakan imputasi yaitu nilai *mean* sebagai pengganti pada data yang bernilai kosong untuk mata kuliah wajib dan mengisi nilai yang kosong dengan nilai 0 untuk mata kuliah pilihan penggantian *missing value* dengan nilai 0 pada mata kuliah pilihan dilakukan karena pada data yang *missing* untuk mata kuliah pilihan berarti tidak mengambil mata kuliah pilihan tersebut. Alur proses *replace missing value* mata kuliah wajib ditunjukkan pada Gambar 4.



Gambar 4. Alur Proses *Replace Missing Value*

3.4 Pemilihan Metode Data Mining

Tahap ini merupakan tahap pemilihan metode *data mining.* Terdapat beberapa metode pada *data mining* diantaranya klasifikasi, klastering, asosiasi, regresi dan lain-lain, Dari data yang diperoleh sudah terdapat target dimana dapat dikategorikan termasuk data *supervised learning* yang bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada dengan data baru. *Supervised learning* merupakan sinonim untuk klasifikasi, *supervised learning* sebuah metode dimana pengetahuan didapat dari label yang terdapat pada set data pelatihan [4]. Metode klasifikasi merupakan salah satu metode *supervised learning* maka dipilih metode klasifikasi untuk membantu menangani status masa studi pada mahasiswa.

3.5 Pemilihan Algoritma Data Mining

Algoritma yang digunakan pada penelitian ini adalah *K-Nearest Neighbor*. Algoritma *K-Nearest Neighbor* merupakan salah satu algoritma dari metode klasifikasi yang bertujuan untuk mengklasifikasikan objek baru berdasarkan jarak antara data latih dan data uji. Algoritma K-NN merupakan algoritma yang digunakan untuk mengklasifikasikan objek baru berdasarkan atribut dan data *training*.

3.6 Implementasi Algoritma Data Mining

Pada tahap ini dilakukan penerapan algoritma *data mining*, dari data yang telah di olah sebelumnya, seperti dilakukannya penanganan *missing value*. Data yang dihasilkan pada proses tersebut kemudian digunakan untuk implementasi menggunakan algoritma *K-Nearest Neighbor*. Berikut adalah tahapan algoritma *K-Nearest Neighbor* :

a. Menggunakan data latih, data uji dan jumlah k.
   Pada tahap pertama yaitu menggunakan data latih, data uji dan jumlah *k* tetangga terdekat K-NN.
b. Menghitung jarak *Euclidean* antara data uji dengan data latih
   Menghitung jarak menggunakan rumus jarak *euclidean distance* pada persamaan 1.
c. Mengurutkan hasil perhitungan jarak *Euclidean* secara *ascending*
   Dari hasil perhitungan jarak, hasil tersebut diurutkan secara *ascending* atau dari terkecil ke terbesar untuk mengetahui kedekatan jarak dengan data uji.
d. Mengambil sejumlah k data dari hasil pengurutan
   Berdasarkan hasil jarak yang telah diurutkan secara ascending diambil sebanyak K tetangga terdekat
e. Menentukan Kelas berdasarkan kelompok mayoritas
   Hasil dari pemilihan tetangga terdekat sebanyak k tetangga menjadi penentu kelas dari data uji.

3.7 Evaluasi dan Interpretasi

Pada tahap ini dilakukan evaluasi dan dilakukan interpretasi terhadap hasil dari klasifikasi masa studi. Pada evaluasi melibatkan perhitungan nilai akurasi, *precision* dan *recall* dari data tersebut dengan menggunakan metode evaluasi *K-fold cross validation* dengan k-*fold* = 10. Tahap interpretasi merupakan tahap visualisasi dari hasil evaluasi yang dilakukan.

3.8 Penggunaan Pengetahuan yang Ditemukan

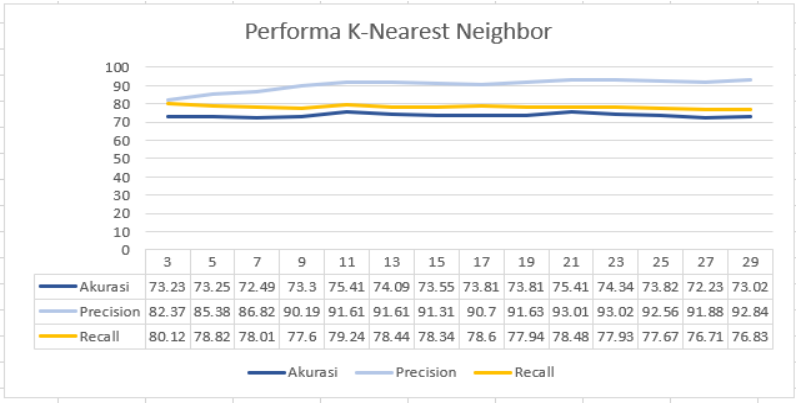
Pada tahap penggunaan *knowledge* merupakan tahap akhir dari model KDD, hasil dari pengembangan algoritma yang telah dilakukan diimplementasikan ke dalam sebuah aplikasi klasifikasi masa studi mahasiswa

4. HASIL DAN PEMBAHASAN

Pada bab ini akan diuraikan hasil penelitian dan pembahasan mengenai penelitian yang telah dilakukan meliputi skenario pengujian yang bertujuan untuk melakukan pengujian terhadap akurasi dari penerapan algoritma *K-Nearest Neighbor* pada dataset yang digunakan. Skenario pengujian meliputi pengujian akurasi dari variasi-variasi atribut yang digunakan. Terdapat 6 skenario yang dilakukan pada penelitian ini untuk mendapatkan akurasi terbaik dari algoritma yang digunakan.

4.1 Skenario Pengujian 1

Pada Pengujian 1 dilakukan untuk mengetahui performa *K-Nearest Neighbor* dengan menggunakan seluruh atribut yang ada pada dataset dengan menggunakan model *k-fold cross validation* dengan nilai k = 10 dan tetangga terdekat sejumlah k=3 sampai dengan k=29 dengan nilai k tetangga bernilai ganjil. Perhitungan yang dilakukan akan menghasilkan nilai akurasi, *precision* dan *recall*. Hasil skenario pengujian 1 dapat dilihat pada Gambar 5.



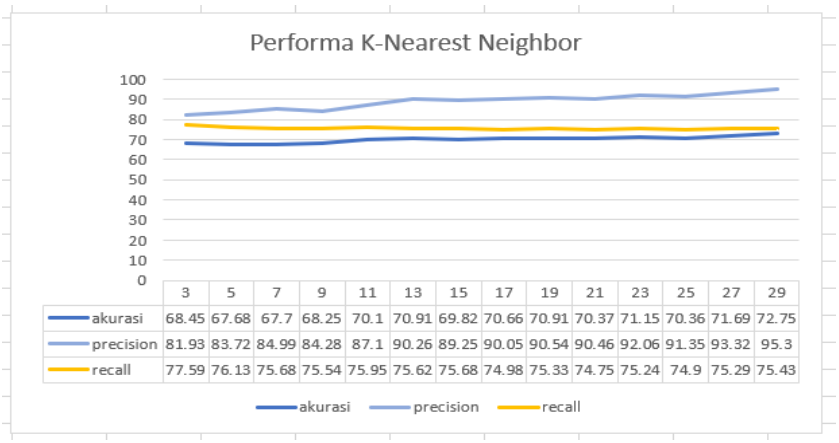
	3	5	7	9	11	13	15	17	19	21	23	25	27	29
Akurasi	73.23	73.25	72.49	73.3	75.41	74.09	73.55	73.81	73.81	75.41	74.34	73.82	72.23	73.02
Precision	82.37	85.38	86.82	90.19	91.61	91.61	91.31	90.7	91.63	93.01	93.02	92.56	91.88	92.84
Recall	80.12	78.82	78.01	77.6	79.24	78.44	78.34	78.6	77.94	78.48	77.93	77.67	76.71	76.83

Gambar 5. Hasil Skenario Pengujian 1

Dari Gambar di atas ditampilkan hasil rata-rata pada akurasi, *precision* dan *recall*. Parameter yang dijadikan performa akurasi. Sehingga didapat *k* terbaik untuk skenario 1 pada *k*=11 dengan performa akurasi 75.41%, *precision* 91.61% dan *recall* 79.24%.

4.2 Skenario Pengujian 2

Pada Pengujian 2 dilakukan untuk mengetahui performa *K-Nearest Neighbor* dengan menggunakan atribut mata kuliah semester 1 sampai dengan semester 4 dengan menggunakan model *k-fold cross validation* dengan nilai k = 10 dan tetangga terdekat sejumlah k=3 sampai dengan k=29 dengan nilai k tetangga bernilai ganjil. Perhitungan yang dilakukan akan menghasilkan nilai akurasi, *precision* dan *recall*. Hasil skenario pengujian 2 dapat dilihat pada Gambar 6.



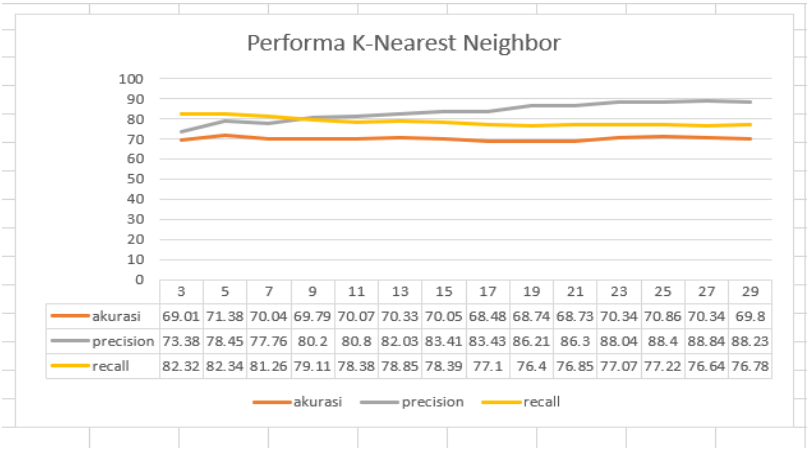
	3	5	7	9	11	13	15	17	19	21	23	25	27	29
akurasi	68.45	67.68	67.7	68.25	70.1	70.91	69.82	70.66	70.91	70.37	71.15	70.36	71.69	72.75
precision	81.93	83.72	84.99	84.28	87.1	90.26	89.25	90.05	90.54	90.46	92.06	91.35	93.32	95.3
recall	77.59	76.13	75.68	75.54	75.95	75.62	75.68	74.98	75.33	74.75	75.24	74.9	75.29	75.43

Gambar 6. Hasil Skenario Pengujian 2

Dari Gambar di atas ditampilkan hasil rata-rata pada akurasi, *precision* dan *recall*. Parameter yang dijadikan performa akurasi. Sehingga didapat *k* terbaik untuk skenario 2 pada *k*=29 dengan performa akurasi 72.75%, *precision* 95.30% dan *recall* 75.43%.

4.3 Skenario Pengujian 3

Pada Pengujian 1 dilakukan untuk mengetahui performa *K-Nearest Neighbor* dengan menggunakan atribut mata kuliah hanya wajib dengan menggunakan model *k-fold cross validation* dengan nilai k = 10 dan tetangga terdekat sejumlah k=3 sampai dengan k=29 dengan nilai k tetangga bernilai ganjil. Perhitungan yang dilakukan akan menghasilkan nilai akurasi, *precision* dan *recall*. Hasil skenario pengujian 3 dapat dilihat pada Gambar 7.



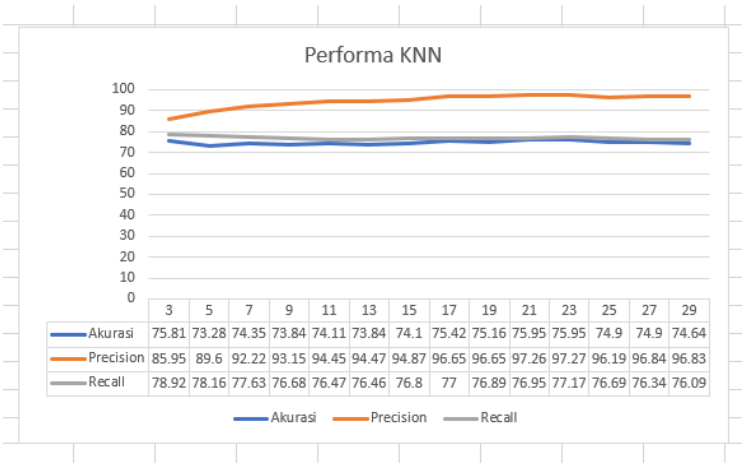
	3	5	7	9	11	13	15	17	19	21	23	25	27	29
akurasi	69.01	71.38	70.04	69.79	70.07	70.33	70.05	68.48	68.74	68.73	70.34	70.86	70.34	69.8
precision	73.38	78.45	77.76	80.2	80.8	82.03	83.41	83.43	86.21	86.3	88.04	88.4	88.84	88.23
recall	82.32	82.34	81.26	79.11	78.38	78.85	78.39	77.1	76.4	76.85	77.07	77.22	76.64	76.78

Gambar 7. Skenario Pengujian 3

Dari Gambar di atas ditampilkan hasil rata-rata pada akurasi, *precision* dan *recall*. Parameter yang dijadikan performa akurasi. Sehingga didapat *k* terbaik untuk skenario 3 pada *k*=5 dengan performa akurasi 71.38%, *precision* 78.45% dan *recall* 82.34%.

4.4 Skenario Pengujian 4

Pada Pengujian 1 dilakukan untuk mengetahui performa *K-Nearest Neighbor* dengan menggunakan atribut semua mata kuliah pilihan dengan menggunakan model *k-fold cross validation* dengan nilai k = 10 dan tetangga terdekat sejumlah k=3 sampai dengan k=29 dengan nilai k tetangga bernilai ganjil. Perhitungan yang dilakukan akan menghasilkan nilai akurasi, *precision* dan *recall*. Hasil skenario pengujian 4 dapat dilihat pada Gambar 8.



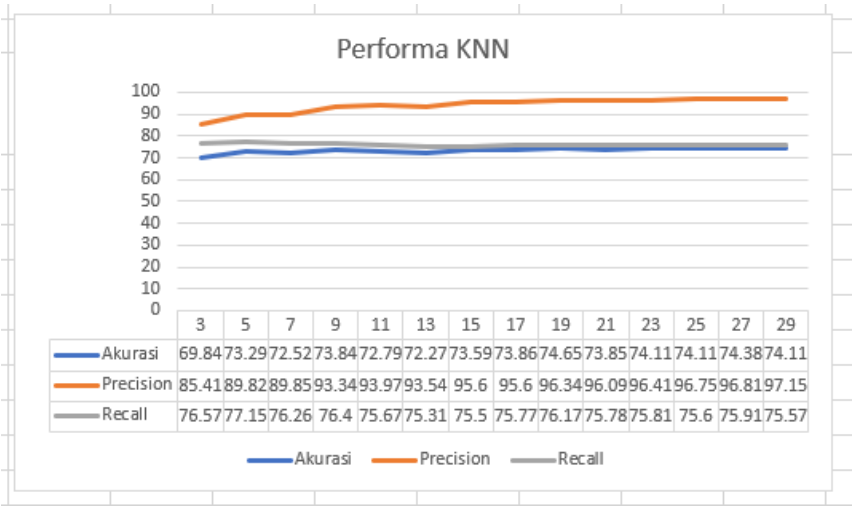
	3	5	7	9	11	13	15	17	19	21	23	25	27	29
Akurasi	75.81	73.28	74.35	73.84	74.11	73.84	74.1	75.42	75.16	75.95	75.95	74.9	74.9	74.64
Precision	85.95	89.6	92.22	93.15	94.45	94.47	94.87	96.65	96.65	97.26	97.27	96.19	96.84	96.83
Recall	78.92	78.16	77.63	76.68	76.47	76.46	76.8	77	76.89	76.95	77.17	76.69	76.34	76.09

Gambar 8. Skenario Pengujian 4

Dari Gambar di atas ditampilkan hasil rata-rata pada akurasi, *precision* dan *recall*. Parameter yang dijadikan performa akurasi. Sehingga didapat *k* terbaik untuk skenario 4 pada *k*=23 dengan performa akurasi 75.95%, *precision* 97.27% dan *recall* 77.17%.

4.5 Skenario Pengujian 5

Pada Pengujian 1 dilakukan untuk mengetahui performa *K-Nearest Neighbor* dengan menggunakan atribut mata kuliah relasi dengan menggunakan model *k-fold cross validation* dengan nilai k = 10 dan tetangga terdekat sejumlah k=3 sampai dengan k=29 dengan nilai k tetangga bernilai ganjil. Perhitungan yang dilakukan akan menghasilkan nilai akurasi, *precision* dan *recall*. Hasil skenario pengujian 5 dapat dilihat pada Gambar 9.



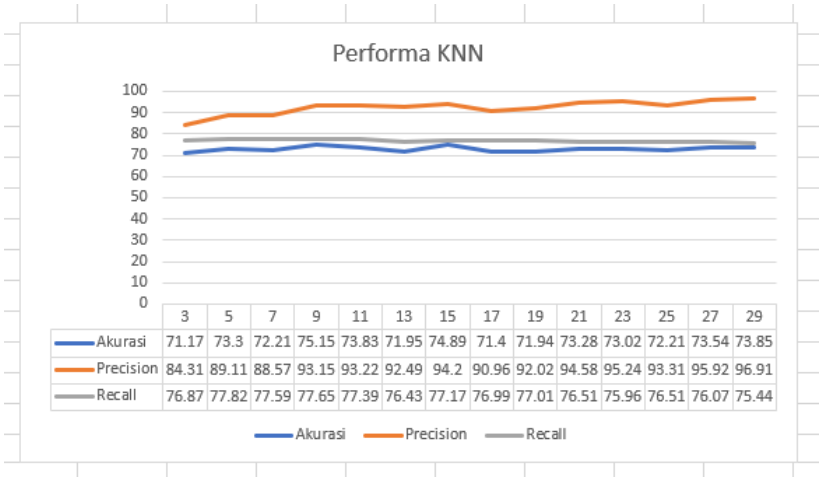
	3	5	7	9	11	13	15	17	19	21	23	25	27	29
Akurasi	69.84	73.29	72.52	73.84	72.79	72.27	73.59	73.86	74.65	73.85	74.11	74.11	74.38	74.11
Precision	85.41	89.82	89.85	93.34	93.97	93.54	95.6	95.6	96.34	96.09	96.41	96.75	96.81	97.15
Recall	76.57	77.15	76.26	76.4	75.67	75.31	75.5	75.77	76.17	75.78	75.81	75.6	75.91	75.57

Gambar 9. Skenario Pengujian 5

Dari Gambar di atas ditampilkan hasil rata-rata pada akurasi, *precision* dan *recall*. Parameter yang dijadikan performa akurasi. Sehingga didapat *k* terbaik untuk skenario 4 pada *k*=19 dengan performa akurasi 74.65%, *precision* 96.34% dan *recall* 76.17%

4.6 Skenario Pengujian 6

Pada Pengujian 1 dilakukan untuk mengetahui performa *K-Nearest Neighbor* dengan menggunakan atribut mata kuliah komdas dengan menggunakan model *k-fold cross validation* dengan nilai k = 10 dan tetangga terdekat sejumlah k=3 sampai dengan k=29 dengan nilai k tetangga bernilai ganjil. Perhitungan yang dilakukan akan menghasilkan nilai akurasi, *precision* dan *recall*. Hasil skenario pengujian 6 dapat dilihat pada Gambar 10.

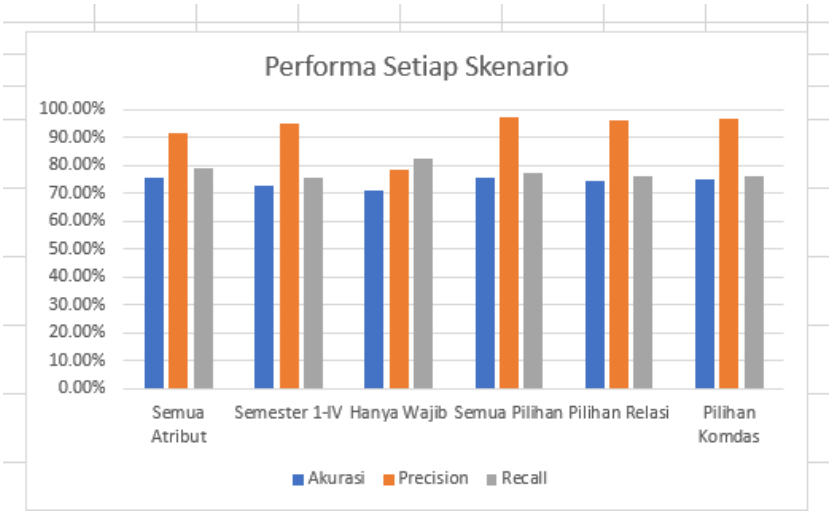


	3	5	7	9	11	13	15	17	19	21	23	25	27	29
Akurasi	71.17	73.3	72.21	75.15	73.83	71.95	74.89	71.4	71.94	73.28	73.02	72.21	73.54	73.85
Precision	84.31	89.11	88.57	93.15	93.22	92.49	94.2	90.96	92.02	94.58	95.24	93.31	95.92	96.91
Recall	76.87	77.82	77.59	77.65	77.39	76.43	77.17	76.99	77.01	76.51	75.96	76.51	76.07	75.44

Gambar 10. Skenario Pengujian 6

Dari Gambar di atas ditampilkan hasil rata-rata pada akurasi, *precision* dan *recall*. Parameter yang dijadikan performa akurasi. Sehingga didapat *k* terbaik untuk skenario 4 pada *k*=9 dengan performa akurasi 75.15%, *precision* 93.15% dan *recall* 77.65%.

Dari ke enam skenario di atas dapat dianalisa bahwa kategori mata kuliah yang memiliki informasi akurat mengenai masa studi adalah mata kuliah pilihan yang memiliki nilai akurasi tertinggi dibanding dengan skenario lainnya yaitu 75.95%. Grafik akurasi terbaik pada setiap skenario dapat dilihat pada Gambar 11.



Gambar 11. Grafik Performa Terbaik Setiap Skenario

5. KESIMPULAN

Kesimpulan yang dapat diambil dari penelitian mengenai klasifikasi masa studi mahasiswa menggunakan algoritma *K-Nearest Neighbor* adalah sebagai berikut:

a. Dengan mengacu proses *data mining Knowledge Discovery Databases* telah dibangun sebuah perangkat lunak yang dapat melakukan klasifikasi masa studi mahasiswa.
b. Dari enam skenario percobaan yang telah dilakukan diperoleh nilai akurasi tertinggi pada skenario yang menggunakan atribut mata kuliah pilihan yaitu 75.95%.
c. Berdasarkan nilai akurasi tertinggi menggunakan semua mata kuliah pilihan dapat disimpulkan bahwa mata kuliah pilihan sangat berpengaruh pada masa studi mahasiswa.

DAFTAR PUSTAKA

[1] BAN-PT, "Buku 3B-Borang Fakultas-Sekolah Tinggi (Versi 08-04-2010)," *Ban-Pt*. 2008.
[2] A. G. Novianti and D. Prasetyo, "Penerapan Algoritma K-Nearest Neighbor (K-NN) untuk Prediksi Waktu Kelulusan Mahasiswa," *Semin. Nas. APTIKOM*, no. November, pp. 108–113, 2017.
[3] Y. S. Nugroho, "DATA MINING MENGGUNAKAN ALGORITMA NAÏVE BAYES UNTUK KLASIFIKASI KELULUSAN MAHASISWA UNIVERSITAS DIAN NUSWANTORO," vol. 75, no. 3 PART A, pp. 1–3, 2014.
[4] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated Web Usage Data Mining and Recommendation System Using K-Nearest Neighbor (KNN) Classification Method," *Appl. Comput. Informatics*, 2016.
[5] S. Jambekar and Z. Saquib, "Application of Data Mining Techniques for Prediction of Crop Production in India," vol. 7, no. 4, pp. 66–69, 2018.
[6] U. Shafique and H. Qaiser, "A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)," *Int. J. Innov. Sci. Res. ISSN*, vol. 12, no. 1, pp. 2351–8014, 2014.
[7] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, vol. 84. 2013.
[8] T. Rismawan, A. W. Irawan, W. Prabowo, and S. Kusumadewi, "Sistem Pendukung Keputusan Berbasis Pocket PC Sebagai Penentu Status Gizi Menggunakan Metode KNN (K-Nearest Neighbor)," *teknoin*, vol. 13, pp. 18–23, 2008.
[9] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Appear. Int. Jt. Conf. Articial Intell.*, vol. 5, pp. 1–7, 1995.
[10] Y. Zhang and S. Wang, "Detection of Alzheimer's Disease by Displacement Field and Machine Learning," *PeerJ*, vol. 3, p. e1251, 2015.
[11] H. Zhou, Z. Deng, Y. Xia, and M. Fu, "A New Sampling Method in Particle Filter Based on Pearson Correlation Coefficient," *Neurocomputing*, vol. 216, pp. 208–215, 2016.
[12] Sugiyono, *Statistik Untuk Penelitian*. 2007.
[13] R. U. Diponegoro, "Peraturan Akademik Bidang Pendidikan Program Sarjana Universitas Diponegoro," 2017.
[14] T. Hendrawati, "Kajian Metode Imputasi dalam Menangani Missing Data," *Pros. Semin. Nas. Mat. dan Pendidik. Mat. UMS*, pp. 637–642, 2015.
[15] E. Acuna and C. Rodriguez, "The Treatment of Missing Values and Its Effect in The Classifier Accuracy," *Classif. Clust. Data Min. Appl.*, no. 639–647, pp. 1–9, 2004.