**Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)** DOI: 10.25126/jtiik.202072493
Vol. 7, No. 6, Desember 2020, hlm. 1237-1244 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

# PREDIKSI TINGKAT INDEKS PRESTASI KUMULATIF AKADEMIK MAHASISWA DENGAN MENGGUNAKAN TEKNIK *DATA MINING*

**Anita Desiani[1], Sugandi Yahdin[2], Desty Rodiah[3]**

[1,2,3]Universitas Sriwijaya
Email:[1]anita_desiani@unsri.ac.id, [2]sugandi@unsri.ac.id, [3]destyrodiah@gmail.com
*Penulis Korespondensi

(Naskah masuk: 21 September 2019, diterima untuk diterbitkan: 26 November 2020)

**Abstrak**

*Educational data mining* (EDM) adalah suatu bidang aplikasi antara pendidikan dan komputer. Salah satu yang dapat dilakukan pada EDM adalah memprediksi tingkat prestasi mahasiswa. Tingkat indeks prestasi kumulatif (IPK) akademik mahasiswa sangat penting karena menentukan tingkat kelulusan dan kualiatas institusi pendidikan. Penelitian ini bertujuan untuk menganalisa atribut-atribut yang mempengaruhi tingkat indeks prestasi kumulatif (IPK) mahasiswa yang berasal dari faktor eksternal pada mahasiswa. Adapun atribut yang digunakan adalah 10 variabel atribut yaitu nilai TOEFL, pendidikan ayah, pendidikan ibu, pekerjaan ayah, pekerjaan ibu, asal daerah, tempat tinggal selama kuliah dan tingkat prestasi akademik yang dicapai. Hasil akurasi pengolahan dengan menggunakan Algoritma C4.5 adalah 75,18% dan *Naive Bayes* 74,47% menunjukkan bahwa model dan atribut yang digunakan baik untuk memprediksi tingkat IPK mahasiswa. Algoritma C4.5 mampu menunjukkan atribut apa yang berpengaruh langsung pada tingkat IPK mahasiswa yaitu Nilai TOEFL, jam belajar, pendidikan ayah, pekerjaan ayah, dan tempat tinggal mahasiswa. Algoritma C4.5 tidak mampu memperhitungkan peluang suatu klasifikasi jika jumlah instan pada klasifikasi tersebut sangat sedikit pada kejadian data. Sebaliknya *Naive Bayes* tetap mampu memperhitungkan peluang kemunculan dan ketepatannya informasi yang dihasilkan meski jumlah instan yang sedikit. Dalam penelitian ini data mahasiswa yang memiliki tingkat IPK *cumlaude* sangat sedikit, namun *Naive Bayes* tetap mampu mengukur *Recall* pada kelas ini sebesar 28,6% dan *Precision* sebesar 40%.

**Kata kunci**: *Prediksi, C4.5, Educational Data mining, Naive Bayes, Tingkat IPK*

# *PREDICTION OF GRADE POINT AVERAGE STUDENT LEVEL USING DATA MINING TECHNIQUE*

***Abstract***

*Educational data mining (EDM) are widely applied to Gain knowledge from educational data. One of EDM is to predict of Grade Point Avarage (GPA) Student Level. The level is very important to determine graduation qualities. In this study, the survey of attributes that affect GPA Student level derived from external factors. The reasearched attributes used 10 variables i.e., TOEFL scores, father and mother educational background, father's and mother's occupation, place of origin, student residences and GPA level. The study uses two methods of classification techniques, The C4.5 and Naive Bayes. The accuracy results are 75.18% by C4.5 algorithm and 74.47% by Naive Bayes. The Naive Bayes showed that the models and variables used both to predict GPA Student level. The C4.5 algorithm shows any attributes that directly affect such as TOEFL score, length of study, father's education, father's occupation, and student residences. C4.5 algorithm was unable to calculate a probability of classification if the number instances in the data is less, e.g cumlaude label. On contrary, Naive Bayes can still get the Recall and the Precision information even though the number of the label is less. In this study, students with cumlaude label is less, however, Naive Bayes is still able to measure the Recall with 28.6% and Precision with 40%.*

**Keywords**: *Prediction, C4.5, Educational data mining, Naive Bayes, GPA Level*

## 1. PENDAHULUAN

*Educational Data mining (EDM)* adalah tren dalam *data mining* yang muncul pada tahun 2005, dimana teknik *data mining* diaplikasikan secara luas untuk memperoleh pengetahuan dalam bidang pendidikan(Aldowah, Al-Samarraie & Fauzy 2019)Tujuan utama dari penelitian pada bidang EDM adalah untuk mendukung pengambilan

keputusan pada institusi pendidikan agar dapat bermanfaat bagi pembuat keputusan dalam bidang pendidikan (Altujjar et al. 2016). EDM dianggap sebagai paradigma yang berorientasi pada perancangan model, tugas, metode, dan algoritma untuk mengeksplorasi data dari bidang pendidikan untuk menemukan pola konten pengetahuan, penilaian, aplikasi domain dan membuat prediksi yang menjadi ciri perilaku dan prestasi peserta didik serta faktor-faktor yang mempengaruhinya (Penaayala 2014).

EDM banyak berperan dalam membantu analisis dan visualisasi data pendidikan sehingga hasil analisisnya dapat digunakan untuk memprediksi kinerja siswa atau mahasiswa, dan mampu menghasilkan rekomendasi untuk pihak-pihak yang terkait dalam bidang pendidikan. Prediksi yang dapat dilakukan seperti mengindentifikasi kondisi pembelajaran suatu matakuliah, mendeteksi perilaku mahasiswa, mengembangkan materi kuliah, menentukan strategi pedagogik dan merencanakan berbagai kegiatan pendidikan lainnya (Wassan 2015). Penerapan teknik *data mining* dalam mengekstraksi pengetahuan bisa dipandang sebagai teknik evaluasi formatif, yaitu teknik untuk mengevaluasi program pendidikan sementara yang masih dalam pengembangan. Tujuan dari teknik tersebut adalah untuk terus memperbaiki program-program pendidikan EDM telah digunakan dalam berbagai penelitian diantaranya memprediksi kinerja mahasiwa tingkat akhir (Wanli et al. 2015), menentukan faktor-faktor yang mempengaruhi tingkat kebahagian mahasiswa (Altujjar et al. 2016), pengenalan pola pengaruh akademik pada alumni setelah 3 tahun masa selesai studi (Adekitan & Salau 2019), pendektesian strategi untuk mengembangkan *project urban* dalam matakuliah arsitektur (Valls et al. 2017). Selain itu beberapa penelitian dalam bidang EDM juga digunakan untuk memperbaiki proses manajemen dan administrasi akademik pada institusi pendidikan tinggi (Aldowah, Al-Samarraie & Fauzy 2019), penerapan teknik *data mining* juga digunakan untuk mengukur kemampuan mahasiswa dalam mendesain dan memahami eksperimen-eksperimen yang dilakukan dalam mata kuliah *systems microworld* (Gobert et al. 2015). (Mayilvaganan & Kalpanadevi 2015) menganalisa *cognitive skill* mahasiswa dengan menggunakan teknik *data mining*, teknik *data mining* juga digunakan untuk memprediksi kinerja dari para instruktur pengajar (Mohamed, Rizaner & Hakan 2016), (Bachtiar, Syahputra & Wicaksono 2019; Farhan et al. 2019) memanfaakan teknik *data mining* untuk membantu mahasiswa dalam memilih matakuliah, sedangkan (Yahdin et al. 2019) menggunakan *data mining* untuk memprediksi masa studi yang berhasil ditempuh mahasiswa.

Prediksi terhadap IPK mahasiswa merupakan salah satu hal yang penting dalam dunia pendidikan, karena hal ini memungkinkan institusi akademik memberikan dukungan atau program yang tepat bagi siswa yang menghadapi kesulitan dalam akademik (Altujjar et al. 2016; Yahdin et al. 2019). IPK merupakan hasil dari kegiatan belajar mahasiswa, dimana biasanya semakin baik usaha belajar yang dilakukan individu, maka semakin baik pula prestasi yang dicapai. IPK dapat menjadi ukuran keberhasilan dan kualitas dari mahasiswa tersebut, sehingga bisa diyakini bahwa mahasiswa tersebut memiliki keterampilan, pengetahuan, dan kemampuan yang mereka butuhkan saat mereka lulus nanti. Setiap universitas membagi tingkatan IPK mahasiswa dalam beberapa tingkatan(Charteris et al. 2016).

Prediksi IPK dengan memanfaatkan teknik *data mining* telah banyak dilakukan untuk berbagai macam paramater antara lain dengan melihat nilai awal yang diperoleh pada setiap mata kuliah, kehadiran kelas, hasil capaian tugas yang diberikan, kuis, project laboratorium, dan perilaku mahasiswa saat di dalam kelas atau dalam proses belajar mengajar(Daud 2017; Hamsa, J. Kizhakkethottam & Indiradevi 2016; López, Guzmán & González 2015; Mohammed et al. 2017; Topîrceanu & Grosseck 2017; Tucker & Pursel 2014). (Altujjar et al. 2016) memprediksi prestasi mahasiswa dengan memperhatikan efek nilai yang diperoleh pada matakuliah-matakuliah tertentu yang dianggap sebagai matakuliah yang krusial. (Ingraham, Davidson & Yonge 2018) meneliti hubungan antara fakultas yang dipilih mahasiswa dan tingkat prestasi akademik mahasiswa. Beberapa penelitian memprediksi tingkat prestasi akademik mahasiswa melalui pembelajaran *online (e-learning)*(Burgos et al. 2017; Rodrigues, Isotani & Zárate 2018).

Kesuksesan dari proses pembelajaran tidak hanya ditentukan oleh faktor di dalam kelas tetapi juga oleh faktor di luar kelas. Kesuksesan mahasiswa atau pelajar ditentukan oleh dua faktor yaitu faktor internal seperti konsentrasi, minat, bakat, intelegensi, motivasi, cita-cita, intensitas belajar, sedangkan faktor eksternal seperti lingkungan fisik, ekonomi, latar belakang pendidikan orang tua, tempat tinggal serta faktor lainnya baik yang secara langsung maupun secara tidak langsung(Aissaoui et al. 2019; Aramburo, Boroel & Pineda 2017; Helal et al. 2018). Pada penelitian ini membahas penerapan *data mining* untuk memprediksi tingkat prestasi akademik mahasiswa dengan melihat faktor eksternal untuk mengtahui faktor apa saja yang mempengaruhi prestasi akademik mahasiswa. Hasil dari prediksi tersebut diharapkan universitas dapat memberikan alternatif pemecahan masalah yang dapat dikembangkan dalam suatu program yang dapat membantu mahasiswa yang memiliki masalah dengan indeks prestasinya.

Permasalahan prediksi banyak memanfaatkan teknik *data mining*. Dalam *data mining* banyak metode yang berkembang untuk prediksi seperti

*Naive Bayes*, C4.5. *K-Nearest Neigborhood* (KNN) dan lain-lain. Beberapa penelitian merekomendasikan metode *Naive Bayes* dan Algoritma C4.5 sebagai metode yang cepat, mudah, kuat dan paling banyak digunakan untuk prediksi terutama pada data set yang memiliki banyak atribut bertipe kategorik atau nominal (Berger 2013; Breiman, L., Friedman, J. H., Olshen, R. A., & Stone 1984; Yahdin et al. 2019; Zhang et al. 2016). Pada penelitian ini menerapkan algoritma C4.5 dan *Naive Bayes* untuk memprediksi tingkat indeks prestasi kumulatif akademik mahasiswa.

## 2. METODE PENELITIAN

### 2.1. Pengumpulan Data

Data diambil dari jurusan Matematika Fakultas MIPA Universitas Sriwijaya yang berasal dari 3 tahun angkatan yaitu 2013, 2014 dan 2015. Jumlah total data yang digunakan adalah sebanyak 200 orang yang diambil dari mahasiswa yang telah lulus. Mahasiswa angkataan 2016, 2017 dan 2018 tidak diambil sebagai data dalam penelitian ini karena belum ada lulusan dari ketiga angkatan tersebut.

### 2.2. Persiapan Data

Dalam penelitian ini diperoleh 13 atribut yang berkaitan langsung dengan mahasiswa, seperti besar penghasilan ayah perbulan, besar penghasilan ibu perbulan, jenis kelamin, pendidikan terakhir ayah, pendidikan terakhir ibu, status hidup ayah, status hidup ibu, pekerjaan ayah, pekerjaan ibu, tempat tinggal asal, tempat tinggal selama kuliah, nilai TOEFL yang digunakan saat mendaftar dan IPK. Atribut-atribut tersebut diseleksi kembali karena ada atribut yang datanya banyak yang tidak lengkap atau hilang. Total atribut yang terlibat dalam penelitian ini ada 10 atribut yaitu jenis kelamin, jam belajar di luar kampus, pendidikan ayah, pekerjaan ayah, pendidikan ibu, pekerjaan ibu, asal daerah, tempat tinggal selama kuliah, nilai TOEFL saat mendaftar dan indeks prestasi akademik akhir. Beberapa atribut diubah kedalam bentuk kategori sehingga algoritma C4.5 dan *Naive Bayes* dapat bekerja optimal. Atribut atribut yang diubah ke dalam tipe kategori adalah nilai TOEFL dinyatakan dalam kurang dan cukup. TOEFL merupakan syarat dari universitas yang harus dapat dicapai selama masa kuliah, jika pada awal masuk TOEFL mahasiswa kurang dari 400 maka mahasiswa wajib mengulang ujian TOEFL selama masa kuliah sampai mencapai skor 400, baru dianggap sebagai cukup, sehingga TOEFL dapat dirubah dalam variabel kategorik dengan nilai cukup dan kurang. Nilai Indeks Prestasi Kumulatif (IPK) akademik mahasiswa akan menjadi label yang diubah dalam bentuk kategori sesuai dengan tingkatan perolehan IPK, yaitu sebagai berikut :

- Jika IPK ≤2,5 maka prestasi tidak memuaskan, disimbolkan sebagai 1
- Jika IPK 2,5,00 - 2,75 maka prestasi memuaskan, disimbolkan sebagai 2
- Jika IPK 2,76 - 3,50 maka Prestasi sangat memuaskan, disimbolkan sebagai 3
- Jika IPK 3,51 - 4,00, maka prestasi adalah cumlaude, disimbolkan sebagai 4.

### 2.3. Penerapan Algorithm C4.5

Algoritma C4.5 bekerja melakukan pencarian dengan menggunakan pohon keputusan. Langkah-langkah yang dilakukan dalam algoritma C4.5 adalah menyeleksi atribut yang menjadi akar, membuat cabang untuk masing-masing atribut, membagi penelusuran dalam cabang berdasarkan *Gain ratio*. Langkah tersebut akan terus diulang untuk masing-masing cabang sampai semua kelas selesai diperiksa.

pemilihan atribut sebagai akar didasarkan pada *Gain ratio* yang tertinggi. Perhitungan *Gain ratio* menggunakan persamaan 1(Yahdin et al. 2019).

$$Gain(S,A) = Entropy(S) - \sum_{1=1}^{N} \frac{|S_I|}{|S|} * Entropy(S_i)) \quad (1)$$

Dimana, S adalah himpunan instant (kejadian), A adalah atribut, n : Jumlah partisi dari atribut A |Si| : Jumlah kasus pada partisi ke i, |S| : Jumlah kasus dalam S, sedangkan nilai *Entropy* sendiri dapat dihitung dengan persamaan 2.

$$Entropy(S) = \sum_{i=1}^{n} -P_i * log_2 P_i \quad (2)$$

Dimana, S : Himpunan Kasus dan n : Jumlah partisi S dan $P_i$ adalah Proporsi dari Si terhadap S

### 2.4. Penerapan *Naive Bayes*

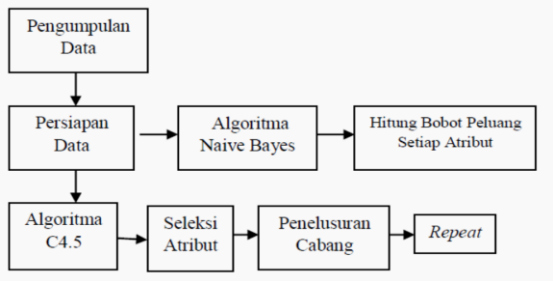
*Naive Bayes* adalah metode dalam *data mining* yang menerapkan konsep peluang *Bayes*. Semua atribut diperlakukan sama dan bebas antara satu atribut dengan atribut lainnya. metode ini menggunakan *Naive Bayes Classifier* untuk menghitung bobot peluang dari setiap atribut. *Naive Bayes Classifier* yang digunakan adalah(Desiani et al. 2019) :

$$C_{NB} = arg \max_{c_k \epsilon C} P(c_k) \prod P(v_j|c_k) \quad (3)$$

Secara umum alur metode yang dilakukan dalam penelitian ini dapat dilihat pada gambar 1. Dari gambar 1 dapat dilihat data yang telah diseleksi akan diklasifikasi dengan menerapkan dua teknik data mining yaitu Algoritma c.4.5 dan *Naive Bayes*.

## 3. HASIL DAN PEMBAHASAN

Dari hasil pengumpulan data yang dilakukan diperoleh 141 data pengamatan dengan atribut sebanyak 13. Dari 13 atribut yang ada hanya dipilih 10 atribut yang memiliki data lengkap (tabel 1). Atribut yang menjadi label klasifikasi adalah tingkat Indeks Prestasi Kumulatif akademik (IPK) terakhir yang diperoleh oleh mahasiswa saat menjelang kelulusan.




Gambar 1. Alur Metode Penelitian

Tingkat IPK memiliki 4 kelas yaitu tidak memuaskan, memuaskan, sangat memuaskan dan *cumlaude*. Metode latihan yang digunakan adalah *K-Cross-Validation.* K yang dipilih adalah 10, artinya data yang ada dibagi dalam 10 kelompok dimana 9 kelompok menjadi data latih dan 1 kelompok menjadi data uji yang dilakukan secara bergantian antara 10 kelompok tersebut. Secara lengkap atribut yang digunakan dalam penelitian ini dapat dilihat pada tabel 1.

Tabel 1. Atribut dan Nilai Atribut

| Nama Atribut | Partisi Nilai |
|---|---|
| Jenis Kelamin | 0: Perempuan; 1: Laki-laki |
| Jam Belajar (JAMblj) | 1: Satu hari menjelang ujian<br>2: 2-3 hari menjelang ujian<br>3: lebih dari 3 hari sebelum ujian |
| Pendidikan Ayah (Pddk_A) | SMA : Sekolah Menengah Atas<br>SD: Sekolah Dasar<br>SMP: Sekolah Menengah Tingkat Pertama<br>D: Diploma<br>S1 : Sarjana<br>S2: Magister ke atas |
| Pekerjaan Ayah (P_A) | BRH: Buruh, seperti petani, Pembantu Rumah Tangga, Pekerja Pabrik, dan seba*Gain*ya<br>SW: Swasta, bekerja pada perusaahaan milik sendiri atau yang membuka jalur usaha mandiri<br>PNS : Pegawai Pemerintahan<br>NonPNS : Pegawai pada perusahaan swasta atau non pemerintah |
| Pendidikan Ibu | SMA : Sekolah Menengah Atas<br>SD: Sekolah Dasar<br>SMP: Sekolah Menengah Tingkat Pertama<br>D: Diploma<br>S1 : Sarjana<br>S2: Magister ke atas |
| Pekerjaan Ibu (P_I) | BRH: Buruh, seperti petani, Pembantu Rumah Tangga, Pekerja Pabrik, dan seba*Gain*ya<br>SW: Swasta, bekerja pada perusaahaan milik sendiri atau yang membuka jalur usaha mandiri<br>PNS : Pegawai Pemerintahan<br>NonPNS : Pegawai pada perusahaan swasta atau non pemerintah<br>IRT : Ibu yang tidak bekerja |
| Daerah Asal | KT: Kota<br>D: Desa |
| Tempat Tinggal Selama Kuliah (TTL) | Kost : tinggal dengan menyewa tempat tinggal<br>Ortu : Tinggal bersama orang tua |
| Indeks Prestasi Akademik(IPK) | 1: tidak memuaskan<br>2: Memuaskan<br>3: Sangat Memuaskan<br>4: Cumlaude |

### 3.1 ALGORITHM C4.5

Hasil dari perhitungan nilai *Gain* pada algoritma c4.5, nilai *Gain* tertinggi dimiliki oleh atribut TOEFL. Nilai *Gain* yang diperoleh menunjukkan jika mahasiswa sudah mampu memenuhi syarat nilai TOEFL pada awal perkuliahan maka dapat diprediksi bahwa mahasiswa tersebut akan memiliki tingkat IPK sangat memuaskan. Sebaliknya jika nilai TOEFL saat awal perkuliahan belum mencukupi syarat maka harus melihat atribut lain untuk memprediksi tingkat indeks prestasi kumulatif yaitu atribut Jam Belajar. Jika mahasiswa tersebut selalu belajar persiapan 3 hari sebelum ujian maka dapat diprediksi mahasiswa tersebut akan memperoleh tingkat IPK yang memuaskan. Secara lengkap aturan yang diperoleh dari perhitungan algoritma C4.5 dapat dilihat pada pohon keputusan gambar 2.

Dari pohon keputusan pada gambar 1, diperoleh aturan linguistik untuk prediksi tingkat indeks prestasi akademik mahasiswa sebagai berikut :

**IF** TOEFL= Cukup **THEN** Prestasi Akademik = Sangat mememuaskan

**IF** Nilai TOEFL= Kurang **AND** Jam Belajar = sebelum 3 hari menjelang ujian **THEN** Prestasi Akademik= sangat memuaskan

**IF** Nilai TOEFL= Kurang **AND** Jam Belajar= 1 hari menjelang ujian **AND** Pendidikan Ayah= SMA **THEN** Prestasi Akademik= memuaskan

**IF** Nilai TOEFL= Kurang **AND** Jam Belajar= 1 hari menjelang ujian **AND** (Pendidikan Ayah= SD **OR** Pendidikan Ayah = SMP) **THEN** Prestasi Akademik Mahasiswa = Tidak memuaskan

**IF** Nilai TOEFL= Kurang **AND** Jam Belajar= 1 hari jelang ujian **AND** Pendidikan Ayah= S1 **AND** Pendidikan Ibu = SMA **THEN** Prestasi Akademik Mahasiswa = Memuaskan.

**IF** Nilai TOEFL= Kurang **AND** Jam Belajar= 1 hari jelang ujian **AND** Pendidikan Ayah= S1 **AND** Pendidikan Ibu = S1 **THEN** Prestasi Akademik Mahasiswa = Sangat Memuaskan.

**IF** Nilai TOEFL= Kurang **AND** Jam Belajar= 2 sampai 3 hari jelang ujian **AND** Pekerjaan Ayah= Buruh **AND** Pendidikan Ayah= SMA **THEN** Prestasi Akademik Mahasiswa= Sangat memuaskan

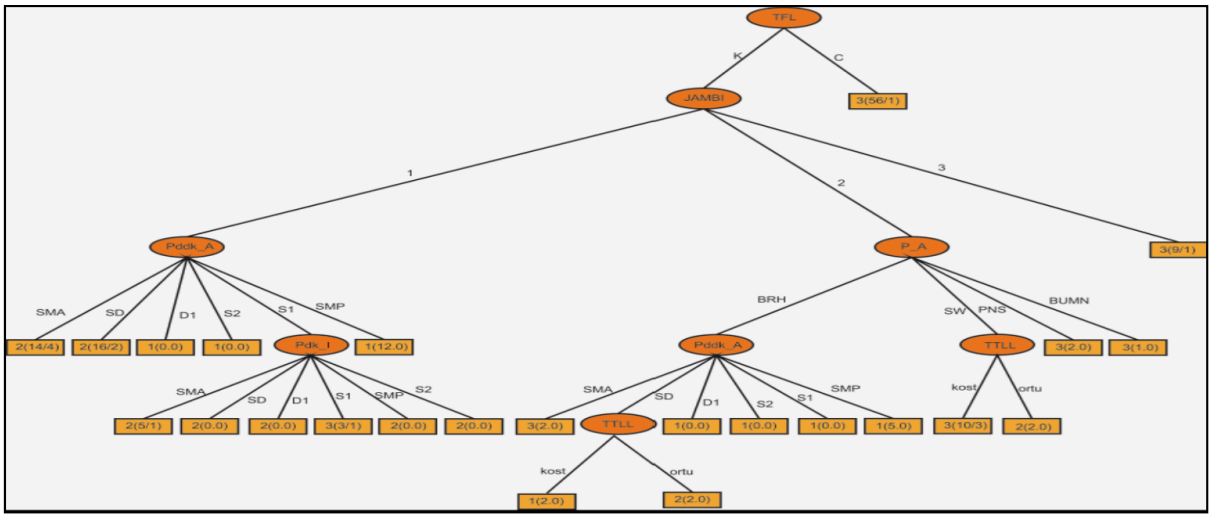
**IF** Nilai TOEFL= Kurang **AND** Jam Belajar= 2 sampai 3 hari jelang ujian **AND** Pekerjaan Ayah= Buruh **AND** Pendidikan Ayah= SMP **THEN** Prestasi Akademik Mahasiswa= Tidak Memuaskan

**IF** Nilai TOEFL= Kurang **AND** Jam Belajar= 2 sampai 3 hari jelang ujian **AND** Pekerjaan Ayah= Buruh **AND** Pendidikan Ayah= SD **AND** Tempat Tinggal selama Kuliah= Kost **THEN** Prestasi Akademik Mahasiswa= Tidak memuaskan
**IF** Nilai TOEFL= Kurang **AND** Jam Belajar= 2 sampai 3 hari jelang ujian **AND** Pekerjaan Ayah= Buruh **AND** Pendidikan Ayah= SD **AND** Tempat Tinggal selama Kuliah= Orang Tua **THEN** Prestasi Akademik Mahasiswa= memuaskan

**IF** Nilai TOEFL= Kurang **AND** Jam Belajar= 2 sampai 3 hari jelang ujian **AND** Pekerjaan Ayah= Swasta **AND** Tempat Tinggal selama Kuliah= Orang Tua **THEN** Prestasi Akademik Mahasiswa= memuaskan

**IF** Nilai TOEFL= Kurang **AND** Jam Belajar= 2 sampai 3 hari jelang ujian **AND** Pekerjaan Ayah= Swasta **AND** Tempat Tinggal selama Kuliah= Kost **THEN** Prestasi Akademik Mahasiswa= sangat memuaskan

**IF** Nilai TOEFL= Kurang **AND** Jam Belajar= 2 sampai 3 hari jelang ujian **AND** (Pekerjaan Ayah= PNS **OR** Pekerjaan Ayah= Non PNS) **THEN** Prestasi Akademik Mahasiswa= sangat memuaskan



Gambar 2. Pohon Keputusan Algoritma C4.5

Setelah diperoleh aturan keputusan dari pohon keputusan algoritma C4.5, proses uji akan mengukur sejauh mana keberhasilan model yang diperoleh dapat digunakan sebagai prediksi dengan menggunakan matriks *Confusion*. Matriks *Confusion* yang diperoleh dari perhitungan algoritma C4.5 adalah sebagai berikut :

Tabel 2. Matriks *Confusion* C4.5

| Kelas | | Label Sebenarnya | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| Label Prediksi | 1 | 31 | 4 | 1 | 0 |
| | 2 | 4 | 16 | 8 | 0 |
| | 3 | 1 | 10 | 59 | 0 |
| | 4 | 0 | 0 | 7 | 0 |

Dari tabel 2 dapat dilihat ada 31 data diprediksi secara benar sebagai kelompok1 (kelompok yang memiliki prestasi akademik tidak memuaskan), 4 data yang harusnya masuk dalam kelompok satu tetapi diprediksi sebagai kelompok 2(memuaskan) dan 1 data pada kelompok 1 diklasifikasikan sebagai kelompok 3( sangat memuaskan). 16 data pada testing ini berhasil diklasifikasikan sebagai kelompok 2, namun 4 data yang harusnya memiliki klasifikasi 2, dikenali sebagai klasifikasi 1(tidak memuaskan, dan 8 data dikenali sebagai kelas 3(sangat memuaskan) yang seharusnya masuk dalam klasifikasi 2.

Dari matriks *Confusion* dapat dihitung nilai akurasi yang diperoleh dari total data yang berhasil diprediksi secara benar sebesar 75,18%. Hasil ini menunjukkan algoritma C4.5 cukup baik dalam memprediksi tingkat indeks prestasi akademik mahasiswa. Hasil pengukuran *Recall* untuk masing-masing tingkat IPK adalah sebagai berikut 81.6% untuk kelompok tingkat IPK yang tidak memuaskan (grup 1), 57.1% untuk kelompok tingkat IPK memuaskan (grup 2), 84.3% untuk kelompok IPK sangat memuasakan (grup 3) dan 0% untuk kelompok tingkat IPK yang *cumlaude* (grup 4). Hasil pengukuran *Precision* untuk masing-masing kelompok adalah sebagai berikut, kelompok 1 (Tingkat IPK yang tidak memuaskan) adalah 86,1%, kelompok kedua adalah 53,3%, untuk kelompok 3 adalah 78,7%, dan 0% untuk kelompok 4.

### 3.2 *NAIVE BAYES*

Pada *Naive Bayes*, metode training yang dipilih adalah *K-Cross validation* dengan nilai *K-fold* yang dipilih adalah 10. Matriks *Confusion* yang dihasilkan adalah sebagai berikut:

Tabel 3. Matriks *Confusion Naive Bayes*

| Kelas | | Label Sebenarnya | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| Label Prediksi | 1 | 33 | 3 | 0 | 0 |
| | 2 | 5 | 15 | 8 | 0 |
| | 3 | 6 | 6 | 55 | 3 |
| | 4 | 0 | 0 | 5 | 2 |

Dari tabel 3 dapat dilihat algoritma *Naive Bayes* dapat memprediksi sebanyak 105 data secara benar, tetapi 36 data diprediksi dalam kelas yang salah. Akurasi yang diperoleh pada *Naive Bayes* adalah 74,47%, lebih kecil dibandingkan hasil akurasi pada algoritma C4.5. Nilai *Recall* yang diperoleh pada *Naive Bayes* untuk masing-masing kelompok adalah 91,7% untuk kelompok 1, 53,6% untuk kelompok 2, 78,6% untuk kelompok 3 dan 28,6% untuk kelompok 4. Nilai *Precision* untuk kelompok 1 adalah 75%, untuk kelompok 2 adalah 62,5%. Nilai *Precision* untuk masing-masing kelompok adalah kelompok 3 adalah 80,9% dan 40% untuk kelompok 4.

### 3.3 PERBANDINGAN HASIL KEDUA ALGORITMA

Hasil prediksi dua algoritma C4.5 dan algoritma *Naive Bayes*, terlihat bahwa kedua metode tersebut bekerja baik dalam memprediksi prestasi akademik mahasiswa. Pada algoritma c4.5 dapat diketahui atribut mana yang saling mempengaruhi prestasi akedemik mahasiswa. Dari hasil pohon keputusan dapat dilihat bahwa mahasiswa yang mampu lulus ujian TOEFL sesuai dengan standar nilai dari universitas, artinya mampu memiliki prestasi akademik sangat memuaskan. Mahasiswa yang nilai TOEFL belum mencukupi syarat universitas, belajar 2-3 hari sebelum ujian, orang tuanya bekerja sebagai buruh latar, dan orang tuanya hanya mengenyam pendidikan SD atau SMP, serta tinggal di kontrakan tidak bersama orang tua lebih riskan memiliki tingkat IPK yang buruk atau tidak memuaskan. Mahasiswa yang memiliki nilai TOEFL yang belum mencukupi syarat, dan belajar hanya 1 hari sebelum ujian jika pendidikan ayah SD atau SMP hasil prestasi akademiknya cenderung masuk dalam kelompok 1.

Kecenderungan mahasiswa yang prestasinya tidak memuaskan datang dari kalangan yang nilai TOEFL masih kurang, orang tuanya bekerja sebagai buruh, pendidikan ayah hanya sebatas SD atau SMP, dan jam belajar yang sedikit, dapat memberikan masukkan kepada pihak jurusan atau universitas memberikan pendampingan program yang dapat memotivasi dan fasilitasi mahasiswa tersebut agar dapat meningkatkan tingkat minat pelajarnya. Pihak jurusan atau universitas dapat bekerja sama dengan dosen penasehat akademik untuk memantau prestasi akademik mahasiswa terutama mahasiswa dengan memiliki atribut-atribut yang masuk dalam kelompok 1. Perbandingan hasil pengukuran kedua algoritma tersebut dapat dilihat pada tabel 4 dan tabel 5.

Tabel 4. Nilai Akurasi dan Presisi Algoritma C4.5 dan *Naive Bayes*

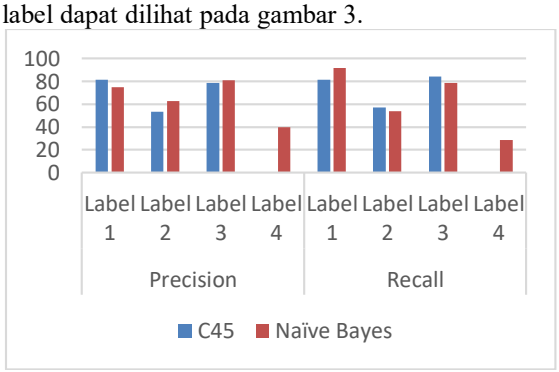
| | Precision(%) | | | | |
|---|---|---|---|---|---|
| | Kelompok | | | | Accuration |
| Algoritma | 1 | 2 | 3 | 4 | (%) |
| C4.5 | 81,6 | 53,3 | 78,7 | 0 | 75,18 |
| Naive Bayes | 75 | 62,5 | 80,9 | 40 | 74,47 |

Tabel 5. Nilai *Recall* Algoritma C4.5 dan Naive Bayes

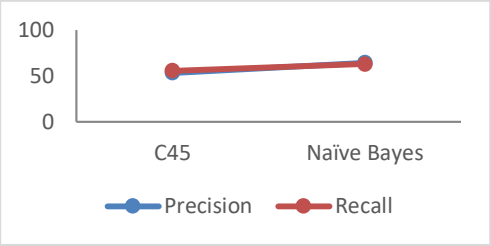
| | Recall(%) | | | |
|---|---|---|---|---|
| | Kelompok | | | |
| Algoritma | 1 | 2 | 3 | 4 |
| C4.5 | 81,6 | 57,1 | 84,3 | 0 |
| Naive Bayes | 91,7 | 53,6 | 78,6 | 28,6 |

Dari tabel 4 dapat dilihat algoritma C4.5 nilai akurasi prediksinya lebih baik dibanding *Naive Bayes*. Hasil Algoritma C4.5 mampu menyajikan secara langsung hubungan-hubungan antar atribut yang mempengaruhi prestasi akademik mahasiswa. Algoritma C4.5 tidak dapat menghitung keberadaan instan yang ada pada kelompok 4, karena instan pada kelompok 4 jumlahnya sangat sedikit. Tingkat keberhasilan sistem dalam menemukan informasi dan tingkat ketepatan antara informasi yang diminta dan jawaban yang diberikan oleh algoritma C4.5 untuk kelompok 4(mahasiswa yang berhasil mencapai prestasi *cumlaude*) adalah sama sekali tidak ada atau 0%. Sebaliknya pada hasil prediksi *Naive Bayes* meski lebih rendah dibanding dari C4.5 namun *Naive Bayes* tetap mampu memperhitungkan kejadian yang masuk pada kelompok 4 meski jumlahnya hanya sedikit. Peluang keberhasilan sistem dalam menemukan informasi kejadian pada kelompok 4 adalah 28,6% dan tingkat ketepatan informasi yang diberikan oleh algoritma berdasarkan informasi yang diminta pengguna untuk kejadian kelompok 4 adalah 40%.

Hasil *Precision* dan *Recall* dari algoritma C45 dan *Naive Bayes* dapat dilihat secara ringkas pada gambar 2, dan untuk nilai rata-rata dari seluruh label dapat dilihat pada gambar 3.




Gambar 2. Nilai *Precision* dan *Recall* Untuk Hasil Prediksi




Gambar 3. Rata-Rata Keseluruhan Nilai *Precision* dan *Recall*

Dari gambar 3 dapat dilihat nilai *Precision* dan *Recall* baik yang dihasilkan oleh algoritma C4.5 ataupun *Naive Bayes* tidak terlalu jauh berbeda. kedua algoritma tersebut cukup baik digunakan sebagai prediksi tingkat prestasi akademik mahasiswa.

## 4. KESIMPULAN

Hasil akurasi dari kedua algoritma C4.5 sebesar 75,18% dan *Naive Bayes* sebesar 74,47%, menjelaskan bahwa kedua algoritma tersebut baik dalam melakukan prediksi prestasi akademik mahasiswa dalam penelitian ini. Hasil C4.5 menunjukkan mampu menunjukkan secara langsung atribut-atribut yang berpengaruh pada prestasi akademik mahasiswa. Hasil C4.5 menunjukkan bahwa mahasiswa yang nilai TOEFL yang belum mencukupi, mahasiswa yang jam belajarnya hanya sedikit, mahasiswa dengan latar belakang orang tua yang bekerja sebagai buruh, dan hanya memiliki pendidikan rendah sebatas SD atau SMP lebih riskan masuk dalam kelompok prestasi akademik yang tidak memuaskan, sehingga diperlukan suatu program dari lembaga pendidik (Universitas) untuk dapat membantu mahasiswa yang memiliki nilai-nilai atribut tersebut. Algoritma C4.5 tidak mampu memperhitungkan kejadian pada kelompok 4 karena jumlahnya terlalu sedikit, namun *Naive Bayes* tetap mampu memperhitungkan kejadian pada kelompok 4 dengan *Recall* sebesar 28,6% dan *Precision* sebesar 40%.

**Ucapan Terimakasih**

Terimakasih kami sampaikan kepada Universitas Sriwijaya atas bantuan pendanaan penelitian ini melalui PNBP 2018

## DAFTAR PUSTAKA

ADEKITAN, A.I. & SALAU, O. 2019, 'The impact of engineering students' performance in the first three years on their graduation result using educational data mining', *Heliyon*, vol. 5, no. 2, p. e01250.

AISSAOUI, O. EL, EL MADANI, Y.E.A., OUGHDIR, L. & ALLIOUI, Y. El 2019, 'Combining supervised and unsupervised machine learning algorithms to predict the learners' learning styles', *Procedia Computer Science*, vol. 148, pp. 87–96.

ALDOWAH, H., AL-SAMARRAIE, H. & FAUZY, W.M. 2019, 'Educational data mining and learning analytics for 21st century higher education: A review and synthesis', *Telematics and Informatics*, vol. 37, no. January, pp. 13–49.

ALTUJJAR, Y., ALTAMIMI, W., AL-TURAIKI, I. & AL-RAZGAN, M. 2016, 'Predicting Critical Courses Affecting Students Performance : A Case Study', *Procedia - Procedia Computer Science*, vol. 82, no. March, pp. 65–71.

ARAMBURO, V., BOROEL, B. & PINEDA, G. 2017, 'Predictive factors associated with academic performance in college students', *Procedia - Social and Behavioral Sciences*, vol. 237, no. June 2016, pp. 945–9.

BACHTIAR, F.A., SYAHPUTRA, I.K. & WICAKSONO, S.A. 2019, 'Perbandingan Algoritme Machine Learning untuk Memprediksi Pengambil Matakuliah', *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 6, no. 5, p. 543.

BERGER, J.O. 2013, *Statistical Decision Theory and Bayesian Analysis*, 2013th edn, Springer-Verlag.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., & STONE, C.J. 1984, *Classification and Regression Trees*, Wadsworth International Group, belmont, CA.

BURGOS, C., CAMPANARIO, M.L., DE, D., LARA, J.A., LIZCANO, D. & MARTINEZ, M.A. 2017, 'action plan to prevent academic dropout R', *Computers & Electrical Engineering*, vol. 66, pp. 1–16.

CHARTERIS, J., QUINN, F., PARKES, M., FLETCHER, P. & REYES, V. 2016, 'e-Assessment for learning and performativity in higher education : A case for existential learning', *Australasian Journal of Educational Technology*, vol. 32, no. 3, pp. 112–22.

DAUD, A. 2017, 'Predicting Student Performance using Advanced Learning Analytics', *International World Wide Web Conference Committee (IW3C2)*, Creative Commons CC, Perth, Australia, pp. 415–21.

DESIANI, A., PRIMARTHA, R., ARHAMI, M. & ORSALAN, O. 2019, 'Naive Bayes classifier for infant weight prediction of hypertension mother', *Journal of Physics: Conference Series PAPER*, vol. 1282.

FARHAN, F., KUMARA, W., SUPIANTO, A.A., Informasi, S., Komputer, F.I., Brawijaya, U., Informatika, T., Komputer, F.I., Brawijaya, U. & Forest, R. 2019, 'Rekomendasi Pengambilan Mata Kuliah Pilihan Untuk Mahasiswa Sistem Informasi Menggunakan Algoritme Decision Tree', *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 6, no. 3, pp. 341–8.

GOBERT, J.D., JEON, Y., SAO, M.A., KENNEDY, M. & BETTS, C.G. 2015, 'systems

microworld', *Thinking Skills and Creativity*, vol. 18, pp. 81–90.

HAMSA, H., J.KIZHAKKETHOTTAM, J. & INDIRADEVI, S. 2016, 'Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm', *Procedia Technology*, vol. 25, pp. 326–32.

HELAL, S., LI, J., LIU, L., EBRAHIMIE, E., DAWSON, S., MURRAY, D.J. & LONG, Q. 2018, 'Predicting academic performance by considering student heterogeneity', *Knowledge-Based Systems*, vol. 161, pp. 134–46.

INGRAHAM, K.C., DAVIDSON, S.J. & YONGE, O. 2018, 'Student-faculty relationships and its impact on academic outcomes', *Nurse Education Today*, vol. 71, pp. 17–21.

LÓPEZ, C.E., GUZMÁN, E.L. & GONZÁLEZ, F.A. 2015, 'A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining', *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, vol. 10, no. 3, pp. 119–25.

MAYILVAGANAN, M. & KALPANADEVI, D. 2015, 'Cognitive Skill Analysis for Students through Problem Solving Based on Data Mining Techniques', *Procedia - Procedia Computer Science*, vol. 47, pp. 62–75.

MOHAMED, A., RIZANER, A. & HAKAN, A. 2016, 'Using data Mining to Predict Instructor Performance', *Procedia - Procedia Computer Science*, vol. 102, no. August, pp. 137–42.

MOHAMMED, D., RAHMAN, A., ABDEL, A., MUSA, N., GAMAL, A. & EL-AZIZ, A. 2017, 'Egyptian Pediatric Association Gazette Incidence , risk factors and complications of hyperglycemia in very low birth weight infants', *Egyptian Pediatric Association Gazette*, vol. 65, no. 3, pp. 1–8.

PENAAYALA, A. 2014, 'Expert Systems with Applications Educational data mining : A survey and a data mining-based analysis of recent works', *Expert Systems with Applications*, vol. 41, no. September, pp. 1432–62.

RODRIGUES, M.W., ISOTANI, S. & ZÁRATE, L.E. 2018, 'Educational Data Mining: A review of evaluation process in the e-learning', *Telematics and Informatics*, vol. 35, no. 6, pp. 1701–17.

TOPÎRCEANU, A. & GROSSECK, G. 2017, 'Decision tree learning used for the classification of student archetypes in online courses', *Procedia Computer Science*, vol. 112, pp. 51–60.

TUCKER, C. & PURSEL, B.K. 2014, 'Mining Student-Generated Textual Data In MOOCS And Quantifying Their Effects on Student Performance and Learning Outcomes Mining Student-Generated Textual Data in MOOCS and Quantifying Their Effects on Student Performance and Learning Outcomes', *121st ASEE Annual Conference & Exposition*.

VALLS, F., REDONDO, E., FONSECA, D., TORRES-KOMPEN, R., VILLAGRASA, S. & MARTÍ, N. 2017, 'Urban Data and Urban Design : A Data Mining Approach to Architecture Education', *Telematics and Informatics*, Elsevier Ltd.

WANLI, X., RUI, G., EVA, P. & SEAN, G. 2015, 'Computers in Human Behavior Participation-based student final performance prediction model through interpretable Genetic Programming : Integrating learning analytics , educational data mining and theory', *COMPUTERS IN HUMAN BEHAVIOR*, vol. 47, pp. 168–81.

WASSAN, J.T. 2015, 'Discovering Big Data Modelling for Educational World', *Procedia - Social and Behavioral Sciences*, vol. 176, pp. 642–9.

YAHDIN, S., DESIANI, A., AMRAN, A. & RODIAH, D. 2019, 'Pattern recognation for study period of student in Mathematics Department with C4 . 5 algorithm data mining technique at the Faculty of Mathematics and Natural Science Universitas Sriwijaya Pattern recognation for study period of student in Mathematics De', *Sriwijaya International Conference on Basic and Applied Science*, pp. 1–6.

ZHANG, L., JIANG, L., LI, C. & KONG, G. 2016, 'Two Feature Weighting Approaches for Naive Bayes Text ClassiÞers', *Knowledge-Based Systems*, vol. 100, pp. 137–44.