




Finding Recommended Feature on Student Enrolment Dataset of University XYZ Using Exploratory Data Analysis (EDA)

Zulkarnaini Zulkarnaini¹, Indra Griha Tofik Isa¹ , Leni Novianti¹,
Febie Elfaladonna¹, and Suzan Agustri²

¹ Politeknik Negeri Sriwijaya, Palembang, South Sumatera, Indonesia
indra_isa_mi@polsri.com

² Universitas Indo Global Mandiri, Palembang, Indonesia

Abstract. One of the success of a modelling is the quality of the analysed data. Exploration Data Analysis is a technique used in understanding data to explore which data has quality which will be used in modelling. The case raised in this study is the student registration dataset at XYZ University, where the ultimate goal is how to predict study program preferences for prospective applicants. However, from this data set with various data, it needs to be studied further to produce quality data that is valid, credible, and supports the modelling of preferences for the choice of study program. An EDA will be implemented as a solution to data analysis by looking at the variety of data from the student enrolment dataset, features that support modelling, recommendations that need to be made for advanced stages in a data science cycle. The stages of the research were Problem Analysis, Data Acquisition, Exploration Data Analysis, Anomaly Interpretation, and Feature Recommendations. The final result is in the form of 14 recommended features from the Student Registration Dataset consisting of Sex, Date of Birth, Study Program, Civil Status, Province, City, Child Order, Number of Siblings, Income, Education stage, Lecturing Program, Jenis Sekolah, Department of School, National Test Score, Year of Graduation.

Keywords: Exploratory Data Analysis (EDA) · Enrolment Student Dataset · Data Understanding

1 Introduction

The Industrial Revolution 4.0 leads to a change in the dynamics of technology into a smart, automated, integrated system, making “past” data an asset that has value for its users. One of the impacts of the industrial revolution 4.0 is the existence part of data science that collaborates on quantitative data learning such as statistics, database programming combined with complex algorithms. Data science provides support in various ways, including in the business world, namely to provide efficiency in production systems and processes. By using several types of analysis, business decision stakeholders can see trends in existing data, so as to create a more efficient and structured process.

© The Author(s) 2023

N. L. Husni et al. (Eds.): FIRST-ESCSI 2022, AHE 14, pp. 407–419, 2023.

https://doi.org/10.2991/978-94-6463-118-0_42

Data science has several stages, starting with a business understanding, which is how to understand the problem and purpose of a case context that will be focused. From this stage, several aspects emerge, these are problem determination, project objectives, solutions from a business perspective to success measurement instruments. After the complete business understanding stage is carried out, then data understanding stage, which is how to describe the data so that implicit information can be extracted which will later strengthen the modeling stage.

Data understanding which in other terms is called exploratory data analysis (EDA) is important because at this stage the selection of data relevant to the context of the problem is carried out. Technically, in data understanding, the process defined by (1) data acquisition and withdrawal is carried out; (2) Data analysis by looking at the correlation between features of a data (hereinafter referred to as dataset), looking at data anomalies (such as redundancy data, missing data, or outliers); (3) Data visualization as a representation of the data to be recommended.

The case focused in this study is the enrolment student dataset at XYZ University, where the ultimate goal is how to predict study program preferences for prospective applicants. However, from these datasets with various data, it is necessary to study further to produce data quality that is valid, credible, and supports the modeling of preferences for study programs.

An EDA will be implemented as a solution for data analysis by looking at the variety of data from the enrollment student dataset, potential features that support the modeling stage, recommendations that need to be made for advanced stages in a data science cycle. So that the formulation of the problem in this study is how to implement Exploratory Data Analysis in producing recommended features in the data science stage. The purpose of this study is to produce recommended features of student enrollment that support the modeling of study program preferences for applicants at XYZ University, Palembang City. The limitation of the problem in this research in terms of the dataset that is processed is the Enrollment Student Dataset with 54 features and 2704 records, the tools used are the Python programming language which is accessed through Google Colab.

2 Literature Review

2.1 State of the Art

EDA is an important part in of data science stages which produces quality data to be involved in the next stage. Several previous studies related to EDA [1], where EDA was carried out on COVID-19 cases in Indonesia using HiveQL and Hadoop Environment. The implementation of EDA in this study resulted in an analysis of correlation values which showed a strong influence between the increase in the number of positive confirmed cases and the number of cases of recovered patients and cases of patients dying of 0.94 and 0.9 respectively. The results of other correlation analyzes found a small correlation value between patient cases. in the treatment of positive confirmed cases, cases of patients dying, and cases of patients recovering. For the sustainability of this research, it is necessary to implement regression analysis because the structured data of COVID-19 cases is like a time series.

Other research, EDA was conducted on the dataset of sales of electronic goods which aims to see how the movement of sales of electronic goods is [2], so that it becomes a consideration in planning business improvement strategies. The final result of this research is that there are products with the highest sales value, namely AAA batteries, the least sales sold are LG Dryer and the most expensive electronic products sold during January - December are Macbook Pro Laptops.

The research of Wahyuni, et al., which aims to explore implied information from fashion product sales data to assist in the pre-processing stage by showing the missing values, imbalance data and outliers [3]. The final result of this study can be concluded that EDA can optimize knowledge of data, which can be used to enrich understanding of evaluation data analysis.

While this research focuses on analyzing data to explore data characteristics, outliers of each feature, correlation values between features, so as to produce features that have the best quality which will later be used in the transformation to modeling stages.

2.2 Data Mining

Data mining terminology is generally associated with KDD, but in fact data mining is a major part of the KDD process [4]. By definition, data mining is a process to find a pattern from a large and complex amount of data. The data sources used can come from databases, data warehouses, websites, other information repositories, or data entered into the system dynamically [5].

A pattern becomes good if it goes through valid test data with a level of data certainty, novelty, and has potential to be useful (for example, the data can be followed up or validated for users who make predictive data) and easy to understand [6]. A good pattern represents knowledge [7]. Data mining can be carried out on various types of data as long as the data is meaningful to the specified target, such as data in databases, data warehouses, transaction data and advanced data types, including sequence data, data streams, spatial and spatiotemporal data, text data and multimedia, graph and network data, web data. [8].

2.3 Exploratory Data Analysis

Chong Ho Yu explained in his journal that EDA has provided a breakthrough in how to handle outliers, where in the past it was said that outliers were detrimental to data analysis because the slope of the regression line could be driven by only one extreme datum point. However, with EDA, outlier handling becomes more accurate, especially when it will be implemented in the data mining or modeling stage of [9]. Exploratory Data Analysis allows analysts to understand the content of the data used, from distribution, frequency, correlation and others. In practice, curiosity is very important in this process, understanding the context of the data is also considered, because it will answer the basic problems of [9].

In general, EDA is carried out in several ways, namely (1) Univariate Analysis — descriptive analysis with one variable; (2) Bivariate Analysis — analysis of the relationship with two variables usually with the target variable; (3) Multivariate Analysis — analysis that uses more than or equal to three variables [10].

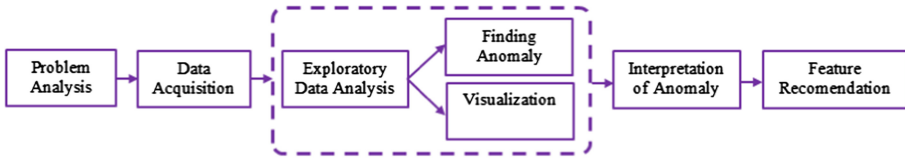


Fig. 1. Research Method

3 Method

The object of this research is the student enrollment dataset of University XYZ with 2704 data records and data processing using Google Colab. The stages of the research can be seen in Fig. 1: deteksi kelebu sungi.

At the research stage, it begins with problem analysis where the problems to be studied are determined, the objectives as solutions to these problems, as well as the mechanism and scope in completing the research that is built. Next is the Data Acquisition which is done by reading the enrolment student dataset with a CSV type file. The raw data that will be processed to produce feature recommendations through EDA. The next stage is the implementation of EDA on the dataset, by finding anomalies from the dataset on each feature, such as imbalance data, missing values, outliers, data characteristics (whether the data is nominal or categorical) as well as the data type for each feature. To facilitate the analysis of data, visualization is carried out with histograms, boxplots and pie charts that are adapted to the context and characteristics of the observed features. After the data is visualized, the next step is the interpretation of anomaly on each observed feature, by providing a follow-up plan for the feature, which is categorized into “Recommend” and “Not Recommend” Furthermore, from this stage, it produces feature recommendations from the enrollment student dataset that have credible values and support the next process, namely transformation and modeling.

4 Result and Discussion

4.1 Problem Analysis

The purpose of this study is to find the recommended features generated from the University XYZ student enrollment dataset, which will later be used in modeling preferences for study program selection for applicants. The dataset is taken from 2018 to 2020 which consists of various numerical and categorical data, various data types, various anomalies such as imbalance/missing value/data redundancy/outliers. So we need Exploratory Data Analysis by looking at how the correlation between features, the extent of the anomaly in each feature, how the visualization is generated from these features.

4.2 Data Acquisition

Data Acquisition is done by implementing the pandas library (`read_csv`) as in Fig. 2 through reading the CSV dataset that has been inserted in the repository. In Fig. 3 is the dataset that appears after the implementation of the pandas library.

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

Fig. 2. Syntax for 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	Desan 3 desa Karang Anyar, Kec. Lingsang Wetan, Kab. Musi Banyuwasin, Sumatera Selatan	-	SUMATERA SELATAN	KAB. MUSI BANYUASIN	Indonesia
L	ISLAM	Palembang	11/12/2002	B	J. DI PANJAITAN Lrg. Poppang	30265	SUMATERA SELATAN	KOTA PALEMBANG	Indonesia
L	ISLAM	Muaradua	24/04/2003	B	J.N SETIUNGGAL KOMPLEK PERGASDA Blok B-73	30114	SUMATERA SELATAN	KOTA PALEMBANG	Indonesia
P	ISLAM	Bandar Lampung	02/03/2002	B	U. Sidahayn Telang Jawa	31714	SUMATERA SELATAN	KAB. MUARA ENIM	Indonesia
L	ISLAM	Lubuk Kilumpang	07/01/2003	B	Lubuk Kilumpang	0	SUMATERA SELATAN	KAB. LAHAT	Indonesia
P	ISLAM	Palembang	11/05/2001	B	J. DI PANJAITAN lrg. Sekaya 3 No.01	30265	SUMATERA SELATAN	KOTA PALEMBANG	Indonesia

Fig. 3. Dataset read using Pandas

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

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

Fig. 4. Features in Dataset

To see further from the various feature data types, print(df.dtypes) is implemented so that all features and data types of these features appear. Figure 4 shows all the features and their data types.

From the exploration of the dataset in Fig. 4, it can be seen that there are several data types with numeric types, namely integer (int64) and float (float64), as well as object data types which are a combination of numeric, letter, ASCII or date characters, which allows it to be adapted to the data type of the character of the feature. If added up, the features with the int64 data type are 11 features, the float64 data type has 9 features and the remaining 36 features.

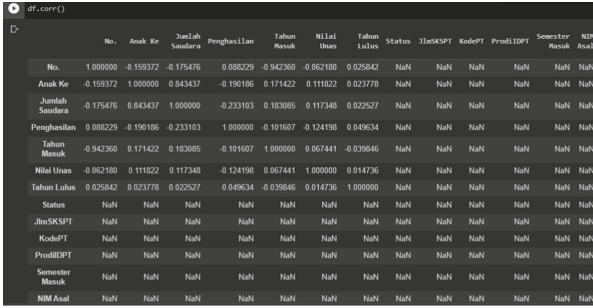


Fig. 5. Feature Correlation (Without Label Transformation)

Furthermore, without transforming or in other words through the original dataset, a correlation test between features/variables is carried out with the `df.corr()` formulation, then seen in Fig. 5 the results of the correlation between these features. It can be seen that most of the features have a low correlation (assuming a threshold of 0.75) and have a NaN (Not a Number) value. However, there are several features that have a high correlation, namely “Jumlah Saudara” with “Anak Ke” which correlates with a value of 0.84. So if it is concluded from this stage it is necessary to transform data for recommended features after the implementation of EDA.

4.3 Exploratory Data Analysis

The implementation of EDA is done by finding anomalies and visualizing data to facilitate data interpretation. Since the main objective of this research is modeling study program preferences for applicants at XYZ University, the class of this dataset is “Program Studi”. So it is necessary to study the data to see how the various classification names and the amount of data from the “Study Program” class are. For this reason, a data analysis was carried out by looking at the number of classifications and data from the “Program Studi” Class, which is represented in Fig. 6.

In Fig. 6 there are 13 study programs with the lowest distribution of study programs being “Keselamatan dan Kesehatan Kerja” with 30 data and the highest being “Manajemen” with 458 data. For the distribution of data with a value of less than 100, there are 4 Study Programs other than “Keselamatan dan Kesehatan Kerja”, namely “Arsitektur”, “Survei dan Mapping” and “Manajemen Informatika”. While the Study Programs with data values above 400 are “Sistem Informasi” and “Manajemen”.

Visualization of the “Program Studi” class is done to make it easier to represent the data. In this case, the visualization is categorized into low, medium and high data distribution, where for low data are the bottom 3 data, high data are the 3 highest data, and the rest are low data distributions. Technically, the visualization implements the `pyplot` library from `matplotlib` by naming the variable “`plt`” in Fig. 7. Then in Fig. 8 is the visualization result of coding Fig. 7 with color effects that make it easier to see the distribution of data.

```
#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

Fig. 6. Classification Class of “Program Studi”

```
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()
```

Fig. 7. Classification Class of “Program Studi”

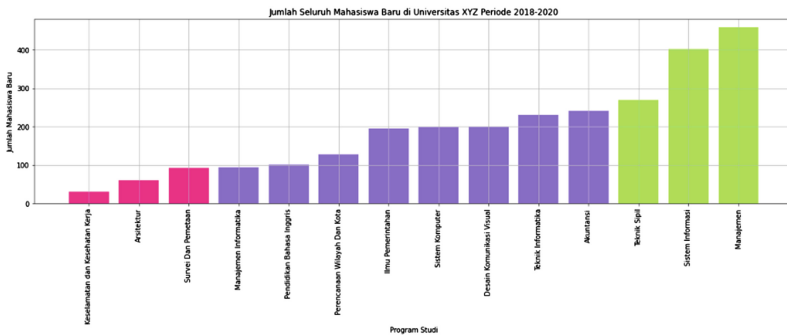


Fig. 8. Result of Class “Program Studi” Visualization

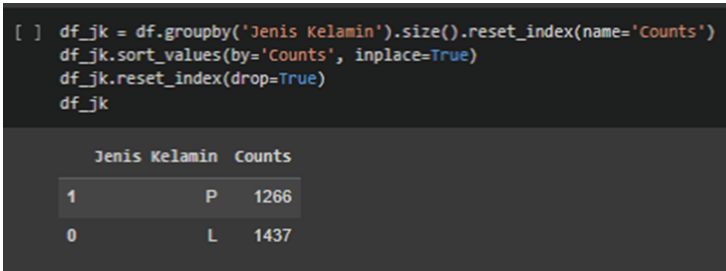


Fig. 9. Count Result for feature “Jenis Kelamin”

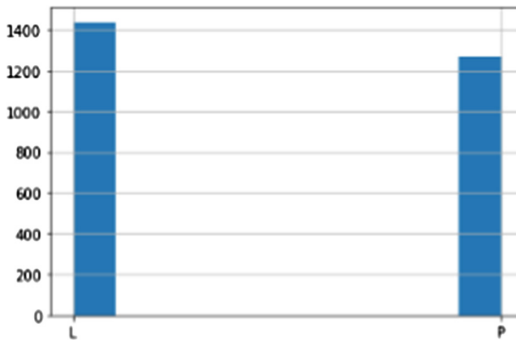


Fig. 10. Visualization feature “Jenis Kelamin”

Next, EDA is carried out on features that are relevant to the context of modeling preferences for study program selection for applicants at XYZ University, meaning that the data or features involved are only data before the registrant becomes a student. In this regard, there are 8 features that are not relevant to the “data pendaftar”, yakni “NIM”, “Status Mahasiswa”, “Pembimbing”, “Batas Studi”, “Jenis Beasiswa”, “NIM Asal”, “Tahun Terakhir KRS”, “IPK”.

Some features have data that tends to be evenly distributed, including “Jenis Kelamin” or called as gender. Gender has a classification of “P” and “L”, through the count function, the resulting “P” is 1266 records and “L” is 1437 records. The implementation of the count for “Jenis Kelamin” can be seen in Fig. 9, while the visualization results for “Jenis Kelamin” are shown in Fig. 10.

Before proceeding to other EDA features, it is necessary to explore missing data to see how much is missing data, whether it affects the balance of the data or not. Figure 11 shows several features that have high missing data (>1000 data records), including “No. Ijazah”, “Jenis Beasiswa”, “JlmSKSPT”, “Semester Masuk”, “NIM Asal”, “Kelas”, “Nama Jenjang”, “NamaPST”, “Telepon Orang Tua”. If you look closely, almost all of these features are features if the registrant’s status is already a student. However, there is a feature that is relevant to the condition of the registrant, namely “Nama Jenjang”. If we look closely again, the number of missing data from the “Nama Jenjang” feature is the entire dataset record, which is 2704 records. This means that the feature is also

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

Fig. 11. Result of Missing Value of Dataset

```
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

Fig. 12. Imbalance data on feature “Status Sipil”

considered irrelevant to the registration context. So if it is concluded that features that have missing data records > 1000 can be omitted, with the following reasons: (1) Not relevant to the Student Enrollment context; and, (2) The “Nama Jenjang” feature of all data content is a missing value.

Data imbalance cases were also found for several features, such as the “Status Sipil” feature where the label “S” has 18 records while “B” has 2683 records, as shown in Fig. 12.

The outliers can be seen in Fig. 13, where there is a label year “201” with 1 record and “0” with 36 data records.

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

Fig. 13. The Outlier data of Feature “Tahun Lulus”

4.4 Interpretation of Anomaly

The various EDA results are then documented and interpreted which has a purpose as a solution plan or follow-up to findings or anomalies in the dataset features, then the result description is categorized as “Recommend” which means the feature is recommended to be involved in the next stage, “Not Recommend” means the feature excluded and “Neutral” which indicates that the feature needs to be reviewed with relevant data and facts. Table 1 shows the interpretation of anomaly.

4.5 Feature Recommendation

From the data in Table 1, there are several features that are recommended for use in the next stage, but data adjustments are needed in relation to the findings of these features such as missing values, outliers and imbalance data. So that necessary adjustments such as labeling, deleting data records and imputation. The recommended features are Sex (Jenis Kelamin), Date of Birth/Age (Tanggal Lahir (Umur), Study Program (Program Studi), Civil Status (Status Sipil), Province (Provinsi), City (Kota), Child Order (Anak Ke), Number of Siblings (Jumlah Saudara), Income (Penghasilan), Education Stage (Jenjang), Lecturing Program (Program Kuliah), Type of School (Jenis Sekolah), Department of School (Jurusan Sekolah), National Test Score (Nilai Unas), Year of Graduation (Tahun Lulus).

Table 1. Interpretation of Anomaly

Features Name	Finding Anomaly	What To Do	Result
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	The data cannot be categorized and is not relevant to the case of Study Program preferences for enrolment student applicants at XYZ University	Does not involve the feature in the next stage	Not Recommend
Status Mahasiswa, Pembimbing, Batas Studi, Tahun Masuk, Jenis Beasiswa, JlmSKSPT, KodePT, ProdiIDPT, Status Pindahan, Semester Masuk, NIM Asal, Asal Jenjang, Kelas, Nama Jenjang, NamaPST, Tahun Terakhir KRS, IPK	The data is not relevant to the case of Study Program preferences for new student applicants at XYZ University	Does not involve the feature in the next stage	Not Recommend
Jenis Kelamin (Sex)	Normal Distribution	The feature is OK and will be involved in the next stage	Recommend
Tempat Lahir	Imbalance Data	Need to balance data	Neutral
Tanggal Lahir (Date of Birth)	Be the basis for determining "umur" feature (Age)	Date of birth as a determination of age can be categorized into a certain range	Recommend
Program Studi (Study Program)	Normal Distribution	The data is OK and will be class	Recommend
Status Sipil, Provinsi, Kota, Anak Ke, Jumlah Saudara, Penghasilan, Jenjang, Program Kuliah, Jenis Sekolah, Jurusan Sekolah, Nilai Unas, Tahun Lulus In English (Civil Status, Province, City, Child order, number of siblings, Income, Level, Study Program, Type of School, School Department, National Test Score, Year Graduated)	There are several outliers, missing data and imbalances in these features	It is necessary to do data balancing, imputation of data to missing data or data transformation	Recommend

5 Conclusion

Based on the implementation of Exploratory Data Analysis (EDA) on the Student Enrolment Dataset at XYZ University, 14 feature recommendations that have relevance to the research context were produced, namely 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. However, from these features there are several anomalies, namely missing values, data imbalances and outliers. So that in the next stage it is necessary to imput, label or delete data.

Acknowledgements. The Author will acknowledge to Research and Community Service Center (P3M) Politeknik Negeri Sriwijaya for Innovative Research Grant.

References

1. T. M. Fahrudin, P. A. Riyantoko, K. M. Hindrayani, and I. G. S. Mas Diyasa, "Exploratory Data Analysis pada Kasus COVID-19 di Indonesia Menggunakan HiveQL dan Hadoop Environment," *Pros. Semin. Nas. Inform. Bela Negara*, vol. 1, pp. 115–123, 2020, <https://doi.org/10.33005/santika.v1i0.32>.
2. M. Radhi, A. Amalia, D. R. H. Sitompul, S. H. Sinurat, and E. Indra, "Analisis Big Data Dengan Metode Exploratory Data Analysis (Eda) Dan Metode Visualisasi Menggunakan Jupyter Notebook," *J. Sist. Inf. dan Ilmu Komput. Prima (JUSIKOM PRIMA)*, vol. 4, no. 2, pp. 23–27, 2022, <https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v4i2.2475>.
3. E. D. Wahyuni, A. A. Arifiyanti, and M. Kustyani, "Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining," *Pros. Nas. Rekayasa Teknol. Ind. dan Inf. XIV Tahun 2019*, vol. 2019, no. November, pp. 263–269, 2019, [Online]. Available: <http://journal.itny.ac.id/index.php/ReTII>
4. 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=jeasci.2017.4102.4107.
5. J. Han, M. Kamber, and J. Pei, *Data Mining: Concept and Techniques*, Second Edi. Waltham: Morgan Kaufmann Publishers, 2006.
6. 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, <https://doi.org/10.21533/pen.v1i2.17>.
7. I. G. T. Isa and F. Elfaladonna, "Penilaian Kinerja Akurasi Metode Klasifikasi dalam Dataset Penerimaan Mahasiswa Baru," *JEPIN (Jurnal Edukasi dan Penelit. Inform.)*, vol. 8, no. 2, pp. 292–298, 2022.
8. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012. <https://doi.org/10.1016/B978-0-12-381479-1.00001-0>.
9. C. Nicodemo and A. Satorra, "Exploratory data analysis on large data sets: The example of salary variation in Spanish Social Security Data," *BRQ Bus. Res. Q.*, vol. 25, no. 3, pp. 283–294, 2020, <https://doi.org/10.1177/2340944420957335>.
10. A. T. Jebb, S. Parrigon, and S. E. Woo, "Exploratory data analysis as a foundation of inductive research," *Hum. Resour. Manag. Rev.*, vol. 27, no. 2, pp. 265–276, 2017, <https://doi.org/10.1016/j.hrmr.2016.08.003>.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

