Chapter 14

Assessment of Ankara University ERASMUS+ Programme Outgoing Students by Using Statistical Methodologies for the 2023-2024 Period

Cafer Yıldırım[1]

Özlem Türkşen[2]

Necdet Ünüvar[3]

İlker Astarcı[4]

Abstract

Data analysis has a crucial importance to realize student profiles within universities. The data should be analyzed with proper statistical methods and obtained results could be taken into consideration to make decision for supporting student activities and success. This study aims to realize the profile of Ankara University ERASMUS+ Programme Outgoing Students for the 2023-2024 period. Data set is obtained from the ERASMUS Office database and data preprocessing is achieved to organize the data. The well-organized data is analysed according to the categorical and numerical data type by using statistical explaratory data analysis, e.g. descriptive statistics, data visualization. Several parametric and non-parametric statistical tests are applied to the data set. In this context, chi square analysis and correspondence analysis are performed to understand whether there is a relationship between categorical data. One-way analysis of variance (ANOVA) is performed to understand whether there is a difference between groups containing numerical data. Additionaly, classification models were created to predict students'

[1] Ankara University, European Union Education Programs Coordinatorship and Project Office, Türkiye, Ankara University, Faculty of Dentistry, Department of Basic Medical Sciences, Türkiye

[2] Ankara University, Faculty of Science, Department of Statistics, Türkiye

[3] Rectorate of Ankara University, Türkiye

[4] Turkish National Agency, Türkiye

https://doi.org/10.58830/ozgur.pub865.c3512

success and failure through machine learning classification algorithms, called Logistic Regression, k-Nearest Neighbors, Support Vector Machine and Random Forest. It can be said from the results that statistical methodologies help to identify patterns of the data to get knowledge for administrative processes at higher education in European level.

INTRODUCTION

In today's data-driven world, effective data analysis is crucial for understanding trends, optimizing processes, and predicting future outcomes in the field of education as in many areas. Just as universities are important in higher education, ERASMUS offices that provide international connections at universities are also important. The ERASMUS Office holds a central position within the university structure as a key unit responsible for coordinating international academic mobility programs, particularly under the ERASMUS+ framework. Data-driven decision making allows to optimize operations, manage budgets and plan for future growth for administrative duties. Moreover, the statistical data analysis plays a critical role by enabling evidence-based decision making across academic, administrative and research activities. It also helps to assess student performance, improve teaching methods, allocate resources efficiently and enhance effectiveness (Ünüvar et al., 2023).

The statistical data analysis process typically includes several stages: data collection, data cleaning, data exploration, data modeling or transformation, and interpretation of results to make clear decision. Each step plays a vital role in ensuring the accuracy and reliability of the insights generated. Statistical methodologies such as descriptive statistics, data visualization, classification analysis, regression analysis and clustering are frequently applied to explore and present data effectively (Türkşen, 2024). Researchers rely on analytical tools to test hypotheses, validate findings and draw meaningful conclusions from large and complex datasets. While data analysis focuses on interpreting existing data to generate insights, data science involves building predictive models and algorithms using advanced mathematics and programming (Geron, 2017). Data scientists often use machine learning methods and data analysts mainly work with descriptive analytics whereas statisticians use both machine learning methods and descriptive analytics. Common tools used in data analysis include Excel, SQL, Python, R, Tableau, and Power BI.

In this study, it is aimed to assess the profile of Ankara University ERASMUS+ Programme Outgoing Students by using statistical methodologies for the 2023-2024 period since the ERASMUS Office plays an important role in supporting international academic mobility and fostering global connections within universities. The main aim of this study is to reveal the current profile of the Ankara University ERASMUS Office and to strength administrative performance with data-based decision making in several parts e.g. collaboration with partner universities across Europe, promoting academic cooperation, joint research projects, and best practices in education. The rest of the paper is organized as follows. Applied statistical methodologies are given in Materials and Methods section. The analysis results are presented in Results section. Finally, conclusion is given in the last section.

MATERIAL AND METHODS

Material

Data set, in which undergraduate students are taken into consideration, is obtained from the ERASMUS Office database. Data preprocessing stage is achieved with checking missing data, data cleaning and feature selection. Features, namely variables, are defined as Faculty, Research Field (Science, Health and Social), Department, Visited University, Country, Zone (West Europe, East Europe and South Europe), Accomation Time, Visiting Time Period and Success Status. The Success Status, which has categorical data, considered as dependent variable while the others are considered as independent variables which are composed with categorical and continuous data. In addition, a new continuous variable, called Success (%), is calculated according to the achieved ECTS score from the raw data. Data exploration is done by calculating summary statistics (descriptive statistics - central tendency and spread) and data visualization (histogram, barplot, pieplot, boxplot etc.). Some non-parametric statistical tests, (i) Chi-Square Analysis, (ii) Correspondence Analysis, and (iii) Kruskal-Wallis test, are applied to the data set. In this study, the problem is considered as classification problem. Then, Machine Learning Classification Algorithms, called Logistic Regression (LR), k-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Random Forest (RF), are applied to predict students' success status. All the calculations are done by using Python libraries, called numpy, pandas, matplotlib, seaborn, scikit-learn.

Methods

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyze datasets for summarizing their main characteristics, often with visual methods. The main goal of the EDA is to understand the data better before applying any formal statistical models or machine learning algorithms. By using plots like histograms, scatter plots, and boxplots, as well as summary statistics such as mean, median, and standard deviation, analysts can gain insights into the underlying structure of the data, identify important variables, and detect outliers or missing values that may affect further analysis.

Non-parametric Statistical Analysis

i. *Chi-Square Analysis*

Chi-Square Analysis, also known as Chi-Squared Test, is a statistical method used to determine if there is a significant association between two categorical variables. It compares the observed frequencies in each category of a contingency table with the expected frequencies if the variables were independent. The main purpose of the Chi-Square Test is to assess whether there is a statistically significant difference between the observed and expected frequencies in one or more categories. The Chi-Square Statistic is calculated using the formula

$$\div^2 = \sum_{i=1}^{r} \frac{\left(O_i - E_i\right)^2}{E_i}$$

where r is the number of category, O_i is observed frequency and E_i is expected frequency. A higher Chi-Square value indicates a greater difference between observed and expected values. The result is compared to a critical value from the chi-square distribution table or evaluated using a p-value to determine significance (Karagöz, 2019). In this study, Chi-Square Test is used to determine if there is a significant association between two categorical variables, called Chi-Square Test of Independence.

ii. *Correspondence Analysis*

Correspondence Analysis is a multivariate statistical method used to visualize and interpret relationships between categorical variables in a contingency table. It reduces the dimensions of the data, allowing researchers to explore associations between rows and columns by projecting them into a lower-dimensional space, typically a 2D plot. The main goal of the Correspondence Analysis is to summarize and visualize patterns in large categorical datasets (Alpar, 2025).

iii. *Kruskal-Wallis Test*

The Kruskal-Wallis test is a non-parametric statistical method used to determine if there are statistically significant differences between the medians of three or more independent groups. It is considered the non-parametric alternative to the one-way ANOVA, and it does not assume that the data follows a normal distribution (Karagöz, 2019).

Machine Learning Classification Algorithms

Classification algorithms in machine learning are supervised learning methods used to predict categorical dependent variables based on indenpendent variables. These algorithms learn patterns from labeled training data and apply that knowledge to classify new, unseen instances into one of the predefined classes (Müller and Guido, 2017).

LR is a classification method. It models the probability of a binary outcome using a logistic function. It works well when the relationship between features and the target is approximately linear.

KNN is a simple instance-based classification algorithm that classifies a new data point based on the majority class among its k nearest neighbors in the feature space.

SVM finds the optimal boundary (hyperplane) that best separates different classes. It works well in high-dimensional spaces and with small datasets.

RF is an ensemble method that builds multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. It's robust and effective for many real-world problems.

The tuning parameters of these algorithms are defined by using expert knowledge. The performance comparison of the classification algorithms is achieved by using several metrics, e.g. Accuracy, Precision, Recall, F1-Score and AUC (Ulu Metin, 2024). If the data set is imbalanced the F1-Score metric can be considered as the most preferable one.

RESULTS

The raw data of the Ankara University ERASMUS Office is well-organized according to the defined data preprocessing process. Exploratory data analysis is applied and obtained results are presented in Figure 1.(a)-(c) for the ERASMUS+ Programme outgoing students.

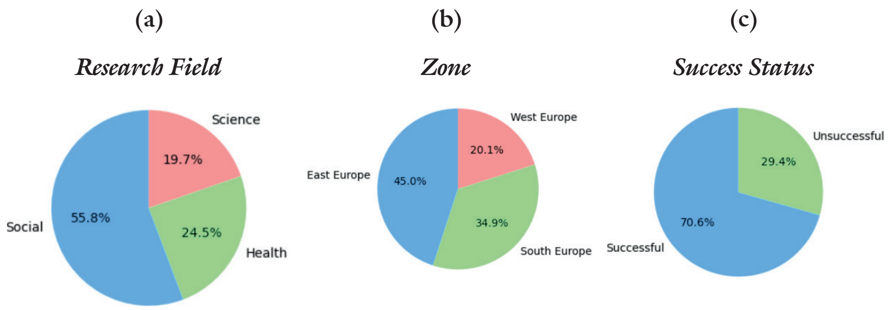
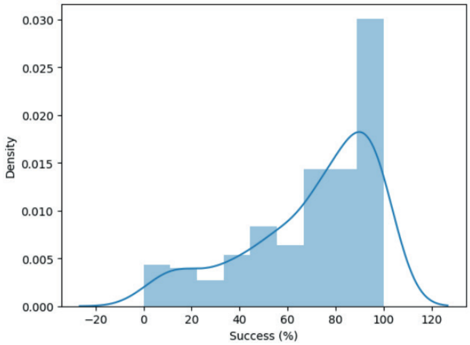


Figure 1. Pie plots of Research Field, Zone and Success Status variables

The distribution of Research Fields can be seen in Figure 1.(a). It can be said from Figure 1.a that the majority of the students are from the social research field. From Figure 1.(b), it is possible to say that the students prefer East Europe. And also, the majority of the students are successful, as presented in Figure 1.(c).

The Success (%) and summary statistics of success quantity can be seen in Figure 2. It is seen from the Figure 2 that the distribution of the Success (%) is left-skewed which means that high-achieving students are the majority. It can be easily said from the summary statistics that number of students, minimum, maximum, mean, median, standart deviation, range and inter quartile range values of the Success (%) are 269, 0, 100, 69.86, 75.76, 27.13, 100 and 36.36, respectively.



Summary Statistics

```
count    269.000000
mean      69.862566
std       27.132149
min        0.000000
25%       54.545455
50%       75.757576
75%       90.909091
max      100.000000
```

Figure 2. The distribution of the Success (%) and the summary statistics

The distribution of data about students Visiting Time Period has two-peaked distribution, presented in Figure 3. It is seen from Figure 3 that the most preferred visiting time periods are 5 and 10 months for the students.

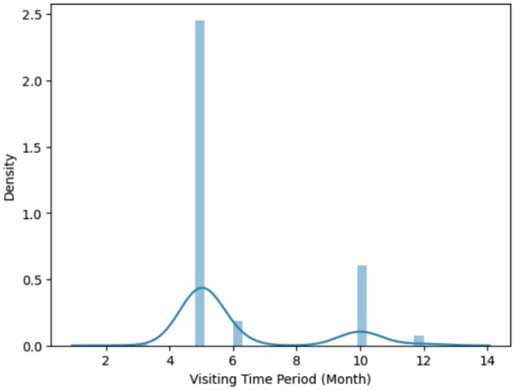


Figure 3. The distribution of the students visiting time period

Chi-Square Analysis is applied to understand whether there is a dependency between success status and Research Field, Success Status and Zone. Contingency tables of Success Status-Research Field and Success Status-Zone can be seen in Table 1 and Table 2, respectively. In Tables 1-2, the frequencies of the categories and expected values, given in parentheses, are presented. Chi-Square statistics is obtained as 6.04 (p-value=0.048) for Success Status and Research Field and the Chi-Square statistics value is equal to 22.98 (p-value=0.0000102) for Success Status and Zone. According to this statistics, Success Status has dependency with Research Field and Zone with %95 confidence.

Table 1. Contingency table of Success Status and Research Field

	Social	Science	Health	Total
Successful	98 (105.95)	38 (37.43)	54 (46.62)	190
Unsuccessful	52 (44.05)	15 (15.57)	12 (19.38)	79
Total	150	53	66	269

Table 2. Contingency table of Success Status and Zone

	East Europe	West Europe	South Europe	Total
Successful	103 (85.46)	34 (38.14)	53 (66.39)	190
Unsuccessful	18 (35.54)	20 (15.86)	41 (27.61)	79
Total	121	54	94	269

The results of Correspondence Analysis are presented in Figure 4. It can be said from Figure 4 that the ERASMUS+ Programme outgoing students from Science, Health and Social research fields are preferred to go to the West Europe, the East Europe and the South Europe, respectively.

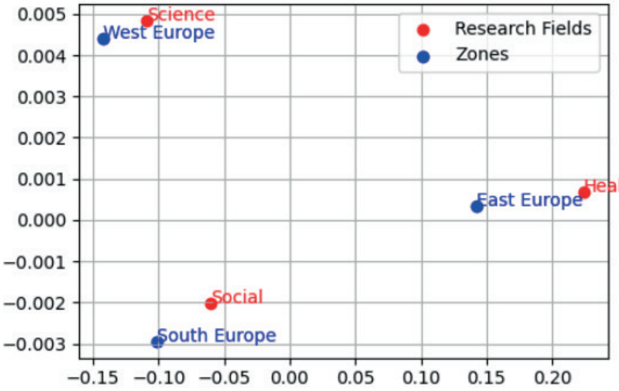


Figure 4. The Correspondence Analysis plot for Research Fields and Zones

It is seen from Figure 5 that the normality assumption is not provided according to the Success (%) in the research fields. Therefore, Kruskal-Wallis test is applied to determine whether there is a difference in Success (%) between the the Research Fields. The Kruskal-Wallis statistics is calculated as 1.6034 and *p*-value is equal to 0.4486. Thus, it is possible to say that there is no difference between the research fields according to the Success (%) with %95 confidence. In addition, this result can be seen from the boxplots clearly as given in Figure 6.

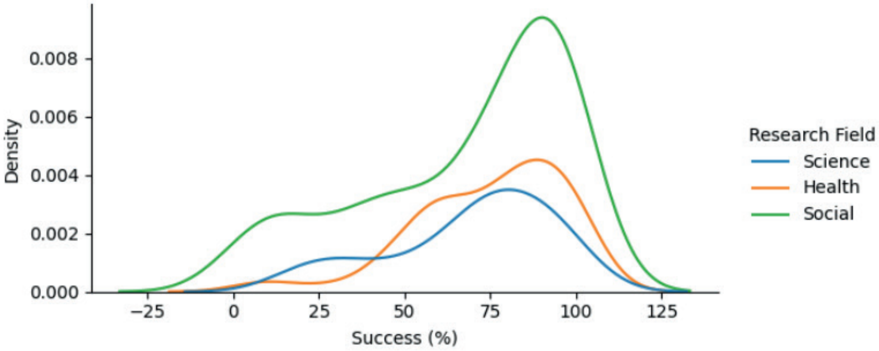


Figure 5. Distribution of the Success (%) for the Research Fields

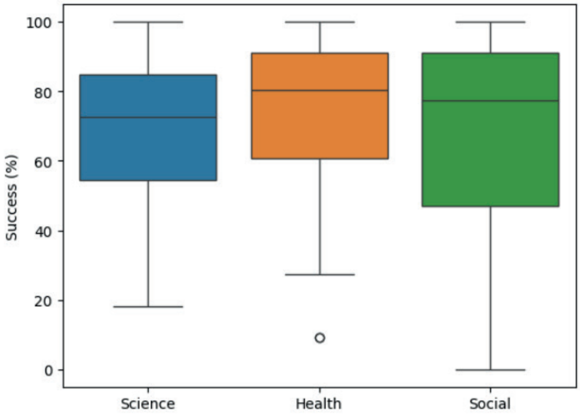
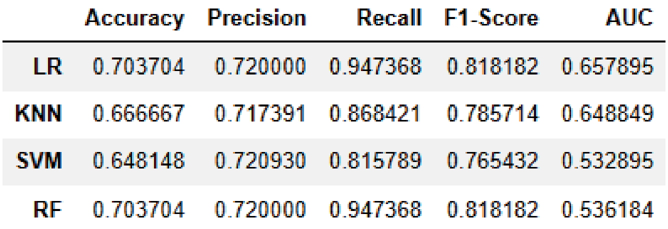


Figure 6. Box-plots of the Success (%) for the research fields

In this study, Success Status is considered as dependent variable with two categories, Successful and Unsuccessful. The values of the dependent variable has imbalanced distribution as can be seen from the pie plot in Figure1.(c). Machine Learning Classification Algorithms are applied to the data set. The obtained performance metrics of the algorithms are given in Table 3. It can be said from the Table 3 that the RF and the LR algorithms have better performance than the KNN and the SVM algorithms for the classification of ERASMUS+ Programme outgoing students according to the their Success Status.

Table 3. Performance metric values of Machine Learning Classification Algorithms



	Accuracy	Precision	Recall	F1-Score	AUC
LR	0.703704	0.720000	0.947368	0.818182	0.657895
KNN	0.666667	0.717391	0.868421	0.785714	0.648849
SVM	0.648148	0.720930	0.815789	0.765432	0.532895
RF	0.703704	0.720000	0.947368	0.818182	0.536184

DISCUSSION AND CONCLUSION

In higher education, the ERASMUS Office, as a key administrative unit within universities, plays a crucial role in facilitating international mobility and cooperation under the ERASMUS+ program, which is one of the most significant educational initiatives of the European Union. Accordingly, it would be appropriate to say that the statistical analysis of the data belonging to the ERASMUS Office is important to track student progress, identify at-risk learners, personalize educational experiences, and data-driven decision making from an administrative perspective.

This study presents the statistical data analysis results of Ankara University ERASMUS Office for Outgoing Students during the 2023-2024 period for the first time. It is seen from the results that majority of the ERASMUS+ Programme outgoing students prefer to go East Europe, majority of them are from social research field and also successful. It is seen from the results that the most preferred visiting time periods are 5 and 10 months for the students. Chi-Squared statistics showed that the success status has dependency with Research Fields and Zones with %95 confidence. The Correspondence Analysis plot helps to realize closeness relationship of the Research Fields and Zones. According to the performance metrics of Machine Learning Classification Algorithms, the LR and the RF can be used for forecasting of Success Status of ERASMUS+ Programme Outgoing Students. It can be also said from the results that the LR is slightly better than the RF for all performance metrics considering the imbalanced data.

References

Alpar R, 2025. Uygulamalı İstatistik ve Geçerlik-Güvenirlik. Detay Yayıncılık, Ankara.

Geron A, 2017. Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly, USA.

Karagöz Y, 2019. SPSS, AMOSA, META Uygulamalı İstatistiksel Analizler. Nobel Yayınevi, Ankara.

Müller AC and Guido S, 2017. Introduction to Machine Learning with Python A Guide for Data Scientists. O'Reilly, USA.

Türkşen Ö, 2024. Veri Analizi ve Madenciliği. Ankara Üniversitesi Yayınevi, Ankara.

Ulu Metin G, 2024. Ar-Ge ve Tasarım Merkezlerinin İstatistiksel Olarak Değerlendirilmesi: Veri Madenciliği Yöntemleri ile Hibrit Karar Verme. Doktora Tezi, Ankara Üniversitesi, Ankara.

Ünüvar N, Apaydın A, Kutlu Ö, Vural MR, Türkşen Ö, Arıca Akkök E, 2023. Ankara Üniversitesi'nde Eğitim: Düşünce Atölyesi. Ankara Üniversitesi Yayınevi, Ankara.

Acknowledgment

This study is the result of research conducted in Ankara University, European Union Education Programs Coordinatorship and Project Office Data founded by Turkish National Agency.

Conflict of Interest

The authors have declared that there is no conflict of interest.

Author Contributions

Cafer Yıldırım: Coordinator, Obtaining data from database, Interpretation, Investigation, Writing – review & editing

Özlem Türkşen: Writing – Original draft, review & editing, Methodology, Statistical Analysis, Data visualization, Interpretation, Investigation

Necdet Ünüvar: Resources, Legal Representative of European Union Projects, Policy maker

İlker Astarcı: Resources, Funding, Conceptualization