JuSIN
Jurnal Sistem Informasi

Website:ojs.itb-ad.ac.id/index.php/JUSIN
E-ISSN: 2797- 8516

IMPLEMENTASI DATA MINING UNTUK MENENTUKAN MINAT SISWA DALAM MENENTUKAN JURUSAN PADA PERGURUAN TINGGI

Saeful Bahri[1] [(*)]

[1]*ITB Ahmad Dahlan, Jakarta*

Abstract

Class XII high school students are students who occupy a period of formal education before entering lectures, students who are in the teenage age range where at this age they have to make decisions for their future. In making these decisions, adolescents are often accompanied by confusion, uncertainty, and even stress so that making decisions that result in regrets in the future. Every year there are many vocational high school students who want to go to a college but do not know what major they want and apply in the world of work according to their talents and interests, so there are still many students who make decisions that are not in accordance with their interests and talents. make decisions based on the opinions of parents, friends or others. For this reason, a model is needed to classify these problems. In this study, three classification algorithms were used: decision tree, nave Bayes, and k-nearest neighbor with data mining techniques to find patterns from the model used, the results of this study are expected to help students determine the majors to be taken in lectures. From the results of this research test, the factor that most influences errors in majoring in college is the variable of majoring based on (self/friends/parents), and of the three algorithms used, the decision tree algorithm is the best algorithm with a high level of accuracy 75.38%.

Kata Kunci: Data mining, Minat Siswa

Januari – Juni 2022, Vol 3 (1) : hlm 23-33

[(*)] Korespondensi: *mr.saeful.bahri@gmail.com (Saeful Bahri)*

PENDAHULUAN

Siswa SMA adalah individu yang sedang mengalami masa remaja akhir (*late adolescence*). Adolesence atau remaja (dalam Hurlock, 1980) berasal dari kata Latin adolescere (kata bendanya, adolescentia yang berarti remaja) yang berarti "tumbuh" atau "tumbuh menjadi dewasa." Istilah adolesence, mencakup kematangan mental, emosi, sosial, dan fisik, perubahan pada biologis, kognitif dan sosial emosional yang terjadi berkisar dari perkembangan fungsi seksual, proses berfikir abstrak sampai pada kemandirian (Setio et al., 2020).

Siswa SMA kelas XII adalah pelajar yang menduduki masa pendidikan formal sebelum memasuki bangku perkuliahan, siswa yang berada pada rentang usia remaja yang mana pada usia ini mereka sudah harus membuat keputusan untuk masa depannya. Dalam pengambilan keputusan tersebut remaja sering disertai kebingungan, ketidakpastian, dan bahkan stress sehingga membuat keputusan yang diambil mengakibatkan penyesalan dikemudian hari. Setiap tahun banyak siswa sekolah menengah kejuruan yang ingin masuk ke suatu perguruan tinggi tetapi tidak tahu jurusan apa yang di inginkan dan di terapkan didunia kerja sesuai dengan bakat dan minat yang dimiliki, sehingga masih banyak siswa yang mengambil keputusan tidak sesuai dengan minat dan bakatnya yang akhirnya membuat pengambilan keputusan berdasarkan pendapat orang tua, teman atau orang lain.

Pengambilan jurusan pada bangku perkuliahan yang tidak tepat menyebabkan proses perkuliahan tidak efektif, prihal ini banyak menyebabkan mahasiswa yang putus perkuliahan di tengah jalan dikarenakan merasa salah jurusan yang tidak sesuai dengan minat dan bakat yang dimiliki. Berdasarkan permasalahan yang disampaikan, dibutuhkan penelitian untuk membantu siswa melakukan klasifikasi terhadap minat dan bakat siswa dalam menentukan jurusan yang akan di ambil di bangku perkuliahan.

Klasifikasi digunakan untuk memprediksi sebuah data baru berdasarkan data-data sebelumnya (Fadma Ristianti, 2019). Salah satu metode yang digunakan adalahal goritma C4.5. Algoritma C4.5 adalah algoritma yang digunakan untuk membentuk pohon keputusan (Decision Tree). Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Pohon keputusan berguna untuk mengekspolari data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target (Setio et al., 2020). Oleh karena itu, penelitian ini memberikan hasil klasifikasi kelompok minat untuk pemilihan jurusan kuliah yang sesuai dengan minat dengan melakukan komparasi algoritma klasifikasi yaitu *Decision tree*, *Naive bayes*, dan *K-NN*.

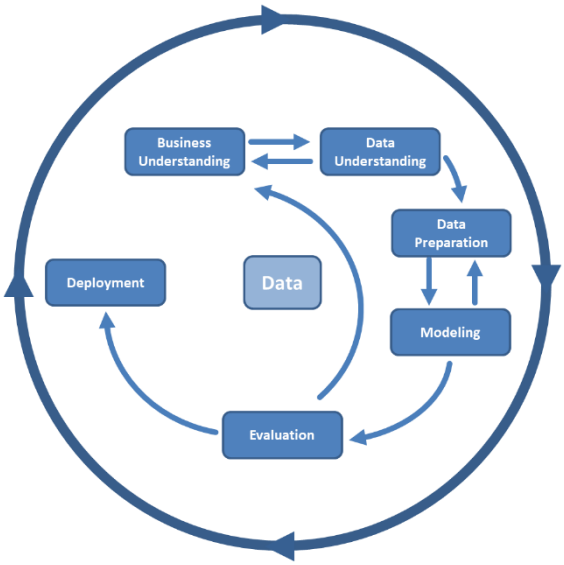
Penelitian terkait yang terdahulu membahas tentang Perbandingan Dan Analisis Metode Klasifikasi Untuk Menentukan Konsentrasi Jurusan. Hasil yang didapat menggunkan Metode yang digunakan yaitu C4.5 dan Naïve Bayes dengan menggunakan aplikasi Rapid Miner sebagai alat bantu untuk mengklasifikasikan penjurusan mahasiswa. Pada penelitian ini diketahui algoritma C4.5 memiliki tinggkat akurasi 48,06 % dan naïve bayes 42,79% (Hidayanti et al., 2020).

Penelitian terkait yang terdahulu membahas tentang Implementasi Metode K-Means Clustering Dalam Menentukan Bidang Studi Perguruan Tinggi Di Smk Negeri 2 Kota Jambi. Metode yang digunakan adalah metode k-means clustering dengan 33 atribut dan 11 cluster yaitu dengan hasil cluster pertama terdapat 25 data siswa dengan hasil persentase 12% masuk

dalam Bidang Seni, cluster kedua terdapat 29 data siswa dengan hasil presentase 14% masuk dalam Bidang Pendidikan, cluster ketiga terdapat 15 data siswa dan hasil presentase 7% masuk dalam Bidang Ekonomi dan Bisnis, cluster keempat terdapat 10 data siswa dan hasil presentase 5% masuk dalam Bidang Perhutanan, cluster kelima terdapat 18 data siswa dan hasil presentase 9% masuk dalam Bidang Kedokteran, cluster keenam terdapat 20 data siswa dan hasil presentase 10% masuk dalam Bidang Olahraga, cluster ketujuh terdapat 27 data siswa dan hasil presentase 13% masuk dalam Bidang Teknik, cluster kedelapan terdapat 24 data siswa dan hasil presentase 12% masuk dalam Bidang Komputer, cluster kesembilan terdapat 13 data siswa dan hasil presentase 6% masuk dalam Bidang Bahasa, cluster kesepuluh terdapat 12 data siswa dan hasil presentase 6% masuk dalam Bidang Pertanian (Prasanti et al., 2020).

METODE

Metode penelitian yang digunakan dalam penelitian ini adalah CRISP-DM (*Cross Industry Standard Process For Data Mining*) merupakan salah satu model dalam proses penambangan data (*data mining*) yang memiliki enam tahapan (Mem et al., 2022):



Gambar 1. CRISP-DM

Enam tahapan pada CRISP-DM (*Cross Industry Standard Process For Data Mining*) (Muttaqin et al., 2022):

1. Pemahaman Bisnis (*Business Undertanding*), memahami tujuan dan kebutuhan dari sudut pandang bisnis, kemudian menterjemahkan pengetahuan ini ke dalam pendefinisian masalah pada data mining, selanjutnya akan ditentukan rencana dan strategi untuk mencapai tujuan tersebut.
2. Pemahaman Data (*Data Understanding*), tahap ini dimulai dengan pengumpulan data yang kemudian akan dilanjutkan dengan proses untuk mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi adanya bagian yang menarik dari data yang dapat digunakan untuk hipotesa untuk informasi yang tersembunyi.

3. Persiapan Data (*Data Preperation*), tahap ini meliputi semua kegiatan untuk membangun dataset akhir (data yang akan diproses pada tahap pemodelan) dari data mentah. Tahap ini dapat diulang beberapa kali. Pada tahap ini juga mencakup pemilihan tabel, *record*, dan atribut-atribut data, termasuk proses pembersihan dan transformasi data untuk kemudian dijadikan masukan dalam tahap pemodelan.
4. Pemodelan (*Modelling*), tahap ini dilakukan pengujian dari variabel yang diguakan untuk mendapatkan nilai yang optimal, dan dilakukan pemilihan model dan penerapan menggunakan teknik data mining. Model yang digunakan dalam penelitian ini adalah:

a. *Decision Tree*

Algoritma C4.5 merupakan salah satu algoritma yang digunakan untuk melakukan klasifikasi atau segmentasi yang bersifat prediktif.. Algoritma C4.5 digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal handal. Pohon keputusan mengubah fakta yang sangat besar menjadi aturan-aturan, dengan begitu aturan tersebut dapat dengan mudah dipahami (Marlina & Bakri, 2021).

Menurut kusrini dalam bukunya, secara umum algoritma C4.5 membangun pohon keputusan terdiri dari beberapa tahapan. Berikut garis besarnya.

a. Pilih atribut sebagai akar
b. Buat cabang untuk tiap-tiap nilai
c. Bagi kasus dalam cabangs
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Memilih atribut sebagai simpul akar Pemilihan atribut yang akan dijadikan akar adalah dengan menghitung nilai *gain* dari semua atribut. Dan yang dipilih mejadi akar pertama adalah yang memiliki nilai *gain* tertinggi. Namun sebelum menentukan nilai *gain*, terlebih dahulu hitung nilai entropy. Penentuan nilai *entropy* menggunakan persamaan berikut.

$$Entropy\ \ (S) = \sum_{j=1}^{n} pi.\log 2\text{pi}$$

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

Setelah itu tentukan nilai gain menggunakan persamaan.

$$Gain\ (S,A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} + entropy(Si)$$

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

b. *Naïve Bayes*

Pengklasifikasi Naive Bayes adalah pengklasifikasi paling sederhana dan paling umum digunakan. Model klasifikasi Naive Bayes menghitung probabilitas posterior suatu kelas berdasarkan distribusi kata dalam dokumen. Hal itu bergantung pada representasi dokumen yangsangat sederhana sebagai *Bag of words*. Model ini bekerja dengan mengekstraksi fitur *bag of words* yang mengabaikan posisi kata dalam dokumen. Ini menggunakan Teorema Bayes untuk memprediksi probabilitas bahwa set fitur yang diberikan milik label tertentu (Yulita et al., 2021).

Naïve Bayes merupakan machine learning yang menggunakan perhitungan probabilitas yang menggunakan konsep pendekatan Bayes. Penggunaan teorema Bayes pada algoritma Naïve Bayes adalah dengan menggabungkan *prior probability* dan conditional *probability* dalam suatu rumus yang dapat digunakan untuk menghitung probabilitas dari setiap kemungkinan klasifikasi (Alvina Felicia Watratan et al., 2020).

c. *K-Nearest Neighboar*

Algoritma K-Nearest Neighbor **(**KNN) merupakan metode pengklasifikasian kumpulan data pembelajaran berdasarkan jarak yang paling dekat dengan object. Algoritma KNN mengelompokkan data baru sesuai dengan nilai k sebagai jarak tetangga terdekat antara data latih dengan data uji. Jauh atau dekatnya tetangga dapat dihitung menggunakan jarak euclidean dengan rumus (Mubarok, 2021).

$$(x, y) = \sqrt{\sum_{i=1}^{p}(xi - yi)}$$

Keterangan :
d(x,y) : Jarak data latih dan uji,
xi : Data latih,
yi : Data uji,
i : Variabel data,
p : Dimensi data.

5. Evaluasi (Evaluation)

Pada tahap ini dilakukan evaluasi untuk mengetahui akurasi dari model yang diusulkan. Selain itu dilakukan validasi dengan teknik 10 *fold Cross Validation, K-fold Cross Validation* yang merupakan teknik validasi dengan membagi data awal secara acak kedalam *k* bagian yang saling terpisah atau "*fold".* Evaluasi dan validasi menggunakan metode

confusion matrix dan kurva ROC. Grafik *Receiver Operating Characteristics* (ROC) adalah teknik untuk memvisualisasikan, mengorganisasikan dan memilih pengklasifikasi berdasarkan kinerja setiap algoritma. Kurva ROC digunakan untuk mengukur nilai *Area Under Curve* (AUC). Nilai akurasi algortima diukur menggunakan *confusion matrix* dan hasil peritungan akan ditampilkan dalam bentuk kurva ROC (Nasution et al., 2019).

6. Penyebaran (*Deployment*)

Pada tahap ini, pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Tahap deployment dapat berupa pembuatan laporan sederhana atau mengimplementasikan proses data mining yang berulang dalam perusahaan. Pada banyak kasus, tahap *deployment* melibatkan konsumen, di samping analis data, karena sangat penting bagi konsumen untuk memahami tindakan apa yang harus dilakukan untuk menggunakan model yang telah dibuat.

HASIL DAN PEMBAHASAN

1. *Business Understanding*

Permasalahan pada penelitian ini, banyaknya mahasiswa yang keluar di tengah perkuliahan yang sedang dilaksanakan dikarenakan merasa salah mengambil jurusan yang tidak sesuai dengan bakat dan minat. Pada penelitian ini dilakukan klasifikasi permasalahan untuk menentukan variabel yang paling mempengaruhi dalam menentukan jurusan pada perguruan tinggi untuk membantu siswa.

2. *Data Understanding*

Data yang digunakan pada penelitian ini diperoleh dari hasil kuisioner pada beberapa perguruan tinggi di wilayah Tangerang.

NAMA	JENIS KELAMIN	JURUSAN SMA/SMK/MA	PRODI	SEMESTER	PENGAMBILAN PRODI	PRODI SESUAI BAKAT	BIAYA KULIAH	BEKERJA	USIA	FASILITAS KAMPUS
Iyan Bachtiar Hadiyatna	Pria	TKR	Teknologi informasi	5	Diri sendiri	Ya	Murah	Tidak	22	Tidak lengkap
Lugyana Salsabila	Perempuan	SMA-IPA	Psikologi	5	Diri sendiri	Ya	Mahal	Tidak	21	Lengkap
Yusuf Adhitia Rachman	Pria	Teknik Kendaraan Ringan	PGSD	8	Orang tua	Ya	Mahal	Tidak	20	Lengkap
Aliyudin	Pria	MA	TI	5	Mengikuti Teman	Tidak	Mahal	Ya	22	Lengkap
Yustinus Laia	Pria	IPS	Teknik Informatika	5	Diri sendiri	Ya	Mahal	Ya	22	Lengkap
Zahra	Perempuan	SMK	Psikologi	5	Diri sendiri	Tidak	Mahal	Tidak	22	Tidak lengkap
Iyan Bachtiar Hadiyatna	Pria	TKR	Teknologi informasi	5	Diri sendiri	Ya	Murah	Tidak	22	Tidak lengkap
Lugyana Salsabila	Perempuan	SMA-IPA	Psikologi	5	Diri sendiri	Ya	Mahal	Tidak	21	Tidak lengkap
Yusuf Adhitia Rachman	Pria	Teknik Kendaraan Ringan	PGSD	8	Orang tua	Ya	Mahal	Tidak	23	Lengkap
Aliyudin	Pria	MA	TI	5	Mengikuti Teman	Tidak	Mahal	Ya	22	Tidak lengkap
Yustinus Laia	Pria	IPS	Teknik Informatika	5	Diri sendiri	Ya	Mahal	Ya	24	Lengkap
Zahra	Perempuan	SMK	Psikologi	5	Diri sendiri	Tidak	Mahal	Tidak	21	Tidak lengkap
Amonio giawa	Pria	SMA-IPA	Teknologi Informasi	5	Diri sendiri	Tidak	Murah	Tidak	25	Tidak lengkap
Hanny Rahmadayanthi	Perempuan	IPA	Manajemen	3	Mengikuti Teman	Tidak	Mahal	Ya	21	Tidak lengkap
Nur Amalia Fajriah	Perempuan	Multimedia	Pendidikan Guru Sekolah Dasar	8	Diri sendiri	Ya	Mahal	Ya	22	Lengkap
Habani halawa	Pria	Ips	Teknik informatika	1	Diri sendiri	Ya	Murah	Tidak	20	Lengkap
Alif	Pria	Sma	Teknologi informasi	5	Diri sendiri	Ya	Murah	Ya	21	Lengkap
La Ode Zoe Tumada	Pria	IPS	Manajemen	7	Diri sendiri	Ya	Murah	Tidak	24	Lengkap
Retno widyastuti	Perempuan	Farmasi	Akuntansi	8	Diri sendiri	Tidak	Mahal	Tidak	26	Lengkap
Adji Sapta Pangestu	Pria	Akuntansi	Ilmu Komunikasi	2	Diri sendiri	Ya	Murah	Tidak	20	Lengkap
RYAN YULIADI	Pria	Sma	Pgsd	8	Diri sendiri	Ya	Murah	Ya	22	Lengkap

Tabel 1. Dataset

3. *Data Preperation*

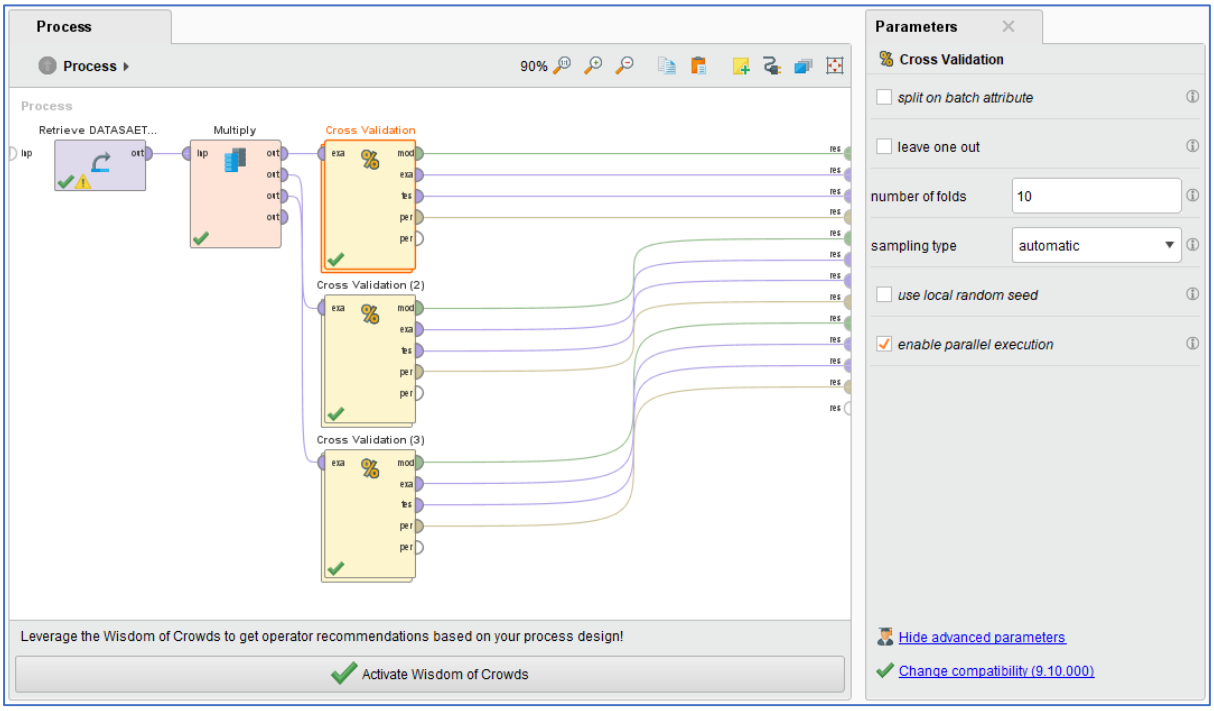
Setelah proses pengumpulan data, data dipersiapkan untuk dilakukan tahap pengujian. Pada tahap ini mencakup pemilihan tabel, record, dan variabel-variabel data, termasuk proses pembersihan variabel yang tidak akan mempengaruhi hasil akhir.

Nama	Jenis	Jurusan	Prodi	Semeste	Pengambilan	Prodi Sesuai bakat	Biaya	Faktor	Usia	Fasilitas	Label
A1	Pria	TKR	Teknologi informasi	5	Diri sendiri	Sesuai Bakat	Murah	Tidak Bekerja	22	Tidak lengkap	Tidak
A2	Perempuan	SMA-IPA	Psikologi	5	Diri sendiri	Sesuai Bakat	Mahal	Tidak Bekerja	21	Lengkap	Tidak
A3	Pria	Teknik Kendaraan Ringan	PGSD	8	Orang tua	Sesuai Bakat	Mahal	Tidak Bekerja	20	Lengkap	Tidak
A4	Pria	MA	TI	5	Mengikuti Teman	Tidak Sesuai Bakat	Mahal	Bekerja	22	Lengkap	Salah Jurusan
A5	Pria	IPS	Teknik Informatika	5	Diri sendiri	Sesuai Bakat	Mahal	Bekerja	22	Lengkap	Tidak
A6	Perempuan	SMK	Psikologi	5	Diri sendiri	Tidak Sesuai Bakat	Mahal	Tidak Bekerja	22	Tidak lengkap	Tidak
A7	Pria	TKR	Teknologi informasi	5	Diri sendiri	Sesuai Bakat	Murah	Tidak Bekerja	22	Tidak lengkap	Tidak
A8	Perempuan	SMA-IPA	Psikologi	5	Diri sendiri	Sesuai Bakat	Mahal	Tidak Bekerja	21	Tidak lengkap	Salah Jurusan
A9	Pria	Teknik Kendaraan Ringan	PGSD	8	Orang tua	Sesuai Bakat	Mahal	Tidak Bekerja	23	Lengkap	Tidak
A10	Pria	MA	TI	5	Mengikuti Teman	Tidak Sesuai Bakat	Mahal	Bekerja	22	Tidak lengkap	Tidak
A11	Pria	IPS	Teknik Informatika	5	Diri sendiri	Sesuai Bakat	Mahal	Bekerja	24	Lengkap	Tidak
A12	Perempuan	SMK	Psikologi	5	Diri sendiri	Tidak Sesuai Bakat	Mahal	Tidak Bekerja	21	Tidak lengkap	Salah Jurusan
A13	Pria	SMA-IPA	Teknologi Informasi	5	Diri sendiri	Tidak Sesuai Bakat	Murah	Tidak Bekerja	25	Tidak lengkap	Tidak
A14	Perempuan	IPA	Manajemen	3	Mengikuti Teman	Tidak Sesuai Bakat	Mahal	Bekerja	21	Tidak lengkap	Tidak
A15	Perempuan	Multimedia	Pendidikan Guru Sekolah Dasar	8	Diri sendiri	Sesuai Bakat	Mahal	Bekerja	22	Lengkap	Tidak
A16	Pria	Ips	Teknik informatika	1	Diri sendiri	Sesuai Bakat	Murah	Tidak Bekerja	20	Lengkap	Salah Jurusan
A17	Pria	Sma	Teknologi informasi	5	Diri sendiri	Sesuai Bakat	Murah	Bekerja	21	Lengkap	Tidak
A18	Pria	IPS	Manajemen	7	Diri sendiri	Sesuai Bakat	Murah	Tidak Bekerja	24	Lengkap	Tidak
A19	Perempuan	Farmasi	Akuntansi	8	Diri sendiri	Tidak Sesuai Bakat	Mahal	Tidak Bekerja	26	Lengkap	Tidak
A20	Pria	Akuntansi	Ilmu Komunikasi	2	Diri sendiri	Sesuai Bakat	Murah	Tidak Bekerja	20	Lengkap	Tidak
A21	Pria	Sma	Pgsd	8	Diri sendiri	Sesuai Bakat	Murah	Bekerja	22	Lengkap	Salah Jurusan
A22	Perempuan	IPA	PGSD	8	Mengikuti Teman	Sesuai Bakat	Mahal	Tidak Bekerja	22	Lengkap	Salah Jurusan
A23	Pria	PGSD	PGSD	8	Diri sendiri	Sesuai Bakat	Mahal	Bekerja	23	Lengkap	Tidak
A24	Perempuan	MA	KEGURUAN ILMU PENDIDIKAN	8	Diri sendiri	Sesuai Bakat	Mahal	Tidak Bekerja	21	Tidak lengkap	Salah Jurusan

Tabel 2. Final Dataset

4. *Modeling*

Proses pemodelan merupakan proses pengujian model *decision tree*, *naïve bayes*, dan *k-nearest neighbor* setelah melalui proses *data preparation, dilanjutkan* proses *Set Role* yang berfungsi untuk menentukan *label*, kemudian menggunakan validasi 10 *fold cross Validation* dalam proses *training*, sedangkan untuk proses *testing* menggunakan *apply model* dan *performance*. Proses *training* dan *testing* untuk mendapatkan *confusion matrix* yaitu nilai tingkat *accuracy*, *class precision*, *class recall*, dan kurva ROC yaitu nilai AUC.



Gambar 2. Pengujian Model

4.1. Hasil pengujian algoritma *Decision Tree*

accuracy: 75.38% +/- 8.94% (micro average: 75.21%)

	true Tidak	true Salah Jurusan	class precision
pred. Tidak	71	16	81.61%
pred. Salah Jurusan	13	17	56.67%
class recall	84.52%	51.52%	

Dari hasil pebgujian model *decision tree* di atas, didapatkan hasil akurasi 75.38% yang artinya tingkat akurasi data sudah baik.

4.2. Hasil pengujian algoritma *Naïve Bayes*

accuracy: 60.98% +/- 12.52% (micro average: 60.68%)

	true Tidak	true Salah Jurusan	class precision
pred. Tidak	53	15	77.94%
pred. Salah Jurusan	31	18	36.73%
class recall	63.10%	54.55%	

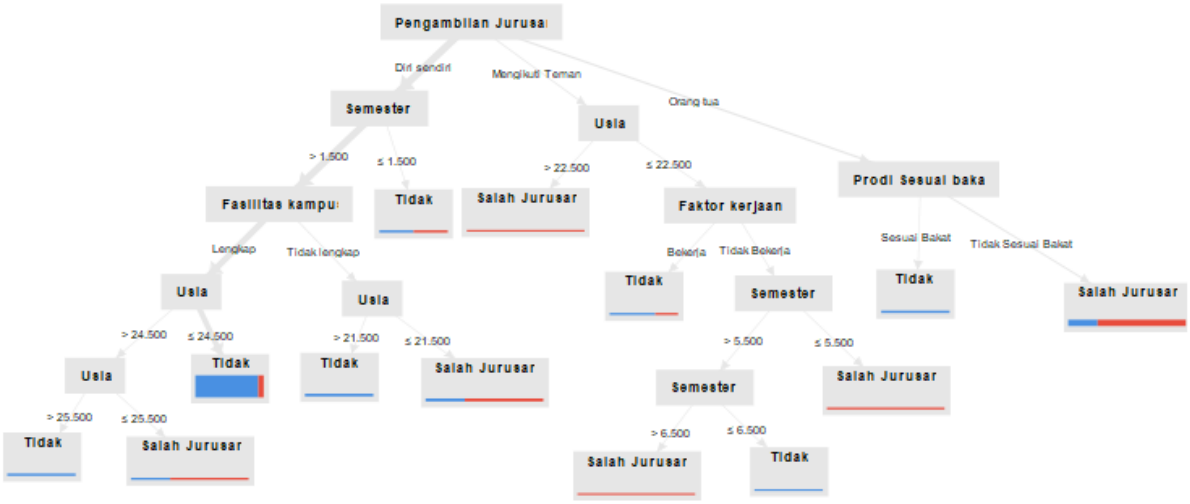
Dari hasil pebgujian model *naïve bayes* di atas, didapatkan hasil akurasi 60.98% yang artinya tingkat akurasi data buruk.

4.3. Hasil pengujian algoritma *k-Nearest Neighbor*

accuracy: 71.67% +/- 9.45% (micro average: 71.79%)

	true Tidak	true Salah Jurusan	class precision
pred. Tidak	76	25	75.25%
pred. Salah Jurusan	8	8	50.00%
class recall	90.48%	24.24%	

Dari hasil pebgujian model *k-Nearest Neighbor* di atas, didapatkan hasil akurasi 71.67% yang artinya tingkat akurasi data sudah baik.



Gambar 3. Pohon Keputusan

Output pohon keputusan yang dihasilkan diatas, menunjukan faktor yang paling mempengaruhi kesalahan dalam pengambila jurusan pada perguruan tinggi adalah variabel "pengambilan jurusan berdasarkan (diri sendiri/teman/orang tua)".

a. Apabila pengambilan jurusan berdasarkan minat diri sendiri, maka akan di tentukan variabel semester, semseter <1.5 maka tidak salah jurusan, apabila semester >1.5 ditentukan fasilitas kampus.
b. Apabila pengambilan jurusan berdasarkan mengikuti teman, maka akan di tentukan variabel usia, usia >22.5 maka tidak salah jurusan, apabila usia <22.5 ditentukan vairabel bekerja atau tidak bekerja.
c. Apabila pengambilan jurusan berdasarkan permintaan orang tua, maka akan di tentukan variabel bakat, apabila bakat sesuai dengan jurusan yang di ambil maka tidak salah jurusan, apabila tidak sesuai bakat makan akan salah jurusan.

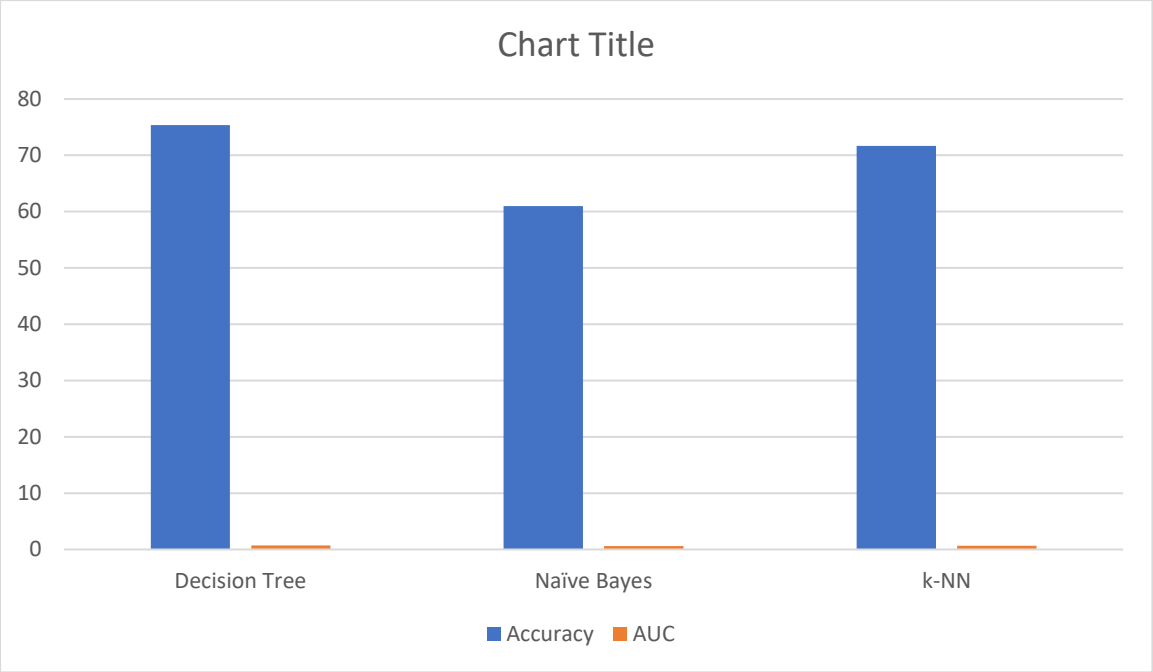
5. *Evaluation*

Perbandingan hasil komparasi *Accuracy* dan AUC dari beberapa model yang telah digunakan adalah sebagai berikut:

No	*Model*	*Accuracy*	AUC
1	***Decision Tree***	**75,38%**	**0.689**
2	*Naïve Bayes*	60,98%	0.625
3	*k-NN*	71,67%	0.641

Tabel 2. Komparasi Model Algoritma

Sumber: analisis literatur, 2022



KESIMPULAN

1. Dari hasil pengujian tiga algoritma yang digunakan, kinerja algoritma *decision tree* menjadi yang terbaik dibandingkan dengan algoritma *k-NN* dan *Naïve bayes* dengan tingkat akurasi 75.38% dan nilai AUC 0.689.
2. Output pohon keputusan yang dihasilkan dari dataset yang dugunakan, menunjukan faktor yang paling mempengaruhi mahasiswa salah dalam memilih jurusan pada perguruan tinggi adalah variabel "Pengambilan Jurusan berdasarkan Minat diri sendiri / mengikuti teman / permintaan orangtua".
3. Dari hasil peneltian ini, diharapkan mampu memberikan solusi bagi kampus, untuk membantu siswa dalam menentukan jurusan yang akan diambil pada pergurua tinggi, dengan melihat faktor-faktor yang mempengaruhi dalam menentukan jurusan.

DAFTAR PUSTAKA

Alvina Felicia Watratan, Arwini Puspita. B, & Dikwan Moeis. (2020). Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia. *Journal of Applied Computer Science and Technology*, *1*(1), 7–14. https://doi.org/10.52158/jacost.v1i1.9

Fadma Ristianti, D. (2019). Komparasi Algoritma Klasifikasi pada Data Mining. *PROCEEDINGS OF THE 1 St STEEEM*, *1*(1), 148–156.

Hidayanti, I., Kurniawan, T. B., & Afriyudi, A. (2020). Perbandingan Dan Analisis Metode Klasifikasi Untuk Menentukan Konsentrasi Jurusan. *Jurnal Ilmiah Informatika Global*,

11(1), 16–21. https://doi.org/10.36982/jig.v11i1.1067

Marlina, D., & Bakri, M. (2021). Penerapan Data Mining Untuk Memprediksi Transaksi Nasabah Dengan Algoritma C4.5. *Jurnal Teknologi Dan Sistem Informasi (JTSI)*, *2*(1), 23–28.

Mem, D., Kelulusan, P., Suatu, P., & Kuliah, M. (2022). *PENENTUAN KLASIFIKASI DENGAN CRISP-DM*. 826–831.

Mubarok, H. (2021). *Penerapan Algoritma K–Nearest Neighbor Untuk Klasifikasi Tingkat Kematangan Tomat Berdasarkan Fitur Warna Red Green Blue*. *April*, 773–782.

Muttaqin, M. R., Hermanto, T. I., Sunandar, M. A., Studi, P., Informatika, T., Tinggi, S., & Wastukancana, T. (2022). *PENERAPAN K-MEANS CLUSTERING DAN CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM) UNTUK*. *19*(1), 38–53.

Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *Computer Engineering, Science and System Journal*, *4*(1), 78. https://doi.org/10.24114/cess.v4i1.11458

Prasanti, R. W., Abidin, D. Z., & ... (2020). Implementasi Metode K-Means Clustering Dalam Menentukan Bidang Studi Perguruan Tinggi Di Smk Negeri 2 Kota Jambi. *Jurnal Ilmiah Mahasiswa …*, *2*(3), 209–221. http://ejournal.stikom-db.ac.id/index.php/jimti/article/view/888

Setio, P. B. N., Saputro, D. R. S., & Bowo Winarno. (2020). Klasifikasi Dengan Pohon Keputusan Berbasis Algoritme C4.5. *PRISMA, Prosiding Seminar Nasional Matematika*, *3*, 64–71.

Yulita, W., Dwi Nugroho, E., Habib Algifari, M., Studi Teknik Informatika, P., Teknologi Sumatera, I., Terusan Ryacudu, J., Huwi, W., Agung, J., & Selatan, L. (2021). Analisis Sentimen Terhadap Opini Masyarakat Tentang Vaksin Covid-19 Menggunakan Algoritma Naïve Bayes Classifier. *Jurnal Data Mining Dan Sistem Informasi*, *2*(2), 1–9. https://ejurnal.teknokrat.ac.id/index.php/JDMSI/article/view/1344