Distribution Analysis of Student Numbers by Gender Using Decision Tree and Data Visualization

Risca Sri Mentari, Sri Wahyuni
[1,2]Master of Information Technology, Universitas Pembangunan Panca Budi, Indonesia

ABSTRACT

Rapid technological developments have brought significant changes in various sectors, including education. In the context of education, data management and analysis are important elements in supporting data-driven decision-making. Data mining, specifically the Decision Tree method, provides valuable insights into analyzing patterns from large data sets. This study uses Decision Tree modeling and data visualization through RapidMiner to analyze the distribution of the number of students based on gender in various classes at SMK Negeri 1 Stabat in the 2023-2024 school year. This research includes data collection, preprocessing, and decision tree modeling to uncover gender-based trends in various skill programs. Visualization using Scatter Plot makes it easier to present data for clearer analysis. The results of the study show that administrative and fashion skills programs are dominated by women, while engineering skills programs, such as TKR and TITL, are dominated by men. Some classes showed a more balanced gender composition. This research provides useful insights for classroom management and decision-making in the educational environment, as well as provides a basis for designing more inclusive learning programs and addressing gender imbalances in certain areas.

Keyword : Data mining; C4.5; Decision Tree.

Corresponding Author:
Risca Sri Mentari,
Master of Information Technology
Universitas Pembangunan Panca Budi
Jl. Jend. Gatot Subroto Km. 4,5 Sei Sikambing 20122, Medan, Indonesia.
Email : risca.mentari18@gmail.com

Article history:
Received Nov 28, 2024
Revised Nov 29, 2024
Accepted Nov 30, 2024

1. INTRODUCTION

Technological developments have brought significant changes to various fields, such as communication, transportation, health, entertainment, and education (Fricticarani et al., 2023). In the context of education, the management and analysis of student data is an important aspect to support data-driven decision-making. Data mining, as a method to find specific patterns in large data sets, provides solutions in automated analysis and retrieval of relevant information (Zafira et al., 2024) (Almufqi & Voutama, 2023). One of the tools used for data mining is RapidMiner, when using RapidMiner, no special programming skills are required, as all models are available. RapidMiner is used for data mining. (Vidiya & Testiana, 2023).

RapidMiner supports data mining, text mining, and predictive analysis, making it a reliable tool for complex data analysis (Lestari & Mulyawan, 2023). One of the data mining methods that is often used is Decision Tree. This method, as part of machine learning, generates a classification in the form of a decision tree based on the tested dataset (Supriyadi, 2023). In the process, Decision Tree leverages statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify information from large databases (Yani et al., 2023).

In addition, data visualization is a key element in accelerating the understanding of patterns resulting from the analysis process. Effective visualization can accelerate informed decision-making (Irmayani, 2021). One relevant visualization method is Scatter Plot, which allows for a better understanding of cluster patterns from the analyzed dataset (Marcelina et al., 2023).

Based on this background, this study aims to analyze the distribution of the number of students based on gender in each class by utilizing the Decision Tree method and data visualization using RapidMiner. This research is expected to provide a comprehensive overview of the distribution of students as well as emerging patterns, which can be used as a basis for more informative decision-making in the educational environment.

2. RESEARCH METHOD

A. Data Collection

At this stage, the researcher collected data from students in grade XII at SMK Negeri 1 Stabat for the 2023-2024 academic year. Data collection is defined as a process or activity carried out by researchers to uncover or capture various phenomena, information or conditions of research locations in accordance with the scope of the research. (Azizah et al., 2023)

B. Preprocessing

Once the data is collected, the next step is preprocessing. Preprocessing is the initial stage of preparing documents or raw data to be ready for processing. (Wahyuningtyas et al., 2023)

C. Modeling and Analysis

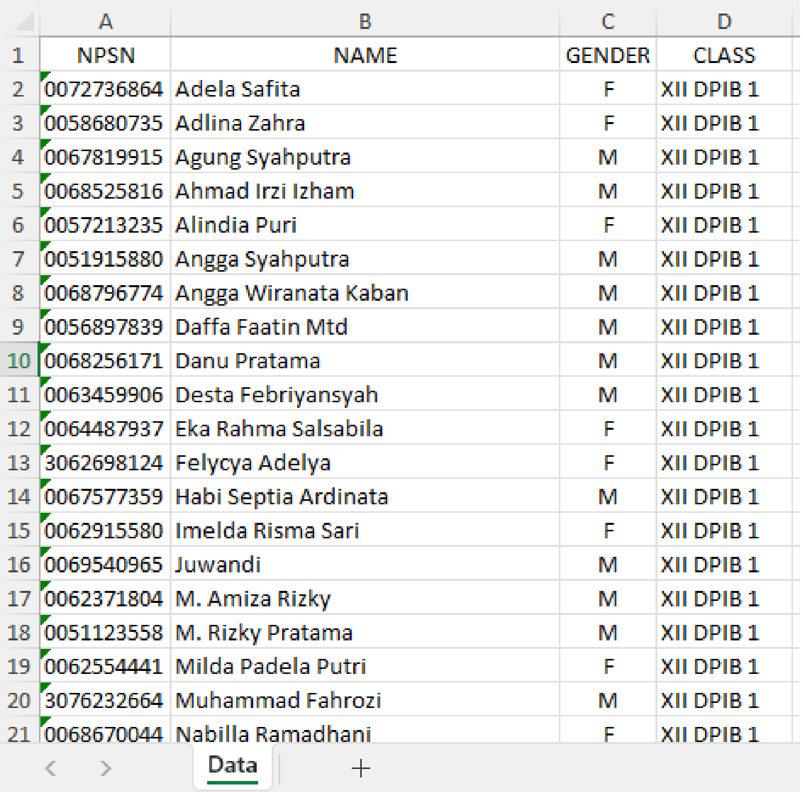
The researcher proceeded to construct a decision tree using RapidMiner. Following the modeling process, an analysis was carried out. This stage of modeling represents a critical step in achieving effective data mining (Leni et al., 2023). Analysis involves examining an object or subject in detail by breaking it down into its constituent components for further exploration and understanding (Alfayed et al., 2023).

C. Visualization and Analysis

After modeling the data using a decision tree, the next step is to visualize it using a Scatter Plot to make the analysis easier.

3. RESULTS AND DISCUSSION

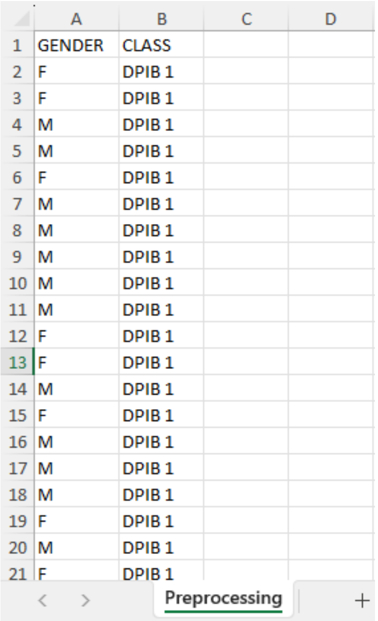
A. Data Collection



	A	B	C	D
1	NPSN	NAME	GENDER	CLASS
2	0072736864	Adela Safita	F	XII DPIB 1
3	0058680735	Adlina Zahra	F	XII DPIB 1
4	0067819915	Agung Syahputra	M	XII DPIB 1
5	0068525816	Ahmad Irzi Izham	M	XII DPIB 1
6	0057213235	Alindia Puri	F	XII DPIB 1
7	0051915880	Angga Syahputra	M	XII DPIB 1
8	0068796774	Angga Wiranata Kaban	M	XII DPIB 1
9	0056897839	Daffa Faatin Mtd	M	XII DPIB 1
10	0068256171	Danu Pratama	M	XII DPIB 1
11	0063459906	Desta Febriyansyah	M	XII DPIB 1
12	0064487937	Eka Rahma Salsabila	F	XII DPIB 1
13	3062698124	Felycya Adelya	F	XII DPIB 1
14	0067577359	Habi Septia Ardinata	M	XII DPIB 1
15	0062915580	Imelda Risma Sari	F	XII DPIB 1
16	0069540965	Juwandi	M	XII DPIB 1
17	0062371804	M. Amiza Rizky	M	XII DPIB 1
18	0051123558	M. Rizky Pratama	M	XII DPIB 1
19	0062554441	Milda Padela Putri	F	XII DPIB 1
20	3076232664	Muhammad Fahrozi	M	XII DPIB 1
21	0068670044	Nabilla Ramadhani	F	XII DPIB 1

Fig 1. Grade XII Student Data

Figure 1 presents data on grade XII students, including several key attributes: NPSN as the school's unique identifier, student names representing individual identities, gender indicating gender information, and class designations reflecting the grouping of students based on their learning levels.



	A	B	C	D
1	GENDER	CLASS		
2	F	DPIB 1		
3	F	DPIB 1		
4	M	DPIB 1		
5	M	DPIB 1		
6	F	DPIB 1		
7	M	DPIB 1		
8	M	DPIB 1		
9	M	DPIB 1		
10	M	DPIB 1		
11	M	DPIB 1		
12	F	DPIB 1		
13	F	DPIB 1		
14	M	DPIB 1		
15	F	DPIB 1		
16	M	DPIB 1		
17	M	DPIB 1		
18	M	DPIB 1		
19	F	DPIB 1		
20	M	DPIB 1		
21	F	DPIB 1		

Fig 2. Grade XII Student Data after Preprocessing

Figure 2 displays the data of grade XII students after undergoing processing, simplified into two primary attributes: gender, identifying the student's gender, and class, categorizing their learning level.

B. Modeling and Analysis

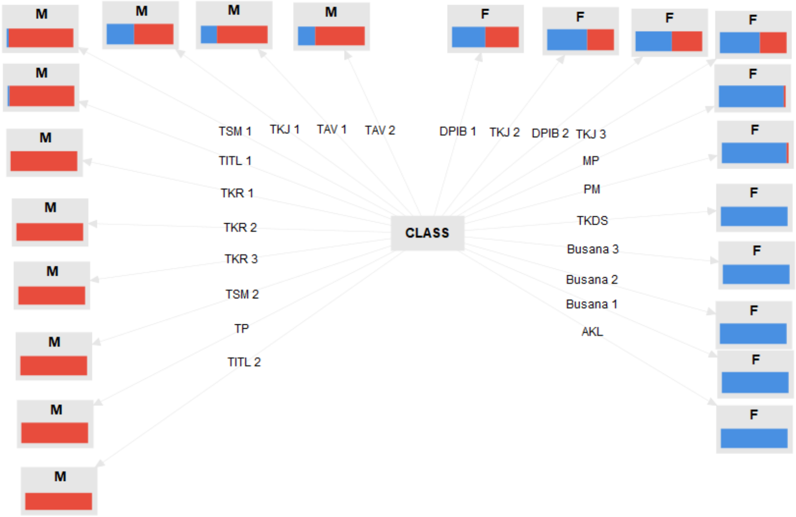
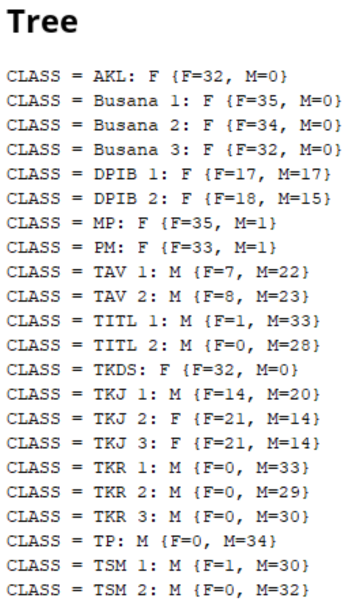


Fig 3. Decision Tree Model

Figure 3 can be explained that the male dominant population class is in classes TITL2, TP, TSM 2, TKR 3, TKR 2, TKR 1, TITL, TSM 1, TKJ 1, TAV 1, TAV 2, while the female dominant population class is in classes AKL, Busana 1, Busana 2, Busana 3, TKDS, PM, MP, TKJ 3, DPIB 2, TKJ 2, DPIB 1.



Tree

```
CLASS = AKL: F {F=32, M=0}
CLASS = Busana 1: F {F=35, M=0}
CLASS = Busana 2: F {F=34, M=0}
CLASS = Busana 3: F {F=32, M=0}
CLASS = DPIB 1: F {F=17, M=17}
CLASS = DPIB 2: F {F=18, M=15}
CLASS = MP: F {F=35, M=1}
CLASS = PM: F {F=33, M=1}
CLASS = TAV 1: M {F=7, M=22}
CLASS = TAV 2: M {F=8, M=23}
CLASS = TITL 1: M {F=1, M=33}
CLASS = TITL 2: M {F=0, M=28}
CLASS = TKDS: F {F=32, M=0}
CLASS = TKJ 1: M {F=14, M=20}
CLASS = TKJ 2: F {F=21, M=14}
CLASS = TKJ 3: F {F=21, M=14}
CLASS = TKR 1: M {F=0, M=33}
CLASS = TKR 2: M {F=0, M=29}
CLASS = TKR 3: M {F=0, M=30}
CLASS = TP: M {F=0, M=34}
CLASS = TSM 1: M {F=1, M=30}
CLASS = TSM 2: M {F=0, M=32}
```

Fig 4. Decision Tree Description

Figure 4, the data shows the distribution of the number of female (F) and male (M) students in different classes based on the skill program. The classes that were completely dominated by women were AKL (32 female students), Busana 1 (35), Busana 2 (34), Busana 3 (32), MP (35), PM (33), and TKDS (32), with no male students at all. On the other hand, the classes with full male dominance are TKR 1 (33 male students), TKR 2 (29), TKR 3 (30), TP (34), TITL 2 (28), and TSM 2 (32), with no female students. Several engineering classes such as TITL 1 (33 males, 1 female), TSM 1 (30 males, 1 female), as well as TAV 1 (22 males, 7 females) and TAV 2 (23 males, 8 females) show a strong male dominance. Classes with a more balanced composition were found in DPIB 1 (17 women, 17 men) and DPIB 2 (18 women, 15 men). Meanwhile, in the TKJ class, there was a variation where TKJ 1 was dominated by men (14 women, 20 men), while TKJ 2 and TKJ 3 had more women with 21 women and 14 men, respectively. Overall, the highest number of female students was recorded in the Busana 1 class (35 students), while the highest number of male students was recorded in the TP class (34 students). This analysis shows that there is a gender bias towards skills programs, where the administrative and fashion fields are dominated by women, while the engineering field is dominated by men, with a strikingly significant gender gap.

D. Visualization and Analysis

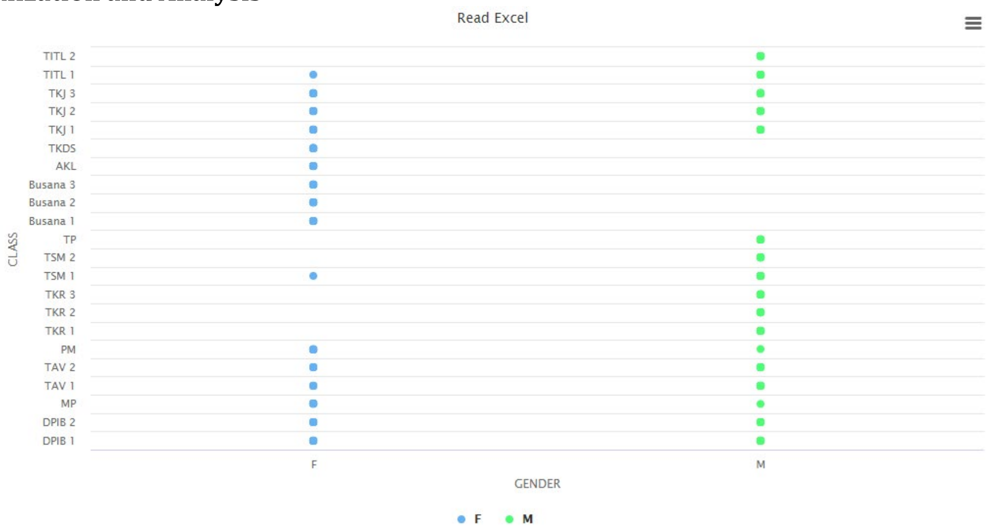


Fig. 5. Data Visualization - Scatter Plot

In Figure 5, the data visualization using Scatter Plot shows that in the classes (TITL 2, TP, TSM 2, TKR 3, TKR 2, TKR 1) there are no male students, while in the classes (TKDS, AKL, Busana 3, Busana 2, Busana 1) there are no female students. Meanwhile, in other classes there are male and female students. With this visualization, analysis can be done more easily.

4. CONCLUSION

Analysis of the distribution of the number of students based on gender using the Decision Tree method succeeded in identifying a clear distribution pattern between male and female students in various classes at SMK Negeri 1 Stabat. The results of the study showed that classes with expertise programs in the field of administration and fashion were dominated by female students, while classes in the field of engineering, such as TKR, TITL, and TSM, were dominated by male students. Some classes, such as TKJ 2 and TKJ 3, show a larger composition of female students than boys. Data visualization using Scatter Plot also facilitates the understanding of this gender distribution, showing that there is a significant gender gap in several skill programs. This research provides useful information for classroom management and decision-making in the educational environment, especially in designing more inclusive and proportional learning programs based on gender distribution.

REFERENCES

Alfayed, E., Ramadeli, L., Agnestasia, R., Amalina, V., Swid, Z. H. O., & Riofita, H. (2023). ANALISIS STRATEGI PEMASARAN DAN PENJUALAN E-COMMERCE PADA TIKTOK SHOP. *Jurnal Ekonomi Manajemen Dan Bisnis*, *1*(2), 195–201.

Almufqi, F. M., & Voutama, A. (2023). PERBANDINGAN METODE DATA MINING UNTUK MEMPREDIKSI PRESTASI AKADEMIK SISWA. *Jurnal Teknika (Jurnal Fakultas Teknik Universitas Islam Lamongan)*, *15*(1), 61–66. https://doi.org/10.30736/jt.v15i1.929

Azizah, A., Rani, Ulum, K., Roni, F., & Reptiningsih, E. (2023). Analisis Penerapan Metode Simpleks Linier Programming Pada Home Industry Martabak. *Journal of Trends Economics and Accounting Research*, *4*(2), 388–395. https://doi.org/10.47065/jtear.v4i2.1059

Fricticarani, A., Hayati, A., Ramdani, Hoirunisa, I., & Rosdalina, G. M. (2023). STRATEGI PENDIDIKAN UNTUK SUKSES DI ERA TEKNOLOGI 5.0. *JURNAL INOVASI PENDIDIKAN DAN TEKNOLOGI INFORMASI*, *4*(1), 56–68.

Irmayani, W. (2021). VISUALISASI DATA PADA DATA MINING MENGGUNAKAN METODE KLASIFIKASI NAÏVE BAYES. *JURNAL KHATULISTIWA INFORMATIKA*, *XI*(1). www.bsi.ac.id

Leni, D., yermadona, H., Usra Berli, A., Sumiati, R., & Haris. (2023). Pemodelan Machine Learning Untuk Memprediksi Tensile Strength Aluminium Menggunakan Algoritma Artificial Neural Network (ANN). *SURYA TEKNIKA*, *10*(1), 625–632.

Lestari, P. D., & Mulyawan. (2023). DATA MINING PADA PENJUALAN AIR BERSIH DI SPAM AKIDAH MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING MENGGUNAKAN RAPIDMINER. *Jurnal Mahasiswa Teknik Informatika*, *7*(1), 412–416.

Marcelina, D., Kurnia, A., & Terttiaavini, T. (2023). Analisis Klaster Kinerja Usaha Kecil dan Menengah Menggunakan Algoritma K-Means Clustering. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(2), 293–301. https://doi.org/10.57152/malcom.v3i2.952

Supriyadi, A. (2023). Perbandingan Algoritma Naive Bayes dan Decision Tree(C4.5) dalam Klasifikasi Dosen Berprestasi. *Generation Journal*, *7*(1), 2580–4952.

Vidiya, E. C., & Testiana, G. (2023). Analisis Pola Pembelian di Lathansa Cafe & Ramen dengan Menggunakan Algoritma FP-Growth Berbantuan RapidMiner. *G-Tech: Jurnal Teknologi Terapan*, *7*(3), 1118–1126. https://doi.org/10.33379/gtech.v7i3.2739

Wahyuningtyas, N. P., Ratnawati, D. E., & Setiawan, N. Y. (2023). Root Cause Analysis (RCA) berbasis Sentimen menggunakan Metode K- Nearest Neighbor (K-NN) (Studi Kasus: Pengunjung Kolam Renang Brawijaya). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *7*(5), 2515–2520. http://j-ptiik.ub.ac.id

Yani, A., Azmi, Z., & Suherdi, D. (2023). Implementasi Data Mining Menganalisa Data Penjualan Menggunakan Algoritma K-Means Clustering. *JURNAL SISTEM INFORMASI TGD*, *2*(2), 315–323. https://ojs.trigunadharma.ac.id/index.php/jsi

Zafira, F., Irawan, B., & Bahtiar, A. (2024). PENERAPAN DATA MINING UNTUK ESTIMASI STOK BARANG DENGAN METODE K-MEANS CLUSTERING. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(1), 156–161.