JEPIN
(Jurnal Edukasi dan Penelitian Informatika)
ISSN(e): 2548-9364 / ISSN(p) : 2460-0741

Vol. 8
No. 2
Agustus
2022

Penilaian Kinerja Akurasi Metode Klasifikasi dalam Dataset Penerimaan Mahasiswa Baru Universitas XYZ

Indra Griha Tofik Isa[#1], Febie Elfaladonna[#2]

[#]*Jurusan Manajemen Informatika, Politeknik Negeri Sriwijaya*
Jl. Srijaya Negara, Palembang, Indonesia

[1]indra_isa_mi@polsri.ac.id
[2]febie_elfaladonna_mi@polsri.ac.id

Abstrak— **Universitas XYZ merupakan salah satu Perguruan Tinggi yang berlokasi di Kota Palembang yang melakukan kegiatan Penerimaan Mahasiswa Baru (PMB) untuk menjaring calon mahasiswa. Data PMB dari tahun ke tahun belum digunakan secara optimal dalam menghasilkan pengetahuan yang memberikan nilai manfaat bagi pengguna, sehingga diperlukan sebuah pemodelan data yang efisien dan tepat untuk menghasilkan akurasi data yang baik. Penelitian yang dilakukan bertujuan untuk menilai kinerja akurasi pemodelan yang terdapat dalam metode klasifikasi yang meliputi pemodelan *k-NN, Decision Tree Classifier, Naive Bayes Classifier, Support Vector Machine* (SVM) dan *AdaBoost* terhadap fitur dalam *dataset* Penerimaan Mahasiswa Baru (PMB) yang digunakan untuk memprediksi preferensi pemilihan program studi. 26 Fitur dalam *dataset* diamati hingga menghasilkan 6 fitur yang memiliki nilai korelasi yang tinggi untuk dilibatkan dalam penilaian kinerja akurasi, yang meliputi 'Jurusan Sekolah', 'Penghasilan', 'Tahun Masuk', 'Tahun Lulus', 'Tipe Sekolah' dan 'Status Sekolah' dengan data *record* sebanyak 2.704 data. Tahapan dilakukan menggunakan *Data Life Cycle* yang meliputi: (1) *Business Understanding* yang terdiri dari Penentuan Masalah, Tujuan Proyek, Solusi dari Perspektif Bisnis, dan Instrumen Pengukuran Keberhasilan; (2) *Data Understanding* dengan penelaahan data; (3) *Data Preparation*; (4) *Modeling*; (5) *Evaluation*. Hasil akhir menunjukkan bahwa *k-NN classifier* memiliki persentasi akurasi tertinggi sebesar 72.2% dan direkomendasikan dalam pemodelan preferensi program studi bagi calon mahasiswa baru di Universitas XYZ Kota Palembang.**

Kata kunci— ***Data Mining*, Klasifikasi, Penerimaan Mahasiswa Baru, Dataset PMB, Penilaian Akurasi**

I. PENDAHULUAN

Perguruan Tinggi merupakan lembaga pendidikan yang menghasilkan sumber daya manusia berkualitas, kompetitif dan siap dalam memenuhi kebutuhan tenaga kerja [1]. Seluruh elemen dalam Perguruan Tinggi bersinergi dari tahapan *input*, tahapan proses hingga tahapan *output*. Dari segi input diawali dengan perekrutan mahasiswa yang diagendakan dalam Penerimaan Mahasiswa Baru (PMB) dimana merupakan kegiatan rutin tahunan Perguruan Tinggi dengan tujuan untuk mencari potensi dari calon mahasiswa [2]. Universitas XYZ merupakan salah satu Perguruan Tinggi yang berlokasi di Kota Palembang yang melakukan kegiatan PMB melalui jalur mandiri dan jalur undangan. Proses PMB di Universitas XYZ dilakukan dengan beberapa tahapan, yakni tahapan seleksi administrasi, seleksi tertulis dengan ujian berbasis komputer, seleksi wawancara hingga penetapan calon mahasiswa menjadi mahasiswa baru. Pendataan PMB dilakukan dengan cara pengarsipan baik secara terkomputerisasi maupun bentuk fisik. Dari tahun ke tahun selama proses PMB, dokumen maupun data tersebut hanya digunakan sebagai laporan rutin tahunan unit PMB kepada Perguruan Tinggi dan belum dioptimalkan sebagai bahan untuk analisis data maupun parameter dalam menentukan kebijakan. Padahal dengan banyaknya data PMB yang dihasilkan tersebut dapat digali informasi yang kemudian menjadi pengetahuan untuk melihat bagaimana tren dari pendaftar, potensi calon mahasiswa, dan aspek teknis lainnya.

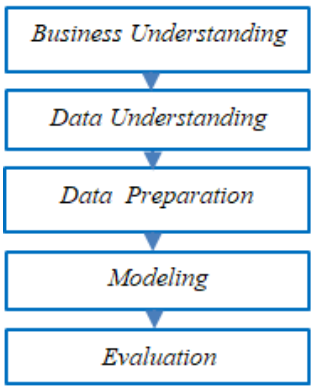
Pemanfaatan data-data historis dapat menjadi sebuah aset pengetahuan baru yang memiliki nilai guna bagi pihak terkait, khususnya untuk menentukan arah kebijakan dengan melihat tren dan potensi yang tergali dari data [3]. Beberapa penelitian sebelumnya yang dilakukan oleh Saini Prianka pada tahun 2014 yang dilatarbelakangi bahwa tingginya minat pendaftar pada perguruan tinggi yang menyebabkan perguruan tinggi di India berkompetisi untuk melakukan strategi dalam menarik minat pendaftar. Sementara dengan tingginya pendaftar tersebut berbanding lurus dengan data pendaftaran yang besar dan kompleks. Sehingga dalam penelitian ini mengimplementasikan metode klasifikasi data mining dengan *decision tree*. Hasil akhir berupa strategi pola pemasaran yang disesuaikan dengan segmentasi dan karakteristik pendaftar [4]. Dalam penelitian [5] menyatakan bahwa data mining diperlukan

dalam manajemen perguruan tinggi yang terbagi ke dalam 3 aspek, yakni mahasiswa, kanselir akademik, dan manajemen pendidikan. Khususnya dari aspek mahasiswa tidak terbatas pada pendaftaran mata kuliah, pembatalan mata kuliah, pengajuan cuti maupun keuangan. Dalam optimalisasi *data mining* dapat dilakukan dengan membandingkan beberapa metode yang sejenis, seperti pada penelitian [6] dengan pendekatan *Educational Data Mining* (EDM) dalam memprediksi kinerja akademik mahasiswa yang bertujuan untuk meningkatkan kualitas pendidikan dan menekan tingkat *drop out* (DO). Analisis dilakukan pada 300 *record* data dengan metode klasifikasi melalui 6 komparasi algoritma, yakni C4.5, *Simple CART, LADTree, Naïve Bayes, Bayes Net with ADTree*, dan *Random Forest*. Hasil akhir menunjukkan Naive Bayes dan Random Forest memiliki nilai akurasi yang lebih baik dibandingkan algoritma lainnya dalam kasus tersebut. Yulia dan Santoso mengimplementasikan data mining pada 55 atribut dan 1665 *record* mahasiswa dengan melakukan klasterisasi menggunakan *K-Means*, lalu analisis lanjutan dengan metode klasifikasi menggunakan *decision tree* dan *naive bayes*. Hasil akhir menunjukkan bahwa algoritma *decision tree* memiliki nilai kinerja lebih baik dalam mengolah data uji dibandingkan dengan naive bayes dalam kasus tersebut [7]. *Data Mining* merupakan salah satu metode yang digunakan dalam penggalian data dari sekumpulan banyak data seperti data PMB Universitas XYZ. *Data mining* umumnya diasosiasikan dengan *Knowledge Discovery in Database* (KDD), dimana faktanya data mining menjadi bagian dari KDD tersebut [8]. *Data Mining* memiliki pengertian yaitu proses untuk menemukan pola tertentu dari sekumpulan atau sejumlah data yang besar dan kompleks. Berbagai sumber data yang digunakan dapat berasal dari basis data, *data warehouse, website*, repositori informasi lainnya, atau data yang diinput [9]. Perkembangan data mining yang menyasar dunia pendidikan memunculkan istilah baru, yakni *Educational Data Mining* (EDM). EDM merupakan kecabangan dari data mining yang mendalami tentang pola yang dihasilkan dalam proses pendidikan, baik tahapan input, proses pembelajaran hingga output. Tujuan dari EDM ini adalah bagaimana menerapkan pola yang cocok, termasuk untuk memetakan potensi yang muncul dalam PMB [10]. Kualitas dunia pendidikan perlu ditingkatkan secara berkelanjutan dan EDM merupakan salah satu tool yang dapat digunakan dalam peningkatan tersebut. Sehingga top level management Perguruan Tinggi dapat memanfaatkan hasil dari eksperimen data mining untuk memahami tren dan perilaku mahasiswa yang dapat mengarah pada desain strategi baru [11].

Beberapa pemodelan yang digunakan dalam metode klasifikasi antara lain: *Naive Bayes, kNN, AdaBoost, Random Forest, Decision Tree, Neural Network*, dan sebagainya [12]. Metode klasifikasi dapat diterapkan untuk mengukur analisis sentimen dalam interaksi antar entitas [13]. Penelitian yang dilakukan bertujuan untuk mengukur kinerja pemodelan/ algoritma yang terdapat dalam metode klasifikasi yang meliputi *k-NN, Decision Tree Classifier, Naive Bayes Classifier, Support Vector Machine (SVM) dan AdaBoost* terhadap fitur dalam PMB yang digunakan untuk memprediksi preferensi pemilihan program studi. Dari analisis tersebut dihasilkan algoritma mana yang memiliki nilai akurasi yang tinggi untuk diterapkan dalam pemetaan PMB Universitas XYZ dan diharapkan dapat menjadi salah satu rekomendasi strategi PMB Universitas XYZ dalam melihat potensi preferensi pemilihan program studi calon mahasiswa berdasarkan fitur-fitur yang diamati melalui *data mining*. Adapun ruang lingkup dari penelitian ini adalah (1) Fitur yang digunakan merupakan data yang terdapat dalam PMB Universitas XYZ; (2) Data berasal dari PMB Universitas XYZ Periode 2018-2020; (3) Tahapan dalam *data life cycle terdiri dari Business Understanding, Data Understanding, Data Preparation, Modeling,* dan Evaluasi; (4) Evaluasi dilakukan dengan pengukuran akurasi kinerja dari pemodelan yang diimplementasikan, dan (5) *Tools* yang digunakan adalah Jupyter Notebook (anaconda3) dan bahasa pemrograman Python.

II. METODE PENELITIAN

Metode penelitian dilakukan dengan pendekatan *Data Life Cycle* yang mengikuti pola *Business Understanding, Data Understanding, Data Preparation, Modeling, dan Evaluation.* Metode yang diamati adalah klasifikasi dengan *pemodelan k-NN, Decision Tree Classifier, Naive Bayes Classifier, Support Vector Machine (SVM) dan AdaBoost.* Fitur yang diamati adalah sejumlah 26 fitur, yang akan menghasilkan beberapa fitur sehingga menjadi parameter dalam preferensi calon mahasiswa baru untuk menentukan pemilihan program studi. Data *record* yang diamati sebanyak 2704 data yang berasal dari data PMB Universitas XYZ Palembang. Gambar 1 berikut merupakan bagan tahapan metode penelitian.



Gambar 1. Tahapan metode penelitian

A. Business Understanding

Untuk mengawali proses *data mining* ataupun *data science* diperlukan pemahaman bisnis yang berorientasi pada pencapaian hasil *(business understanding).* Terdapat aspek yang dilihat dalam tahapan *business understanding*, yakni (1)Penentuan Masalah; (2) Tujuan Proyek; (3) Solusi dari Perspektif Bisnis; (4) Instrumen Pengukuran Keberhasilan [14].

B. *Data Understanding*

Tahapan lebih lanjut merupakan *data understanding* atau pemahaman data. Hal yang dilakukan berupa pengambilan serta menelaah data dengan tujuan untuk mendapatkan gambaran utuh atas data-data yang diperoleh sebagai bahan solusi pemecahan masalah dari permasalahan bisnis [15]. Di dalam tahapan sebelumnya *(business understaning)* memungkinkan terjadinya kesalahan dalam menentukan formulasi bisnis sehingga menjadi tidak tepat sasaran. Oleh karenanya, tahapan *data understanding* memungkinkan pengulangan tahapan *business understanding* untuk memperbaiki definisi atau ruang lingkup permasalahan bisnis tersebut. Pada penelitian yang dilakukan, data diambil dari *spreadsheet* PMB Universitas XYZ Kota Palembang periode 2018, 2019 dan 2020. Jumlah *record* data yang diamati adalah 2704 record dengan 26 fitur dan 1 kelas. Tabel 1 berikut menunjukkan karakteristik dari fitur yang diamati:

TABEL I
KARAKTERISTIK FITUR YANG DIAMATI

No	Fitur	Tipe Data
1	ID	Int64
2	Jenis Kelamin	Object
3	Agama	Object
4	Tempat Lahir	Object
5	Provinsi	Object
6	Kota	Object
7	Anak Ke	Int64
8	Jumlah Saudara	Int64
9	Penghasilan	Int64
10	Jenjang	Object
11	Program Kuliah	Object
12	Program Studi	Object
13	Status Mahasiswa	Object
14	Pembimbing	Object
15	ID PA	Object
16	Batas Studi	Object
17	Tahun Masuk	Int64
18	Jenis Sekolah	Object
19	Nama Sekolah	Object
20	Jurusan Sekolah	Object
21	Nilai Unas	Float64
22	Tahun Lulus	Float64
23	Jenis Beasiswa	Object
24	Kota Orangtua	Int64
25	Tahun Terakhir KRS	Object
26	IPK	Object

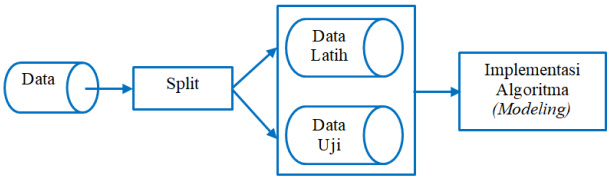
Dalam tahapan ini, dilakukan uji korelasi pada 26 fitur tersebut sehingga dihasilkan beberapa fitur yang memiliki nilai korelasi tinggi yang akan digunakan pada tahapan berikutnya.

C. *Data Preparation*

Data Preparation atau dikenal dengan *pre-processing* data merupakan proses transformasi untuk membersihkan data. Transformasi yang dilakukan dengan mengubah data mentah menjadi format yang dipahami untuk dianalisis sehingga memberikan informasi. Proses *data preparation* dilakukan dengan pembersihan, transformasi dan konsolidasi data[16]. Didalam penelitian ini dilakukan *data preparation* karena beberapa faktor, antara lain: (1) Data perlu diformat sesuai dengan kebutuhan perangkat lunak Jupyter Notebook; (2) Data mentah cenderung 'kotor', seperti data tidak lengkap, terdapat nilai yang outlier, inkonsistensi nilai antar atribut terkait, tertukar antara kolom dan baris; (3) Dalam satu kolom yang sama terdapat banyak variabel.

D. *Modeling*

Tahapan *modeling* merupakan implementasi algoritma dalam metode klasifikasi [17]. Dalam tahap ini dilakukan pembagian data latih dan data uji. Data latih digunakan untuk mengembangkan model dan data uji digunakan untuk mengukur kinerja dari model yang diimplementasikan. Adapun model yang diimplementasikan untuk dilihat manakah yang memiliki kinerja terbaik dalam penelitian ini yaitu *k-Nearest Neighbor (k-NN) Classifier, Decision Tree Classifier, Naive Bayes Classifier, Support Vector Machine (SVM), dan Adaptive Boosting (AdaBoost).* Gambar 2 menunjukkan tahapan dalam proses *modeling*:



Gambar 2. Tahapan *modeling*

E. *Evaluasi*

Hasil implementasi metode klasifikasi dengan beberapa algoritma pada tahapan *modeling* dapat dilihat pada tahapan evaluasi. Tahapan ini dilakukan pengukuran kinerja dengan mengukur nilai akurasi. Dalam evaluasi dapat dilihat apakah model atau algoritma yang diimplementasikan layak atau tidak untuk digunakan. Nilai Akurasi didapatkan dengan membandingkan jumlah prediksi yang tepat terhadap sejumlah prediksi yang dilakukan dari data uji kepada data latih. Nilai akurasi dinyatakan dalam bentuk nilai yang telah dikonversi ke dalam persentasi [18].

III. HASIL DAN PEMBAHASAN

A. *Business Understanding*

Business Understanding diterjemahkan ke dalam aspek awal yang mendefinisikan tujuan dan ruang lingkup domain yang dicermati, meliputi penentuan masalah, tujuan proyek, solusi dan perspektif bisnis. Tabel 2 berikut

menunjukkan secara ringkas dari 4 aspek dalam *business understanding*:

TABEL II
ASPEK *BUSINESS UNDERSTANDING*

No	Aspek	Deskripsi
1	Penentuan Masalah	a. PMB merupakan salah satu bagian vital dalam Perguruan Tinggi khususnya Universitas XYZ Kota Palembang b. Data PMB semakin banyak tiap tahun dan tidak dimanfaatkan secara optimal sebagai referensi strategi bagi PMB Universitas XYZ c. Diperlukan pemanfaatan data PMB melalui *data mining*, sehingga terjadi optimalisasi data historis PMB tersebut. d. Data PMB merupakan data terstruktur sehingga diperlukan metode *supervised learning* melalui klasifikasi. e. Beberapa metode klasifikasi perlu diuji dan dianalisis untuk mengetahui efektifitas akurasi
2	Tujuan Proyek	Pemanfaatan data PMB Universitas XYZ Kota Palembang melalui pengukuran efektifitas akurasi beberapa algoritma dalam metode klasifikasi untuk mengetahui preferensi pemilihan program studi calon mahasiswa
3	Solusi dari Perspektif Bisnis	Implementasi *data mining* pada data PMB Universitas XYZ Kota Palembang dengan metode klasifikasi untuk memberikan rekomendasi preferensi pemilihan program studi oleh pendaftar atau calon mahasiswa berdasarkan kelas tertentu
4	Instrumen Pengukuran Keberhasilan	Terdapat model preferensi pemilihan program studi yang memiliki tingkat akurasi terbaik melalui pengukuran akurasi

B. Data Understanding dan Data Preparation

Guna memudahkan analisis data maka dilakukan penelaahan dengan melihat keseluruhan atribut (yang selanjutnya disebut dengan fitur) yang dimulai dengan pembacaan dataset untuk melihat komposisi 26 fitur data PMB Universitas XYZ. Hasil dari pembacaan data dapat dilihat pada tabel 3 berikut:

TABEL III
KOMPOSISI 26 FITUR DATA PMB UNIVERSITAS XYZ KOTA PALEMBANG

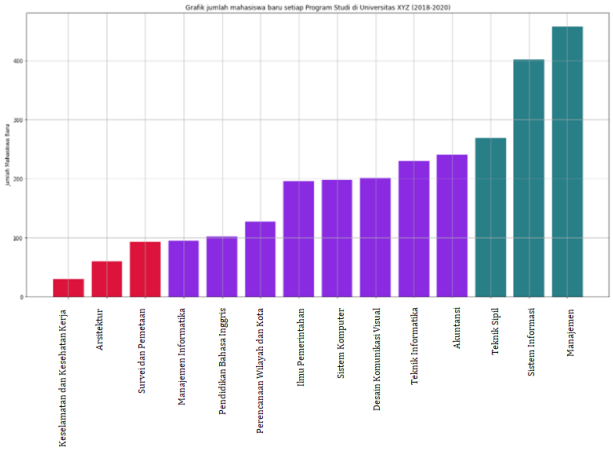
No	ID	Jenis Kelamin	...	Tahun Masuk	Jenis Sekolah	Jurusan Sekolah
0	181	L	...	2018	SMK NEGERI	IPA
1	182	L	...	2018	SMK NEGERI	T.ELEKTRONIKA
2	183	P	...	2018	SMU NEGERI	IPS
3	184	L	...	2018	SMU SWASTA	IPA
4	185	L	...	2018	SMK NEGERI	TEKNIK GAMBAR BANGUNAN
5	186	L	...	2018	SMU NEGERI	IPA
6	187	P	...	2018	SMU NEGERI	IPA
7	188	P	...	2018	SMU NEGERI	IPA
8	189	L	...	2018	SMK NEGERI	TEKNIK KEN-DARAAN RINGAN
9	1810	L	...	2018	SMU NEGERI	IPA
...	...	...	...	...	...	...

Dataset pada tabel 3 di atas berdasarkan klasifikasi "Program Studi" terdapat 14 data *record unique* yang terdiri dari : (1) Desain Komunikasi Visual; (2) Teknik Informatika; (3) Sistem Informasi; (4) Manajemen Informatika; (5)Arsitektur; (6) Survei dan Pemetaan; (7) Perencanaan Wilayah dan Kota; (8) Ilmu Pemerintahan; (9) Manajemen; (10)Akuntansi; (11) Keselamatan dan Kesehatan Kerja; (12) Pendidikan Bahasa Inggris; (13) Sistem Komputer; (14) Teknik Sipil. Selanjutnya kelas "Program Studi" dilakukan visualisasi untuk melihat jumlah data dari masing-masing *record unique* tersebut, dan didapatkan jumlah data tertinggi adalah "Manajemen" dengan total 458 data dan data terendah adalah "Keselamatan dan Kesehatan Kerja" sejumlah 30 data seperti yang ditunjukkan pada gambar 3:

	Program Studi	Counts
0	Keselamatan dan Kesehatan Kerja	30
1	Arsitektur	60
2	Survei Dan Pemetaan	93
3	Manajemen Informatika	95
4	Pendidikan Bahasa Inggris	102
5	Perencanaan Wilayah Dan Kota	128
6	Ilmu Pemerintahan	196
7	Sistem Komputer	198
8	Desain Komunikasi Visual	201
9	Teknik Informatika	230
10	Akuntansi	241
11	Teknik Sipil	269
12	Sistem Informasi	402
13	Manajemen	458

Gambar 3. Hasil *sorting* jumlah data "program studi"

Hasil visualisasi grafik pada "Program Studi" dibuat dengan grafik batang. Penyesuaian skema warna sesuai dengan kategori rendah, sedang dan tinggi. Untuk kategori rendah berada pada range < 95 data, kategori sedang pada range 95 hingga 268 data, dan kategori tinggi pada range > 268 data seperti yang divisualisasikan pada gambar 4.



Gambar 4. Visualisasi grafik "program studi"

Dari fitur yang ada, tidak seluruhnya dapat digunakan untuk implementasi pemodelan. Sehingga perlu dianalisis korelasi yang memiliki nilai keterhubungan yang tinggi dengan kelas "Program Studi". Gambar 5 berikut memberikan gambaran korelasi antar fitur dalam dataset PMB Universitas XYZ Kota Palembang.

	ID	Anak Ke	Jumlah Saudara	Penghasilan	Tahun Masuk	Nilai Unas	Tahun Lulus
ID	1.000000	0.011252	0.010079	-0.040218	0.118250	0.042453	-0.027439
Anak Ke	0.011252	1.000000	0.843437	-0.190186	0.171422	0.111815	0.023778
Jumlah Saudara	0.010079	0.843437	1.000000	-0.233103	0.183085	0.117340	0.022527
Penghasilan	-0.040218	-0.190186	-0.233103	1.000000	-0.101607	-0.124192	0.049634
Tahun Masuk	0.118250	0.171422	0.183085	-0.101607	1.000000	0.067405	-0.039846
Nilai Unas	0.042453	0.111815	0.117340	-0.124192	0.067405	1.000000	0.014741
Tahun Lulus	-0.027439	0.023778	0.022527	0.049634	-0.039846	0.014741	1.000000

Gambar 5. Hasil korelasi antar fitur *dataset* PMB Universitas Kota Palembang

Gambar di atas hanya memperhitungkan korelasi untuk data numerikal, maka diperlukan perhitungan korelasi untuk data-data yang bersifat teks maupun kategorikal seperti data Jurusan, Pekerjaan Orang Tua, Jenis Kelamin dan sebagainya sehingga dapat dilakukan analisis korelasi antara fitur-fitur tersebut. Pada tahapan berikutnya adalah membuat pengkodean dengan angka pada fitur-fitur tersebut sebagai bahan dalam analisis korelasi. Setelah dilakukan analisis korelasi terhadap seluruh fitur yang terdapat dalam *dataset* PMB Universitas XYZ Kota Palembang, nilai skor yang tinggi berada pada nilai >0.7[19], maka didapatkan 6 fitur yang memiliki nilai korelasi yang tinggi terhadap kelas "Program Studi" sebagaimana terdapat dalam tabel 4:

TABEL IV
FITUR DENGAN NILAI > 0.7

No	Nama Fitur	Nilai Korelasi
1	Jurusan Sekolah	0.760
2	Penghasilan	0.713
3	Tahun Masuk	0.731
4	Tahun Lulus	0.716
5	Tipe Sekolah	0.743
6	Status Sekolah	0.701

Dari hasil seleksi fitur pada dataset tersebut dilakukan pengecekan *data record* untuk melihat apakah terdapat *outlayer* atau *missing value* [20]. Beberapa data record pada fitur dan kelas tersebut didapatkan *missing value* yang direpresentasikan dengan data yang bernilai "nan" pada fitur "Tahun Lulus", "Tipe Sekolah", "Status Sekolah" dan kelas "Program Studi".

Karena kelas "Program Studi" menjadi target dalam klasifikasi preferensi pilihan program studi pada *dataset* PMB Universitas XYZ Kota Palembang, maka dilakukan perhitungan data untuk melihat sebaran data dari kelas "Program Studi".

Dari hasil sebaran kelas "Program Studi" didapatkan bahwa data tidak seimbang, dimana pada gambar 6 (kiri) menunjukkan data nilai yang jauh berbeda terutama antara kelas "Program Studi" 8, 3, 12, 14 dan 2 terhadap kelas "Program Studi" 9. Sehingga dilakukan penyeimbangan data *(Balancing)* dengan menggunakan *library* SMOTE yang bertujuan agar sebaran data merata dan meningkatkan kualitas dalam implementasi pemodelan yang dilakukan pada tahap selanjutnya. Gambar 6 (kanan) merupakan hasil dari balancing data dengan data record yang merujuk pada kelas "Program Studi" 9 dengan jumlah data 458 di masing-masing kelas "Program Studi".

```
9     458
5     402
0     269
13    241
1     230
11    201
4     198
10    196
6     128
7     102
8      94
3      93
12     60
14     30
2       1
Name: ProgramStudi, dtype: int64
```

```
0     458
4     458
8     458
12    458
1     458
5     458
9     458
13    458
2     458
6     458
10    458
3     458
7     458
11    458
Name: ProgramStudi, dtype: int64
```

Gambar 6. (kiri) Sebelum *balancing data;* (kanan) Setelah *balancing data*

C. *Implementasi Pemodelan Klasifikasi*

Data yang sudah dilakukan *balancing* selanjutnya dilakukan pembagian data latih dan data uji dengan bobot persentasi data latih sebesar 80% dan data uji sebesar 20%. Data yang dihasilkan dari hasil *balancing* kelas "Program Studi" menghasilkan sejumlah 7452 data, lalu dilakukan pembobotan data latih dan data uji yang menghasilkan 5961 data latih dan 1491 data uji. Fitur pada data latih disebut dengan "X_train" dan kelas "Program Studi" disebut dengan "y_train", sedangkan fitur pada data uji

disebut dengan "X_test" dan kelas "Program Studi" disebut dengan "y_test".

Dari hasil pembobotan data latih dan data uji tersebut selanjutnya dilakukan implementasi pemodelan yang terdapat dalam metode klasifikasi. Dalam kasus penelitian ini dilakukan pengimplementasikan pemodelan *k-Nearest Neighbor (k-NN) Classifier, Decision Tree Classifier, Naive Bayes Classifier, Support Vector Machine, dan Adaptive Boosting.* Hasil implementasi pemodelan tersebut menghasilkan nilai akurasi, dimana dalam nilai akurasi dihasilkan pemodelan k-NN yang ditunjukkan pada gambar 7 memiliki nilai 0.722:

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics

knn = KNeighborsClassifier()

knn.fit(X_train, y_train)
y_pred = knn.predict(X_test)
score = metrics.accuracy_score(y_test, y_pred)
print("Akurasi dengan menggunakan Nearest Neighbor: ", score)
```

```
Akurasi dengan menggunakan Nearest Neighbor: 0.7220187061574435
```

Gambar 7. Pemodelan *k-nearest neighbor* (k-nn) *classifier*

Dalam pemodelan *Decision Tree Classifier* pada gambar 12 dilakukan import module dengan "DecisionTreeClassifier", atribut max_depth dengan nilai "None" dan min_samples_split = 2 yang menghasilkan nilai akurasi 0.334.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics

dt = DecisionTreeClassifier(
  max_depth = None,
  min_samples_split = 2
)

dt.fit(X_train, y_train)
y_pred = dt.predict(X_test)
score = metrics.accuracy_score(y_test, y_pred)
print("Akurasi dengan menggunakan Decision Tree: ", score)
```

```
Akurasi dengan menggunakan Decision Tree: 0.3343725643023162
```

Gambar 8. Pemodelan *k-nearest neighbor* (k-nn) *classifier*

Pemodelan *Naive Bayes Classifier* diimplementasikan dengan menggunakan naive_bayes.BernoulliNB() dimana model yang ditetapkan diwakili oleh vektor atribut biner yang menunjukkan data yang muncul dan tidak muncul dalam dokumen. Gambar 9 adalah hasil dari pemodelan *Naive Bayes Bernoulli* dimana menghasilkan nilai akurasi 0.607

```
from sklearn import naive_bayes
from sklearn import metrics

nb = naive_bayes.BernoulliNB()

nb.fit(X_train, y_train)
y_pred = nb.predict(X_test)
score = metrics.accuracy_score(y_test, y_pred)
print("Akurasi dengan menggunakan Naive Bayes: ", score)
```

```
Akurasi dengan menggunakan Naive Bayes: 0.607560405300077795
```

Gambar 9. Pemodelan *naive bayes classifier*

Support Vector Machine (SVM) diimplementasikan dengan nilai kernel = 'rbf' atau "radial basis function"; nilai C = 1 yang merupakan parameter regularisasi SVM untuk seluruh kernel; dan parameter gamma sebagai indikator koefisien kernel untuk 'rbf' dan 'poly'. Hasil nilai akurasi dari SVM pada dataset menghasilkan nilai 0.617 pada gambar 10.

```
from sklearn.svm import SVC
from sklearn import metrics

svm = SVC(
  kernel = 'rbf',
  C = 1,
  gamma = 0.01
)

svm.fit(X_train, y_train)
y_pred = svm.predict(X_test)
score = metrics.accuracy_score(y_test, y_pred)
print("Akurasi dengan menggunakan Support Vector Machine: ", score)
```

```
Akurasi dengan menggunakan Support Vector Machine: 0.61745908028059235
```

Gambar 10. Pemodelan *support vector machine* (SVM)

AdaBoost merupakan bagian dari *Boosting Algorithm* yang merupakan salah satu strategi dalam mentransformasikan model 'lemah' menjadi model 'kuat', selain dari 2 pemodelan atau algoritma lainnya yakni *Gradient Tree Boosting* dan XGBoost. Implementasi AdaBoost dalam dataset menghasilkan nilai akurasi 0.624 seperti pada gambar 11.

```
from sklearn.ensemble import AdaBoostClassifier
from sklearn import metrics

ab = AdaBoostClassifier(
        n_estimators = 50,
        learning_rate=1
    )

ab.fit(X_train, y_train)
y_pred = ab.predict(X_test)
score = metrics.accuracy_score(y_test, y_pred)
print("Akurasi dengan menggunakan AdaBoost: ", score)
```

```
Akurasi dengan menggunakan AdaBoost: 0.624707716289945444
```

Gambar 11. Pemodelan *Adaptive Boosting* (AdaBoost)

D. Evaluasi

Pengukuran kinerja akurasi dilakukan untuk melihat bagaimana kualitas dari algoritma dalam metode klasifikasi diimplementasikan pada kasus *dataset* PMB Universitas XYZ Kota Palembang dengan fitur "Jurusan Sekolah", "Penghasilan", "Tahun Masuk", "Tahun Lulus", "Tipe Sekolah", "Status Sekolah" dan Kelas "Program Studi". Dari hasil implementasi tersebut dalam bobot persentasi dihasilkan algoritma kNN memiliki nilai tertinggi yakni 72.2% dan algoritma *Decision Tree Clasifier* merupakan nilai terendah, dengan persentasi akurasi sebesar 33.4%. Sedangkan untuk algoritma lainnya, yakni *Naive Bayes Classifier* didapatkan persentasi 60.7%, *Support Vector Machine* (SVM) sebesar 61.7% dan *Adaptive Boosting* (AdaBoost) sebesar 62.4%. Secara lengkap nilai akurasi dan persentasi akurasi dalam algoritma/ pemodelan dapat dilihat pada tabel 5.

TABEL V
NILAI AKURASI ALGORITMA / PEMODELAN

No	Algoritma / Pemodelan	Nilai Akurasi	Persentasi Akurasi
1	k-Nearet Neighbor (k-NN) Classifier	0.722	72.2%
2	Decision Tree Classifier	0.334	33.4%
3	Naive Bayes Classifier	0.607	60.7%
4	Support Vector Machine	0.617	61.7%
5	Adaptive Boosting	0.624	62.4%

IV. KESIMPULAN

Dari hasil penilaian kinerja algoritma dalam metode klasifikasi pada dataset PMB Universitas XYZ Kota Palembang didapatkan persentasi akurasi tertinggi dengan algoritma *k-Nearest Neighbor* (k-NN) *Classifier* dengan persentasi akurasi sebesar 72.2%. Sedangkan persentasi akurasi terendah pada algoritma *Decision Tree Classifier* dengan nilai 33.4%. Hasil persentasi akurasi menyimpulkan bahwa k-NN dapat direkomendasikan dalam pemodelan preferensi pemilihan program studi pada *dataset* PMB Universitas XYZ Kota Palembang dengan mempertimbangkan fitur "Jurusan Sekolah", "Penghasilan", "Tahun Masuk", "Tahun Lulus", "Tipe Sekolah" dan "Status Sekolah". Dalam pengembangan selanjutnya guna meningkatkan persentasi akurasi perlu dilakukan *parameter tuning* terhadap algoritma yang diimplementasikan, juga dilakukan proses evaluasi yang lebih kompleks dengan menggunakan *confusion matrix*.

REFERENSI

[1] S. A. Nulhaqim, R. D. Heryady, R. Pancasilawan, and M. Fedryansyah, "Peranan Perguruan Tinggi Dalam Meningkatkan Kualitas Pendidikan Di Indonesia Untuk Menghadapi Asean Community 2015," Share Soc. Work J., vol. 6, no. 2, pp. 197–219, 2015.

[2] K. Law, T. Li, and S. Geng, "Student enrollment, motivation and learning performance in a blended learning environment: The mediating effects of social, teaching, and cognitive presence," Comput. Educ., vol. 136, no. September, pp. 1–12, 2019, doi: 10.1016/j.compedu.2019.02.021.

[3] D. Bukhari, "Data Science Curriculum: Current Scenario," Int. J. Data Min. Knowl. Manag. Process, vol. 10, no. 3, pp. 1–13, 2020, doi: 10.5121/ijdkp.2020.10301.

[4] P. Saini, "Building a Classification Model for Enrollment in Higher Educational Courses using Data Mining Techniques," ArXiv, vol. 1405, no. 3729, pp. 696–697, 2014, doi: 10.1021/ja01318a049.

[5] D. A. A. AlHammadi and M. S. Aksoy, "Data Mining in Higher Education," Period. Eng. Nat. Sci., vol. 1, no. 2, pp. 1–4, 2013, doi: 10.21533/pen.v1i2.17.

[6] S. Alturki and N. Alturki, "Using Educational Data Mining to Predict Students' Academic Performance for Applying Early Interventions," J. Inf. Technol. Educ. Innov. Pract., vol. 20, pp. 121–137, 2021. doi: 10.28945/4835

[7] L. W. Santoso and Yulia, "The Analysis of Student Performance Using Data Mining," Adv. Intell. Syst. Comput., vol. 924, no. June, pp. 559–573, 2019, doi: 10.1007/978-981-13-6861-5_48.

[8] I. G. T. Isa, "Aplikasi Asesmen Calon Debitur menggunakan Naive Bayes di Koperasi Mitra Sejahtera SMK Negeri 1 Kota Sukabumi," J. Sisfokom (Sistem Inf. dan Komputer), vol. 10, no. 1, pp. 31–39, 2021, doi: 10.32736/sisfokom.v10i1.1013.

[9] I. G. T. Isa and D. Jhoansyah, "Implementasi Association Rules Dalam Menentukan Posisi Gerobak (Studi Kasus: Foodcourt Universitas Muhammadiyah Sukabumi)," Inform. Mulawarman J. Ilm. Ilmu Komput., vol. 13, no. 2, p. 65, 2019, doi: 10.30872/jim.v13i2.1273.

[10] F. Marisa, "Educational Data Mining (Konsep dan Penerapan)," J. Teknol. Inf., vol. 4, no. 2, pp. 91–93, 2013.

[11] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 40, no. 6, pp. 601–618, 2010, doi: 10.1109/TSMCC.2010.2053532.

[12] C. C. Aggarwal, Data Classification Algorithms and Application. New York: CRC Press, 2015.

[13] M. K. Anam, B. N. Pikir, and M. B. Firdaus, "Penerapan Na¨ıve Bayes Classifier, K-Nearest Neighbor (KNN) dan Decision Tree untuk Menganalisis Sentimen pada Interaksi Netizen danPemeritah," MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput., vol. 21, no. 1, pp. 139–150, 2021, doi: 10.30812/matrik.v21i1.1092.

[14] J. Hurwitz and D. Kirsch, Machine Learning for Dummies. New Jersey: John Wiley & Sons, Inc, 2018.

[15] M. Brackett, Data Resource Understanding. Utilizing the Data Resource Data. New Jersey: Technics Publication, 2015.

[16] O. Masmoudi, M. Jaoua, A. Jaoua, and S. Yacout, "Data Preparation in Machine Learning for Condition-based Maintenance," J. Comput. Sci., vol. 17, no. 6, pp. 525–538, 2021, doi: 10.3844/JCSSP.2021.525.538.

[17] S. Suthaharan, Machine Learning Models and Algorithms for Big Data Classification. North Carolina: Springer, 2016.

[18] M. Hossin, "A Review on Evaluation Metrics for Data Classification Evaluations," Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

[19] P. Schober and L. A. Schwarte, "Correlation Coefficients: Appropriate Use and Interpretation," Anesth. Analg., vol. 126, no. 5, pp. 1763–1768, 2018, doi: 10.1213/ANE.0000000000002864.

[20] A. Suad A. and B. Wesam S., "Review of data preprocessing techniques in data mining.pdf," J. Eng. Appl. Sci., vol. 12, no. 16, pp. 4102–4107, 2017, doi: doi=jeasci.2017.4102.4107.