JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

ANALISIS PREDIKSI KETEPATAN MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA NAIVE BAYES CLASSIFIER DAN *FEATURE SELECTION*

MUHAMMAD NUR AKBAR[1], HARIANI[2], ASEP INDRA SYAHYADI[3]
[1,2,3] Teknik Informatika, Fakultas Sains dan Teknologi, UIN Alauddin Makassar
email: muhammad.akbar@uin-alauddin.ac.id[1], hariani.kasim@uin-alauddin.ac.id[2], asep@uin-alauddin.ac.id[3]

ABSTRAK

Lama masa studi mahasiswa merupakan salah satu poin penilaian dalam akreditasi suatu program studi pada institusi perguruan tinggi. Pendeteksian dini keterlambatan masa studi dapat dilakukan dengan memanfaatkan teknik *data mining*. Pada penelitian ini diterapkan algoritma Naïve Bayes Classifier (NBC) dan teknik *feature selection* menggunakan Informaton Gain (IG) dan Correlation Attribute (CA) dengan tujuan membangun model prediksi yang akurat dan *menganalisis atribut yang berpengaruh dalam menentukan lama masa studi s*ehingga dapat membantu perguruan tinggi dalam membuat kebijakan akademis agar dapat mengoptimalkan tingkat kelulusan mahasiswa pada tahun-tahun berikutnya. *Hasil uji coba pada dataset diperoleh akurasi tertinggi yaitu NBC+CA sebesar 81.2%, meningkat 12% dibandingkan NBC tanpa feature selection.*

Kata kunci: Correlation Attribute, *feature selection*, Information Gain, masa studi mahasiswa, Naïve Bayes Classifier, prediksi

I. PENDAHULUAN

Akreditasi program studi pada suatu universitas merupakan daya tarik tersendiri dalam proses penerimaan mahasiswa baru. Oleh karena itu setiap program studi dituntut untuk dapat meningkatkan akreditasinya, tidak terkecuali pada Program Studi Teknik Informatika UIN Alauddin Makassar.

Program Studi Teknik Informatika merupakan program studi di bawah naungan Fakultas Sains dan Teknologi UIN Alauddin Makassar sejak tahun 2005 dan menjadi salah satu program studi yang paling diminati oleh calon mahasiswa. Tercatat Program Studi Teknik Informatika UIN Alauddin Makassar menempati peringkat kedua terfavorit pada Ujian Masuk Perguruan Tinggi Keagamaan Islam Negeri (UM-PTKIN) tahun 2021 yaitu dengan 983 pendaftar dan kuota 26 orang.

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

Banyaknya jumlah pendaftar tersebut tidak lepas dari peran akreditasi yang terus meningkat. Namun capaian tersebut masih perlu ditingkatkan hingga memperoleh akreditasi A atau Unggul sesuai dengan visinya. Salah satu poin yang dianggap perlu dibenahi yaitu lama masa studi mahasiswa sebab menjadi poin penilaian dalam akreditasi (Khasanah & Harwati, 2017).

Berbagai upaya yang telah dilakukan, pemanfaatan teknologi informasi khususnya implementasi *data mining* pada aplikasi deteksi dini mahasiswa yang berpotensi mengalami keterlambatan lulus dapat menjadi salah satu alternatif solusi. Pendeteksian dini dapat dilakukan dengan memanfaatkan teknik *data mining* serta ketersediaan data pada Pusat Teknologi Informasi dan Pangkalan Data (PUSTIPAD) UIN Alauddin Makassar yang menyimpan berbagai data akademik mahasiswa terkait perkuliahan, nilai, data kelulusan, serta data profil mahasiswa.

Data mining dalam dunia pendidikan lebih dikenal dengan istilah *Educational Data Mining* (EDM) (Baker & Yacef, 2009). EDM dimanfaatkan dalam mengektraksi pengetahuan atau menemukan pola tersembunyi untuk menganalisis lebih jauh terkait kebiasaan belajar, mengidentfikasi mahasiswa yang membutuhkan dukungan, dan memprediksi hasil perkuliahan mahasiswa.

Ketepatan waktu lulus dapat diprediksi dengan menggunakan teknik klasifikasi. Klasifikasi adalah proses pencarian model yang dapat membedakan kelas data dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu obyek yang belum diketahui kelasnya. Penelitian ini menggunakan Naïve Bayes Classifier (NBC) dalam melakukan klasifikasi mahasiswa lulus tepat waktu atau tidak. NBC sangat cocok digunakan sebagai algoritma klasifikasi dengan beberapa kelebihan antara lain, sederhana, cepat, dan berakurasi tinggi (Pintoko & Muslim L., 2018).

Penelitian ini tidak hanya berfokus pada performa algoritma klasifikasinya saja, namun juga pada teknik pemilihan atribut atau fitur yang dianggap penting yang dikenal dengan istilah *feature selection*. Hasil dari analisis ini diharapkan dapat membantu perguruan tinggi dalam membuat kebijakan akademis agar dapat mengoptimalkan tingkat kelulusan mahasiswa pada tahun-tahun berikutnya.

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

II. METODE PENELITIAN

Alur pada penelitian ini menggunakan metode *Cross-Industry Standard Process for Data Mining* atau CRISP-DM. CRISP-DM merupakan standarisasi proses *data mining* sebagai strategi pemecahan masalah secara umum dari bisnis atau unit penelitian (Larose & Larose, 2014). Proses *data mining* berdasarkan CRISP-DM terdiri dari 6 tahapan, yaitu *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*.

Gambar 1. Proses CRISP-DM (Chapman et al., 2000)

A. *Business Understanding*

Business understanding atau pemahaman domain (penelitian). Pada tahap ini dibutuhkan pemahaman tentang substansi dari kegiatan data mining yang akan dilakukan, kebutuhan dari perspektif bisnis.

Pokok permasalahan pada penelitian ini terdiri dari 2, yaitu bagaimana membuat model yang akurat dalam memprediksi ketepatan masa studi dengan studi kasus mahasiswa Teknik Informatika UIN Alauddin Makassar dan menganalisis atribut apa saja yang paling berpengaruh dalam menentukan lama masa studi. Penelitian ini hanya berfokus dalam menganalisis 2 kondisi atau kelas pada data yaitu mahasiswa yang lulus tepat waktu (lulus dengan masa studi ≤ 8 semester) dan mahasiswa yang lulus terlambat (lulus dengan masa studi lebih > 8 semester), kondisi lain seperti mahasiswa drop out dll. tidak dimasukkan.

B. *Data Understanding*

Data Understanding atau pemahaman data adalah tahap mengumpulkan data awal, mempelajari data untuk bisa mengenal data yang akan dipakai.

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

Ada berbagai macam data yang tersimpan dalam sistem basis data PUSTIPAD UIN Alauddin Makassar namun hanya digunakan beberapa atribut dari 3 data yang berasal dari tabel berbeda yang dianggap berpotensi mempengaruhi lama masa studi mahasiswa, yaitu data profil mahasiswa (PM), data riwayat pengambilan mata kuliah (PMK), dan data IPS serta total SKS per semester (IPS). Data diambil dalam rentang tahun 2013-2018.

Tabel 1. Contoh data profil mahasiswa

NIM	Nama	Asal sekolah	Jenis kelamin	Jalur seleksi	Lama studi semester
60200113***	Nirwana****	MAN Pangkep	P	SNMPTN-PRESTASI	10
60200118***	Yuni *******	SMTA Lain-lain	P	UM-PTKIN	8

Tabel 2. Contoh data riwayat pengambilan mata kuliah

NIM	Angkatan	Kode MK	Nama MK	Tahun Ambil	Semester Ambil	Nilai
60200113***	2013	TIN2332	Basis Data	2014	2	A
60200118***	2018	TIN1318	Fisika	2018	1	A

Tabel 3. Contoh data IPS dan total SKS per semester

NIM	Angkatan	Semester Id	SKS	IPS
60200113***	2013	20132	22	3.73
60200118***	2018	20191	24	3.67

C.Data Preparation

Data preparation atau persiapan data. Aktivitas yang dilakukan antara lain memilih *table* dan *field* yang akan ditransformasikan ke dalam *database* baru untuk bahan *data mining*.

a) Transformasi data profil mahasiswa

1. Mengubah atribut “Asal Sekolah” menjadi “kategori_sekolah” Melakukan pengelompokan data menjadi 3 kategori “MA”, “SMA”, dan “SMK”.
2. Standarisasi nilai atribut “Jalur Seleksi” menjadi “jalur_seleksi” Terdapat perbedaan penamaan jalur seleksi selama tahun 2013-2018 sehingga beberapa nilai atribut butuh diseragamkan.

```
“SNMPTN-PRESTASI” → “SNMPTN”,
```

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

```
"SPMBPTAIN-PRESTASI" → "SPAN-PTKIN"
"UM-PTAIN" → "UM-PTKIN"
```

3. Membuat label "class_status"
   Membuat atribut *class target* prediksi yaitu "class_status" menggunakan atribut "Lama studi semester" dengan aturan:

```
IF (Lama studi semester <= 8) AND (Lama studi semester <> 0) THEN
      class_status ← "TEPAT"
ELSE IF (Lama studi semester > 8) THEN
      class_status ← "TERLAMBAT"
ELSE
      class_status ← "TIDAK DIKETAHUI"
ENDIF
```

Tabel 4. Contoh hasil transformasi data profil mahasiswa (PM)

Id	kategori_sekolah	jenis_kelamin	jalur_seleksi	class_status
60200113***	MA	P	SNMPTN	TERLAMBAT
60200118***	SMA	P	UM-PTKIN	TEPAT

b) Transformasi data riwayat pengambilan mata kuliah

Dilakukan fungsi *transpose* yaitu mengubah baris nilai atribut "Nama MK" menjadi kolom. Terdapat 41 mata kuliah yang digunakan yang ditawarkan pada semester 1-4 dalam rentang tahun 2013-2018.

Tabel 5. Contoh hasil transformasi data riwayat pengambilan mata kuliah (PMK)

Id	akidah akhlak	algoritma pemrograman	*...38 atribut lainnya...*	teori bahasa dan automata
60200113***	A	A	…	A
60200118***	A	A	…	A

c) Transformasi data IPS dan SKS per semester

1. Membuat atribut "Semester" dan melakukan *filtering*
   Dengan menghitung selisih antara "Semester Id" dan atribut "Angkatan". Setelah itu dilakukan *filtering* data untuk mengambil data pada semester 1-4.
2. *Transpose* atribut "IPS" dan "SKS"
   Dilakukan transformasi berupa fungsi *transpose* yaitu mengubah baris nilai atribut "IPS" dan "SKS" menjadi kolom seperti yang terlihat pada tabel 6.

Tabel 6. Contoh hasil transformasi data IPS dan SKS per semester (IPS)

Id	SKS1	IPS1	SKS2	IPS2	SKS3	IPS3	SKS4	IPS4
60200113***	19	3.68	22	3.55	24	3.63	22	3.32
60200118***	23	3.43	22	3.45	24	3.67	23	3.87

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

d) Melakukan fungsi *JOIN* terhadap tabel data PM + PMK + IPS

Data PM, PMK, dan IPS kemudian digabungkan menggunakan *query* fungsi *JOIN* dengan "Id" sebagai *key* penghubung antar tabel. Hasil transformasi berupa *dataset* yang terdiri dari 54 atribut dan 267 baris.

Tabel 7. Deskripsi *dataset* hasil transformasi

No	Atribut	Sumber Data	Deskripsi	Tipe Data
1	Id	PM	nomor induk mahasiswa	numerik
2	kategori_sekolah	PM	kategori sekolah {"MA", "SMA", "SMK"}	kategorik
3	jenis_kelamin	PM	jenis kelamin {"L", "P"}	kategorik
4	jalur_seleksi	PM	jalur seleksi {"SPAN-PTKIN", "UMM", "SNMPTN", "SBMPTN", "UM-PTKIN"}	kategorik
5	akidah_akhlak	PMK	nilai mata kuliah Akidah Akhlak {"A", "B", "C", "D", "E"}	kategorik
…	…	PMK	*nilai 40 mata kuliah lainnya {"A", "B", "C", "D", "E"}*	kategorik
46	IPS1	IPS	Indeks Prestasi Semester 1. *min* = 0.0, *max* = 4.0	numerik
47	IPS2	IPS	Indeks Prestasi Semester 2. *min* = 0.0, *max* = 4.0	numerik
48	IPS3	IPS	Indeks Prestasi Semester 3. *min* = 0.0, *max* = 4.0	numerik
49	IPS4	IPS	Indeks Prestasi Semester 4. *min* = 0.0, *max* = 4.0	numerik
50	SKS1	IPS	Jumlah SKS yang diambil pada semester 1. *min* = 0, *max* = 24	numerik
51	SKS2	IPS	Jumlah SKS yang diambil pada semester 2. *min* = 0, *max* = 24	numerik
52	SKS3	IPS	Jumlah SKS yang diambil pada semester 3. *min* = 0, *max* = 24	numerik
53	SKS4	IPS	Jumlah SKS yang diambil pada semester 4. *min* = 0, *max* = 24	numerik
54	**class_status**	PM	Label / *class* target prediksi {"TEPAT", "TERLAMBAT"}	kategorik

D.Modeling

Modeling adalah tahap menentukan teknik *data mining*, *tools*, serta menentukan parameter dengan nilai yang optimal. Tahap *modeling* diawali dengan melakukan *feature selection* menggunakan metode Information Gain dan Correlation Attribute untuk mengevaluasi atribut apa saja yang mempunyai pengaruh relevan dalam menentukan lama masa studi mahasiswa. Kemudian 2 metode tersebut dikombinasikan dengan algoritma Naïve Bayes Classifier untuk

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

melakukan prediksi "class_status" pada *dataset*, lalu dievaluasi untuk menemukan kombinasi yang menghasilkan performa terbaik. Tahap *modeling* dilakukan menggunakan aplikasi WEKA sebagai *tools*.

a) *Feature Selection*

Feature selection adalah salah satu teknik *data mining* yang umum digunakan untuk mengurangi kompleksitas atribut pada tahap *preprocessing*. *Feature selection* membantu memilih fitur yang berpengaruh (fitur optimal) dan mengesampingkan fitur yang tidak berpengaruh sehingga mempercepat proses pemodelan (Han et al., 2012).

1. Information Gain (IG)

   IG adalah salah satu metode dari *feature selection* yang akan merangking fitur pada dataset yang diberikan dengan menghitung *entropy* dari salah satu kelas sebelum dan setelah melakukan proses pengamatan terhadap fitur yang ada pada satu data yang sama (Purbasari et al., 2013).

2. Correlation Attribute (CA)

   CA mengevaluasi atribut-atribut sehubungan dengan kelas target. Metode korelasi Pearson digunakan untuk mengukur korelasi antara setiap atribut dengan *class* target. Metode ini mempertimbangkan atribut nominal dalam basis nilai dan setiap nilai bertindak sebagai indikator (Gnanambal et al., 2018).

b) Naïve Bayes Classifier (NBC)

Naïve Bayes Classifier (NBC) adalah algoritma pengklasifikasian probabilitas sederhana yang mengaplikasikan Teorema Bayes dengan asumsi tidak ada ketergantungan (independent) yang tinggi. Karena diasumsikan sebagai variabel independen maka hanya varians dari variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians (Han et al., 2012). Berikut ini merupakan persamaan umum Naïve Bayes:

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

$$p\,C\,F1,\ldots,Fn = \frac{p\,C\;p\,F1,\ldots,Fn\,C}{p\;F1,\ldots,Fn} \quad (1)$$

Dimana C merepresentasikan kelas, sementara F1...Fn merepresentasikan berbagai karakteristik petunjuk yang dibutuhkan dalam melakukan klasifikasi. Maka persamaan tersebut menjelaskan bahwa peluang masuknya sebuah data dengan karakteristik tertentu ke dalam kelas C adalah peluang munculnya kelas C sebelum masuknya data tersebut dikalikan dengan peluang munculnya berbagai karakteristik data pada kelas C dibagi dengan peluang kemunculan karakteristik data secara global (*evidence*).

E.Evaluation

Evaluation adalah tahap interpretasi terhadap hasil *data mining* yang ditunjukan dalam proses pemodelan pada tahap sebelumnya. Pengujian model pada penelitian ini merupakan pengujian data terhadap algoritma. Akurasi dan *mean absolute error* (MAE) algoritma dihitung menggunakan *confusion matrix* dengan 10 folds cross-validation. Cross-validation dipilih karena dianggap mampu memberikan hasil yang lebih baik dibandingkan dengan metode validasi lainnya (Horvat et al., 2020).

F.Deployment

Deployment atau penyebaran adalah tahap penyusunan laporan atau presentasi dari pengetahuan yang didapat pada evaluasi pada proses *data mining*.

III. HASIL DAN PEMBAHASAN

A. *Feature Selection*

Eksperimen dalam implementasi teknik feature selection menggunakan rangking dalam jumlah fitur tertentu (n) yaitu 5, 10, dan 15. Adapun hasilnya dapat dilihat pada tabel 8. Hasil feature selection baik menggunakan IG maupun CA sebagian besar menghasilkan fitur yang serupa, terlihat dari jumlah fitur IG∩***CA.*** Selain itu diketahui pula bahwa tidak satupun fitur terpilih berasal dari tabel profil mahasiswa, dengan kata lain data profil dianggap kurang berpengaruh dalam menentukan lama masa studi.

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

Tabel 8. Hasil *feature selection*

Jumlah Fitur (*n*)	Fitur / Atribut Terpilih		
	Information Gain (IG)	Correlation Attribute (CA)	Jumlah IG ∩ CA
5	sks4, teknologi_dan_desain_web, matematika_diskrit, pemrograman_berorientasi_objek, pengantar_teknologi_informasi	matematika_diskrit, teknologi_dan_desain_web, ips1, pengantar_teknologi_informasi, ips3	3
10	sks4, teknologi_dan_desain_web, matematika_diskrit, pemrograman_berorientasi_objek, pengantar_teknologi_informasi, basis_data, sks1, fisika, ips3, ips1	matematika_diskrit, teknologi_dan_desain_web, ips1, pengantar_teknologi_informasi, ips3, ips4, pemrograman_berorientasi_objek, sks1, etika_profesi, kewirausahaan	7
15	sks4, teknologi_dan_desain_web, matematika_diskrit, pemrograman_berorientasi_objek, pengantar_teknologi_informasi, basis_data, sks1, fisika, ips3, ips1, sistem_operasi_komputer, teknologi_informasi, kewirausahaan, probabilitas_dan_statistik, ilmu_fikih	matematika_diskrit, teknologi_dan_desain_web, ips1, pengantar_teknologi_informasi, ips3, ips4, pemrograman_berorientasi_objek, sks1, etika_profesi, kewirausahaan, sks4, ilmu_fikih, ips2, kecerdasan_buatan, matematika_komputer	10

B. *Hasil Prediksi*

Hasil seleksi fitur akan dikombinasikan dan diuji pada algoritma NBC untuk mengetahui performa akurasi dan MAE yang dihasilkan dari masing-masing teknik feature selection. Hasil uji coba prediksi "class_status" pada dataset dapat dilihat pada tabel 9.

Tabel 9. Hasil uji akurasi dan MAE algoritma NBC dan *feature selection*

Algoritma	***NBC***	***NBC+IG (n=5)***	***NBC+IG (n=10)***	***NBC+IG (n=15)***	***NBC+CA (n=5)***	***NBC+CA (n=10)***	***NBC+CA (n=15)***
Akurasi	***69.2%***	***79.3%***	***73.3%***	***74.8%***	***81.2%***	***74.8%***	***72.9%***
MAE	***0.2903***	***0.2559***	***0.2699***	***0.2673***	***0.2621***	***0.2781***	***0.2847***

Berdasarkan hasil uji coba diketahui bahwa feature selection secara umum mampu meningkatkan akurasi NBC. Namun kombinasi NBC+CA memiliki akurasi lebih baik dibandingkan NBC+IG. Akurasi terbaik diperoleh pada NBC+CA dengan jumlah fitur (n) = 5 sebesar 81.2%, meningkat 12% dibandingkan NBC tanpa feature selection; Sedangkan yang terendah pada NBC+CA dengan jumlah fitur (n) = 15 sebesar 72.9%. Tingkat kesalahan prediksi dari IG dan CA relatif rendah dan tidak jauh berbeda dilihat dari angka mean absoute error berkisar antara 0.26-0.28**.**

JURNAL INSTEK
INFORMATIKA SAINS DAN TEKNOLOGI
Volume 7, Nomor. 2, Oktober 2022
P –ISSN : 2541-1179, E-ISSN : 2581-1711
OJS :http://journal.uin-alauddin.ac.id/index.php/instek/index
Email : instek@uin-alauddin.ac.id

IV. KESIMPULAN

Dari uji coba yang telah dilakukan dapat disimpulkan ***bahwa feature selectio***n secara umum mampu meningkatkan akurasi NBC. Kombinasi NBC+CA memiliki akurasi lebih baik dibandingkan NBC+IG. Akurasi terbaik diperoleh pada NBC+CA dengan jumlah fitur (n) = 5 sebesar 81.2%, meningkat 12% dibandingkan NBC tanpa feature selection. Selain itu diperoleh pula fitur atau atribut yang dianggap berpengaruh dalam menentukan masa studi dari hasil proses feature selection yang dapat dijadikan bahan evaluasi dalam membuat kebijakan akademis agar dapat mengoptimalkan tingkat kelulusan mahasiswa pada tahun-tahun berikutnya.

DAFTAR PUSTAKA

Baker, R. S. J. d., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–17. https://doi.org/10.5281/zenodo.3554657

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guides.

Gnanambal, Dr. S., Thangaraj, Dr. M., Meenatchi, Dr. V. T., & Gayathri, Dr. V. G. (2018). Classification Algorithms with Attribute Selection: An evaluation study using WEKA. *Int. J. Advanced Networking and Applications*, *9*(6), 3640–3644.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (Third Edition). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00018-6

Horvat, T., Havaš, L., & Srpak, D. (2020). The Impact of Selecting a Validation Method in Machine Learning on Predicting Basketball Game Outcomes. *Symmetry*, *12*(3). https://doi.org/10.3390/sym12030431

Khasanah, A. U. & Harwati. (2017). A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques. *IOP Conference Series: Materials Science and Engineering*, *215*, 012036. https://doi.org/10.1088/1757-899X/215/1/012036

Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining, Second Edition. John Wiley & Sons, Ltd.

Pintoko, B. M., & Muslim L., K. (2018). Analisis Sentimen Jasa Transportasi Online pada Twitter Menggunakan Metode Naïve Bayes Classifier. *e-Proceeding of Engineering*, *5*(3), 8121.

Purbasari, I. Y., Nugroho, B., & Madya, J. R. (2013). Benchmarking Algoritma Pemilihan Atribut Pada Klasifikasi Data Mining. 8.