Exploratory Data Analysis (EDA) dalam Dataset Penerimaan Mahasiswa Baru Universitas XYZ Palembang

Indra Griha Tofik Isa[*1], Zulkarnaini[2], Leni Novianti[3], Febie Elfaladonna[4], Suzan Agustri[5]
[1,2,3,4,5]Jurusan Manajemen Informatika, Politeknik Negeri Sriwijaya
E-mail: *[1]**indra_isa_mi@polsri.ac.id**, [2]zulkarnaini@polsri.ac.id, [3]leni@polsri.ac.id,
[4]febie_elfaladonna_mi@polsri.ac.id, [5]zuzanoid@uigm.ac.id

Abstrak

Keberhasilan suatu pemodelan salah satunya dipengaruhi oleh kualitas dari dataset yang dianalisis. Exploratory Data Analysis merupakan teknik yang digunakan dalam data understanding untuk mengeksplorasi data mana saja yang memiliki kualitas yang nantinya digunakan dalam tahapan pemodelan. Kasus yang diangkat dalam penelitian ini adalah dataset penerimaan mahasiswa baru di Universitas XYZ, dimana untuk tujuan akhirnya adalah bagaimana memprediksikan preferensi program studi bagi calon pendaftar. Namun dari dataset tersebut dengan beragam data perlu dikaji lebih lanjut untuk mencermati kualitas data yang valid, kredibel, mendukung dalam pemodelan preferensi pilihan program studi. Sebuah EDA akan diimplementasikan sebagai solusi dari penelaahan data dengan melihat ragam data dari dataset penerimaan mahasiswa baru, potensi fitur yang mendukung dalam tahap pemodelan, rekomendasi yang perlu dilakukan untuk tahapan lanjut dalam sebuah siklus data sains. Tahapan penelitian dilakukan dengan Analisis Permasalahan, Akuisisi Data, Exploratory Data Analysis (EDA), Interpretasi Anomali, Rekomendasi Fitur. Hasil akhir berupa 14 rekomendasi fitur dari dataset penerimaan mahasiswa baru yang terdiri dari Jenis Kelamin, Tanggal Lahir (Umur), Program Studi, Status Sipil, Provinsi, Kota, Anak Ke, Jumlah Saudara, Penghasilan, Jenjang, Program Kuliah, Jenis Sekolah, Jurusan Sekolah, Nilai Unas, Tahun Lulus

Kata Kunci*—Exploratory Data Analysis (EDA), Penerimaan Mahasiswa Baru, Data Understanding*

1. PENDAHULUAN

Revolusi Industri 4.0 mengarahkan pada perubahan dinamika teknologi menjadi sistem cerdas, terotomatisasi, terintegrasi, menjadikan data-data "masa lalu" menjadi asset yang memiliki nilai manfaat bagi penggunanya [1]. Salah satu dampak dengan adanya revolusi industry 4.0 adalah dengan adanya cabang keilmuan data sains yang mengkolaborasikan pembelajaran data kuantitatif seperti statistika, pemrograman basis data yang dipadukan dengan algoritma yang kompleks. Data sains memberikan dukungan dalam berbagai hal, diantaranya dalam dunia bisnis yaitu untuk memberikan efisiensi dalam sistem dan proses produksi[2] ataupun memberikan rekomendasi keputusan dalam kebijakan [3]. Dengan menggunakan beberapa tipe analisis, stakeholder keputusan bisnis dapat melihat tren data yang ada, sehingga dapat menciptakan sebuah proses yang lebih efisien dan terstruktur.

Data sains memiliki beberapa tahapan yang diawali dengan business understanding yakni bagaimana memahami permasalahan dan tujuan dari sebuah konteks kasus yang akan diangkat [4]. Dari tahapan ini memunculkan beberapa aspek, yaitu penentuan masalah, tujuan proyek, solusi dari perspektif bisnis hingga instrumen pengukuran keberhasilan. Setelah tahapan business understanding secara utuh dilakukan, kemudian dilanjutkan dengan tahapan data understanding yang merupakan bagaimana mendeskripsikan data sehingga dapat tergali informasi implisit yang nantinya memperkuat dalam tahapan pemodelan.

Data understanding dimana dalam istilah lain disebut dengan eksploratory data analysis (EDA) menjadi penting karena dalam tahapan ini dilakukan pemilihan data yang relevan dengan konteks permasalahan. Secara teknis dalam data understanding dilakukan proses (1) Akuisisi dan penarikan data; (2) Penelaahan data dengan melihat korelasi antar fitur dari suatu data (yang selanjutnya disebut dengan dataset), melihat data anomali (seperti data redudansi, missing data, maupun outlier); (3) Visualisasi data sebagai representasi dari data yang akan direkomendasikan.

Kasus yang diangkat dalam penelitian ini adalah Dataset Penerimaan Mahasiswa Baru di Universitas XYZ, dimana untuk tujuan akhirnya adalah bagaimana memprediksikan preferensi program studi bagi calon pendaftar. Namun dari dataset tersebut dengan beragam data perlu dikaji lebih lanjut untuk menghasilkan kualitas data yang valid, kredibel, mendukung dalam pemodelan preferensi pilihan program studi.

Sebuah EDA akan diimplementasikan sebagai solusi dari penelaahan data dengan melihat ragam data dari Dataset Penerimaan Mahasiswa Baru, potensi fitur yang mendukung dalam tahap pemodelan, rekomendasi yang perlu dilakukan untuk tahapan lanjut dalam sebuah siklus data sains [5]. Sehingga rumusan masalah dalam penelitian ini adalah bagaimana implementasi Exploratory Data Analysis dalam menghasilkan fitur yang direkomendasikans dalam tahapan data sains. Adapun tujuan dari penelitian ini adalah menghasilkan fitur yang direkomendasikans dari enrolment student yang mendukung dalam pemodelan preferensi pemilihan program studi bagi pendaftar di Universitas XYZ Kota Palembang. Batasan masalah dalam penelitian ini dari segi dataset yang diolah adalah Dataset Penerimaan Mahasiswa Baru dengan 54 fitur dan 2704 record, tools yang digunakan adalah bahasa pemrograman Python yang diakses melalui Google Colab.

EDA menjadi salah satu bagian penting dalam serangkaian tahapan data sains dimana menghasilkan data-data yang berkualitas untuk dilibatkan pada tahapan berikutnya [6]. Beberapa penelitian terdahulu terkait EDA dimana diimplementasikan EDA pada Kasus COVID-19 di Indonesia menggunakan HiveQL dan Hadoop Environment [7]. Dari implementasi EDA pada penelitian ini menghasilkan analisis nilai korelasi dimana menunjukkan pengaruh kuat antara pertambahan jumlah kasus terkonfirmasi positif dengan jumlah kasus pasien sembuh dan kasus pasien meninggal sebesar 0.94 dan 0.9 masing-masing. Hasil analisis korelasi yang lain ditemukan nilai korelasi yang kecil antara kasus pasien dalam perawatan terhadap kasus terkonfirmasi positif, kasus pasien meninggal, dan kasus pasien sembuh. Untuk keberlanjutan penelitian ini, perlu diimplementasikan analisis regresi karena bentuk data terstruktur kasus COVID-19 ini seperti time series.

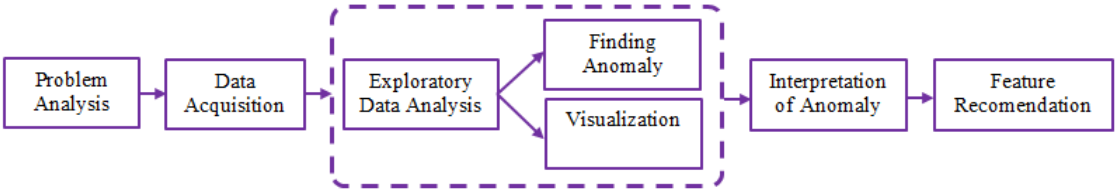
Pada penelitian selanjutnya, EDA dilakukan pada dataset penjualan barang elektronik yang bertujuan untuk melihat bagaimana pergerakan penjualan barang elektronik [8], sehingga menjadi pertimbangan dalam merencanakan strategi peningkatan usaha. Hasil akhir dari penelitian ini didapatkan bahwa terdapat produk dengan nilai penjualan tertinggi yakni produk baterai AAA, penjualan paling sedikit terjual adalah LG Dryer dan produk elektronik yang paling mahal terjual sepanjang bulan Januari – Desember adalah Macbook Pro Laptop.

Wahyuni dkk mengobservasi EDA dengan tujuan untuk menggali informasi tersirat dari data penjualan produk fashion untuk membantu pada tahapan preprocessing dengan menampilkan keberadaan missing value dan juga outlier [9]. Hasil akhir dari penelitian ini dapat disimpulkan bahwa EDA dapat mengoptimalkan pengetahuan mengenai data, yang dapat digunakan untuk memperkaya pemahaman atas analisis data evaluasi.

Sedangkan pada penelitian ini berfokus pada penelaahan data untuk menggali karakteristik data, outlier dari setiap fitur, nilai korelasi antar fitur, sehingga menghasilkan fitur yang memiliki kualitas terbaik yang nantinya akan digunakan pada tahapan pemodelan.

2. METODE PENELITIAN

Objek penelitian adalah dataset penerimaan mahasiswa baru of University XYZ dengan record data sebanyak 2704 record dengan pengolahan data menggunakan Google Colab. Adapun tahapan penelitian dapat dilihat pada gambar 1 berikut:



Gambar 1. Tahapan Penelitian

Pada tahapan penelitian diawali dengan *problem analysis* dimana ditentukan permasalahan yang akan dikaji, tujuan sebagai solusi atas permasalahan tersebut, serta mekanisme dan ruang lingkup dalam penyelesaian penelitian yang dibangun.

Berikutnya adalah *Data Acquisition* yang dilakukan dengan pembacaan Dataset Penerimaan Mahasiswa Baru dengan file bertipe CSV. Data tersebut merupakan data mentah yang akan diolah untuk menghasilkan fitur rekomenation melalui EDA [10]. Tahapan selanjutnya adalah implementasi EDA terhadap dataset, dengan *finding anomali* dari dataset pada setiap fitur, seperti imbalance data, missing value, outlier, karakteristik data (apakah data tersebut nominal atau kategorikal) maupun tipe data pada setiap fitur. Untuk memudahkan dalam penelaahan data, visualization dilakukan dengan histogram, boxplot maupun pie chart yang disesuaikan dengan konteks dan karakteristik dari fitur yang diamati.

Setelah data divisualisasikan, selanjutnya adalah *interpretation of anomali* pada setiap fitur yang diamati, dengan memberikan rencana tindak lanjut terhadap fitur tersebut, yang dikategorikan menjadi "Rekomen", "Dipertimbangkan" dan "Tidak Rekomen". Untuk selanjutnya dari tahapan ini menghasilkan *feature rekomendation* dari Dataset Penerimaan Mahasiswa Baru yang memiliki nilai kredibel dan mendukung pada proses selanjutnya yakni transformation dan modelling.

3. HASIL DAN PEMBAHASAN

3.1. Problem Analysis

Tujuan dari penelitian ini adalah bagaimana menemukan fitur yang direkomendasikan yang dihasilkan dari dataset penerimaan mahasiswa baru University XYZ, yang nantinya akan digunakan dalam pemodelan preferensi pemilihan program studi bagi pendaftar. Dataset diambil dalam kurun waktu 2018 hingga 2020 yang didalamnya terdiri dari ragam data numerical dan categorical, ragam tipe data, ragam anomali seperti imbalance/ missing value/ data redundancy/ outlier. Sehingga diperlukan Exploratory Data Analysis dengan melihat bagaimana korelasi antar fitur, sejauhmana anomali di setiap fitur, bagaimana visualisasi yang dihasilkan dari fitur tersebut

3.2. Data Acquisition

Data Acquisition dilakukan dengan mengimplementasi library pandas (read_csv) seperti pada gambar 2 melalui pembacaan dataset CSV yang sudah diinsertkan di dalam repository. Di dalam gambar 3 merupakan dataset yang muncul setelah pengimplementasian library panda

```
df = pd.read_csv('/content/All Data PMB 2018-2020.csv', sep=";")
df
```

Gambar 2. Syntax CSV data

	Jenis Kelamin	Agama	Tempat Lahir	Tanggal Lahir	Status Sipil	Alamat	Kode Pos	Provinsi	Kota	Negara
	L	ISLAM	Karang Anyar	29/12/2002	B	Dusun 3 desa Karang Anyar, Kec. Lawang Wetan, Kab. Musi Banyuasin, Sumatra Selatan	-	SUMATERA SELATAN	KAB. MUSI BANYUASIN	Indonesia
	L	ISLAM	Palembang	11/12/2002	B	Jl. DI PANJAITAN Lrg. Pegagan	30265	SUMATERA SELATAN	KOTA PALEMBANG	Indonesia
	L	ISLAM	Muaradua	24/04/2003	B	JLN SETUNGGAL KOMPLEK PERSADA Blok B-23	30114	SUMATERA SELATAN	KOTA PALEMBANG	Indonesia
	P	ISLAM	Bandar Lampung	02/03/2002	B	Lk. Sidoharjo Talang Jawa	31714	SUMATERA SELATAN	KAB. MUARA ENIM	Indonesia
	L	ISLAM	Lubuk Kelumpang	07/01/2003	B	Lubuk Kelumpang	0	SUMATERA SELATAN	KAB. LAHAT	Indonesia
	P	ISLAM	Palembang	11/05/2001	B	Jl. DI PANJAITAN Irg. Sriraya 3 No.01	30265	SUMATERA SELATAN	KOTA PALEMBANG	Indonesia

Gambar 3. Pembacaan Dataset dengan Library Pandas

Untuk melihat lebih jauh dari ragam tipe data fitur maka dilakukan implementasi print(df.types) sehingga memunculkan seluruh fitur beserta tipe data dari fitur tersebut. Gambar 4 menampilkan seluruh fitur beserta tipe datanya.

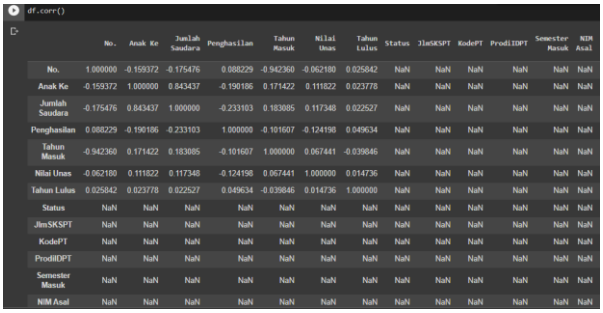
```
print(df.dtypes)
```

```
No.                 int64
NIM                object
Nama               object
Jenis Kelamin      object
Agama              object
Tempat Lahir       object
Tanggal Lahir      object
Status Sipil       object
Alamat             object
Kode Pos           object
Provinsi           object
Kota               object
Negara             object
Telepon            object
HP                 object
Email              object
Anak Ke             int64
Jumlah Saudara      int64
Penghasilan         int64
Jenjang            object
Program Kuliah     object
Program Studi      object
Status Mahasiswa   object
Pembimbing         object
Batas Studi        object
Tahun Masuk         int64
Jenis Sekolah      object
Nama Sekolah       object
Jurusan Sekolah    object
Nilai Unas        float64
Tanggal Lulus      object
Tahun Lulus       float64
No Ijazah          object
Tanggal Masuk      object
Status              int64
Jenis Beasiswa     object
JlmSKSPT          float64
KodePT              int64
ProdiIDPT           int64
Status Pindahan    object
Semester Masuk    float64
NIM Asal          float64
Asal Jenjang        int64
Kelas             float64
Nama Jenjang      float64
NamaPST           float64
Ayah               object
Ibu                object
Alamat Orang Tua   object
Kota Orang Tua      int64
Kode Pos Orang Tua  int64
Telepon Orang Tua  float64
HP Orang Tua       object
P.A                object
Tahun Terakhir KRS object
IPK                object
dtype: object
```

Gambar 4. Fitur pada Dataset

Dari explorasi dataset pada gambar 4 dapat dilihat bahwa terdapat beberapa data dengan tipe numeric yakni integer (int64) dan float (float64), serta tipe data object dimana merupakan kombinasi dari karakter angka, huruf, ASCII ataupun tanggal, yang memungkinkan untuk disesuaikan dengan tipe data dari karakter fitur tersebut. Jika dijumlahkan, fitur dengan tipe data int64 sejumlah 11 fitur, tipe data float64 sejumlah 9 fitur dan sisanya sebanyak 36 fitur.

Selanjutnya, tanpa melakukan transformasi atau dengan kata lain melalui dataset asli, dilakukan uji korelasi antar fitur / variabel dengan formulasi df.corr(), maka dilihat pada gambar 5 hasil dari korelasi antar fitur tersebut. Dapat dilihat bahwa hampir sebagian besar fitur memiliki korelasi yang rendah (dengan asumsi threshold 0.75) serta memiliki nilai NaN (Not a Number). Namun ada beberapa fitur yang berkorelasi tinggi, yakni “Jumlah Saudara” dengan “Anak Ke” dimana berkorelasi dengan nilai 0.84. Sehingga jika disimpulkan dari tahapan ini diperlukan transformasi data untuk fitur yang direkomendasikans setelah dilakukan implementasi EDA.



```
df.corr()
```

	No.	Anak Ke	Jumlah Saudara	Penghasilan	Tahun Masuk	Nilai Unas	Tahun Lulus	Status	JlmSKSPT	KodePT	ProdiIDPT	Semester Masuk	NIM Asal
No.	1.000000	-0.159372	-0.175476	0.088229	-0.942360	-0.062180	0.025842	NaN	NaN	NaN	NaN	NaN	NaN
Anak Ke	-0.159372	1.000000	0.843437	-0.190186	0.171422	0.111822	0.023778	NaN	NaN	NaN	NaN	NaN	NaN
Jumlah Saudara	-0.175476	0.843437	1.000000	-0.233103	0.183085	0.117348	0.022527	NaN	NaN	NaN	NaN	NaN	NaN
Penghasilan	0.088229	-0.190186	-0.233103	1.000000	-0.101607	-0.124198	0.049634	NaN	NaN	NaN	NaN	NaN	NaN
Tahun Masuk	-0.942360	0.171422	0.183085	-0.101607	1.000000	0.067441	-0.039846	NaN	NaN	NaN	NaN	NaN	NaN
Nilai Unas	-0.062180	0.111822	0.117348	-0.124198	0.067441	1.000000	0.014736	NaN	NaN	NaN	NaN	NaN	NaN
Tahun Lulus	0.025842	0.023778	0.022527	0.049634	-0.039846	0.014736	1.000000	NaN	NaN	NaN	NaN	NaN	NaN
Status	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
JlmSKSPT	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
KodePT	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ProdiIDPT	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Semester Masuk	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NIM Asal	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Gambar 5. Korelasi Antar Fitur (Tanpa Perubahan Label)

3.3. Exploratory Data Analysis (EDA)

Dalam implementasi EDA dilakukan dengan *finding anomali* dan visualisasi data untuk memudahkan dalam interpretasi data. Sehubungan dengan tujuan besar dari penelitian ini adalah pemodelan preferensi program studi bagi pendaftar di Universitas XYZ, maka class dari dataset ini adalah "Program Studi". Sehingga diperlukan penelaahan data untuk melihat bagaimana ragam nama klasifikasi serta jumlah data dari class "Program Studi. Untuk itu dilakukan telaah data dengan melihat berapa jumlah klasifikasi dan data dari Class "Program Studi", yang direpresentasikan pada gambar 6:

```
#Classification Data Based on Program Studi
df_prodi = df.groupby('Program Studi').size().reset_index(name='Counts')
df_prodi.sort_values(by='Counts', inplace=True)
df_prodi.reset_index(drop=True)
```

	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	231
10	Akuntansi	241
11	Teknik Sipil	269
12	Sistem Informasi	402
13	Manajemen	458

Gambar 6. Class "Program Studi"

Pada gambar 6 terdapat 13 program Studi yang memiliki sebaran Program Studi terendah adalah "Keselamatan dan Kesehatan Kerja" sebanyak 30 data dan tertinggi "Manajemen" sebanyak 458 data. Untuk sebaran data dengan nilai kurang dari 100 terdapat 4 Program Studi selain "Keselamatan dan Kesehatan Kerja", yakni "Arsitektur", "Survei dan Pemetaan " dan "Manajemen Informatika". Sedangkan Program Studi dengan nilai data diatas 400 adalah "Sistem Informasi" dan "Manajemen".

Visualisasi dari class "Program Studi" dilakukan untuk memudahkan dalam merepresentasikan data. Dalam kasus ini, visualisasi dikategorikan ke dalam sebaran data rendah, sedang dan tinggi, dimana untuk data rendah adalah 3 data terbawah, data tinggi adalah 3 data tertinggi, dan sisanya merupakan sebaran data rendah. Secara teknis, visualisasi mengimplementasikan library pyplot dari matplotlib dengan penamaan variabel "plt" pada

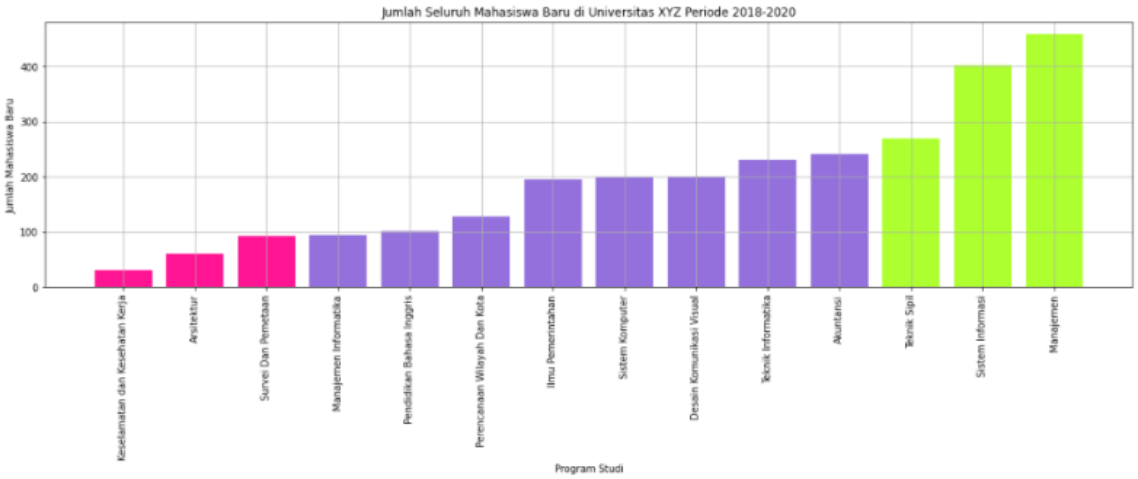
P-ISSN: 2089-676X
E-ISSN: 2549-0796

gambar 7. Lalu pada gambar 8 adalah hasil visualisasi dari coding gambar 7 dengan efek warna yang memudahkan dalam melihat sebaran data.

```
colors = ['#9370DB' for _ in range(len(df_prodi['Program Studi']))]
colors[:3] = ['#FF1493' for _ in range(3)]
colors[-3:] = ['#ADFF2F' for _ in range(3)]

x_coords = np.arange(len(df_prodi))
plt.figure(figsize=(20,5))
plt.bar(x_coords, df_prodi['Counts'], tick_label=df_prodi['Program Studi'], color=colors)
plt.xticks(rotation=90) #rotates text for x-axis labels
plt.title('Jumlah Seluruh Mahasiswa Baru di Universitas XYZ Periode 2018-2020')
plt.xlabel('Program Studi')
plt.ylabel('Jumlah Mahasiswa Baru')
plt.grid()
plt.show()
```

Gambar 7. Klasifikasi Class of “Program Studi”



Gambar 8. Hasil Visualisasi dari Class “Program Studi”

Berikutnya EDA dilakukan pada fitur yang relevan dengan konteks pemodelan preferensi pemilihan program studi bagi pendaftar di Universitas XYZ, artinya data atau fitur yang dilibatkan hanya data sebelum pendaftar menjadi mahasiswa. Berkaitan dengan hal tersebut, terdapat 8 fitur yang tidak relevan dengan “data pendaftar”, yakni “NIM”, “Status Mahasiswa”, “Pembimbing”, “Batas Studi”, “Jenis Beasiswa”, “NIM Asal”, “Tahun Terakhir KRS”, “IPK”.

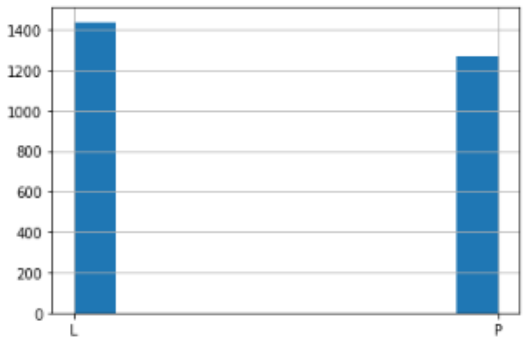
Beberapa fitur memiliki sebaran data cenderung merata, diantaranya adalah “Jenis Kelamin”. Jenis kelamin memiliki klasifikasi “P” dan “L”, melalui fungsi count dihasilkan “P” sejumlah 1266 record dan “L” sejumlah 1437 record. Implementasi count untuk “Jenis Kelamin” dapat dilihat pada gambar 9, sedangkan hasil visualisasi “Jenis Kelamin” ditunjukkan pada gambar 10.

```
[ ] df_jk = df.groupby('Jenis Kelamin').size().reset_index(name='Counts')
    df_jk.sort_values(by='Counts', inplace=True)
    df_jk.reset_index(drop=True)
    df_jk
```

	Jenis Kelamin	Counts
1	P	1266
0	L	1437

Gambar 9. Hasil Visualisasi dari fitur “Jenis Kelamin”

P-ISSN: 2089-676X
E-ISSN: 2549-0796



Gambar 10. Visualisasi fitur “Jenis Kelamin”

Sebelum melanjutkan ke EDA fitur lainnya, diperlukan eksplorasi *missing value* untuk melihat seberapa banyak missing value, apakah mempengaruhi terhadap keseimbangan data atau tidak. Di dalam gambar 11 dapat dilihat beberapa fitur yang memiliki missing value yang tinggi (>1000 record data), diantaranya “No. Ijazah”, “Jenis Beasiswa”, “JlmSKSPT”, “Semester Masuk”,”NIM Asal”, “Kelas”, “Nama Jenjang”, “NamaPST”, “Telepon Orang Tua”. Jika dicermati, hampir seluruh fitur tersebut merupakan fitur jika status pendaftar sudah menjadi mahasiswa. Namun terdapat fitur yang relevan dengan kondisi pendaftar, yakni “Nama Jenjang”. Jika dicermati kembali, jumlah missing data dari fitur “Nama Jenjang” adalah seluruh record dataset, yaitu 2704 record. Artinya fitur tersebut juga dianggap tidak relevan dengan konteks pendaftaran. Sehingga jika disimpulkan fitur yang memiliki record *missing value* > 1000 dapat dihilangkan, dengan alasan : (1) Tidak relevan dengan konteks penerimaan mahasiswa baru; dan, (2) Fitur “Nama Jenjang” seluruh konten data adalah *missing value.*

```
mv=df.isna().sum()
mv
```

```
No.                  0
NIM                  0
Nama                 0
Jenis Kelamin        1
Agama                2
Tempat Lahir         1
Tanggal Lahir        0
Status Sipil         3
Alamat               0
Kode Pos            13
Provinsi             1
Kota                 1
Negara               0
Telepon             65
HP                  18
Email                3
Anak Ke              0
Jumlah Saudara       0
Penghasilan          0
Jenjang              0
Program Kuliah       0
Program Studi        0
Status Mahasiswa     0
Pembimbing         185
Batas Studi         16
Tahun Masuk          0
Tahun Masuk          0
Jenis Sekolah        1
Nama Sekolah        18
Jurusan Sekolah     18
Nilai Unas           0
Tanggal Lulus        0
Tahun Lulus          1
No Ijazah         1297
Tanggal Masuk        0
Status               0
Jenis Beasiswa    1258
JlmSKSPT          2702
KodePT               0
ProdiIDPT            0
Status Pindahan     11
Semester Masuk    2704
NIM Asal          2704
Asal Jenjang         0
Kelas             2704
Nama Jenjang      2704
NamaPST           2704
Ayah                 4
Ibu                  4
Alamat Orang Tua     4
Kota Orang Tua       0
Kode Pos Orang Tua   0
Telepon Orang Tua 2704
HP Orang Tua         8
P.A                  0
Tahun Terakhir KRS   3
IPK                  0
dtype: int64
```

Gambar 11. Data Missing Value dari Dataset

Kasus imbalance data juga ditemukan terhadap beberapa fitur, seperti fitur “Status Sipil” dimana label “S” memiliki 18 record sedangkan “B” memiliki 2683 record, seperti pada gambar 12

```
df_status_sipil = df.groupby('Status Sipil').size().reset_index(name='Counts')
df_status_sipil.sort_values(by='Counts', inplace=True)
df_status_sipil.reset_index(drop=True)
df_status_sipil
```

	Status Sipil	Counts
1	S	18
0	B	2683

Gambar 12. Imbalance data pada fitur “Status Sipil”

Adapun outlier dapat dilihat pada gambar 13, dimana terdapat label tahun “201” sebanyak 1 record dan “0” sebanyak 36 record data.

	Tahun Lulus	Counts
11	2008.0	1
1	201.0	1
2	1993.0	1
3	1999.0	1
5	2002.0	1
7	2004.0	1
15	2012.0	1
13	2010.0	1
4	2001.0	2
6	2003.0	2
9	2006.0	2
10	2007.0	2
8	2005.0	3
14	2011.0	6
12	2009.0	6
16	2013.0	9
17	2014.0	16
0	0.0	36
18	2015.0	43
19	2016.0	102
20	2017.0	305
23	2020.0	513
22	2019.0	751
21	2018.0	897

Gambar 13. Outlier data pada fitur “Tahun Lulus”

3.4. Interpretation of Anomali

Ragam hasil EDA selanjutnya didokumentasikan dan diinterpretasikan dimana memiliki tujuan sebagai rencana solusi atau tindak lanjut terhadap temuan atau anomali pada fitur dataset, lalu keterangan hasil dikategorikan sebagai “Rekomen” yang artinya fitur tersebut direkomendasikan untuk dilibatkan pada tahap berikutnya, “Tidak Rekomen” artinya fitur tersebut tidak dilibatkan and “Netral” yang menunjukkan bahwa fitur tersebut perlu ditinjau kembali dengan data dan fakta yang relevan. Tabel 1 meunjukkan interpretasi anomaly dari dataset:

Tabel 1. Interpretasi Anomali dari Dataset

Nama Fitur	Temuan Anomali	Rencana Tindak Lanjut	Hasil
No, NIM, Nama, Ayah, Ibu, Alamat, Kode Pos, Telepon, HP, Email, No Ijazah, Tanggal Masuk, Status, Alamat Orang Tua, Kota Orang Tua, Kode Pos Orang Tua, Telepon Orang Tua, HP Orang Tua, Agama, Nama Sekolah	Data tidak bisa dikategorikan dan tidak relevan dengan kasus preferensi Program Studi bagi pendaftar mahasiswa baru di Universitas XYZ	Tidak Melibatkan Fitur tersebut pada tahap berikutnya	Tidak Rekomen

Nama Fitur	Temuan Anomali	Rencana Tindak Lanjut	Hasil
Status Mahasiswa, Pembimbing, Batas Studi, Tahun Masuk, Jenis Beasiswa, JlmSKSPT, KodePT, ProdilDPT, Status Pindahan, Semester Masuk, NIM Asal, Asal Jenjang, Kelas, Nama Jenjang, NamaPST, Tahun Terakhir KRS, IPK	Data tidak relevan dengan kasus preferensi Program Studi bagi pendaftar mahasiswa baru di Universitas XYZ	Tidak Melibatkan Fitur tersebut pada tahap berikutnya	Tidak Rekomen
Jenis Kelamin	Sebaran Merata	Sudah OK dan akan dilibatkan pada tahapan berikutnya	Rekomen
Tempat Lahir	Terdapat imbalance	Perlu dilakukan penyeimbangan data	Netral
Tanggal Lahir	Menjadi dasar untuk menentukan umur	Tanggal lahir sebagai penentuan umur dapat dikategorikan ke dalam rentang tertentu	Rekomen
Program Studi	Sebaran Merata	Sudah OK dan menjadi class	Rekomen
Status Sipil, Provinsi, Kota, Anak Ke, Jumlah Saudara, Penghasilan, Jenjang, Program Kuliah, Jenis Sekolah, Jurusan Sekolah, Nilai Unas, Tahun Lulus	Terdapat beberapa outlier, missing data dan imbalance di dalam fitur tersebut	Perlu dilakukan balancing data, imputasi data terhadap missing data ataupun transformasi data	Rekomen

3.5. Fitur Rekomendation

Dari data tabel 1, terdapat beberapa fitur yang direkomendasikan untuk digunakan pada tahap berikutnya, namun diperlukan penyesuaian data sehubungan dengan temuan-temuan terhadap fitur tersebut seperti adanya *missing value*, *outlier* maupun *imbalance data*. Sehingga diperlukan penyesuaian seperti *labelling*, *deleting record data* maupun imputasi. Adapun fitur yang direkomendasikan yakni Jenis Kelamin, Tanggal Lahir (Umur), Program Studi, Status Sipil, Provinsi, Kota, Anak Ke, Jumlah Saudara, Penghasilan, Jenjang, Program Kuliah, Jenis Sekolah, Jurusan Sekolah, Nilai Unas, Tahun Lulus.

4. KESIMPULAN

Berdasarkan implementasi Exploratory Data Analysis (EDA) terhadap Student Enrolment Dataset Universitas XYZ dihasilkan rekomendasi 14 feature yang memiliki relevansi

dengan konteks penelitian, yakni Jenis Kelamin, Tanggal Lahir (Umur), Program Studi, Status Sipil, Provinsi, Kota, Anak Ke, Jumlah Saudara, Penghasilan, Jenjang, Program Kuliah, Jenis Sekolah, Jurusan Sekolah, Nilai Unas, Tahun Lulus. Namun dari feature tersebut terdapat beberapa anomaly, yakni missing value, imbalance data dan outlier. Sehingga dalam tahapan berikutnya perlu dilakukan imputasi, labelling ataupun deleting data.

DAFTAR PUSTAKA

[1] Simon Kemp, 5 April 2022, Digital 2022 Indonesia :Internet use in Indonesia 2022, https://datareportal.com/reports/digital-2022 indonesia?rq=indonesia%202022.

[2] N. Rohman, R. Luviana Musyarofah, E. Utami, and S. Raharjo, "Natural Language Processing on Marketplace Product Review Sentiment Analysis," in 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS), 2020, pp. 1–5, doi: 10.1109/ICORIS50180.2020.9320827.

[3] X. Wang, T. Zhou, X. Wang, and Y. Fang, "Harshness-aware sentiment mining framework for product review," Expert Systems with Applications, vol. 187, p. 115887, 2022, doi: https://doi.org/10.1016/j.eswa.2021.115887.

[4] J.-W. Bi, Y. Liu, and Z.-P. Fan, "Representing sentiment analysis results of online reviews using interval type-2 fuzzy numbers and its application to product ranking," Information Sciences, vol. 504, pp. 293–307, 2019, doi: https://doi.org/10.1016/j.ins.2019.07.025.

[5] Q. Wang, W. Zhang, J. Li, F. Mai, and Z. Ma, "Effect of online review sentiment on product sales: The moderating role of review credibility perception," Computers in Human Behavior, vol. 133, p. 107272, 2022, doi: https://doi.org/10.1016/j.chb.2022.107272.

[6] Kevin, V. et al. (2020) "Analisis Sentimen Transportasi Online Menggunakan Support Vector Machine Berbasis Particle Swarm Optimization (Online Transportation Sentiment Analysis Using Support Vector Machine Based on Particle Swarm Optimization)," Jurnal Nasional Teknik Elektro dan Teknologi Informasi, 9(2), hal. 162–170FLEXChip Signal Processor (MC68175/D), Motorola, 1996.

[7] Y. Yennimar and R. Rizal, "Comparison of Machine Learning Classification Algorithms in Sentiment Analysis Product Review of North Padang Lawas Regency," SinkrOn, vol. 4, p. 268, 2019, doi: 10.33395/sinkron.v4i1.10416

[8] Sihombing, L., Hannie, H. dan Dermawan, B. (2021) "Sentimen Analisis Customer Review Produk Shopee Indonesia Menggunakan Algortima Naïve Bayes Classifier," Edumatic: Jurnal Pendidikan Informatika, 5, hal. 233–242. doi: 10.29408/edumatic.v5i2.4089

[9] M. T. Akter, M. Begum, and R. Mustafa, "Bengali Sentiment Analysis of E-commerce Product Reviews using K-Nearest Neighbors," in 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), 2021, pp. 40–44, doi: 10.1109/ICICT4SD50815.2021.9396910

[10] S. F. N. H. R. JAYADI, "Sentiment Analysis Of Indonesian E-Commerce Product Reviews Using Support Vector Machine Based Term Frequency Inverse Document," vol. 99, no. 17, pp. 4316–4325, 2022