

# **Enrollment Rate Prediction for Foundation Universities in Turkey Using Machine Learning**

**Yasin Göçgün**

Associate Professor

Industrial Engineering Department

İstanbul Medipol University

İstanbul, Turkey

yasin.gocgun@medipol.edu.tr

**Merve Ece Akat, Esra Akgün, Nerimana Bulut, Esma Engin, and Hiba Sadioğlu**

Graduate

Industrial Engineering Department

İstanbul Medipol University

İstanbul, Turkey

merve.akat@std.medipol.edu.tr, esra.akgun@std.medipol.edu.tr,

nerimana.bulut@std.medipol.edu.tr, esma.engin@std.medipol.edu.tr,

hebah.abualais@std.medipol.edu.tr

## **Abstract**

We investigate the application of machine learning techniques to predict student enrollment rates at Turkish foundation universities. In this study, prediction was made with three different algorithms: multiple linear regression, decision tree, and random forest. The model used historically collected data to accurately forecast future enrollment rates and delivers useful findings for optimizations within the area of university institutions. Key factors taken into consideration by foundation universities in Turkey in the process of determining student registrations and capacity of departments were examined. A number of variables affect the success criteria, and benchmarks set to evaluate how well these variables perform. In general, the practical use of machine learning in enrollment estimation was discussed in the project and an enrollment forecast was made for the 2024-2025 academic year. It was emphasized that foundation universities should benefit from machine learning to increase efficiency, optimize resource allocation, and enable relevant managers to make better decisions and develop strategies.

## **Keywords**

Enrollment rate prediction, Machine Learning.

## **1. Introduction**

Student enrolment rate is an important performance indicator of colleges' long-term academic and financial stability. The main aim of this research project is to develop a stable and accurate machine learning model that can predict the enrollment rates of private universities in Turkey. It allows universities to identify enrollment trends over time and identify potential bottlenecks, guiding the institution toward more efficient and effective student enrollment and institutional management.

Upon a successful project accomplishment, subsequent objectives listed below are fulfilled:

- Improved prediction of university student enrollments: creating a machine learning model that can accurately predict future enrollment at foundation universities based on historical data to provide useful information about expected student enrollment.
- Optimized resource allocation: foundational universities can deploy their assets more strategically through predicting upcoming enrollment projections. An accurate forecast of enrollment rates empowers institutions to predetermine which majors/programs are likely to experience increased interest. Projections made by the program model can help university administrations when planning student capacities, setting enrollment quotas, staff recruitment, and the development of infrastructure necessary to accommodate demand effectively.
- Strengthen decision-making processes within the education sector through the provision of a decision-making tool that is able to foresee variations in educational enrollment rates. Modeling predictions serve as a roadmap for the development of policies aimed at improving the school sector's overall quality.
- Engage long-term strategic and financial planning capabilities through the visualization of possible changes in registration behavior. Such a proactive approach facilitates flexibility and responsiveness to changing educational needs and sustainable growth.

## **2. Literature Review**

We next provide a review of the recent literature.

Calvo et al. (2020) carried out a predictive investigation regarding the process of student enrollment at Cebu University of Technology. An enrollment forecast has been formulated for the academic years spanning from 2019 to 2025, utilizing the ARIMA (p, d, q) model and historical data collected between 2011 and 2019. Data mining techniques such as ACF were also applied in the study. In another study, Mia et al. (2019) used machine learning to prevent various difficulties related to university student registration in densely populated countries like Bangladesh. Diverse machine learning algorithms have been employed to forecast enrollment, dropout rates, as well as final grades, among other factors. The Least Squares Method has been employed to predict new registrations, yielding an accuracy rate of 97.8%. Further, Ujkani et al. (2021) used data such as students' high school grades, Matura exam results and university entrance exam scores to predict enrollment using machine learning, data taken from the University of Mitrovica. A total of sixteen algorithms have been employed, utilizing Weka software, Naive Bayes, Logistic regression, K-nearest neighbor, and decision trees techniques.

In Yang et al. (2021), machine learning techniques were used to predict freshman enrollment. The authors explored the application of machine learning techniques, including Decision Tree, Random Forest, and Backpropagation neural network, utilizing historical admission record data obtained from a university in Guangzhou for the purpose of conducting this analysis. In another study, Shilbayeh and Abonamah (2021) considered models made with machine learning established to predict students' enrollment behavior. The data set, which includes the years 2013 - 2018 and contains information of 1600 students, was taken from Abu Dhabi School Administration. As a result, it has been observed that students have a high risk of dropping out of school. Further, Akinode and Bada (2021) investigated the impact of the multiple pre-enrollment parameters such as WAEC grades, JAMB Scores etc. that might affect the enrollment of students in a Federal Polytechnic in South west Nigeria. The study was conducted on 560 students who enrolled at Federal Polytechnic in South-West Nigeria between 2017-2018.

Shao et al. (2022) performed enrollment prediction of a course (General Chemistry) by applying conditional probability analysis using the San Diego State University student data. The authors combined student's demographic and academic information into algorithms to investigate the effect on improving prediction accuracy. In another study, Basu et al. (2019) used machine learning techniques to predict whether someone will accept the offer. The study was implemented using four-year data from a university with 11,001 students in California. In a recent study, Kye (2023) employed four distinct machine learning algorithms to forecast enrollment in United States educational institutions: Artificial neural network, support vector machine, decision tree, and logistic regression algorithms. The accuracy, sensitivity, specificity, and precision metrics were computed for each of the four models.

Canada et al. (2021) considered the application of machine learning to forecast enrollment, it is discussed that logistic regression can be employed to determine the elements affecting the probability of a student registering in a program. This way, it can estimate the number of people enrolled in programs by analyzing and using demographics, pass rates, and program trends as input data. Further, Patungan and Francia (2022) utilized machine learning for enrollment prediction at the esteemed University of Santo Tomas. The investigation employed data derived from the Admissions

Office and the Office of the Secretary General at the aforementioned university, with a particular focus on the admission details of students throughout the past five academic years. The dataset comprises 24 attributes, consisting of 23 input variables for the label and one output variable denoting registration status.

### 3. Methods

Managing the project's goals and resources properly is of high importance when determining the scope and boundaries of the project. The time constraint of this study is 28 weeks. In this study, numerical data regarding the Higher Education Institutions Examination (YKS) placement results are used. This research covers 78 foundation universities in Turkey. The data of foundation universities was limited to the period 2020-2023 and four-year data was collected. In this data, the names of the universities, department information in the universities, total capacity, year and total occupancy information in the departments are included as independent variables. All this data was collected and organized. Then, the enrollment rate variable was calculated separately for each university and department.

Utilizing the use of machine learning techniques, a prediction model was developed from historical university enrollment, capacity, and student population data. The "enrollment rate" is chosen as a target variable, and a specific model will be developed to split the required data in training as well as testing sets after it has been collected and cleaned. Scikit-learn is a favored library for machine learning implemented with the Python programming language that may be utilized to build models that manage dataset complexities. Machine learning models like multiple linear regression, decision tree and random forest methods are implemented. It can make predictions on the test set after it has been tied to the training set. Universities can plan and manage resources more effectively in this way.

For data preprocessing and preparation, a properly organized data set for analytical use was needed. For this reason, the data file is changed several times for different reasons. In addition, the features selected that make a relevant contribution to the model's prediction strength; and the features were considered to be university name, city, faculty, program name, program language, scholarship, educational type, smallest score, and year.

Further, data is improved, cleaned, and preprocessed to avoid data complexity, noise, duplication and missing data. As a result, Multiple Linear Regression is implemented as a first part of the research. The code written for data preprocessing is shown in Figure 1.

```
# Function to preprocess the data
def preprocess_data(data):
    # Convert percentiles and numeric values to the correct format
    data['enrollment_rate'] = data['enrollment_rate'].str.rstrip('%').replace('---', pd.NA).astype(float) / 100.0
    data['enrollment_rate'] = data['enrollment_rate'].fillna(data['enrollment_rate'].mean())
    # Convert scholarship status to numeric values
    data['scholarship'] = data['scholarship'].apply(lambda x: 0 if 'Burslu' in x else (1 if '50 İndirimli' in x else 2))
    # Convert education type to numeric values
    data['educational_type'] = data['educational_type'].apply(lambda x: 1 if 'ÖRGÜN' in x else 0)
    # Convert the smallest scores to numerical values and fill in missing values
    data['smallest_score'] = data['smallest_score'].str.replace('.', '').str.replace(',', '.')
    data['smallest_score'] = pd.to_numeric(data['smallest_score'], errors='coerce')
    data['smallest_score'] = data['smallest_score'].fillna(data['smallest_score'].mean())
    # Convert year to numeric value
    data['year'] = pd.to_numeric(data['year'], errors='coerce')
    return data
```

Figure 1. Data Preprocessing Code

Similarly, decision tree and random forest algorithms decided to be implemented and evaluation of each method chosen to be with the same performance criteria which are considered to be mean squared error (MSE) and mean absolute error (MAE) and eventually compared with each other to find out the best performer method of the research.

### 4. Results and Discussion

As the first demo plan Multiple Linear Regression application was accomplished successfully, the average predicted enrollment rate for all departments in 2024 is found as 81.16% with average mean absolute error for all departments as 0.30 which can be considered as success for a range of one to a hundred (enrollment rates being 0% to 100%). In Figure 2 prediction demonstration results are presented for the universities' average enrollment rate prediction and performance criteria.

```
Average Predicted Enrollment Rate for all departments in 2024: 81.16%
Average Mean Squared Error for all departments: 9.816154228196924
Average Mean Absolute Error for all departments: 0.3007186590097876
```

Figure 2. Representation of Multiple Linear Regression Output for the All-University’s Average Enrollment Rate Prediction and Performance Criteria Measures in the First Demo Plan

After the first demo plan is created, the years are added to the dataset as independent variables and it is improved the accuracy of the data from MSE 9.82 to MSE 0.12 which is approximately 98.78% improvement. This dramatic improvement is not only dependent on the new variable but also increased data quality, enhanced data preprocessing etc.

The objective of the study is achieving accurate enrollment rate predictions for the year 2024 based on the previous data between 2020-2023. This data includes university name, city, faculty, program name, language, educational type, scholarship, smallest score, year and enrollment rate. Three supervised machine learning techniques applied to the dataset using Python programming language and for each technique performance evaluation and comparison is made by application of mean squared error (MSE), and mean absolute error (MAE). These techniques are determined as multiple linear regression, decision tree and random forest.

For the multiple linear regression, the below formula is used to show the correlation between the independent variables and target value.

$$Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4$$

The Y in the formula represents the “enrollment rate”,  $\beta_0$  represents the intercept,  $\beta_1, \beta_2, \beta_3, \beta_4$  represents the slope, and  $x_1, x_2, x_3, x_4$  represents the independent variables which are educational type, smallest score, scholarship, and year. According to multiple linear regression method, the average predicted enrollment rate is found to be 0.91 for 2024. Moreover, average predicted enrollment rates for 2023, 2022, 2021, and 2020 are found as 0.95, 0.89, 0.84 and 0.97 respectively. As the second step decision tree is implemented, the results are found to be as average predicted enrollment rate for 2024 as 0.91. Furthermore, average predicted enrollment rates for 2023, 2022, 2021, and 2020 are found as 0.95, 0.91, 0.83 and 0.97 respectively. Lastly, random forest method is applied and the results are as follows; average predicted enrollment rate for 2024 as 0.93, average predicted enrollment rates for 2023, 2022, 2021, and 2020 are found as 0.96, 0.93, 0.88, and 0.98 respectively. Each enrollment rate prediction for each program of each university is gathered in a separate txt file, as shown in Figure 3.

```
116 BAHCESEHIR UNIVERSITESI - Ağız ve Diş Sağlığı (2023): Predicted: 1.0, Actual: 0.8, MSE: 0.03999999999999998, MAE: 0.19999999999999996
117 BAHCESEHIR UNIVERSITESI - Anestezi (2023): Predicted: 1.0, Actual: 1.0, MSE: 0.0, MAE: 0.0
118 BAHCESEHIR UNIVERSITESI - Çocuk Gelişimi (2023): Predicted: 1.0, Actual: 0.67, MSE: 0.10889999999999997, MAE: 0.32999999999999996
119 BAHCESEHIR UNIVERSITESI - Diş Protez Teknolojisi (2023): Predicted: 1.0, Actual: 0.8, MSE: 0.03999999999999998, MAE: 0.19999999999999996
120 BAHCESEHIR UNIVERSITESI - Fizyoterapi (2023): Predicted: 1.0, Actual: 0.86, MSE: 0.019600000000000003, MAE: 0.14
121 BAHCESEHIR UNIVERSITESI - İlk ve Acil Yardım (2023): Predicted: 1.0, Actual: 0.86, MSE: 0.019600000000000003, MAE: 0.14
122 BAHCESEHIR UNIVERSITESI - Optisyonluk (2023): Predicted: 1.0, Actual: 0.83, MSE: 0.028900000000000012, MAE: 0.17000000000000004
123 BASKENT UNIVERSITESI - Anestezi (2023): Predicted: 1.0, Actual: 0.9, MSE: 0.009999999999999995, MAE: 0.09999999999999998
124 BASKENT UNIVERSITESI - Fizyoterapi (2023): Predicted: 0.9875303106854135, Actual: 1.0, MSE: 0.0006549766408717271, MAE:
0.023223518311680902
125 BASKENT UNIVERSITESI - Fizyoterapi (2023): Predicted: 0.9239773473087753, Actual: 0.89, MSE: 0.0006549766408717271, MAE:
0.023223518311680902
126 BASKENT UNIVERSITESI - İlk ve Acil Yardım (2023): Predicted: 0.9584236769519774, Actual: 1.0, MSE: 0.00202813436847808, MAE:
0.04491116972553244
127 BASKENT UNIVERSITESI - İlk ve Acil Yardım (2023): Predicted: 0.9482460164030423, Actual: 0.9, MSE: 0.00202813436847808, MAE:
0.04491116972553244
128 BASKENT UNIVERSITESI - Tıbbi Görüntüleme Teknikleri (2023): Predicted: 0.9084427811077462, Actual: 0.9, MSE: 7.128055283331634e-05, MAE:
0.008442781107746211
```

Figure 3. Representation of Multiple Linear Regression Output for the All-University’s Enrollment Rate Prediction and Performance Criteria Measures as txt File

For the representation of the output of the model Tkinter which is a GUI toolkit is used. As a result of application Average Predicted Enrollment Rate for each university, and as performance indicators; Mean Squared Error and Mean Absolute Error for each University is obtained. The overall performance evaluation criteria average mean squared error (MSE) and mean absolute error (MAE) for all departments for each model is calculated and results are shown in Figure 4. Furthermore, average MSE and MAE results are found for each year from 2020 to 2023 as well. For the 2024 MSE and MAE values as follows:

1. Multiple Linear Regression MSE: 0.12, MAE: 0.19
2. Decision Tree MSE: 0.11, MAE: 0.18
3. Random Forest MSE: 0.11, MAE: 0.18

As a result of performance evaluations, Random Forest and Decision Tree methods gave the most accurate results for 2024 as shown in Figure 4.

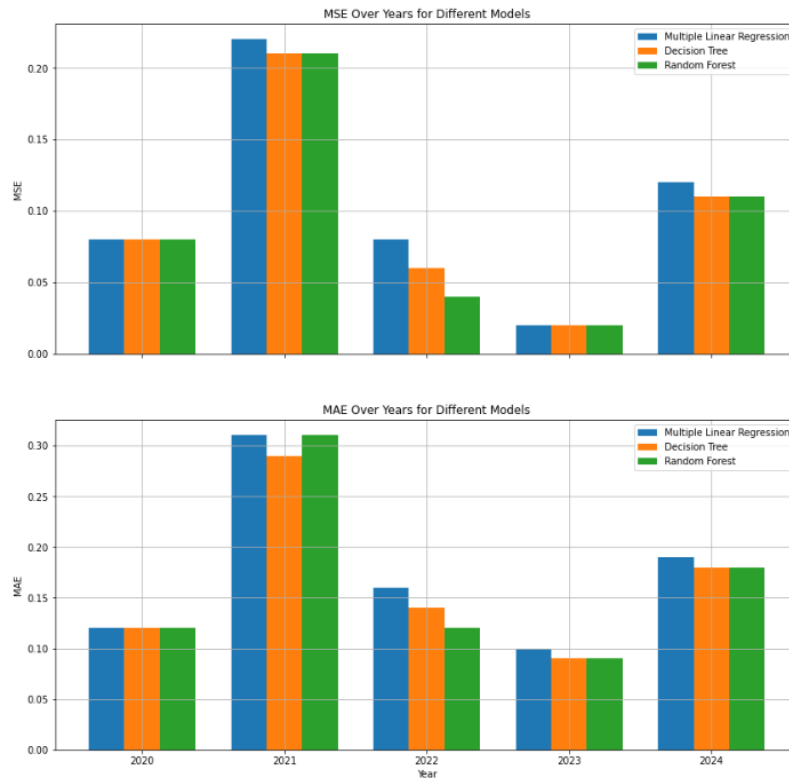


Figure 4. Representation of Each Methods Performance in Clustered Column Chart

#### 4.1. Discussion

Our primary focus is on predicting enrollment rates at Turkish universities through machine learning techniques, with a focus solely on foundation universities for research purposes.

For an accurate prediction, it is essential to have access to historical information. Taking this into consideration, this research recognizes this by highlighting the requirement of at least a four-year period of historical records. Through an extensive dataset of historic records, the ML model manages to understand the patterns, trends and drivers that shape the enrollment rates over the years, resulting in better forecasting accuracy and the ability to precisely estimate the enrollment rate of foundation universities in Türkiye for the following academic year.

The data was taken from the Student Selection and Placement Center (ÖSYM) institution in Turkey. Numerical data on the grading results of the University Examination (YKS) was used. This study includes 78 foundation universities demographically focusing on Turkey. The scope of the foundation universities' records has been restricted to years

2020-2023 and subsequently quadrennial data have been collected. Following the gathering of the data from the sources named, the dataset was analyzed, cleansed and organized for the preparation for project realization by supervised machine learning algorithms with Python.

The data set created includes the name of the university, the city, the faculty, the name of the study program (departments), the language of the study program, the scholarship, the type of score, the type of learning, the largest and the smallest score, the total capacity and the total occupancy as a whole. The dependent or target variable chosen for the analysis was "enrollment rate", which was calculated individually by dividing the total occupancy by the total capacity values for each university/faculty/department.

A total of nine independent variables were then defined:

1. University Name
2. City (Location)
3. Faculty
4. Program Name (Department)
5. Program Language
6. Scholarship
7. Smallest Score
8. Educational Type
9. Year

The Python code had successfully combined the previously mentioned dependent and independent variables to improve and enhance enrollment rate predictions. By using machine learning techniques, specifically linear regression algorithms, the code predicted enrollment rates based on a variety of independent variables. The dataset most included historical enrollment data and associated attributes for each university.

The model's output has included predicted enrollment rate percentages for each particular university as well as metrics: mean square error (MSE) and mean absolute error (MAE), which are crucial for assessing the accuracy and reliability of the predictions. These percentages usually provide valuable insights into the factors influencing/affecting student enrollment at various universities.

Interestingly, the predicted rate values for some several universities ranged widely, from 1% to 100%. These variations could be attributed/referred to factors such as university reputation, ranking, program popularity, distinct location, and scholarship availability. Understanding these variations is critical for comprehending/justifying the complex dynamics that influence enrollment decisions. Identifying the most influential variables enables universities to improve and enhance their strategies for attracting more students, such as focusing marketing efforts on popular programs or locations.

Despite the results, prediction accuracy is usually highly dependent on the dataset's quality, accuracy and representativeness. Missing data, outliers, and biases can all have an impact on model performance.

In order to improve predictive accuracy/quality and address the limitations of the 3 algorithms launched: linear regression, Decision Tree and Random Forest models were used. To prevent and avoid overfitting in Decision Trees, the Random Forest model was implemented, which generates an ensemble of Decision Trees and averages their predictions. This method increases robustness and predictive accuracy by simply reducing the variance observed in individual Decision Trees. In this study, the Random Forest model outperformed the other models by effectively balancing bias and variance, resulting in improved generalization performance. Thus, near optimal predictions.

The comparison and contrast of the three models had revealed an important insight into their predictive accuracy, quality and reliability:

**Average Predicted Enrollment Rate output:**

The Random Forest algorithm came up with the highest average predicted enrollment rate (0.9304), demonstrating its superior and strong ability to accurately capture enrollment trends.

**Mean Squared Error (MSE):** The Random Forest model got the lowest average MSE (0.1059), followed by Decision Tree (0.1092) and Linear Regression (0.1192). The smaller MSE value operation integrates, the better and near accurate prediction the model provides. Lower MSE values indicate that the Random Forest and Decision Tree models are better at minimizing error variance.

**Mean Absolute Error (MAE):** The Decision Tree got the lowest average MAE (0.1782), followed closely by Random Forest (0.1802) and Linear Regression (0.1888). MAE measures average error magnitude, indicating that the Decision Tree and Random Forest models are more accurate overall.

#### **Yearly predictions and errors outputs:**

For 2023, all models got a low MSE and MAE values, with Random Forest and Decision Tree slightly outperforming Linear Regression. (linear regression got highest error values)

Random Forest consistently got the lowest errors in 2022 and 2020, demonstrating its robustness and lowest bias over the years.

In 2021, despite all models having higher error rates, Decision Tree and Random Forest outperformed Linear Regression similarly with 2023.

Based on a comparison and contrast of average predicted enrollment rates and error metrics (MSE and MAE), the Random Forest algorithm was found to be the best model for predicting enrollment rates in Turkey's foundation universities. Its higher average predicted successfully estimated and targeted enrollment rate, combined with lower MSE and MAE values, demonstrates its superior ability to model complex relationships in the data. Furthermore, the Random Forest model's consistency and accuracy across years reinforces its reliability as a predictive tool.

Additionally, when compared to other models used, decision tree and linear regression, the Random Forest approach was the most effective, lowest MSE, near accurate targeted predictions with higher accuracy and stability in predicting enrollment rates. The Decision Tree model was insightful and interpretable, but it lacked the robustness of the Random Forest and contained low bias. Ensemble methods, particularly the Random Forest, improve predictive performance dramatically by combining the strengths of multiple weak learners. These machine learning models offer valuable and precious insights into the factors that influence student enrollment, allowing universities to tailor their strategies more effectively.

## **5. Conclusion**

The goal of this research is to create a machine learning model that can forecast enrollment rates in Turkey's foundation universities. Universities will be better able to manage student enrollment and institutional resources by using this information to identify enrollment trends and possible bottlenecks. Optimizing predictions of university enrollments, managing resources as efficiently as possible, strengthening decision-making procedures, and utilizing long-term financial and strategic planning capacities are the objectives of the project. Additionally, the model will support hiring staff, developing infrastructure, and organizing student capacity.

Several studies that have forecast enrollment in schools have been thoroughly examined, providing insight into the various methods used in the literature review section. The review showed the implementation of several complex algorithms, including advanced neural networks to traditional regression models, along with various data sets, some comprising studies with a wide range of criteria, from socioeconomic aspects to historical record trends.

Our results reveal that the Random Forest algorithm is the best-performing model for predicting enrollment rates in foundation universities in Turkey, offering higher accuracy and stability. Its durability across different years and ability to model complex relationships within data make it a valuable predictive tool. Methods, particularly Random Forest, enhance predictive performance by leveraging multiple learners, making it recommended for effective educational planning and resource allocation.

## **References**

- Akinode, J. L., and Bada, O., Student Enrollment Prediction using Machine Learning Techniques. Presented applicants at University of Santo Tomas. In *AIP Conference Proceedings* (Vol. 2472, No. 1), 2022.  
at the 5th National Conference of the School of Pure & Applied Sciences Federal Polytechnic, Ilaro, Nigeria,  
Basu, K., Basu, T., Buckmire, R., and Lal, N., Predictive models of student college commitment decisions using

- Calvo, P., Arroyo, J. C. T., and Delima, A. J. P., Higher education institution (HEI) enrollment forecasting using data Canada, J. I., Amorado, R. V., Sarmiento, J. S., and Melo, P. M. B., Machine Learning Estimation for Course Enrollment using Logistic Regression. *IEEE 9th Conference on Systems, Process and Control (ICSPC Four Machine Learning Algorithms (Logistic Regression, Decision Tree, Support Vector Machine, Artificial Kye, A., Comparative Analysis of Classification Performance for US College Enrollment Predictive Modeling Using learning in the context of Private University of Bangladesh. International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 9(01), 2594-2600, 2019.
- machine learning, *Data*, 4(2), 65, 2019.
- Mia, M. J., Biswas, A. A., Sattar, A., and Habib, M. T., Registration status prediction of students using machine mining technique, *International Journal*, 9(2), 2020.
- Neural Network* (Doctoral dissertation, Loyola University Chicago), 2023.
- of Supercomputing*, 77, 11853-11865, 2021.
- Patungan, A. J. and Francia, M. L. M., A machine learning modeling prediction of enrollment among admitted college Prediction, *Strategic enrollment management quarterly*, 10(2), 2022.
- Shao, L., Jeong, M., Levine, R. A., Stronach, J., and Fan, J., Machine Learning Methods for Course Enrollment Shilbayeh, S. and Abonamah, A., Predicting student enrollments and attrition patterns in higher educational institutions Ujkani, B., Minkovska, D., and Stoyanova, L., A machine learning approach for predicting student enrollment in the University, *XXX International Scientific Conference Electronics (ET)* (pp. 1-4), 2021.
- using machine learning, *Int. Arab J. Inf. Technol.*, 18(4), 562-567, 2021.
- Yang, L., Feng, L., Zhang, L. and Tian, L., Predicting freshmen enrollment based on machine learning, *The Journal*

## **Biographies**

**Yasin Göçgün** received his B.S. degree and M.S. degree from the Industrial Engineering Department at Bilkent University in 2003 and 2005, respectively. After completing his doctoral studies in the Industrial and Systems Engineering Department at the University of Washington in 2010, Dr. Gocgun worked as a postdoctoral fellow in Canada between 2010 and 2014. Prior to joining the Industrial Engineering Department at Medipol University, Dr. Gocgun worked as an assistant professor in the Industrial Engineering Department at Istinye University between 2020 and 2022.

**Merve Ece Akat, Esra Akgün, Nerimana Bulut, Esmâ Engin, and Hiba Sadioğlu** received their B.S. degrees from the Industrial Engineering Department at Istanbul Medipol University in June, 2024. They are currently in the job market.