

Exploratory Data Analysis on the Process of Determining the Relationship between Student Interest and Talent Variables

Febie Elfaladonna¹, Devi Sartika², Andre Mariza Putra³

^{1,2} Department of Informatics Management, Sriwijaya State Polytechnic
Jl. Srijaya Negara, Bukit Lama, Kec. Ilir Barat. I, Kota Palembang, Sumatera Selatan 30128
Email: febie_elfaladonna_mi@polsri.ac.id

ABSTRACT

One of the determinants of student success in the future is the maximum management of interests and talents. To support the classification of interests and talents in schools, it can be done by using exploratory data analysis techniques. Exploratory data analysis (EDA) is one of the important stages in the data science cycle. The EDA process carried out on elementary school student datasets comes from student interest and talent questionnaires filled out by parents. EDA processing in this research uses several python programming libraries. The purpose of this research is to find relationships or interrelationships between items on the dependent variable and the independent variable which will later be carried out in model building. The main methods used in the implementation of EDA are chi square and T-test. The results of this study show the final variable analysis to determine interest and talent, so that it can be used as a basis for developing interest and talent assessment applications.

Keywords: *exploratory data analyses, interests, talents, primary school, variables.*

Introduction

The development of students in primary school is inseparable from the selection of appropriate interests and talents. With the right exploration of interests and talents, it can create a strategy for student self-development at school. It also aims to prepare students in determining career choices and matters related to their future. Interest and talent are two different things. Interest is a sense of preference or interest in something that exists without attraction or coercion [1]. Interest is very influential on the learning process because with interest it can increase student enthusiasm in the learning process. Meanwhile, talent is something that has been seen in human behaviour towards a certain skill [2]. With talent, the potential that exists in a person can be developed for the better. Talent is more constant because it is given by God. If utilised and developed properly, interests and talents can be the path to one's success.

Identifying children's interests and talents is a vital aspect. This is due to the need for each child to have an educational program that suits their interests and talents, so that they can maximize the development and utilization of these interests and talents. Some previous research on interest and talent analysis conducted for early childhood is about the description of planning, implementation, and evaluation of the development of students' interests and talents at Kak Seto Solo homeschooling. The research revealed that there are obstacles in carrying out the development of interests and talents caused by the lack of openness of parents in providing information. Most parents tend to force children to follow a certain profession without considering the child's talent or interest. If parents routinely force children to do activities that are not in accordance with their interests, this can cause children's discomfort in carrying out these activities [3], further research was conducted to provide knowledge about homeschooling involvement that can provide children with greater opportunities to explore and develop their interests and talents. The time flexibility of homeschooling, compared to formal schooling, gives children more time to focus on self-development. Some children decide not to continue school because they experience culture shock or are often bullied by classmates, so they refuse to attend school for a year. In such situations, parents in response provide the best alternative by choosing homeschooling. It is proven that homeschooling not only protects children from bullying experiences but is also able to accelerate the development of their interests and talents [4]. Research conducted in elementary schools explains that gifted children are those who, by professional definition, have exceptional abilities that enable them to achieve highly. They consistently excel in one or more areas, including general intellectual, creative, artistic/kinetic, and psychosocial or leadership [5].

This research was conducted at XYZ State Elementary School in Palembang City involving 144 students. This school does not have facilities to analyse students' specialisation. Whereas the process of

determining interests and talents from an early age is done so that later students in elementary school can recognise their potential and hone it so that the potential can continue to be developed until adulthood. This certainly helps them to be able to plan and not be confused later when entering the next stage of education. This study aims to see the extent to which the implementation of specialisation in the school is successful. Usually, the counselling teacher at XYZ State Primary School divides the students' interests and talents according to their preferences. If the student likes to play football, then every time there is a football match, the student will be appointed to participate in the match. This is certainly not very effective because it could be that students participate in football matches because they are invited by their friends. Therefore, this research is intended to classify students interests and talents so that later teachers can direct students in school activities based on their interests and talents. In this research, the interest and talent classification process are carried out using exploratory data analysis techniques based on datasets obtained from schools. Interest and talent datasets are obtained from filling out questionnaires by parents of students with reference to several variables that determine interests and talents. The results of filling out the questionnaire will be processed using several python libraries and to see the extent of the relationship between one variable and another in determining the classification of interests and talents.

Research Methods

The research conducted was qualitative research with questionnaire data collection. The questionnaire was created using google form and disseminated to 144 parents of first grade to 3rd grade elementary school students. Parents filled in the google form with Likert scale answers. Likert scale is a very familiar psychometric scale used for filling out questionnaires and is most often used for survey-related research activities [6]. There are 5 types of Likert scale answers on the questionnaire which can be seen in table 1. The person who gives the scale to the answer choices is the counseling guidance teacher. There are five answer choices with confidence levels ranging from 0 to 100. The purpose of giving this range is so that parents can ascertain whether the statements in the questionnaire are in accordance with the interests or talents of the child.

Table 1. Likert Scale for Questionnaire Filling

Answer Scale	Information
0	Do not know
20	Likely
40	Most Likely
60	Almost Certain
80	Certain
100	Very Sure

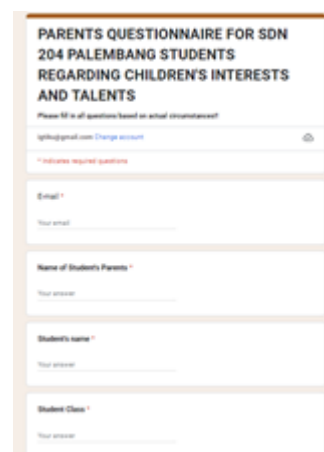


Figure 1. Parent questionnaire

Item No.	Statement	Don't know	Possible	Most likely	Almost sure	Certain
No. 1	Critical thinking	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
No. 2	Likes reading and calculating	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
No. 3	Likes Certain Types of Sports	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
No. 4	Likes playing musical instruments or singing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 2. Questionnaire with Likert Scale

After getting the results from the parent questionnaire, the next step is to process the respondent's data into tabulated data. Data tabulation can use a variety of equipment and usually starts with recording either manually in a book or by using electronic media such as computers [7]. Ms. Excel or better known as Microsoft Office Excel, is a very famous spreadsheet program and is ranked at the top of its category. Microsoft Excel offers a wide range of features, including complex and flexible formula calculation capabilities, powerful database management capabilities, and graphics processing expertise [8]. Excel is one of the tools most often used for the data tabulation process with various equipment, ranging from manual recording in ledgers to the use of computers as an auxiliary medium for data storage and processing. Data tabulation will facilitate analysis activities in research. There are two types of variables in the data processing process, namely dependent variables and independent variables. Included in the dependent variable are interest labels and talent labels, while the independent variable itself consists of fifteen questionnaire items. The dependent variable is the dependent variable that is influenced by the independent variable while the independent variable is the variable that has a strong influence on the dependent variable [9]. Later this tabulation data processing is used to select which items from the independent variable have the highest relationship to the dependent variable. Analysis of the relationship between variables is processed with the python language with the jupyter notebook tool.

Python was first developed in the late 1980s by Guido van Rossum from the Netherlands. Implementation of this programming language began in late 1989. In 2000, Python 2.0 was released with major new features. Since much of the code could not be easily ported to Python 3, the original end date of Python 2 was extended to 2020. In 2022, the release of Python 3.10.4 and 3.9.12 was accelerated due to numerous security issues. Today, Python has become one of the most widely learned and used programming languages [10]. The use of the Python programming language can increase convenience and efficiency in the data analysis process because Python is equipped with various libraries and functions that facilitate data management for visualization purposes. In addition, Python also offers a variety of visualization options that can be accessed through the library [11]. The Jupyter Notebook app introduced in 2015, has become a top choice for data scientists working with big data analysis. Jupyter Notebook stands for Julia (Ju), Python (Py), and R. It is a freely accessible web-based platform for creating and sharing code, computation results, analysis visualizations, and text in the creation of computational narratives [12]. Jupyter Notebook has been used to introduce open science to present and analyze data more easily. With its various features, this application can be used for collaboration with other application programs, such as Google Drive, to draw accurate conclusions [13]

Results and Discussion

All items in the interest and talent dataset are carried out exploratory data analysis with chi square and T-Test. The t-test is a commonly used statistical method to test whether there is a significant difference between the means of variables. The t-test research results show whether there is a significant difference between the variables being compared [14]. In general, Chi Square is part of testing with statistical properties with the aim of determining the relationship or relationship between several variables [15]. Chi square is denoted by X^2 . T-Test is also part of statistical testing that supports testing the unrelatedness or difference between several variables to get a conclusion whether the variable relationship is acceptable or not. There are 15 items in the independent variable, namely: 'Student's Class', 'Student's Age', 'Critical Thinking', 'Likes to Trade and Calculate', 'Likes Certain Types of Sports', 'Likes to Play Musical Instruments or Sing', 'Likes Cooking Activities', 'Likes to Write, Read or Tell Stories', 'Likes to Draw, Paint or Color', 'Often participates

in coloring, drawing or painting competitions', 'Likes to participate in storytelling, speech or poetry competitions', 'Likes to participate in singing or acting competitions', 'Good at singing, playing music', 'Good at drawing, painting or coloring', 'Good at acting activities'. And two labels on the dependent variable, namely: 'Interest', and 'Talent'. After determining the variables involved in the research, the next step is to apply python programming to find the relationship between one variable and another using the chi square and T-Test tests. The libraries used for data processing with python are panda's library, sciPy library, numPy Library, Sklearn Library. In carrying out the process on exploratory data analysis, the data used must have a .csv format. The process of calling the .csv dataset using the Pandas library with the alias "pd" through the "read_csv" function so that it displays the dataset represented as shown below.

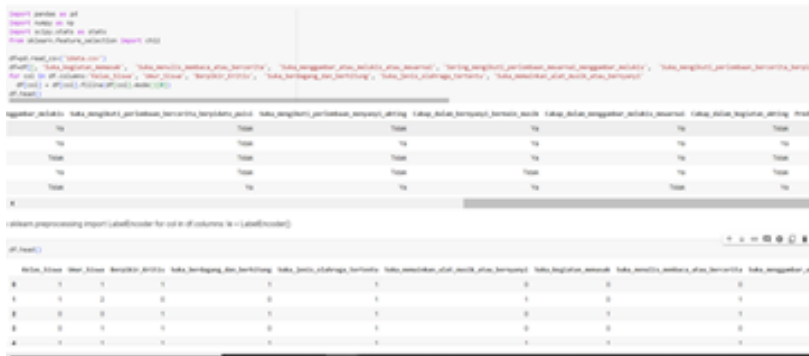


Figure 3. Python Library in Research

In the program, chi square is implemented using the sklearn. Feature_selection function where there are two variables, X as the independent variable and Y as the dependent variable. The following is a picture of the chi square implementation in python along with the results.

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

chi_score = chi2(x, y)

[ ] chi_score

(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]),
 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]))

chi_values = pd.Series(chi_score[0], index=x.columns)
chi_values.sort_values(ascending=False, inplace=True)
```

Figure 4. Chi Square Implementation in Python

The results of the first trial Chi Square test obtained a graph with 15 items of the independent variable as follows.

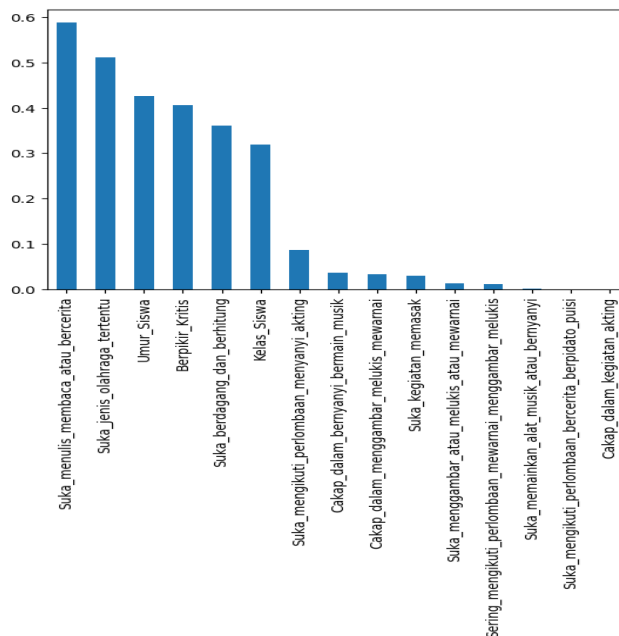


Figure 5. First Chi Square Test Graph

From the picture above, the table can be described as follows.

Table 2. First Chi Square Test Results

excellent in acting	17.79284731
often joining writing, reading or story telling	15.46748504
playing musical instrument and singing	9.09129968
often joining colouring, drawing, or painting competition	6.24648434
drawing, painting, or colouring	6.17494196
cooking activity	4.68226551
excellent in drawing, painting, and colouring	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 story telling	0.29389977

Furthermore, chi square testing was carried out again for the second experiment by setting a threshold value of 0.5. This is done so that later an item attachment is formed that is suitable to be used as a reference in the classification of Interests and Talents. To ensure how the relationship between the independent variable and the dependent variable has a high influence value, a T-Test is conducted.

```

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

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

chi_score = chi2(x, y)

chi_score

(array([ 0.69202329,  0.83476502,  9.09129968,  4.68226551,  6.17494196,
         6.24648434, 15.46748504,  2.93558343,  4.31730334,  4.55054406,
        17.79284731]),
 array([4.05477036e-01, 3.60890283e-01, 2.56828298e-03, 3.04755318e-02,
        1.29572161e-02, 1.24440052e-02, 8.39368115e-05, 8.66473532e-02,
        3.77266641e-02, 3.29082881e-02, 2.46307450e-05]))

chi_values = pd.Series(chi_score[0], index=x.columns)
chi_values.sort_values(ascending=False, inplace=True)
chi_values.plot.bar()

-----
NameError                                Traceback (most recent call last)
<ipython-input-3-2ab24a7d9b64> in <cell line: 1>()
----> 1 chi_values = pd.Series(chi_score[0], index=x.columns)
      2 chi_values.sort_values(ascending=False, inplace=True)
      3 chi_values.plot.bar()
    
```

Figure 6. Second Experiment of Chi Square Test

The results of the T-Test can be seen in the figure below, where in the T-test the accepted value is if the results show less than 0.5. Of the 15 test items, only 11 items are used as reference items for interest and talent classification.

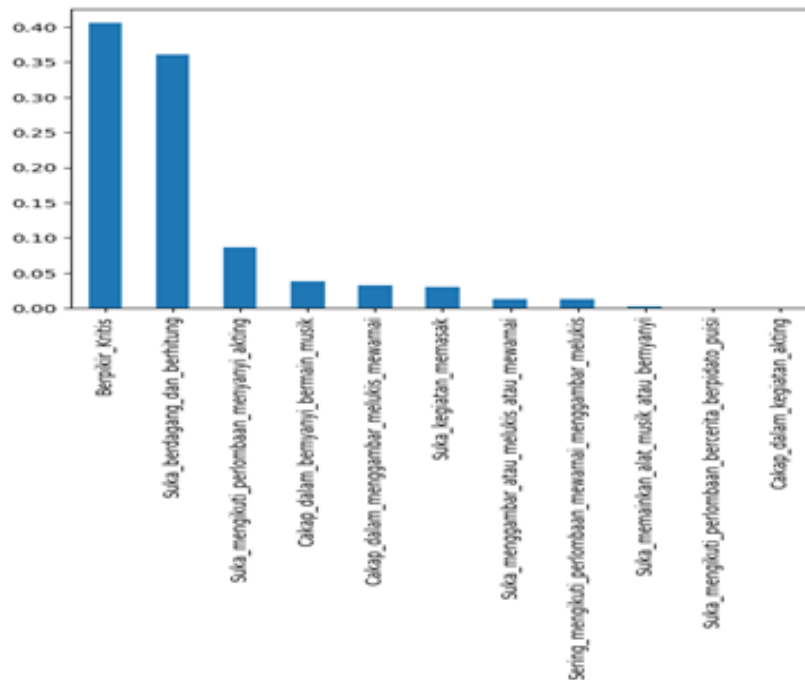


Figure 7. Final Chi Square and T-Test Results

Conclusion

Exploratory Data Analysis (EDA) was conducted to determine the variable "interest" or "talent" of XYZ Palembang primary school students. Tests were conducted with chi square and T-test on 15 independent variable items with one dependent variable with two labels of interest and talent. From the chi square test results, the highest value is the variable "good at acting" 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 sports activities". The "medium" influence category with a range of 0.3~0.4, namely the variables "student class", "trading and calculating", "critical thinking", "student age". Meanwhile, the "strong" influence variables with a value of less than 0.1 consisted of 9 variables: "good at acting", "often participate in writing, reading or storytelling activities", "playing musical instruments and singing", "often participate in coloring, drawing or painting competitions", "drawing, painting or coloring", "cooking activities", "good at drawing, painting and coloring", "good at singing while playing musical instruments", "often participate in singing and acting competitions". As a future research plan, the 11 recommended variables will be implemented in modeling to predict how "interest" or "talent", especially in elementary school students. The variables that have been determined by testing can be used as a reference to perform calculations with machine learning algorithms before making an application for determining interests and talents in future research.

References

- [1] W. Warsito, "Peningkatan Minat Belajar Matematika Kelas Iv Melalui Alat Peraga Layang-layang," *Jurnal Sinetik*, vol. 2, no. 2, pp. 242-248, 2019.
- [2] Yusfandaria, "Upaya Mengembangkan Kemampuan Bakat Melalui Layanan Bimbingan Karir Dengan Strategi Problem Solving Peserta Didik Kelas X IPS.2 SMA Negeri 18 Palembang," *Juang : Jurnal Wahana Konseling*, vol. 2, no. 1, pp. 60-69, 2019.
- [3] M. N. Mahfud and Utama, "Pengelolaan pengembangan minat dan bakat anak didik di," *Jurnal Akuntabilitas Manajemen Pendidikan*, vol. 9, no. 2, pp. 113-124, 2021.
- [4] R. M. Siregar, "Homeschooling: Alternatif Pendidikan yang Menjanjikan untuk Pengembangan Minat Bakat Anak Sejak Dini," 20 Agustus 2019.
- [5] C. Aciakatura, I. Magdalena and A. Zahra, "Analisis Pengembangan Minat Dan Bakat Siswa Pada Siswa Sekolah Dasar," *Cerdika: Jurnal Ilmiah Indonesia*, vol. 1, no. 2, pp. 89-94, 2020.
- [6] D. Taluke, Ricky and Amanda, "Analisis Preferensi Masyarakat Dalam Pengelolaan Ekosistem Mangrove Di Pesisir Pantai Kecamatan Loloda Kabupaten Halmahera Barat," *Jurnal Spasial*, vol. 6, no. 2, pp. 531-540, 2019.
- [7] E. Rahayu, H. Tantri and R. Dewi, "Sosialisasi Pengolahan Tabulasi Data Administrasi Perkantoran Menggunakan Aplikasi Microsoft Excel Pada Perangkat Desa Sei Mencirim," *Wahana Inovasi*, vol. 10, no. 1, pp. 111-116, 2021.
- [8] I. W. Nuarsa, *Jalan Pintas Menguasai Microsoft Excel XP*, Yogyakarta: Andi, 2003.
- [9] 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.
- [10] R. Putri Kurnia and Y. Adi Atma, "Analisis Rekomendasi Film Dari Data Imdb Dengan Python," *Devic : Journal Of Information System, Computer Science And Information Technology*, vol. 3, no. 2, pp. 23-28, 2022.
- [11] L. Regina and dkk, "Penggunaan Bahasa Pemrograman Python dalam Menganalisis Hubungan Kualitas Kopi dengan Lokasi Pertanian Kopi," *Jurnal Publikasi Teknik Informatika (JUPTI)*, vol. 2, no. 2, pp. 100-109, 2023.
- [12] P. Panyahuti and Y. Yadi, "Pengembangan Aplikasi E-Assessment Skill Programming berbasis Web," *Edumatic: Jurnal Pendidikan Informatika*, vol. 6, no. 1, pp. 78-87, 2022.
- [13] R. R. Putri Asyrofi and R. Asyrofi, "Implementasi Aplikasi Jupyter Notebook Sebagai Analisis Kriteria Plagiasi dengan Teknik Semantik," *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 2, pp. 627-637, 2023.
- [14] I. C. N. d. A. Prabowo, "Penggunaan Ujia Chi Square Untuk Mengetahui Pengaruh Tingkat Pendidikan dan Umur Terhadap Pengetahuan Penasun Mengenai HIV-AIDS di Provinsi Jakarta," in *Prosiding Seminar Nasional Matematika dan Terapannya 2018*, Jakarta, 2019.
- [15] D. Wahyudi, J. Idris and Z. Abidin, "Tren Dan Isu Penelitian Uji-T Dan Chi Kuadrat Dalam Bidang Pendidikan," *Journal of Mathematics Education*, vol. 4, no. 2, pp. 182-196, 2023.