Predicting First-Year Engineering Student Dropout Using Data Mining: A Case Study at Salahaddin University-Erbil

Nian Khidr Aziz[1,*]

[1,*]College of Engineering, Salahaddin University-Erbil,Kurdistan Region, Iraq

[*]Correspondent Author, email: nian.aziz@su.edu.krd

https://doi.org/10.31972/iceit2024.010

Abstract

High dropout rates among newly registered students in engineering programs significantly impact the academic plans and long-term strategies of many schools and universities. Student dropouts and low enrollment rates in engineering programs are major concerns for higher education institutions and governmental decision-makers in both developed and developing countries, however the reasons for these issues vary. This study utilizes a data mining approach to predict the features influencing dropout decisions. The method was applied to pre-enrollment and first-semester official data from the engineering college at Salahaddin University-Erbil. A dataset was built, preprocessed, modeled, and analyzed. The research yielded valuable results and suggested redesigning the application forms for engineering programs to further improve prediction rates and plan for student admittance who will complete the program successfully.

Keywords: Higher education Performance, Student Dropout, Data Mining, Dataset

1. Introduction

Students drop out and low-rate enrollment in engineering programs is a real concern of higher education institutions and governmental decision makers in many countries (NCES, 2024; Tocto et al., 2023) . As engineering is highly coupled with industry, energy, environment, quality of life and defense, an adequate number of new blood engineers are needed in both national and international level. To find the factors affect the low enrollment and high dropouts, research have been conducted and funded worldwide during the last two decades (Khan et al., 2023; Mouton et al., 2020; Won et al., 2023). Researchers found different reasons in different areas in the world for low enrollment and dropout in Engineering programs in special and in STEM programs in general (Aina et al., 2024; Desai

& Neuhold, 2017; Rajlakshmi Ghosh, 2023; Sharon A. Jones et al., 2021). In developing countries, the situation and the reasons are different. In Kurdistan Region of Iraq, the student enrollment for BSc. degree in engineering programs (Architectural Engineering programs are excluded) is lower than desired. Also, the number of students who postpone their study for more than one academic year and dropouts are within a risky level. However, Kurdistan Regional Government implemented some steps to encourage young people to enter engineering programs and to support engineering professions in general, for example the compulsory engineering allowance and hazard payments added by law to the wages of all engineers working in Kurdistan region, but these actions did not solve the problem completely. So detecting the reasons of these phenomena in an early stage is essential for keeping these students from dropping out. This research field is part of a broadly field of the Academic Performance Evaluation (Nian Kh. Aziz, 2016; Villanueva & Moreno, 2018).

In this study ten features from students profile at the college of engineering, Salahaddin university-Erbil, Kurdistan Region of Iraq have been selected and a data has been collected through students dossier and direct communication with students. The dataset is used to develop a model for the drop out phenomena. The model is evaluated and verified through. WEKA software (Weka 3, 2017) has been used as the data mining and analysis tool in this study.

This paper is organized as follows: The next section provides a review of the related works. Following this, the methodology and data collection processes are detailed. The dataset description section presents an insight to the dataset used in the study. In the subsequent section, modeling and analysis, the methods used to analyze the data and the results obtained are discussed. The paper concludes with a summary of findings, implications of the study, and suggestions for future work. Finally, references are listed in the references section.

2. Related Works

Academic performance evaluation is widely studied (Villanueva & Moreno, 2018), but using artificial intelligence (AI) and Data Mining (DM) techniques are somehow new (Rabelo and Zárate, 2024; Romero and Ventura, 2007) made a survey about the literature of data mining in educational systems. They concluded that real research to address educational issues using DM are mostly conducted in the nineties of the last century. One of the first studies was published in 1995 by Sanjev and Zytkow (1995). They studied retrievals of enrollment knowledge using university's database system. Rai, Jain, and Saini (2014) used decision tress algorithm to study the students drop out in electrical engineering

department at Eindhoven University of Technology. They developed a model for student drop out, they used ID3 decision tree algorithm in their study However their study focused mostly in the evaluation of students performance in the offered courses. Meena et. al. (2014) used k-means clustering algorithm and rule base classification to study attributes like attendance and assignment marks. They also gave a general consideration to the role of datamining methods in the educational systems. A more updated research to utilize data mining and machine learning approaches to educational systems using retrieved datasets from Moodle, which reflects student-instructor academic activities can be found in (Burgos et al., 2018; Dascalu et al., 2021; Heidrich et al., 2018)

Interviews with students have served as a key source of information and data in numerous studies aimed at uncovering the real reasons behind the high dropout rates. For instance, two studies by Casanova (Casanova et al., 2018, 2021) identified various factors influencing student retention in engineering programs. These factors include vocational training, starting a job, a perceived lack of fit with the program, and even the mother's educational level.

The Educational Data Initiative of America, in its 2023 report, revealed that between the fall semesters of 2021 and 2022, 28.9% of all full-time first-year students dropped out of college, with public schools experiencing the highest dropout rates. The report is abundant in statistics regarding the demographics, ethnicity, gender, and reasons for dropout (Hanson, 2024).

Since various reasons contribute to dropout rates in different societies, this study aims to examine the phenomenon in Iraqi Kurdistan, specifically focusing on dropout rates of first year students in engineering programs.

3. Methodology

Due to the wide availability of huge amount of data and the need for converting such data into knowledge, data mining has attracted a great deal of attention to researchers and practitioners in many domains including higher education datasets. In this research, a well-known machine learning methodology called Cross Industry Standard Process for Data Mining (CRISP-DM) has been used. The life cycle of this technique consists of six phases (Blokdyk, 2021):

i. Problem Description

The initial phase involves understanding project objectives with business perspective and transforming this information into data mining problem description.

ii. Data and Business Understanding

The data understanding phase begins with identifying the sources of data and obtaining an initial set of data to access the information scope of the data for the business problem.

iii. Data Preparation

Consists of pre-processing, cleaning, converting to a format suitable for the DM algorithm.

iv. Modeling

Developing different models using comparable analytical techniques, selecting the appropriate model, setting the parameters to the optimal values for the selected technique.

v. Evaluation

Consists of evaluating and accessing the validity of the developed model , mostly a portion of the data set is used.

vi. Deployment

Involves deploying the model in the process of decision making.

The CRISP_DM structure and lifecycle is shown in figure 1.

Several analytical methods are utilized in data mining (Moscoso-Zea et al., 2019). The main types include classification, clustering, association rules, and decision trees. In this study, we propose using the CRISP-DM framework alongside the classification approach to predict the relationships between factors that influence students' dropout decisions.

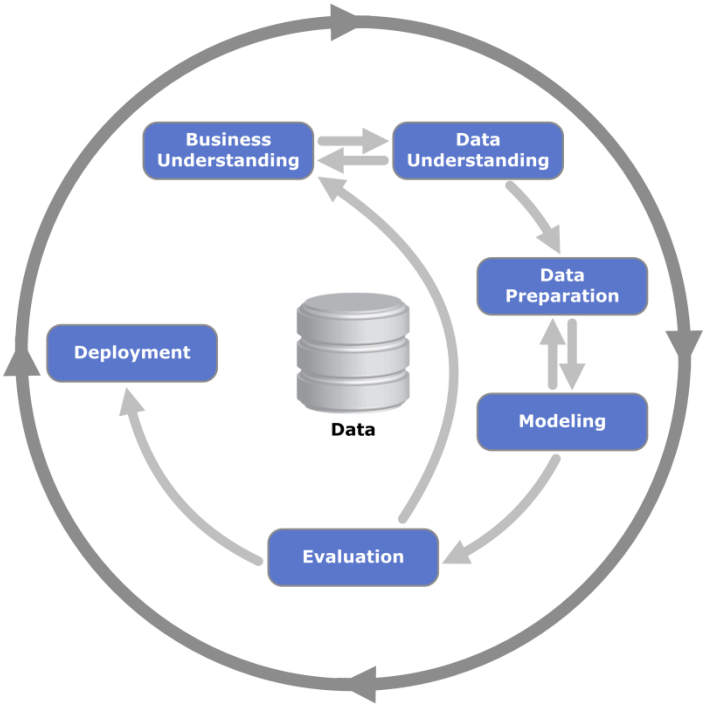


Fig.1 CRISP-DM Structure

4. Data Collection

The dataset used in this study was collected from two sources. The first source is the students' profiles, which are available in the registrar's office at the engineering college. Although these data are confidential, the college released part of it for this research, with the restriction that sensitive information such as students' names and national identity numbers do not appear in any documentation, in accordance with ethical regulations at Salahaddin University-Erbil. This part of the data includes the eco-social and academic background of the students.

The second source of data is a questionnaire designed specifically for this study, which was distributed to the students with the help of the registrar's office. The questionnaire included four questions about the reasons for dropout, categorized into the following domains:

- Financial issues (e.g., parents' monthly income, student works part-time, parents cannot afford support)
- Home-University distance (e.g., student lives in another city)
- Better alternative programs (e.g., studying in another college or university, private universities)
- Better alternative life conditions (e.g., relocation to another country, immigration)

- Lack of interest in the program (e.g., postponing, taking a gap year, waiting for the next intake to re-apply for a new program)

The features, along with their descriptions and data types, are listed in Table 1.

Table 1. Attributes and Their Domains

Attribute and Target	**Description**	**Domain**	**Data type (Values in the algorithm)**
Address	Only the center of five governorates is considered cities.	City, village	Numerical (1,2)
Family Monthly income	Combined income from all sources.	<500, 500<>1000, 1000<>1500, >1500	Numerical (1,2,3,4)
Parents Educational Degree	Education level of either parent.	No Degree , Diploma or BSc (BA), MSc.(MA) or PhD.	Numerical (1,2,3)
English Language Skills	Student's grade in high school English.	<70, 70<>85, >85	Numerical (1,2,3)
Ethnicity	Ethnicity of local students.	Kurd, Turkmen, Assyrian, Arab	Numerical (1,2,3)
Religion	Religion of local students.	Muslim, Christian, Yezidi, Others	Numerical (1,2,3,4)
Target:Reason for Dropout	Based on students' questionnaire responses.	Part time work, Parents cannot afford	Numerical (1,2)

		Better Alternative Program	Nominal, (Yes, No)
		Better living condition	Immigration, study abroad (1,2)
		Address (dependent variable to address attribute)	Address (1,2,3)
Classification	Indicates whether the student stays or drops out	Stay/dropout	Numerical (1,2)

This study focuses on first-year BSc. Engineering students at Salahaddin University-Erbil for the period 2018-2020, utilizing 635 records. Initially, the data were stored in an MS-Excel file and then converted to Attribute-Relation File Format (ARFF) for use in the WEKA data mining application. This dataset is publicly available at https://doi.org/10.17632/ddmhbb7xgr.1 under the name Erbil Engineering Student Dropout (Aziz, 2024).

5. Results and Analysis

In this section, SPSS Software is used to visualize the dataset, then I use the results to write my insights and analysis. Table 2 displays the addresses of 150 students from the four provinces in the Kurdistan Region of Iraq: Erbil, Sulaymani, Duhok, and Halabja. The table categorizes locations into major cities and other areas, which are classified as towns or villages. Table 3 presents statistics based on three religions: Muslim, Christian, and Ezidies. The attribute names in both tables are sorted alphabetically.

Table 2. Home Address Data layout

Address	Frequency	Percent	Valid Percent	Cumulative Percent
Duhok	21	14.0	14.0	14.0
Erbil	24	16.0	16.0	30.0
Halabja	17	11.3	11.3	41.3
Sulaymani	45	30.0	30.0	71.3
Village	43	28.7	28.7	100.0
Total	150	100.0	100.0	

Table 3. Religion Demographic Data

Religion	Frequency	Percent	Valid Percent	Cumulative Percent
Christian	12	8.0	8.0	8.0
Ezedi	2	1.3	1.3	9.3
Muslim	136	90.7	90.7	100.0
Total	150	100.0	100.0	

Figure 2 presents five charts illustrating the reasons for student dropouts, focusing on those leaving for better living conditions or to study abroad. The charts highlight the influence of five different factors:

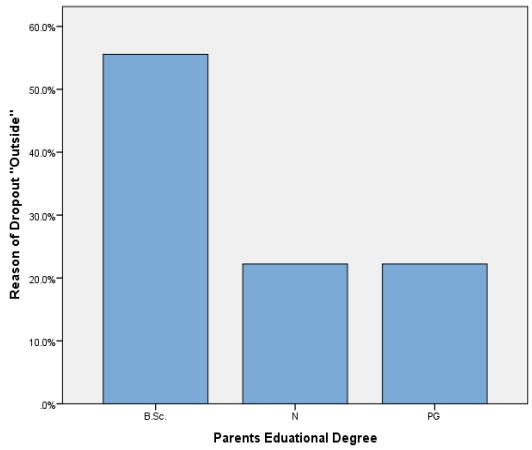
(a) Parent's educational degree: Data shows that parents with a B.Sc degree are more likely to send their children abroad compared to non-graduate parents.

(b) Students' family income: Students from families with an income exceeding 1,500,000 are more likely to study abroad.

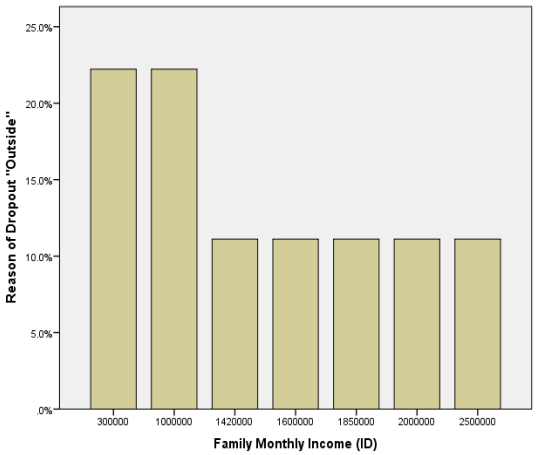
(c) English proficiency: A higher percentage of students with excellent or very good English skills are inclined to drop out and study abroad.

(d) Regional differences: Students from Erbil and Sulaimani exhibit similar dropout rates, while around 45% of students from Duhok go abroad to pursue their degrees.

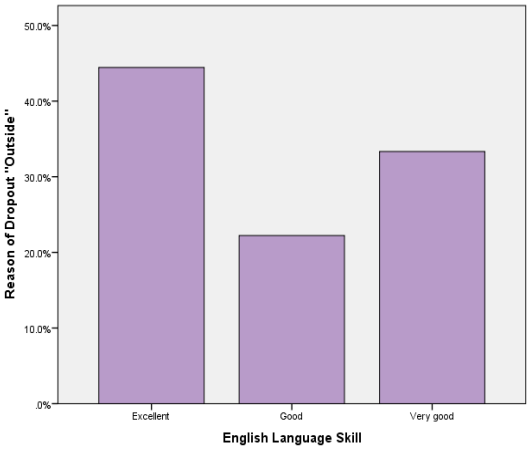
(e) Religious background: Almost all Christian students tend to study abroad.



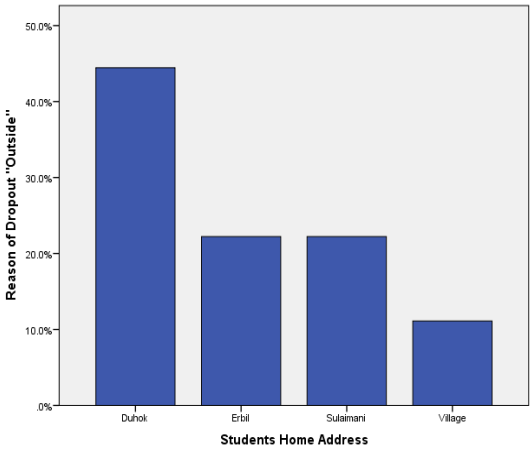
(a)



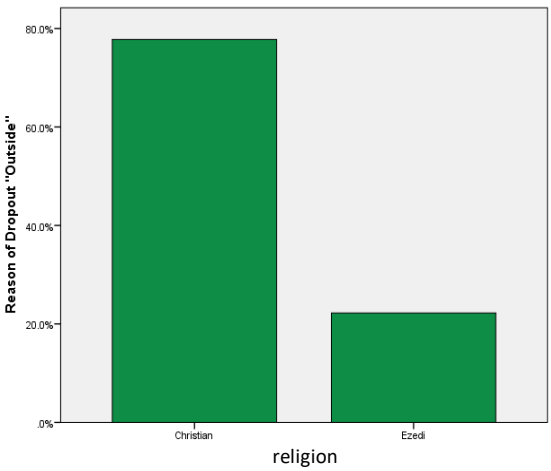
(b)



(c)



(d)



(e)

Fig. 2 Dropout Reasons Related to Better Living Conditions Analyzed by Five Factors: (a)Parents' Educational Degree, (b) Family Monthly Income, (c) English Language Skills, (d) Home Address, (e) Religion.

The major reasons for students changing their major based on five are presented in figure 3. The changing major of the dropped out students are illustrated by five different features. The column bars reveal that students are influenced by these features in their decision to change their major:

(a) Parent's educational degree: According to the data, parents with a B.Sc. degree are more likely to seek alternative universities where their children could be more successful.

(b) Family income: The data indicates that 20% of students with a family income of 1,000,000 are more likely to change their major.

(c) English proficiency: A higher percentage of students with good English skills are inclined to transfer to different universities.

(d) Regional differences: More than 35% of students from Duhok and Sulaimani drop out from Salahaddin University-Erbil and transfer to universities in their home regions to complete their degrees.

(e) Religious background: Nearly all Muslim students who dropped out changed their university.

Figure 4 illustrates the reasons for student dropouts due to postponing or taking a gap year, analyzed by five different features. The charts reveal how these features influence students' decisions to postpone their studies:

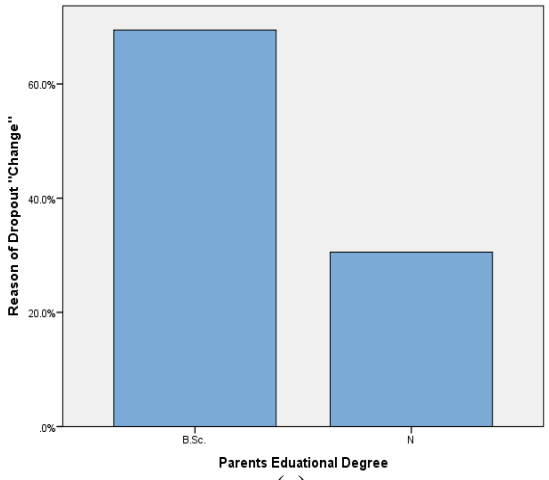
(a) Parents' educational degree: The data shows that over 50% of students with non-graduate parents, just above 40% of students with parents holding a B.Sc. degree, and under 5% of students with postgraduate parents postpone their studies.

(b) Family income: Students from families with an income of about 1,000,000 are more likely to postpone their studies.

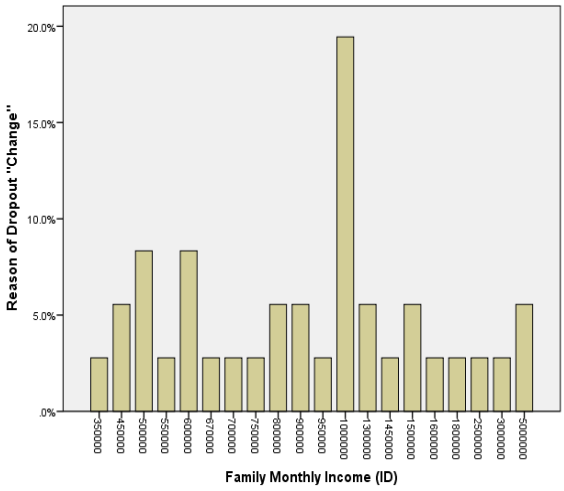
(c) English proficiency: A significant percentage of students with good English skills are inclined to postpone their studies.

(d) Home address: The data indicates that more than 60% of students from rural areas postpone their studies, compared to under 20% from cities within the region.

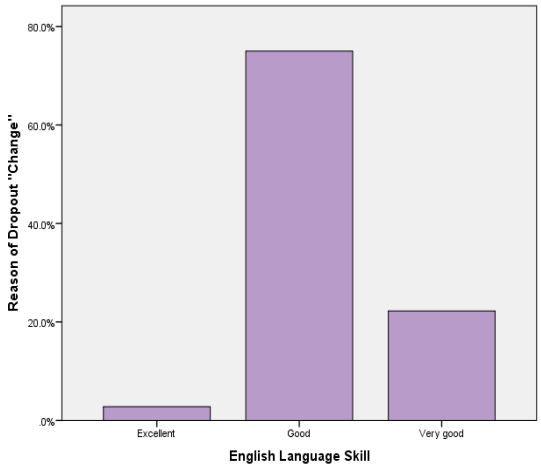
(e) Religious background: The data shows that under 5% of Christian students and nearly 100% of Muslim students are likely to postpone their studies



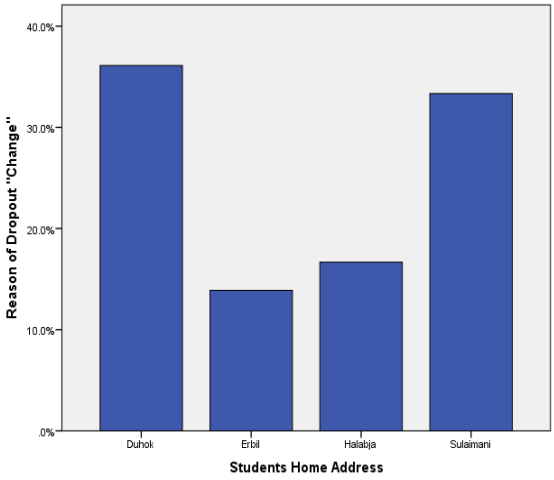
(a)



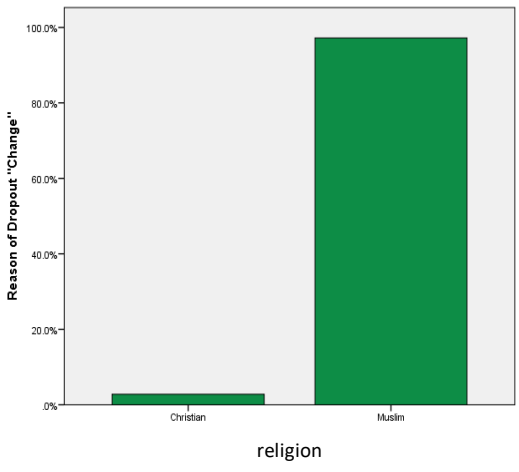
(b)



(c)

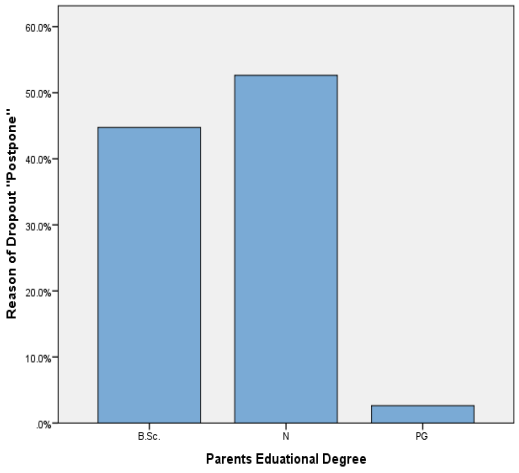


(d)

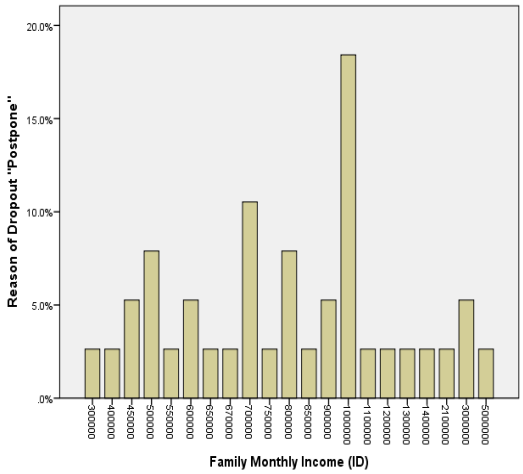


(e)

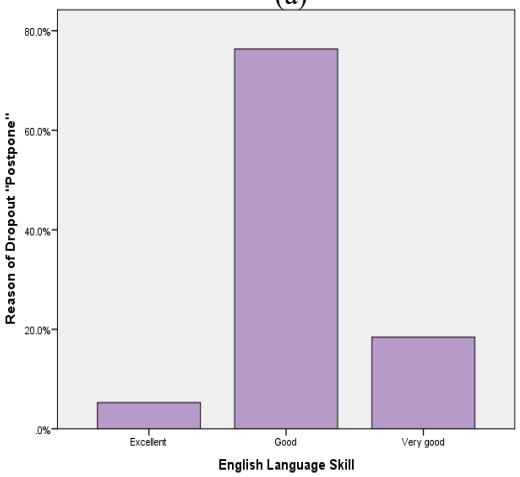
Fig. 3 Major Reasons for Students Changing Their Major Based on Five Factors (a) Parents' Educational Degree, (b) Family Monthly Income, (c) English Language Skills, (d) Home Address, (e) Religion.



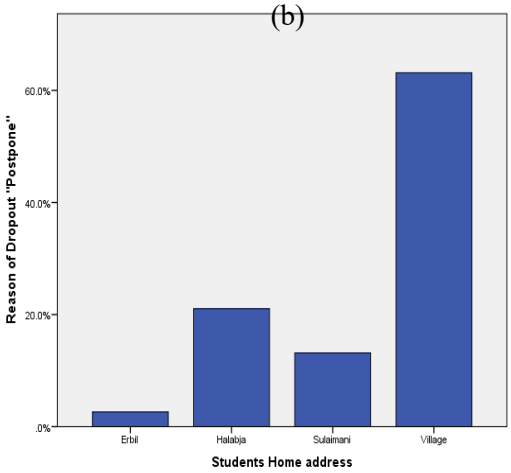
(a)



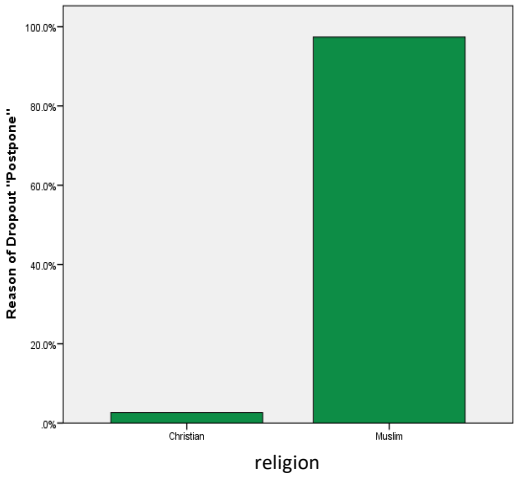
(b)



(c)



(d)



(e)

Fig.4 Dropout Reasons Due to Postponing or Taking a Gap Year. (a)Parents' Educational Degree, (b) Family Monthly Income, (c) English Language Skills, (d) Home Address, (e) Religion.

Figure 5 presents the dropout data layout due to students changing their major or university, illustrated by the aforementioned five different featuers:

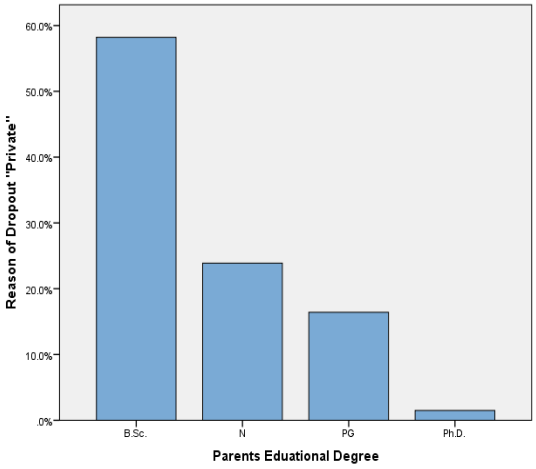
(a) Parents' educational degree: The data shows that about 55% of students with parents holding a B.Sc. degree, just above 20% of students with non-graduate parents, and only 15% of students with postgraduate parents attend private universities.

(b) Family income: Students from families with an income exceeding 1,000,000 are most likely to study at private universities.

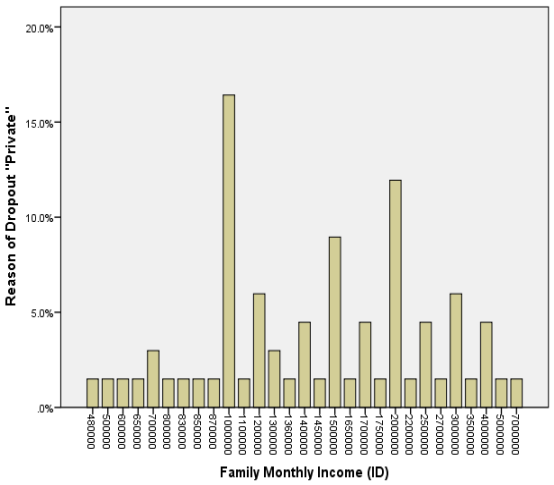
(c) English proficiency: Over 55% of students with good English skills and 30% of students with very good English skills choose to finish their studies at private universities.

(d) Home address: Under 40% of students from Sulemani, more than 25% of students from villages and Erbil, and under 10% of students from Halabja and Duhok attend private universities.

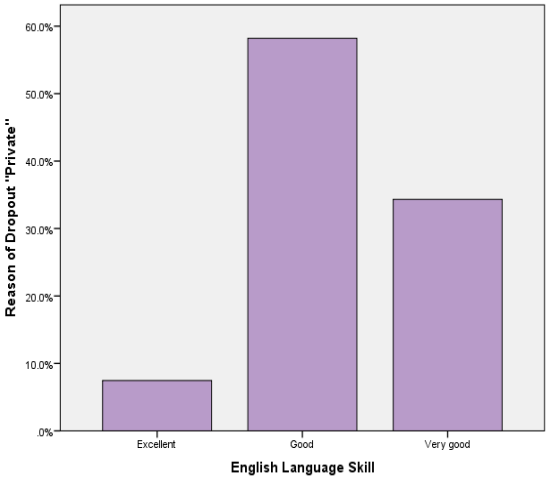
(e) Religious background: The data shows that nearly 55% of students with parents holding a B.Sc. degree tend to study abroad compared to those with non-graduate parents. Additionally, students from families with an income exceeding 1,500,000 are more likely to study abroad. A significant percentage of students with excellent and very good English skills also choose to study abroad. The dropout rates are similar for students from Erbil and Sulemani, while approximately 45% of students from Duhok study abroad. Nearly all Christian students dropped out to study abroad, whereas only 20% of Ezedi students did so to complete their degrees.



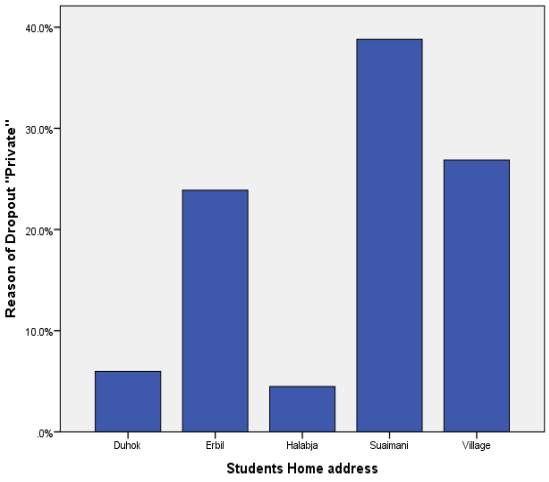
(a)



(b)



(c)



(d)

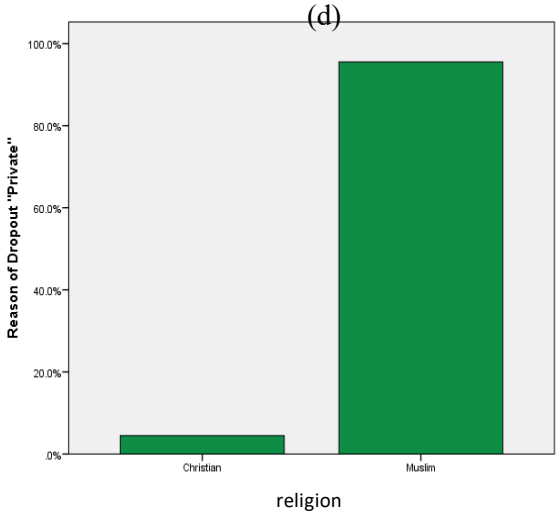


Fig.5 Dropout Data Layout Due to Major and University Changes. (a)Parents' Educational Degree, (b) Family Monthly Income, (c) English Language Skills, (d) Home Address, (e) Religion.

(e)

6. Discussion

The non-parametric Kruskal-Wallis H test, which is a distribution-free test, was used as a statistical tool because the assumptions for one-way ANOVA were not met. Unlike ANOVA, this test does not require the data to be normally distributed with homogeneous variance among participant groups (Neuhauser, 2011).

The monthly income data of dropout students' parents did not follow a Gaussian distribution, as evidenced by the Kolmogorov-Smirnov and Shapiro-Wilk tests (see Table 4 and 5), both of which indicated a statistically significant non-normal distribution. The Lilliefors Significance Correction further confirmed these results.

Given these findings, the Kruskal-Wallis H test was crucial for examining the reasons for leaving college across multiple groups.

According to Table 6, SPSS labeled the Kruskal-Wallis H test by Chi-squared, but it is correctly referred to as the H test. The results indicated a significant difference between the four participant groups, with a statistically significant P-value. The H test is asymptotically Chi-Squared with degrees of freedom equal to k-1, which supports the null hypothesis of equality between groups.

Table 4. Tests of Normality

	Reason of Dropout	Kolmogorov-Smirnova			Shapiro-Wilka		
		Statistic	df	Sig.	Statistic	df	Sig.
Family Monthly Income	postpone	.312	38	.000	.620	38	.000
	private	.198	67	.000	.805	67	.000
	outside	.137	9	.200*	.952	9	.712

	change	.281	36	.000	.658	36	.000

*. This is a lower bound of the true significance.

Table 5. Mean Rank

	Reason of Dropout	N	Mean Rank
Family Monthly Income	postpone	38	52.51
	private	67	95.28
	outside	9	79.50
	change	36	61.94
	Total	150	

Table 6. Kruskal Wallis Test

	Family Monthly Income
Chi-Square	28.305
df	3
Asymp. Sig.	.000

7. Conclusion

The dropout of engineering students in their first year is a concern in both developed and developing countries, but for different reasons. This paper studies the features affecting the dropout rates of first-

year engineering students. The study includes data collection, preprocessing, and modeling using Weka, followed by analysis with SPSS. The results have been deeply analyzed, leading to the following conclusions:

The analysis reveals key factors influencing student dropouts, major changes, and the choice of private universities. Students with parents holding a B.Sc. degree and higher family income are more likely to study abroad. English proficiency and regional differences also significantly impact dropout rates, with students from rural areas and specific regions more inclined to postpone their studies. Nearly all Muslim students and a significant percentage of Christian students drop out to study abroad. The Kruskal-Wallis H test confirmed significant differences between groups, emphasizing the importance of parental education, family income, and regional factors in students' educational decisions. Addressing these factors can help reduce dropout rates and support students effectively.

Ethical Statement

The sensitive parts of the data have been removed, and the data collection was conducted in accordance with the regulations of Salahaddin University-Erbil.

References

Aina, C., Aktaş, K., & Casalone, G. (2024). Effects of workload allocation per course on students' academic outcomes: Evidence from STEM degrees. *Labour Economics*, *90*, 102559. https://doi.org/10.1016/J.LABECO.2024.102559

Aziz, N. Khidr. (2024). *Erbil Engineering Student Dropout*. *1*. https://doi.org/10.17632/DDMHBB7XGR.1

Blokdyk, G. (2021). *Cross-industry standard process for data mining: A Clear and Concise Reference.* 5STARCooks Publishers.

Burgos, C., Campanario, M. L., Peña, D. de la, Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, *66*, 541–556. https://doi.org/10.1016/J.COMPELECENG.2017.03.005

Casanova, J. R., Cervero, A., Núñez, J. C., Almeida, L. S., & Bernardo, A. (2018). Factors that determine the persistence and dropout of university students. *Psicothema*, *30*(4), 408–414. https://doi.org/10.7334/PSICOTHEMA2018.155

Casanova, J. R., Vasconcelos, R., Bernardo, A. B., & Almeida, L. S. (2021). University Dropout in Engineering: Motives and Student Trajectories. *Psicothema*, *33*(4), 595–601. https://doi.org/10.7334/PSICOTHEMA2020.363

Dascalu, M. D., Ruseti, S., Dascalu, M., McNamara, D. S., Carabas, M., Rebedea, T., & Trausan-Matu, S. (2021). Before and during COVID-19: A Cohesion Network Analysis of students' online participation in moodle courses. *Computers in Human Behavior*, *121*, 106780. https://doi.org/10.1016/J.CHB.2021.106780

Desai, N., & Neuhold, L. (2017). Understanding the reasons behind the decreasing enrollment numbers in engineering programs in the United States. *ASEE Zone II Conference*.

Heidrich, L., Victória Barbosa, J. L., Cambruzzi, W., Rigo, S. J., Martins, M. G., & dos Santos, R. B. S. (2018). Diagnosis of learner dropout based on learning styles for online distance learning. *Telematics and Informatics*, *35*(6), 1593–1606. https://doi.org/10.1016/J.TELE.2018.04.007

Khan, S., Shiraz, M., Shah, G., & Muzamil, M. (2023). Understanding the factors contributing to low enrollment of science students in undergraduate programs. *Cogent Education*, *10*(2). https://doi.org/10.1080/2331186X.2023.2277032

Kumari Meena, Nabi Abdul, & Priyanka Puppal. (2014). Educational Data Mining and its role in Educational Field. *International Journal of Computer Science and Information Technologies, 5*(2), 2458–2461.

Hanson M. (2024). *College Dropout Rate : by Year + Demographics*. Education Data. https://educationdata.org/college-dropout-rates

Moscoso-Zea, O., Saa, P., & Luján-Mora, S. (2019). Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining. *Australasian Journal of Engineering Education*, *24*(1), 4–13. https://doi.org/10.1080/22054952.2019.1601063

Mouton, D., Zhang, H., & Ertl, B. (2020). German university student's reasons for dropout. Identifying latent classes. *Journal for Educational Research Online*, *12*(2).

NCES. (2024). *US College Enrollment Decline – 2024 Facts & Figures*. https://www.collegetransitions.com/blog/college-enrollment-decline/

Neuhauser, M. (2011). Nonparametric Statistical Tests : A Computational Approach. *Nonparametric Statistical Tests*. https://doi.org/10.1201/B11427

Nian Kh. Aziz. (2016). Student Academic Performance Using Artificial Intelligence | Request PDF. *Zanco Journal for Pure and Applied Sciences*, *28*(2), 56–69.

Rabelo, A. M., & Zárate, L. E. (2024). A Model for Predicting Dropout of Higher Education Students. *Data Science and Management*. https://doi.org/10.1016/j.dsm.2024.07.001

Rajlakshmi Ghosh. (2023). *Engineering enrolments decline due to lack of job prospects in core Engineering streams - Times of India*. https://timesofindia.indiatimes.com/education/news/engineering-enrolments-decline-due-to-lack-of-job-prospects-in-core-engineering-

streams/articleshow/97795387.cms?utm_source=contentofinterest&utm_medium=text&utm_campaign=cppst

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146. https://doi.org/10.1016/J.ESWA.2006.04.005

Sanjeev, A. P., & Iytkowt, J. M. (1995). *Discovering Enrollment Knowledge in University Databases*. www.aaai.org

Sharon A. Jones, Caitlin Cairncross, & Tammy Vandegrift. (2021). Persistence of Students who Begin Engineering Programs in Precalculus. *Advances in Engineering Education*, *9*(4).

Tocto, P., Huamaní, G. T., & Zuloaga, L. (2023). Machine learning application in university management: Classification model Dropping out of engineering students in Peru. *Proceedings of the LACCEI International Multi-Conference for Engineering, Education and Technology*, *2023-July* 17-21, 2023 (Hybrid) – Buenos Aires, Argentina. https://doi.org/10.18687/laccei2023.1.1.1332

Villanueva, A., & Moreno, L. G. (2018). Data mining techniques applied in educational environments: Literature review Andrés Villanueva Manjarres Data mining techniques applied in educational environments: Literature review. In *Salinas Digital Education Review-Number* (Vol. 33). http://greav.ub.edu/der/

Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.). Retrieved July 30, 2024, from https://ml.cms.waikato.ac.nz/weka

Won, H. S., Kim, M. J., Kim, D., Kim, H. S., & Kim, K. M. (2023). University Student Dropout Prediction Using Pretrained Language Models. *Applied Sciences 2023, Vol. 13, Page 7073*, *13*(12), 7073. https://doi.org/10.3390/APP13127073