Extracting the Recommended Features from the Elementary School Student Dataset through Exploration Data Analysis (EDA)

Devi Sartika[1], Febie Elfaladonna[2*], Indra Griha Tofik Isa[3] , Andre Mariza Putra[4]

1,2,3,4 Politeknik Negeri Sriwijaya, Palembang, South Sumatera, Indonesia
`febie_elfaladonna_mi@polsri.com`

Abstract. Exploratory data analysis (EDA) is an important stage in a data science cycle. In this research, the EDA process is carried out on the elementary school student dataset derived from the student "interest" and "talent" questionnaires. The purpose of this research is to find recommended features that will be used in the modeling stage. The main methods used in the implementation of EDA are chi square and T-test on the dependent variable, "class" and fifteen dependent variables. The stages were carried out by (1) analyzing the documents, data, and participants; (2) developing the questionnaire; (3) implementing the Likert and Yes/No questions; (4) formatting the data into tabular data; (5) coding and exploratory data analysis; (6) interpreting the findings and conclusion. From the results of chi square testing, the highest value was obtained in the "excellent in acting" variable with a value of 17.79284731, while the lowest result was found in the "writing, reading or storytelling" variable with a value of 0.29389977. Through the T-test, 3 categories of variable influence were obtained, i.e., "strong", "moderate", and "weak".

Keywords: Exploratory Data Analysis (EDA), Elementary School Student Dataset, Chi Square, T-test

1   Introduction

In the field of education, especially primary schools, the aspect of interests and talents is important because it has a significant impact on the development of students [1]. Through the exploration of interests and talents, the school can easily gain insights such as personal development strategies that will be applied to students according to their specialization [2]. Increase student motivation, facilitate the identification of strengths, early exposure in potential career exploration, facilitate students in the process of problem solving and adaptability, and provide an impact of inclusivity for students.

The implementation of interest and talent tracking can be applied at State Elementary School (SDN) 204 Palembang, which is one of the elementary schools located in Kertapati District, Palembang. It currently has 600 students spread across grades one to six, has 30 teachers, and 3 staff. Currently, the assessment process for exploring interests and talents is done through a normative process by asking students questions. However, the question is that in the search process, the information obtained from the

a student is really an interest or a talent. Because with the school knowing whether the information is an interest or talent, it can certainly help the school in providing the best treatment for students. So that the output of student participation in every activity carried out inside or outside the school will get optimal results.

One of the techniques that can be used to support the exploration of students' interests and talents is through the implementation of machine learning through data extracted based on students' experiences, which ultimately becomes new knowledge for stakeholders [3] [4]. However, the fundamental thing that needs to be considered in machine learning modeling is the features that have a significant impact on the observed class, where in this case the class is divided into "interest" and "talent". The data search process is carried out by developing questionnaires given to parents and students themselves.

So the purpose of this research is how recommended features are generated from elementary school student data, which can later be implemented in modeling the determination of interests or talents of elementary school students. The contribution of this research is to analyze and extract potential features from student data using statistical analysis tools, namely chi-square and T-test. The results of the research conducted will then become recommendation features that will be implemented in the next stage of data modeling and application development. The data used in the research is primary school student data from grade one to grade three in the 2023-2024 school year with a total of one hundred and fifty data.

2   Literature Review

2.1   State of The Art

EDA is one of the important stages in data science or machine learning. A lot of data science research never misses the Exploratory Data Analysis (EDA) process.

Exploratory Data Analysis is a stage for initial testing aimed at recognising patterns, finding anomalies, and testing initial assumptions. EDA is very helpful for early identification of errors and outliers, knowing the relationship of variables, and of course useful for analyzing activities in the field of statistics. There are many studies related to EDA such as mass spectrometry analysis of marine world analogues. The analysis from this study has features with correlations lower than 0.9. A restriction on the repetition of measurements for each experiment was made by limiting the number of features where the correlation value was below 0.9. This basic dimensionality reduction algorithm shows that it is better at clustering than the raw data. [5]. The next research discusses the analysis of large data sets with cases of variations in Spanish social security. This research is carried out with A discussion of some box plot and scatter plot analysis techniques by reviewing the increasing use of modern graphical computing tools whose use is illustrated with Spanish Social Security and security data. The research explores how earnings vary across several factors such as age, gender, type of employment, and contract,

and in particular to the gender gap in salaries visualized in different dimensions related to job type. Exploratory data analysis methods were also applied to assess and refine competing regressions by plotting residual values against corresponding values.

The methods discussed will be useful for researchers to assess heterogeneity in the data, intergroup variation, and classical diagnostic plots of the residuals from alternative models of fit. [6]. In addition, the following research explains exploratory data mining
   techniques could be simultaneously or sequentially employed. For example, because both neural networks and classification trees are capable of selecting important predictors, they could be run side by side and evaluated by classification agreement and ROC curves. On other occasions, a sequential approach might be more appropriate. For instance, if the researcher suspects that the observations are too heterogeneous to form a single population, clustering could be conducted to divide the sample into sub-samples. Next, variable selection procedures could be run to narrow down the predictor list for each sub-sample. Last, the researcher could focus on the inter-relationships among just a few variables using pattern recognition methods. The combinations and possibilities are virtually limitless. Data detectives are encouraged to explore the data with skepticism and openness. [7].

2.2   **Data Science**

Data science is the method or science used to extract knowledge from data. This includes how to retrieve data and use data to gain knowledge that will be used in making decisions, making predictions for the future, understanding the past and helping to make innovations to creations in the industrial field. In general, data science can be thought of as concepts historically proposed in the statistics and math communities with a focus on data analysis. mathematics community with a focus on data analysis. Today, data science has expanded beyond statistics and math and has encompassed fields such as data mining and machine learning. and mathematics and has encompassed fields such as data mining and machine learning. It can now be considered a new interdisciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge-including statistics, mathematics informatics, computing, communications, data management, sociology, to study data domains that follow "data science thinking." Sometimes, data science can be thought of as the scope of principles, problem definitions, algorithms, and processes for non-trivial extraction and discovery of useful patterns from large data sets. It involves the process of collecting, preparing, managing, analyzing, explaining, and disseminating data and analysis results [8]. Data Science is also intended as a forum for data processing that can be used with people who have skills in programming, statistics, and business. In business, Data Science is utilized for forecasting activities to predict something that can be done in the short, medium, and long term [9].

2.3   **Exploratory Data Analysis**

According to John W. Tukey Exploratory Data Analysis (EDA) is defined as "detective work -numerical detective work -or counting detective work -or graphical detective work", by being a detective or doing detective work to discover what the data can provide without being prejudiced and relying on facts. JohnW. Turkey invented Confirmatory Data Analysis (CDA) which is a contrast to EDA, focusing on the scope of data analysis related to statistical hypothesis testing, confidence intervals, and

estimation. EDA and CDA should be used together in a complementary way to discover patterns and structures that lead to hypotheses and models [10].

The purpose of EDA is to find data patterns. This is in line with the concept of data mining which is also used to explore patterns in data. With the onslaught of large amounts of data in the big data era, exploring data patterns has become more complicated due to the large data dimensions. EDA is used for the purpose of reducing data dimensionality or enriching the understanding of data analysis through data visualization. EDA is also used, among others, to optimize knowledge about the data, generate important variables, detect outliers and anomalies in the data, and test initial assumptions [11].

3   Method

In this research, systematic stages were carried out, starting with the initiation process and ending with the interpretation of the observed data. Specifically, the research stages are represented in figure 1.
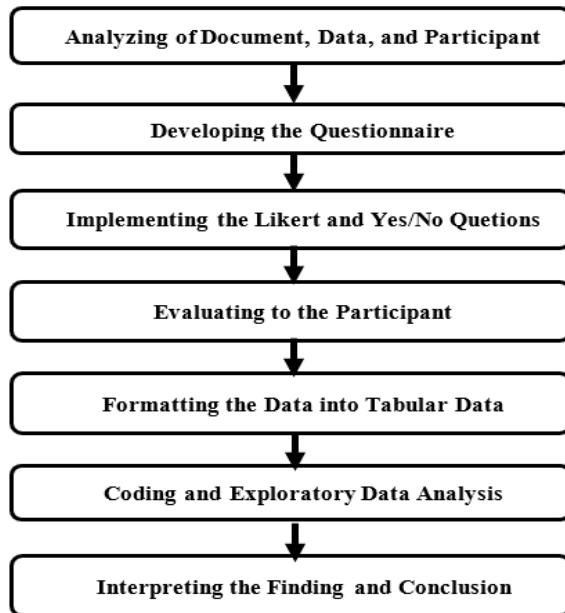

Fig. 1. Research Method

At the research stage, it begins with data analysis where the data used in the research is questionnaire data obtained from parents of SDN 204 Palembang students intended for grades one to three. The data was created by the school using google form and distributed to parents. The answer to the statement is an ordinal answer that shows the level of confidence in the answer to the statement. The statements contained in this google form discuss matters related to students' interests and talents that have been monitored by parents. After distributing questionnaires and being filled in by parents,

the number of respondents was 144 parents. The respondent data was then processed by the school manually to classify between interests and talents based on the knowledge of the counseling teacher. References to statements were obtained from various sources including teacher knowledge and journals. From these references, the process of sorting statements that become interest or talent criteria is then carried out. Statements that have the same meaning are put together so that the total statements or criteria contained in the questionnaire are seventeen with the classification results divided into interest and talent. Table 1 below shows the list of questionnaires:

Table 1. Questionnaire Statement

Id	Statements
Q1	Student class
Q2	Age of student
Q3	Critical thinking
Q4	Trading and calculating
Q5	Certain sport activity
Q6	Playing musical instrument and singing
Q7	Cooking activity
Q8	Writing, reading or storytelling
Q9	Drawing, painting or coloring
Q10	Often joining coloring, drawing or painting competition
Q11	Often joining writing, reading or story telling competition
Q12	Often joining singing and acting competition
Q13	Able in singing while playing musical instrument
Q14	Excellent in drawing, painting and coloring
Q15	Excellent in acting

The technique used to answer questions is to use a likert scale. The likert scale is the most widely used psychometric scale in questionnaires and is more often found in survey research. This was also done to fill out the questionnaire. The person who gives the scale to the answer choices is the counseling teacher. There are five answer options with a confidence level that starts from 0 to 100. The purpose of giving the range is so that parents can ascertain whether the statements on the questionnaire are in accordance with the child's interests or talents.

Table 2. Answer Scale

Answer scale	Information
0	Don't know
20	Likely
40	Most likely
60	Almost certain
80	Certain
100	Very sure

The questionnaire distributed by SDN 204 Palembang has been conducted for about a year and is intended for new students from grade one to grade three. Filling out the questionnaire can be done wherever parents are because it is distributed online. The data for the study used data that had just been distributed at the beginning of the school semester of the 2023/2024 school year. The time span for filling in the questionnaire varies because not all parents can fill in at the same time. Filling out the questionnaire also found many obstacles because some of the parents did not understand how to fill out the questionnaire. Based on data obtained from the school, every day 20 to 40 respondents' data were collected so that in approximately one week almost all parents had filled out the questionnaire.

After distributing questionnaires and obtaining data on students' interests and talents from parents as respondents, the next step is to process the survey data into tabular data. Data tabulation is the processing of questionnaire data into tabular form with the aim of facilitating the analysis process. Because the form of the respondent's answer is a Likert scale, the data processing process is not difficult. There are two variables in the data processing process (1) dependent variables, namely classes consisting of labels Interest and Talent and (2) independent variables consisting of fifteen questionnaire items when processed with python. The dependent variable is the dependent variable that is affected because of the independent variable while the independent variable is the variable that has a strong influence on the dependent variable. The application used to tabulate data is Ms Excel along with its format. This tabular data processing will be used to select which criteria have the most influence on the dependent and independent variables with Jupyter Notebook.

Exploratory Data Analysis (EDA) is a step of analyzing and displaying data with the intention of getting a better understanding of the data. There are several steps used in the implementation of Exploratory Data Analysis, namely:

Focus on the data and look for missing values. This stage is carried out by the author by looking and observing whether the existing data is in accordance with the research needs. It is intended that the results that will be obtained will be maximized (minimal error).

Categorizing the data. This stage is useful for equalizing the final perception of the analysis results. In this study, the authors used non-numerical (categorical) data. This categorical data is the original data obtained from respondents.

Identifying the relationship between each variable. After performing the above two stages, the author found the relationship between variables that affect class classification.

Finding anomalies in the data. Anomalies are data that deviate from the existing data set. This data does not mean it is not important but must be deleted because it does not meet certain value requirements (threshold). This data will later be selected several times to get variables that are really suitable and have a relationship with interest or talent classes.

4   Result and Discussion

4.1   The Questionnaire

As described in section 2, the dataset collected is in the form of questionnaire data given to 144 participants. The questionnaire form consists of (1) informant summary data which includes Name of Student's Parents, Student's name, Student Class, Current Student Age (years); (2) the fifteen of questionnaires. The following figures are visualizations of the questionnaires given to participants.

Fig. 2. Short Informant Data in Questionnaire

Fig. 3. The 15 of Questionnaire

4.2   Exploratory Data Analysis

The next stage is exploration data analysis by involving all features in the elementary school interest and talent dataset. The analysis uses chi square and T test, where theoretically chi square is part of statistical testing which aims to see how significant the relationship between two variables is, i.e. the independent variable and the dependent variable. Chi square is symbolized by X2. While the T-test is also part of statistical testing, it focuses on how significant the difference between these variables is, so that it can produce a hypothesis whether the relationship variable is acceptable or not.

In this study, the dependent variable is "prediction" which consists of the labels "interest" or "talent", while the independent variables consist of "student class", "age of student", "critical thinking", "trading and calculating", "certain sport activity", "playing musical instrument and singing", "cooking activity", "writing, reading or storytelling", "drawing, painting or coloring", "often joining coloring, drawing or painting competition", "often joining writing, reading or story telling competition", "often joining singing and acting competition", "able in singing while playing musical instrument", "excellent in drawing, painting and coloring", "excellent in acting".

After determining the dependent and independent variables, the next step is the implementation of python coding using the Jupyter Notebook application which is described in detail in section 2. The first step is to import the supporting libraries used in the exploration data analysis (EDA) process, namely pandas, numpy, scipy and sklearn. In particular, the sklearn library is used to perform the chi square and T-test processes. Figure 4 represents the coding flow of the library importing process.

```
import pandas as pd
import numpy as np
import scipy.stats as stats
from sklearn.feature_selection import chi2
```

Fig. 4. Library Importing Process in Python

In carrying out the EDA process, a dataset source is needed to be analyzed where in this study using a dataset with comma separated value (CSV) extension with the filename "datamentah.csv". The process of calling the dataset using the pandas library with the alias "pd" through the "read_csv" function so that it displays the dataset represented in figure 5 as follows below:

	Kelas_Siswa	Umur_Siswa	Berpikir_Kritis	Suka_berdagang_dan_berhitung
0	2	7	Ya	Ya
1	2	8	Tidak	Tidak
2	1	6	Ya	Ya
3	1	7	Ya	Tidak
4	2	7	Ya	Ya

Fig. 5. The Representation of Dataset "datamentah.csv"

The labels in the "datamentah.csv" dataset are mostly categorical textual data where the format does not support the implementation of chi square and T-test, so it is necessary to transform the data into numerical categorical data through the "LabelEncoder" function from the sklearn.preprocessing library. Technically, the implemented coding structure is shown in Figure 6.

```
from sklearn.preprocessing import LabelEncoder
for col in df.columns:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    df.head()
```

Fig. 6. Coding Structure of "LabelEncoder" function

The implementation result of the "LabelEncoder" function transforms text data into numeric data, according to the labels contained in each feature, as shown in Figure 7.

	Kelas_Siswa	Umur_Siswa	Berpikir_Kritis	Suka_berdagang_dan_berhitung
0	1	1	1	1
1	1	2	0	0
2	0	0	1	1
3	0	1	1	0
4	1	1	1	1

Fig. 7. Transformation Result using "LabelEncoder" Function

The implementation of chi square is through the "chi2" function contained in the "sklearn.feature_selection" library, where there is an "X" variable as the independent variable, and a "y" variable as the dependent variable. Technically, the coding structure implemented in chi square is represented in Figure 8.

```
from sklearn.feature_selection import chi2
x = df.drop(columns=['Prediksi'], axis=1)
y = df['Prediksi']


chi_score = chi2(x, y)
```

Fig. 8. Coding Structure of Chi Square in Python

The result of the chi square coding structure is a chi score with the results represented in Figure 9.

```
(array([ 0.99209243,  0.63224066,  0.69202329,  0.83476502,  0.43042221,
         9.09129968,  4.68226551,  0.29389977,  6.17494196,  6.24648434,
        15.46748504,  2.93558343,  4.31730334,  4.55054406, 17.79284731]),
```

Fig. 9. Result of Chi Score

In addition to the chi score output, which is the implementation of "chi2", there is also an output in the form of T-test results that show how significantly the independent variable affects the dependent variable, with the results represented in Figure 10 below.

```
array([3.19231504e-01, 4.26534613e-01, 4.05477036e-01, 3.60898283e-01,
       5.11781849e-01, 2.56828298e-03, 3.04755318e-02, 5.87732233e-01,
       1.29572161e-02, 1.24440052e-02, 8.39368115e-05, 8.66473532e-02,
       3.77266641e-02, 3.29082881e-02, 2.46307450e-05]))
```

Fig. 10. Result of T-test

The technical explanation of the result of chi score and T-test will be discussed further in section 4.3.

4.3   Interpreting the Findings and Discussion

Table 3 is a detailed representation of the chi score and its variables. The variable "student class" produces a chi score of 0.99209243, "cooking activity" produces a chi score of 4.68226551. If examined in detail, the highest chi score results in the variable "excellent in acting" with a value of 17.79284731, the second of highest score belongs to "often joining writing, reading or storytelling" variable with the score of 15.46748504. Meanwhile, the lowest chi score is belong to "writing, reading or storytelling" with the value of 0.29389977. In this case, the higher the chi score, the better the correlation between the variables. Table 4 shows the results of sorted chi-square with the highest to the lowest value.

Table 3. Unsorted Result of Chi Square

Variable	Value
Student class	0.99209243
Age of student	0.63224066
Critical thinking	0.69202329
Trading and calculating	0.83476502
Certain sport activity	0.43042221
Playing musical instrument and singing	9.09129968
Cooking activity	4.68226551
Writing, reading or storytelling	0.29389977
Drawing, painting or coloring	6.17494196
Often joining coloring, drawing or painting competition	6.24648434
Often joining writing, reading or story telling competition	15.46748504
Often joining singing and acting competition	2.93558343
Able in singing while playing musical instrument	4.31730334
Excellent in drawing, painting and coloring	4.55054406
Excellent in acting	17.79284731

Table 4. Sorted Result of Chi Square

Variable	Value
Excellent in acting	17.79284731
Often joining writing, reading or storytelling	15.46748504
Playing musical instrument and singing	9.09129968
Often joining coloring, drawing or painting competition	6.24648434
Drawing, painting or coloring	6.17494196
Cooking activity	4.68226551
Excellent in drawing, painting and coloring	4.55054406
Able in singing while playing musical instrument	4.31730334
Often joining singing and acting competition	2.93558343
Student class	0.99209243
Trading and calculating	0.83476502
Critical thinking	0.69202329
Age of student	0.63224066
Certain sport activity	0.43042221

Writing, reading or storytelling	0.29389977

To ensure how the relationship between the independent variable and the dependent variable has a high influence value, a T-test is conducted. The results of the T-Test can be seen in Figure 15, where in the T- test the accepted value is if the results show less than 0.5. If observed, there are 2 variables that have a value greater than 0.5, namely "writing, reading or storytelling" and "certain sport activities" and it is concluded that these variables have a "weak" influence. Then there are 4 variables that have a value in the range of 0.3 to 0.4, namely "student class", "trading and calculating", "critical thinking", "age of student" and are included in variables that have a "moderate" influence. While there are 9 variables that are classified as having a "strong" influence where the value is less than 0.1, namely "excellent in acting", "often joining writing, reading or storytelling", "playing musical instrument and singing", "often joining coloring, drawing or painting competition", "drawing, painting or coloring", "cooking activity", "excellent in drawing, painting and coloring", "able in singing while playing musical instrument", "often joining singing and acting competition".
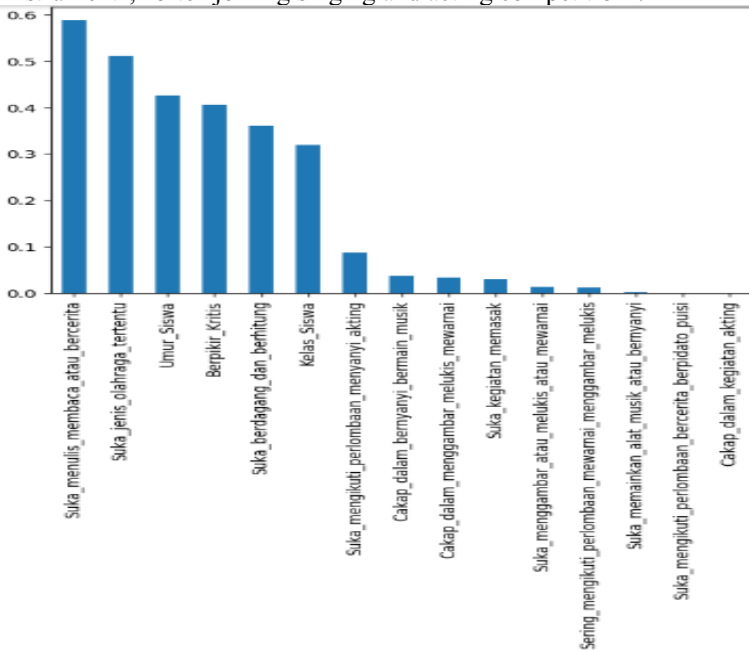

Fig. 11. Graphic Result of T-test

5   Conclusion

Exploratory data analysis (EDA) was conducted to determine the "interest" or "talent" status of elementary school students of SDN 24 Palembang. Tests were conducted with chi square and T-test on 15 independent variables with dependent variables. From the chi square test results, the highest value is "excellent in acting" variable with a value of 17.79284731 and the lowest value is "writing, reading or storytelling" with a value of

0.29389977. Meanwhile, the T-test results produced three categories of influence where the weak influence (>0.5) was found in the variables "writing, reading or storytelling" and "certain sport activities". The "moderate" influence category with a range of 0.3 ~ 0.4, namely the variables "student class", "trading and calculating", "critical thinking", "age of student". Meanwhile, the "strong" influence variable with value less than 0.1 consists of 9 variables: "excellent in acting", "often joining writing, reading or storytelling", "playing musical instrument and singing", "often joining coloring, drawing or painting competition", "drawing, painting or coloring", "cooking activity", "excellent in drawing, painting and coloring", "able in singing while playing musical instrument", "often joining singing and acting competition".

Acknowledgements

References

1. H.-C. Huang, H.-H. Hsieh, C.-M. Chang, Y.-C. Lu and W.-C. Mui, "An Exploration of Dance Learning Stress Sources of Elementary School Dance Class Students with Artistic Abilities: The Influences of Psychological Capital and Self Concepts," Int. J. Environ. Res. Public Health, pp. 1-13, 2022, doi: 10.3390/ijerph19127398.
2. B. Y, A. Losh, N. Hamsho, A. Eisenhower and J. Blacher, "General Education Teachers' Perceptions of Autism, Inclusive Practices, and Relationship Building Strategies," J. Autism Dev Disord, vol. 52, no. 9, pp. 3977-3990, 2022, doi: 10.1007/s10803-021-05266-4.
3. I. G. T. Isa, Zulkarnaini. L. Novianti, F. Belladonna and S. Agustri, "Exploratory Data Analysis (EDA) dalam Dataset Penerimaan Mahasiswa Baru Universitas XYZ Palembang," Smart Comp, vol. 12, no. 3, pp. 600-609, 2023, doi:10.30591/smartcomp.v12i3.4125.
4. I. G. T. Isa and F. Elfaladonna, "Penilaian Kinerja Akurasi Metode Klasifikasi dalam Dataset Penerimaan Mahasiswa Baru," JEPIN (Jurnal Edukasi dan Penelitian Informatika), vol. 8, no. 2, pp. 292-298, 2022, doi: 10.26418/jp.v8i2.54316J.
5. V. Da Polan, T. Bethany, L. Clough, B. McKinney, J. Major, J. Chen and S. Hörst, "Exploratory Data Analysis (EDA) Machine Learning Approaches for Ocean World Analog Mass Spectrometry," Frontiers in Astronomy and Space Sciences, pp. 01-17, 2023, doi: 10.3389/fspas.2023.1134141.
6. C. Nicodemo and A. Satorra, "Exploratory Data Analysis on Large Data Sets: The example of salary Variation in Spanish Social Security Data," Business Research Quarterly, vol. 25, no. 3, p. 283–294, 2022, doi: 10.1177/2340944420957335.
7. C. Ho Yu, "Exploratory Data Analysis in the Context of Data Mining and Resampling," International Journal of Psychological Research, vol. 3, no. 1, pp. 9-22, 2010, doi: 10.21500/20112084.819.
8. C. K. Leung, G. Pasi and L. Wang, "Theoretical and practical data science and analytics: challenges and solutions," International Journal of Data Science and Analytics , vol. 16, p. 403–406, 2023, doi: 10.1007/s41060-023-00465-x.
9. Damianus DatonWitak, "Analisis Data Science Pada Struktur Data Kepadatan Penduduk Kota Malang," KURAWAL Jurnal Teknologi, Informasi dan Industri , vol. 3, no. 2, pp. 191-201, 2020, doi: 10.33479/kurawal.v3i2.353.

10. S. B. Rafael Stevi Oktavian, "Analisis Dataset Google Playstore Menggunakan Metode Exploratory Data Analysis," Jurnal Strategy, Vol. 2, No. 2, Pp. 636-649, 2020, doi: 10.13140/RG.2.2.14192.12808.
11. 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.