Education and Information Technologies
https://doi.org/10.1007/s10639-021-10655-4

Predicting the percentage of student placement: A comparative study of machine learning algorithms

Erman Çakıt[1] · **Metin Dağdeviren**[1]

Received: 26 April 2021 / Accepted: 24 June 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
In recent years, there has been an increase in the demand for higher education in Turkey, where the demand, as in most other countries, exceeds what is available. The main purpose of this research is to develop machine learning algorithms for predicting the percentage of student placement based on the data related to the university's academic reputation, opportunities of the city where the university is located, facilities and cultural opportunities of the university. When the model accuracy was evaluated on the basis of performance metrics, the Extreme Gradient Boosting (XGBoost) algorithm showed greater predictive accuracy than other machine learning approaches. A sensitivity analysis was performed using the extreme gradient boosting machines algorithm to identify the degree to which the input variables contribute to the determination of the output variable. Five input variables, namely the percentage of student placement at year *t-1,* the university scientific document score, university Phd programme score, university faculty member/student score, and the percentage of student placement at year *t-2* were found to be the most effective parameters. Prediction and sensitivity analysis results obtained in this study can be used in many different ways, such as determining the quotas for universities, allocating resources, and making new regulations.

Keywords Student placement · Machine learning · Prediction · University quotas · Higher education

1 Introduction

Students in many nations have the opportunity to choose the university which they want to study (Bringula & Basa, 2011). There are some challenges aiming to attend a university. In addition to getting a high school diploma, students are expected to

✉ Erman Çakıt
ecakit@gazi.edu.tr; ecakit@gmail.com

[1] Department of Industrial Engineering, Gazi University, 06570 Ankara, Turkey

receive preferred scores from the general central exams every year in order to pursue higher education and choose their preferred higher education institution. In recent years, there has been an increase in the demand for higher education in Turkey, where the demand, as in most other countries, exceeds what is available.

In each year, the Measurement, Selection and Placement Center announced that almost two million individuals have taken the student selection examination over the last 10 years (Esen, 2019). Over the years, there are many reasons for increasing the number of students who wish to participate in higher education. Firstly, there has been a rise in the number of high school graduates, especially after compulsory primary education was implemented for a period of eight years in 1997. Secondly, the pool of applicants is growing cumulatively, including those from previous years who could not do well in their initial entrance tests. Lastly, a large number of students who managed to be accepted to an academic programme re-take the entrance test many times in order to change their academic programme. Each year, these three candidate groups reflect a significant "snowball effect." Table 1 summarizes the student placement rate between 1980 and 2020. Recently, public and particularly private universities in Turkey have seen substantial increases; thus, universities face fierce competition.

As of 2021, Turkey has two hundred and nine higher education institutions in different cities, one hundred and thirty-one of them are public universities and the other seventy-eight are private universities. This increase drives universities into a competitive environment and causes them to transfer resources to many areas. In addition, universities are looking for the means to attract more students in response to the increasingly competitive environment. For this purpose, this research performs the application of machine learning algorithms that directly benefit from their accuracy in estimation models to investigate the effects of parameters grouped according to higher education institutions' attractiveness on the percentage of student placement.

Table 1 Student placement rate between 1980 and 2020 (Olcme, Secme ve Yerlestirme Merkezi (OSYM), 2020)

Year	Total Number of Applicants	Total Number of Placement	Rate (%)
1980	466,963	41,574	8.9
1990	892,975	196,253	22.0
2000	1,407,920	414,647	29.5
2010	1,587,866	763,516	48.1
2011	1,759,403	789,169	44.9
2012	1,895,479	865,482	45.7
2013	1,924,550	877,784	45.6
2014	2,086,115	922,275	44.2
2015	2,126,670	983,090	46.2
2016	2,256,377	961,864	42.6
2017	2,265,844	825,397	36.4
2018	2,381,412	857,240	36.0
2019	2,528,031	904,176	35.8
2020	2,436,958	921,886	37.8

In addition to the nonlinear machine learning algorithms, we included linear models including regression analysis results to assess the predictive accuracy of the nonlinear machine learning approaches in comparison to linear models.

The remainder of this article is structured as follows. A summary of the literature is given in Section 2. The study design and a detailed introduction to the use of machine learning algorithms are presented in Section 3. Section 4 outlines the application of the selected machine learning algorithms and presents a comparative discussion of the results. Finally, the conclusions and suggestions for future research are discussed in Sections 5.

2 Background

Higher education is defined as a commodity from a marketing point of view with intangibility features, concurrence, and diversity (Brown et al., 2009; Simões & Soares, 2010). University selection is typically a one-off choice that leads to an extended consumption experience of four to five years, and the decision will affect the career of the "consumers" over a longer period (Moogan et al., 1999; Walsh et al., 2015). Park and Kim (2020) have recently identified the factors affecting high school graduates' career decision-making on their choice of university (re-) entrance in South Korea. Based on the findings of the study, new enrollees were primarily affected by the relationship between parents and re-enrollees, while social consciousness macrosystems, as well as human skill and capacity estimates of individual structures, have affected re-enrollees. Recently, Mostafa and Beshir (2021) conducted a university selection model that classified candidates' opinions' based on the candidate's ecological system, whether individual, microsystems, or ecosystem. Pal and Pal (2013) proposed model based on classification approach to determine the relations between academic achievement of students and their placement in campus selection. Manvitha and Swaroopa (2019) aimed to predict whether the student he/she gets placed or not in campus recruitment. They used the data that includes the academic history of student like overall percentage, backlogs, credits. Guleria and Sood (2015) applied Bayesian theorem on educational dataset including attendance, GPA, reasoning, quantitative, communication and technical skills to predict placement results.

The academic reputation of the university tends to be a strong influence on the students' university choices (Briggs, 2006). Academic reputation (Dao & Thorpe, 2015; Simões & Soares, 2010; Briggs, 2006; Veloutsou et al., 2004; Soutar & Turner, 2002), programs offered (Pampaloni, 2010; Veloutsou et al., 2004; Moogan et al., 1999), academicians' profiles (Bringula & Basa, 2011), and reputations of the departments (Simões & Soares, 2010) are among the university image attractors related to the university's quality. These variables are among the quality components of attracting students to universities, such as university's academic curriculum, personnel, education and facilities for the student, scientific research performed at the university, and graduate qualifications. In 2009, the University Ranking and Academic Performance (URAP) (2020) Research Laboratory created the URAP-TR

(The University Ranking by Academic Performance in Turkey) ranking framework for ranked Turkish universities.

The ranking is based on the data obtained from both national and international databases, such as Web of Science, The Council of Higher Education, and Measurement, Selection and Placement Center. The level of research carried out by Turkish higher education institutions is measured by metrics in terms of the number of articles, the number of citation records, articles per staff, citations per staff, the total number of publications, the number of doctoral students studying, the proportion of doctoral students to the total number of students, and the ratio between the total number of students and the total number of academic staff (Kalkan et al., 2015).

On the other hand, some scholars have discussed the cultural opportunities of the university. The attractiveness of social life in the university (Briggs, 2006; Capraro et al., 2004) and the cultural life of the university (Ramasubramanian et al., 2002) are being considered by potential university students. Also, university student clubs (Veloutsou et al., 2004) and university sports facilities (Tatar & Oktay, 2006) have been addressed by other studies as affecting the decision-making of potential university students. Some other parameters affect university preferences. For instance, the city where the university is located provides opportunities that affect students' university interests (Absher & Crawford, 1996; Anılan et al., 2008; Briggs, 2006).

Although higher education institutions' attractiveness within many categories of university selection has been well studied, there is a dearth of information in modeling the relationship between these categories and the percentage of student placement at universities. To our best knowledge, no studies have been conducted on the use of any machine-learning algorithm particularly for this purpose in estimating the percentage of student placement and identifying the degree to which parameters contribute to the determination of the percentage of student placement. This fact provided the impetus for our present investigation. The present research will therefore lead to the field of higher education.

3 Materials and methods

3.1 Dataset used

The variables used in this study were determined based on the literature review in section 2. The related data were collected using the sources at The Council of Higher Education (Yuksek Ogretim Kurulu (YOK), (2020)) and the University Ranking and Academic Performance (URAP) (2020). In this study, we considered five engineering departments in public universities including computer, industrial, electrical and electronics, mechanical, and civil engineering departments.

For this purpose, in the scope of academic reputation; the percentage of student placement at year *t-1* (i.e., the previous year), *t-2* (two years ago), *t-3* (three years ago), and *t-4* (four years ago), university foundation year, the number of incoming and outgoing erasmus students, total number of faculty members in the departments, total number of faculty members at university, education language, programme accreditation, international diploma supplement, number of labs at university,

university article scores, university citation scores, university scientific document scores, university Phd programme scores, university faculty member/student scores, and university total performance scores were considered. In the scope of the opportunities of the city where the university is located, the number of movie theaters, the number of theater halls, and the number of museums were considered. In the scope of cultural opportunities and facilities of the university, the number of sports facilities, the number of student societies, accommodation oppurtunities were considered.

All these variables were used to estimate the percentage of student placement in year *t*. Thus, the data included a total of 29 inputs and 1 output reflecting student placement, and the corresponding format of the data is shown in Table 2.

Table 2 Input and output variables used in model development

Input variable (29 Inputs)	Output variable
University foundation year	The percentage of student placement at year *t*
Number of incoming erasmus students (2017–2018)	
Number of incoming erasmus students (2018–2019)	
Number of outgoing erasmus students (2017–2018)	
Number of outgoing erasmus students (2018–2019)	
Total number of faculty members at department	
Total number of faculty members at university	
Education language (100% Turkish)	
Education language (30% Turkish)	
Education language (100% English)	
Is a program accredited? (Yes = 1, No = 0)	
University article score	
University citation score	
University scientific document score	
University Phd programme score	
University faculty member/student score	
University total performance score	
Number of movie theater where the university is located at	
Number of theater hall where the university is located at	
Number of museums where the university is located at	
Is there an international diploma supplement? (Yes = 1, No = 0)	
Number of sports facilities at university	
Number of student societies at university	
Is there a dorm opportunity? (Yes = 1, No = 0)	
Number of lab at university	
The percentage of student placement at year *t-1*	
The percentage of student placement at year *t-2*	
The percentage of student placement at year *t-3*	
The percentage of student placement at year *t-4*	

Of the total 523 records, 209 records had missing data in them; hence, they were removed. The remaining 314 records were grouped for training using 80% of the data (251 records) and testing using the remaining 20% of the data (63 records). Various programming languages can be used for machine learning such as Java, Python, R, JavaScript, Scala etc., but for the purpose of this study, we performed Python programming language. Among data scientists and software developers, the Python programming language is becoming very popular (Robinson, 2017). All data analyses for each model were carried out with Python language, version 3.4. The steps in this study were carried out in the order shown in Fig. 1.

3.2 Machine learning algorithms

Machine learning requires a series of methodologies where computers are able to "learn" the relation between numerical data representations and certain target values (Hao & Ho, 2019).

Machine learning is broadly classified into supervised (trains known inputs and outputs to predict future outputs) and unsupervised learning (finding patterns or structures in the input data) (Swamynathan, 2019) (Fig. 2). In various implementations, machine learning techniques have been commonly used, and these approaches address a broad range of subjects that far exceed the scope of this research. A variety of outstanding texts provide a thorough coverage of machine learning (Marsland, 2015; Mohri et al., 2018; Alpaydin, 2020; Bishop, 2006; Witten and Frank, 2002).

The machine learning algorithms used in our study were linear *("multiple linear regression, ridge regression, lasso regression, and elastic net regression"*) and

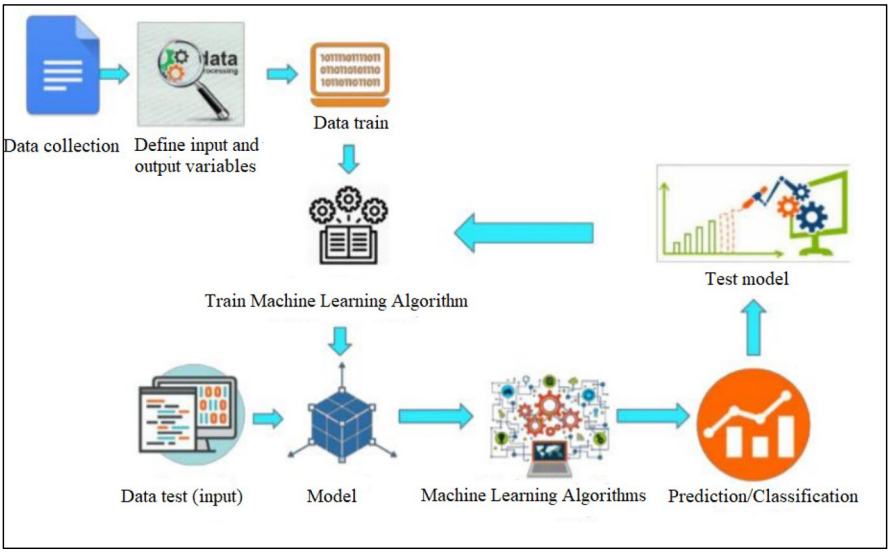


Fig. 1 Study steps for machine learning algorithms(Adapted from Swamynathan, 2019)

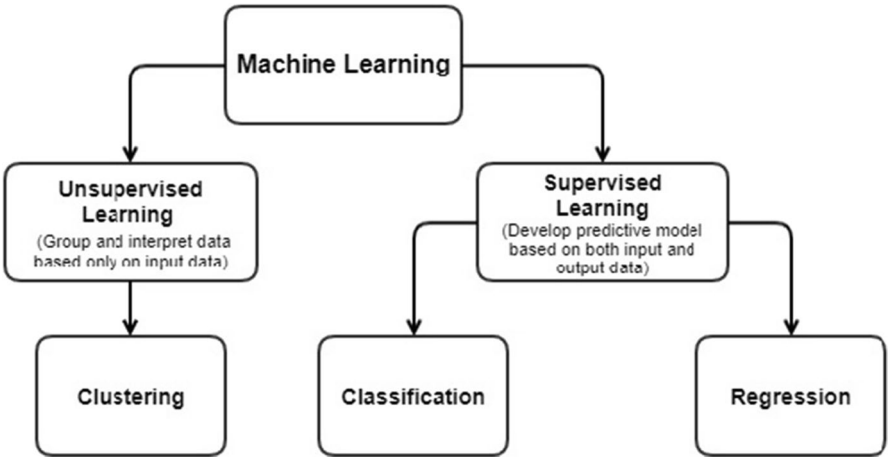


Fig. 2 Types of Machine Learning

nonlinear models (*"k-nearest neighbors, random forests, artificial neural network, support vector regression, classification and regression tree (CART), gradient boosting machines (GBM), extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), and categorical boosting (CatBoost)").* These algorithms are among the most commonly used algorithms in recent literature. Each one of them reflects a separate "family" algorithm, each with fundamentally different internal algorithm architectures (Fernández-Delgado et al., 2014). All models were written in Python/Jupyter Notebook, and we used scikit-learn packages to incorporate all of the machine learning algorithms used in this study (Scikit-Learn: Machine Learning in Python, 2021).

3.3 Performance metrics

In order to calculate the difference between observed and predicted values (error) in the model, various performance metrics were applied (Çakıt et al., 2020; Çakıt & Karwowski, 2017a). In this analysis, in order to measure the efficiency of machine learning algorithms, the accuracy of the model was evaluated on the basis of root mean square error (RMSE), mean square error (MSE), mean absolute error (MAE), and coefficient of determination (R^2) values between the predicted and observed values. The lower RMSE, MSE, and MAE values represent more accurate model results. The higher R^2 values represent a better fit between the observed and estimated values. The following equations were used for these calculations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i - A_i\right)^2} \tag{1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(P_i - A_i\right)^2 \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|e_i\right| \tag{3}$$

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n}\left(P_i - A_i\right)^2}{\sum_{i=1}^{n} A_i^2}\right) \tag{4}$$

where A_i and P_i were the actual and predicted values, respectively,

e_i is the prediction error for each record.
n total number of testing records.

$$i = 1, 2, 3, \ldots\ldots., n \tag{5}$$

4 Results and discussion

4.1 Model results

4.1.1 Multiple regression analysis

A multiple regression analysis determines the relationship between variables such as the dependent variable and other independent variables. In particular, multiple regression analysis aims to understand changes in dependent variable y through movements in k (independent) explanatory variables $x_1, x_2, \ldots\ldots\ldots, x_k$.

The general form of the multiple regression analysis is expressed in the following way:

$$\begin{gathered} y_i = f\left(x_{i1}, x_{i2}, \ldots\ldots\ldots., x_{ik}\right) + \in_i \\ y_i = \beta_0 + \beta_1\, x_{il} + \beta_2\, x_{i2} + \ldots\ldots\ldots. + \beta_k\, x_{ik} + \in_i \end{gathered} \tag{6}$$

where β_k are parameters, x_{ik} are known constants, $\in_i$ is the residual, and n is equal to the number of predictors. The RMSE, MSE, MAE, and R^2 statistics for the developed multiple regression model were provided respectively, and model outputs have demonstrated that the real values are closely followed. (Fig. 3).

4.1.2 Ridge regression

Ridge regression is a conventional linear regression, using an error function and a regularization technique in the sum of squares to monitor the bias variance interchange, with the goal of detecting the linear structures embedded in the data. (Zhang

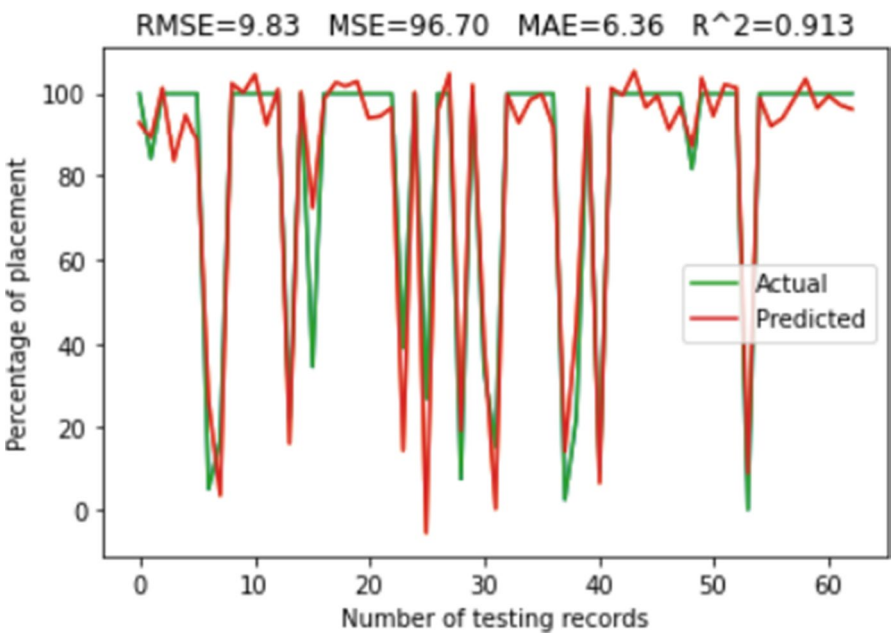


Fig. 3 Comparison of the actual and proposed Multiple regression outputs

et al., 2015). Ridge regression implements *L2* regularization, which penalizes the coefficients by including the square of the magnitude of the coefficient to the cost function:

$$J = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\theta_j \right)^2 + \lambda \sum_{j=1}^{p} \theta_j^2 \tag{7}$$

where $\lambda \sum_{j=1}^{p} \theta_j^2$ is the regularization component, and λ is a regularization factor, which can be optimized by examining validation error. The comparison of predicted and ridge regression output is illustrated in Fig. 4. The ridge regression outputs have demonstrated that the real values are closely followed.

4.1.3 Least absolute shrinkage and selection operator (LASSO) regression

The LASSO regression is presented by Tibshirani (1996), and it is a modification of linear regression. Similar to ridge regression, LASSO regression model penalizes the magnitude of coefficients to avoid overfitting (Vrontos et al., 2021; Qin et al., 2018). The LASSO regression performs *L1* regularization:

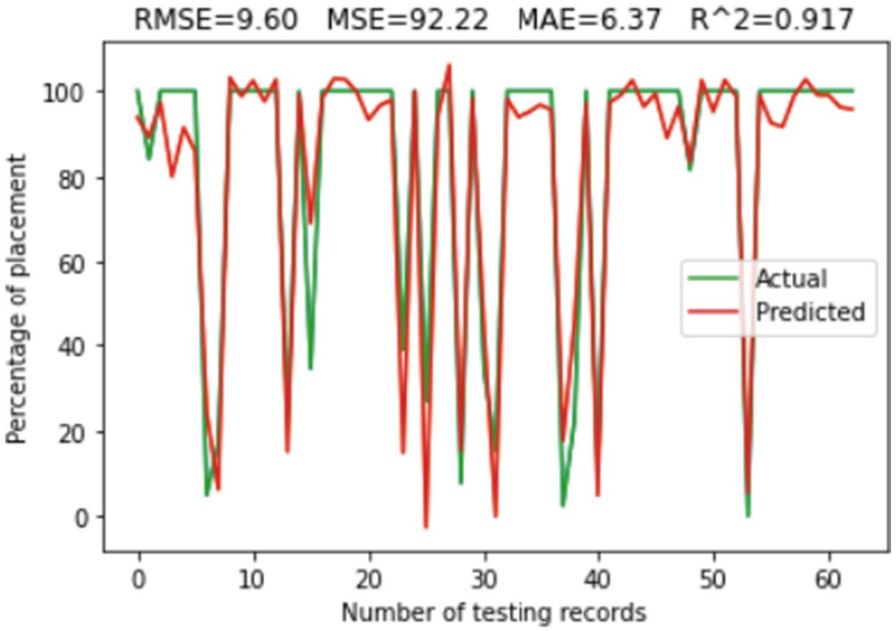


Fig. 4 Comparison of the actual and proposed Ridge regression outputs

$$J = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\theta_j \right)^2 + \lambda \sum_{j=1}^{p} \left|\theta_j\right| \tag{8}$$

where $\lambda \sum_{j=1}^{p} \left|\theta_j\right|$ is the regularization component, which is a summation of the absolute values of the feature coefficients. The RMSE, MSE, MAE, and R^2 statistics for the developed lasso regression model were provided respectively, and the lasso regression model outputs have demonstrated that the real values are closely followed (Fig. 5).

4.1.4 Elastic-net regression

The elastic-net algorithm linearly combines the penalties of the ridge and lasso methods (Richardson et al., 2021; Zou & Hastie, 2005). The elastic-net algorithm performs a fading between the lasso and ridge algorithms through a parameter $0 \leq \alpha \leq 1$. For $\alpha = 1$; the elastic-net is equal to the lasso algorithm, and for $\alpha = 0$ it is equal to the ridge algorithm. It is a compromise between the two approaches when $0 < \alpha < 1$. Thus, elastic-net algorithm generalizes the ridge and lasso methods. Based on the results obtained, the RMSE, MSE, MAE, and R^2 values for the developed elastic-net regression model were noted to be 9.64, 93.08, 6.33, and 0.916, respectively (Fig. 6).

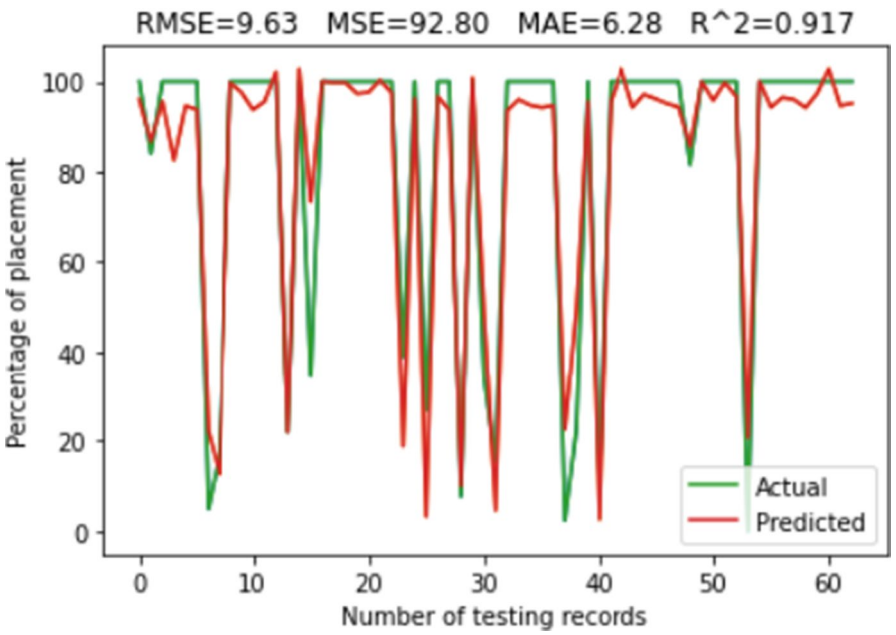


Fig. 5 Comparison of the actual and proposed Lasso regression outputs

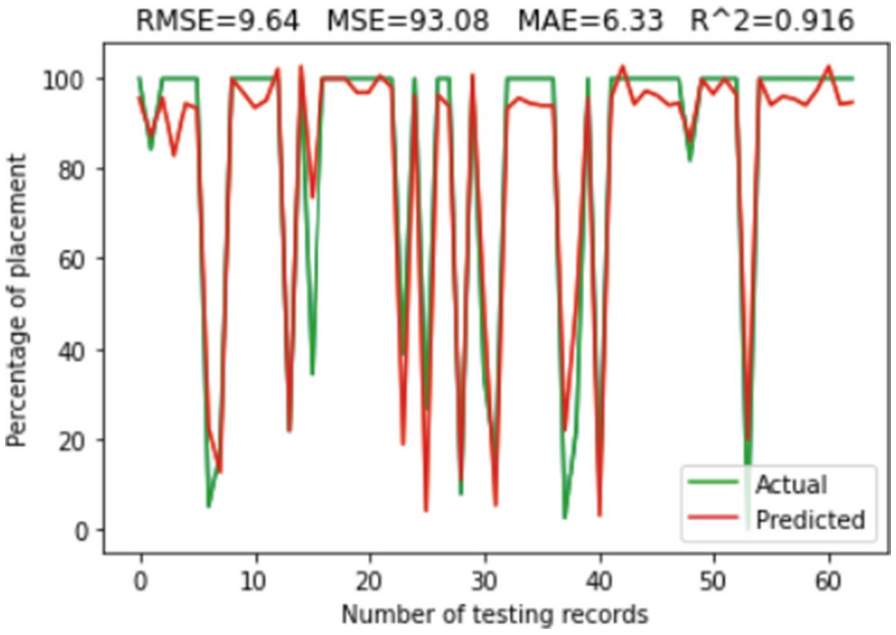


Fig. 6 Comparison of the actual and proposed Elastic Net regression outputs

4.1.5 Random forests

The Random Forest method was introduced by Breiman (2001) as a machine learning algorithm (Iatrellis et al., 2021; Pliakos et al., 2019; Beaulac & Rosenthal, 2019). It has been applied as integrated statistical learning classification and regression algorithm. The output is the average of the random forest resulting from all decision tree. Assume that the training set is derived seperately from the random vector distribution. The results of the model estimation are the mean of k regression trees. Three main parameters for random forest algorithms are the number of trees, the minimum number of observations in the terminal node, the number of suitable features for splitting. The random forests can be found in the literature with extensive mathematical explanations (Breiman, 2001; Genuer et al., 2017). Based on the results obtained, the RMSE, MSE, MAE, and R^2 values for the developed random forests model were noted to be 7.95, 63.31, 4.60, and 0.944, respectively (Fig. 7).

4.1.6 K-nearest neighbors

K-nearest neighbors *(K*NN) is a technique of non-parametric pattern recognition that utilizes the average of the nearest *K* observations in the training set to provide an approximation. *K*NN can be applied to classification (Cover & Hart., 2006; Dudani, 1976) and regression problems (Yousafzai et al., 2020). In order to

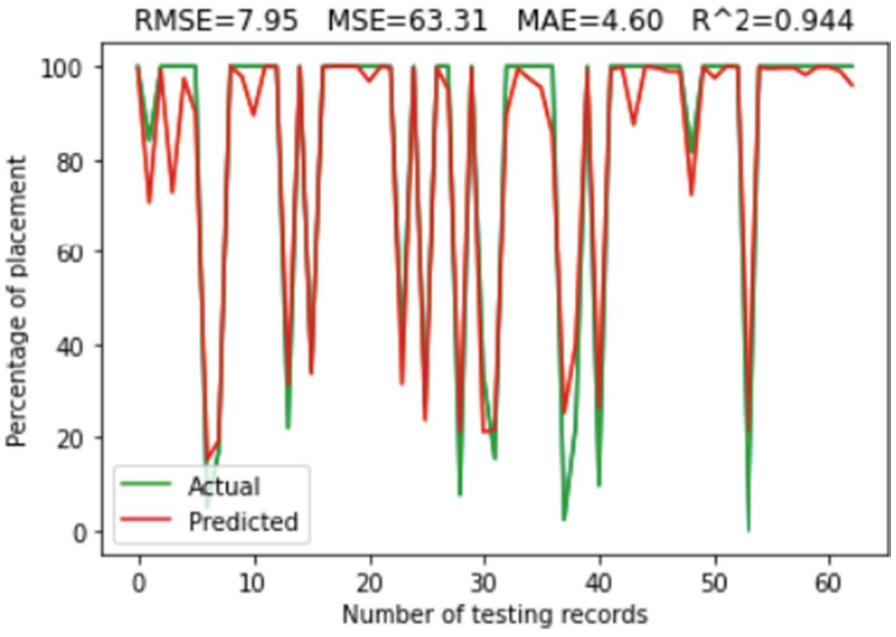


Fig. 7 Comparison of the actual and proposed RF outputs

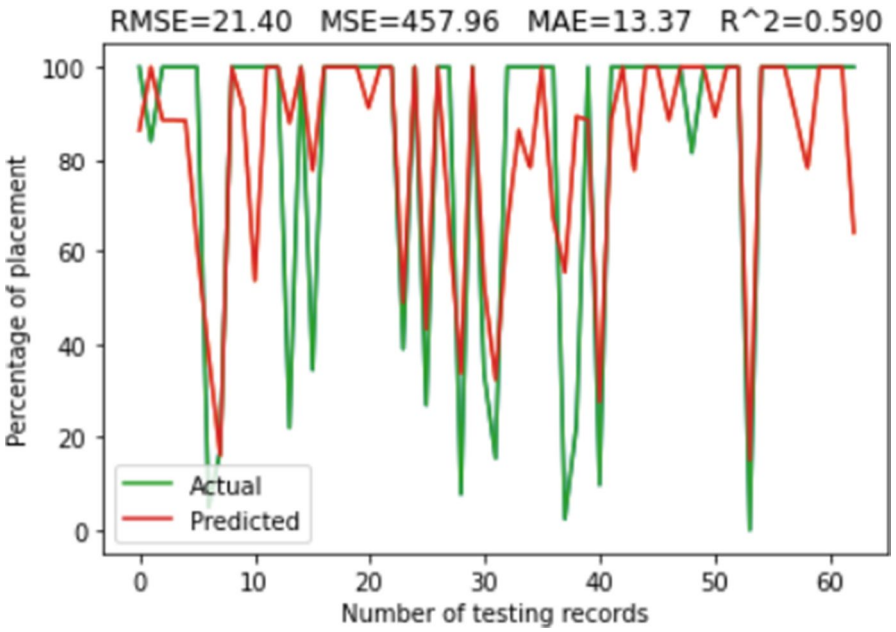


Fig. 8 Comparison of the actual and proposed KNN outputs

determine whether an observation is close in the feature space, a distance metric is applied. In addition to the distance metric, *K*, you can freely pick the number of neighbors that are considered to approximate the response variable. One technique for finding the optimum value of *K* is to compare various subsets of training results. Based on the results obtained, the RMSE, MSE, MAE, and R^2 values for the constructed lightgbm model were noted to be 21.40, 457.96, 13.37, and 0.590, respectively (Fig. 8). The *K*NN model outputs have demonstrated that the real values are closely followed.

4.1.7 Artificial neural networks (ANNs)

Artificial neural networks (ANNs) are simple mathematical models of the brain that process information and make predictions (Mason et al., 2018). They basically do the non-linear mapping or identification of patterns. ANNs help to model complex and non-linear structures (Çakıt et al., 2014; Çakit et al., 2015; Çakıt & Karwowski, 2017b). Many different resources are available in the literature for further explanations of ANNs (Zurada, 1992; Gurney, 1997; Fausett, 2006; Haykin, 2007). A key decision for precise prediction is the proper choice of neural network architecture. (Sheela & Deepa, 2013). Multilayer Perceptron (MLP) is a feed-forward multi-layer network architecture that uses the principle of back propagation.

In the present study, we performed the MLP from the scikit-learn library. The parameters were selected as follows:*("alpha=0.02, beta_1=0.9, beta_2=0.999, epsilon=1e-08, hidden layer sizes=(100,100), learning_rate_init = 0.001, max_iter*

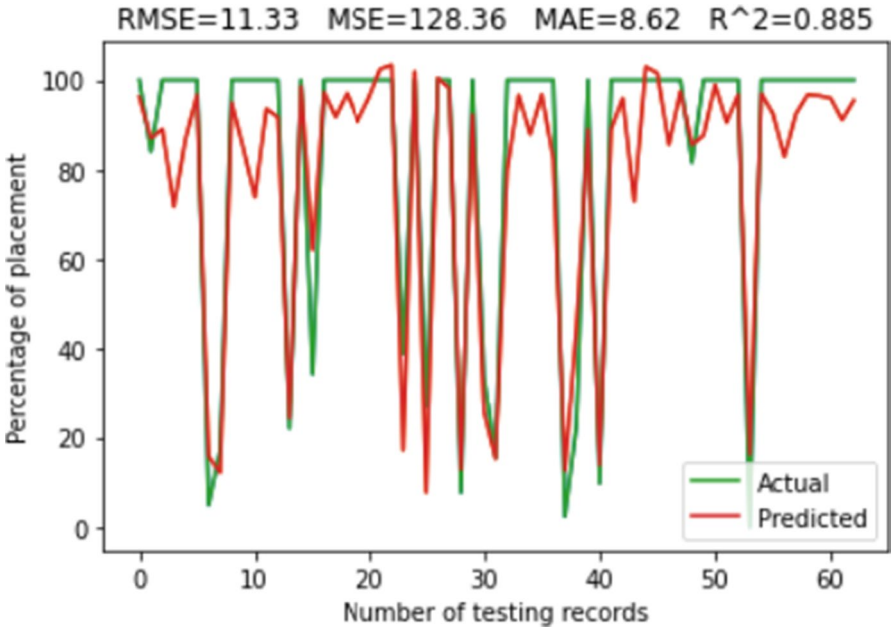


Fig. 9 Comparison of the actual and proposed ANN outputs

= 200, momentum=0.9, n_iter_no_change = 10,"). Based on the results obtained, the RMSE, MSE, and MAE values for the developed ANN model were noted to be 11.33, 128.36, and 8.62, respectively. R^2 (0.885) is another clear indication that supports the below inference (Fig. 9).

4.1.8 Support vector regression

The support vector regression is the support vector machine application for regression analysis (Smola & Schölkopf, 2004; Lykourentzou et al., 2009). Vapnik's ε-insensitive cost function is used as the standard support vector regression formulation:

$$\zeta_{\varepsilon}(e) = Cmax(0, |e| - \varepsilon), \qquad C > 0, \tag{9}$$

in which an error $e = y - \bar{y}$ up to ε is not penalized, otherwise it will incur in a linear penalization. Three hyperparameters have to be specified in the support vector regression model: the C penalization factor, the insensitive zone and the kernel parameter. It is not easy to tune all these parameters. In this study, we applied a cross-validation procedure to do this. More discussions on the support vector regression can be found on (Drucker et al., 1996). In this study, the support vector regression model is imported from support vector machines class of sklearn python library, and the model parameters were selected as (C = 0.1, ε = 0.1, and kernel = 'rbf'). Based on the results obtained, the RMSE, MSE, MAE, and R^2 values

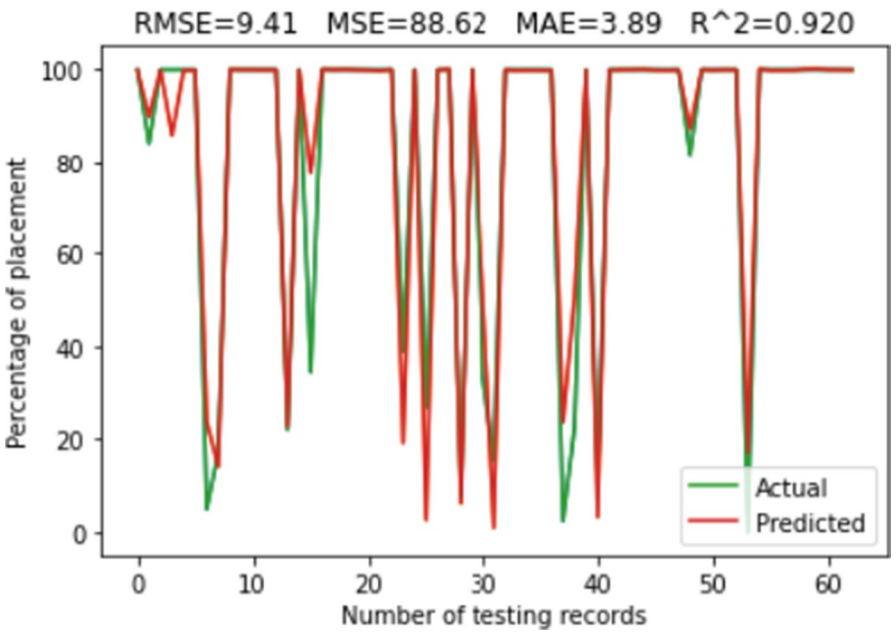


Fig. 10 Comparison of the actual and proposed SVM outputs

for the developed support vector regression model were noted to be 9.41, 8.62, 3.89, and 0.920, respectively (Fig. 10).

4.1.9 Classification and regression tree (CART)

CART model was developed by Breiman et al. (1984). Based on the data type, the algorithm can be applied as classification tree as well as regression tree. This algorithm aims to find the best possible split, and it can handle huge continuous variables. In order to find the best split point, the procedure divides the test. Assuming that "S" is the call split point, then all the cases with value below "S" go to the left, forming the child node "(X < S)". Otherwise, instances are sent to the right child node "(X > S)". When the perfect set of indices has been completed, the best forecast results can be obtained. The RMSE, MSE, MAE, and R^2 statistics for the developed CART algorithm were provided, respectively and the CART algorithm outputs have demonstrated that the real values are closely followed (Fig. 11).

4.1.10 Gradient boosting machines (GBM)

GBM algorithm was developed by Friedman (2001). The key heuristic behind the approach is the development of new base-learners to be as strongly correlated as possible with the negative gradient of the loss function associated with the whole

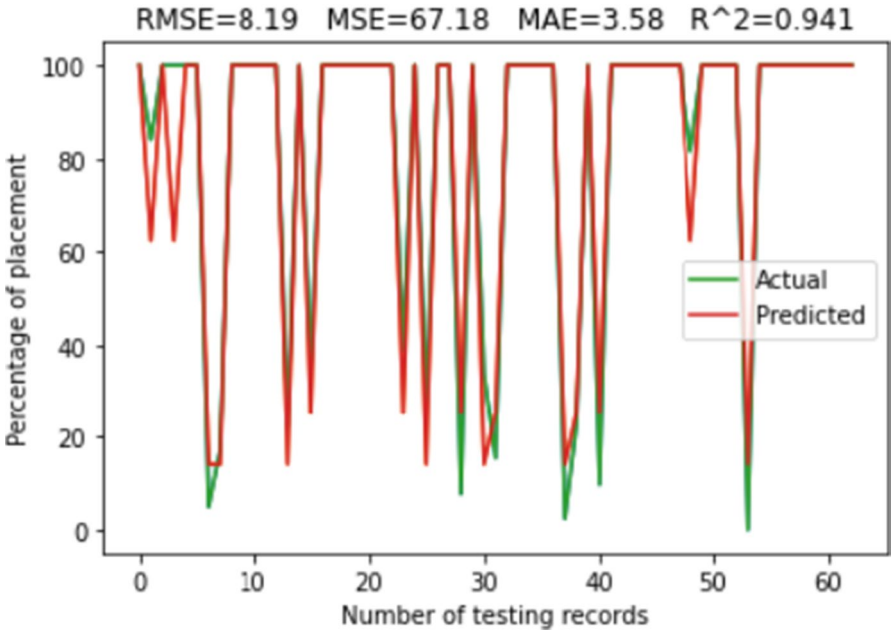


Fig. 11 Comparison of the actual and proposed CART outputs

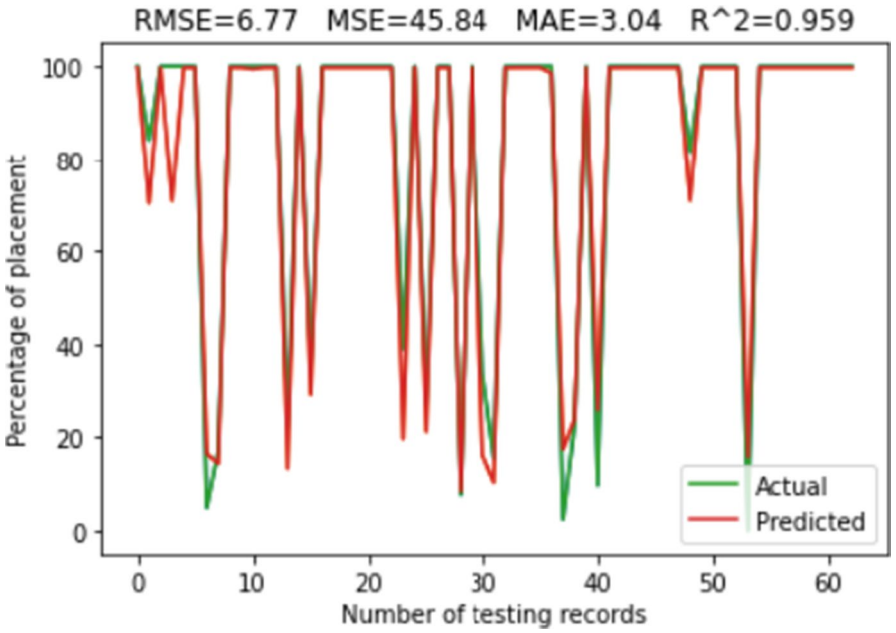


Fig. 12 Comparison of the actual and proposed GBM outputs

ensemble (Lagrari et al., 2018). The GBM algorithm provides numerous hyperparameters that should, and perhaps must, be tuned for a specific dataset.

In this research, we identified the generic parameters used in the development of the "gbm" model. The following parameters were selected as best parameters for our model development: ("learning rate": 0.01, "loss": "ls", "max depth": 3, "n estimators": 500, "subsample": 1). Default settings were left for the remaining parameters. We then used assigned parameters to initiate the "gbm" development. The RMSE, MSE, MAE, and R^2 statistics for the developed gbm model were provided respectively, and the GBM model outputs have demonstrated that the real values are closely followed (Fig. 12).

4.1.11 Extreme gradient boosting (XGBoost)

XGBoost algorithm (Chen & Guestrin, 2016) follows the principle of gradient boosting machine. XGBoost needs several prediction parameters, but model success often depends on the best parameter combination. The way the XGBoost algorithm works is as follows: assume a dataset (DS) that has m features and an n number of instances $DS = \{(x_i, y_i) : i = 1 \ldots\ldots n, x_i \epsilon \mathbb{R}^m, y_i \epsilon \mathbb{R}\}$. We must find the best combination of functions by minimizing the loss and regularisation objective.

$$\mathcal{L}(\phi) = \sum_i l\left(y_i, \phi\left(x_l\right)\right) + \sum_k \Omega\left(f_k\right) \tag{10}$$

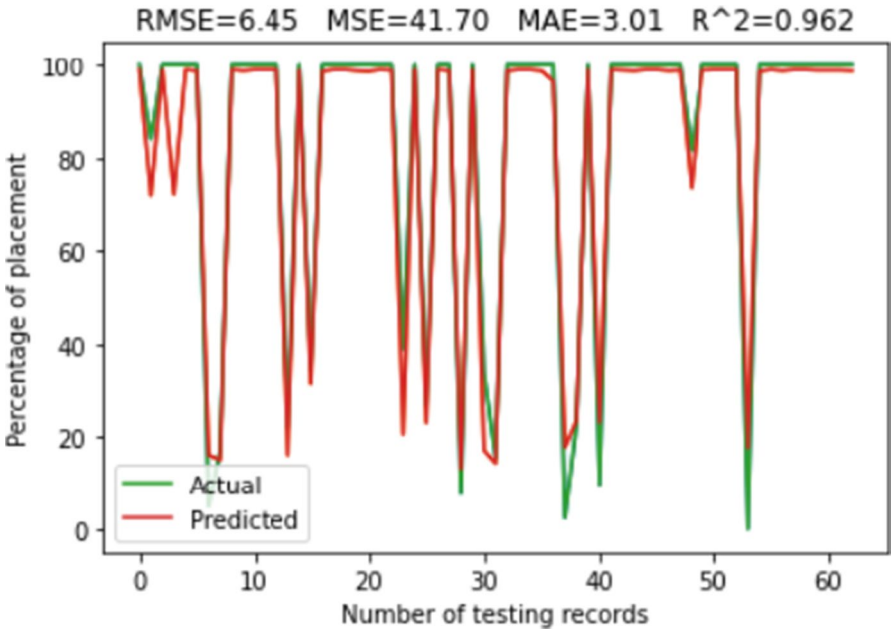


Fig. 13 Comparison of the actual and proposed XGBoost outputs

where l represents the loss function (the difference between the predicted output $\hat{y}_i$ and the observed output y_i), f_k represents the (k-th tree), to solve the above equation, while Ω is a measure of how complex the model is, this tends to prevent model over-fitting.

We applied 10-fold cross-validation for parameter tuning. In this study, we initiated the "xgb" development with the assigned parameters. The developed "xgb" model was selected based on the best parameters ("colsample bytree": 1, "learning rate": 0.01, "max depth": 2, "n estimators": 500). Based on the results obtained, the RMSE, MSE, MAE, and R^2 values for the developed xgboost model were noted to be 6.45, 41.70, 3.01, and 0.962 respectively (Fig. 13).

4.1.12 Light gradient boosting machine (LightGBM)

LightGBM is a high efficiency, fast open source tool that is utilized for ranking, classification and regression problems (Ke et al., 2017). For decision tree learning, the LightGBM uses histograph-based algorithms to reduce computational complexity (Wang & Liu, 2020). LightGBM has two approaches, namely Gradient-based one-side sampling (GOSS) and leaf-wise growth for training and splitting the data (Chen et al., 2019). Several model parameters (boosting type, max depth, learning rate, and number of leaves) need to be tuned for better prediction results (Ustuner & Balik Sanli, 2019).

For model development, we initiated the "lightgbm" development with the assigned parameters. In this study, we selected GOSS as a boosting type. The

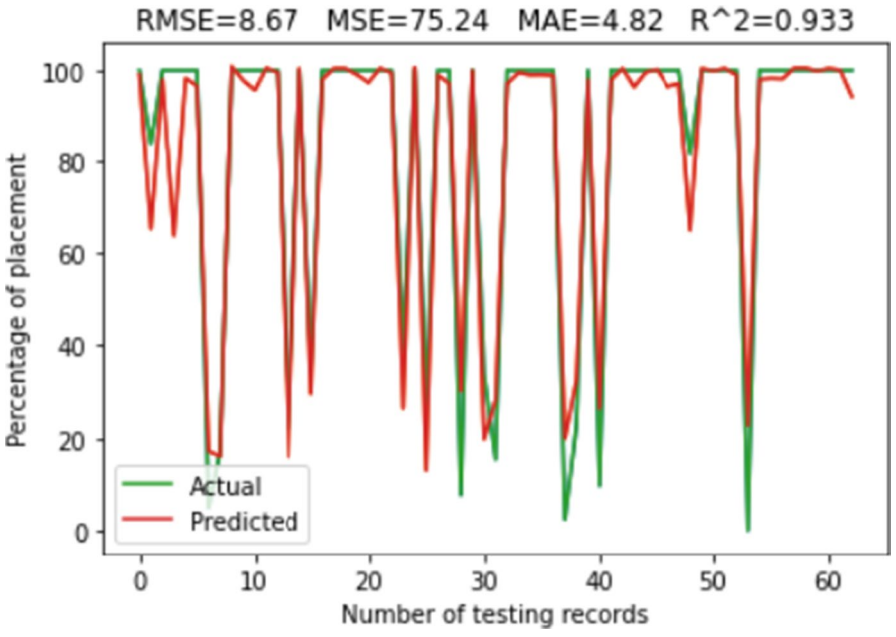


Fig. 14 Comparison of the actual and proposed LightGBM outputs

parameters were tuned by the trial and error method, and the developed "lightgbm" model was selected based on the best parameters ("learning rate":0.1, "max depth":2, "n_estimators":40). Based on the results obtained, the RMSE, MSE, MAE, and R^2 values for the developed "lightgbm" model were noted to be 8.67, 75.27, 4.82, and 0.933 respectively (Fig. 14).

4.1.13 Categorical boosting

Categorical boosting (CatBoost) (Prokhorenkova et al., 2018) applies gradient boosting on decision trees. Compared to the other machine learning algorithms, the algorithm can accommodate multiple data types such as categorical features and requires a small data training. One of the main benefits of this approach is that overfitting is minimized. Many different resources are available in the literature for further explanations of CatBoost algorithm (Azizi & Hu, 2019).

Several model parameters (the learning rate, the maximum depth of the tree, etc.) were tuned for CatBoost algorithm. In this study, we performed the "CatBoost" algorithm with the assigned parameters. The developed "CatBoost" model was selected based on the best parameters ("depth": 3, "iterations": 500, "learning rate": 0.01). Based on the results obtained, the RMSE, MSE, MAE, and R^2 values for the developed CatBoost model were noted to be 6.96, 48.43, 3.59, and 0.956, respectively (Fig. 15).

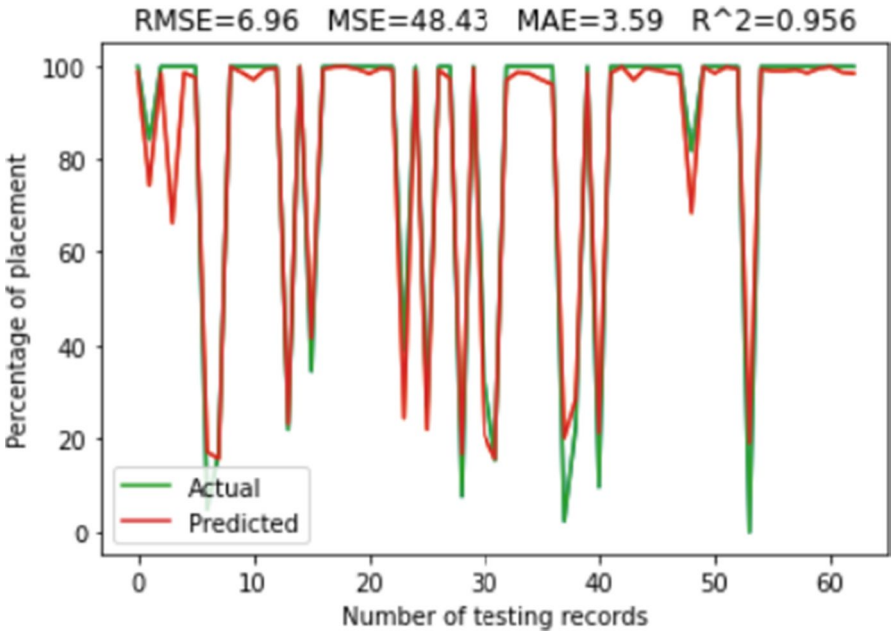


Fig. 15 Comparison of the actual and proposed CatBoost outputs

Table 3 Performance comparison of machine learning algorithms

Testing data ($n=63$)				
	RMSE	MSE	MAE	R^2
Multiple Regression	9.83	96.70	6.36	0.913
Ridge Regression	9.60	92.22	6.37	0.917
Lasso Regression	9.63	92.80	6.28	0.917
Elastic Net Regression	9.64	93.08	6.33	0.916
Random Forests	7.95	63.31	4.60	0.944
K-nearest Neighbors	21.40	457.96	13.37	0.590
Artificial Neural Network	11.33	128.36	8.62	0.885
Support Vector Regression	9.41	88.62	3.89	0.920
Classification and Regression Tree	8.19	67.18	3.58	0.941
Gradient Boosting Machines	6.77	45.84	3.04	0.959
Extreme gradient boosting	**6.45**	**41.70**	**3.01**	**0.962**
Light gradient boosting machine	8.67	75.24	4.82	0.933
Categorical Boosting	6.96	48.43	3.59	0.956

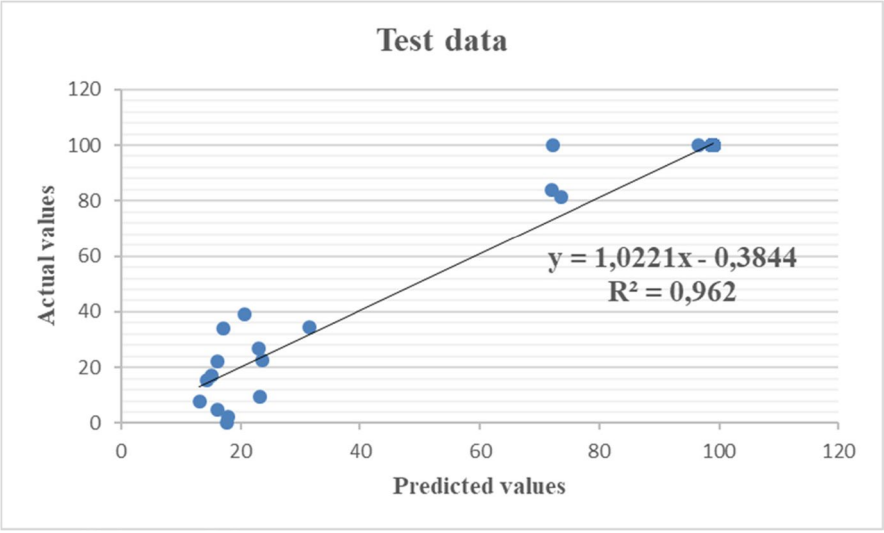


Fig. 16 Extreme gradient boosting predicted and actual values

4.2 Performance comparison of algorithms

To assess the effectiveness of modeling approaches in predicting the percentage of student placement and to identify the most accurate approach among the machine learning algorithms applied in this study, performance measurements were used to compare the algorithms on the same basis. Based on the performance metrics comparison, the extreme gradient boosting machines algorithm demonstrated better predictive accuracy than other machine learning approaches

for predicting the percentage of student placement (Table 3). The comparison of predicted and extreme gradient boosting output is shown in Fig. 16. According to the Fig. 16, the extreme gradient boosting outputs closely follow the actual values. The finding from pairwise $t-$ test analysis indicates that there was no significant difference at the alpha value of 0.05 ($p=0.083$) between observed and predicted student placement testing data based on the extreme gradient boosting algorithm.

Just following the extreme gradient boosting machines algorithm, and focusing on RMSE results, gradient boosting machines algorithm comes in second concerning performance. Random forests algorithm ranks in fourth, surpassed by the categorical boosting algorithm. The.

classification and regression tree algorithm was unable to exceed random forests performance. The performance of most models is quite close, which implies that the machine learning algorithms on those platforms were able to identify the extent of the hidden relationship between the percentage of student placement and the data related to the university's academic reputation, opportunities of the city where the university is located, facilities and cultural opportunities of the university. The results of the K-nearest Neighbors machine learning algorithm carried out in this study show that there is a relationship between the percentage of student placement and the parameters grouped according to higher education institutions' attractiveness but this relationship is weak as revealed by the accuracies of the predictive algorithms.

The results show that there is a strong influence of the university's academic reputation, opportunities of the city where the university is located, facilities and cultural opportunities of the university on the percentage of student placement. On the basis of the results obtained, with the help of machine learning algorithms, we can predict the the percentage of student placement easily and accurately based on various parameters grouped according to higher education institutions' attractiveness.

Higher education institutions can apply machine learning algorithms and they can turn the results of this study into the right strategies to gain a significant competitive advantage. These comparison results shed light not only on the decision-making of new enrollers, but also on the decision-making of re-enrollers who prepare for another year of university entrance exams in Turkish society. In addition, by comparing the performance of machine learning algorithms, the study elucidated the predictors that affect the students' preparations for university entrance exams.

4.3 Comparison with previous studies

Previous studies applied classification models using historical academic data (Mostafa & Beshir, 2021; Pal & Pal, 2013; Manvitha & Swaroopa, 2019; Guleria & Sood, 2015). Differently from the previous studies, we conducted thirteen machine-learning algorithms in predicting the percentage of student placement based on the data related to the university's academic reputation, opportunities of the city where the university is located, facilities and cultural opportunities of the university. In addition, the output variables were boolean data (0, incorrectly classified and

1, correctly classified)) in previous studies. Differently from the previous studies, we predicted the percentage of student placement (between 0 and 100) as an output variable. For this reason, it is not possible to compare our study with previous studies on the same basis. The results of this study therefore emphasize the need to re-evaluate the higher education institutions' attractiveness in Turkey and the need to consider other non-academic factors that may be indicators of students' choice.

4.4 Sensitivity analysis

In the previous section, the extreme gradient boosting machines algorithm demonstrated better predictive accuracy than other machine learning approaches. Therefore, sensitivity analysis was conducted based on the extreme gradient boosting machines algorithm under Python language, version 3.4 to identify the degree to which the input variables contribute to the determination of the output variable (the percentage of student placement at year *t*). Based on the results in Fig. 17, "the percentage of student placement at year *t-1"* was found to be the most effective parameter. Four input variables, namely the university scientific document score, university Phd programme score, university faculty member/student score, and the percentage of student placement at year *t-2* were found to be other effective parameters.

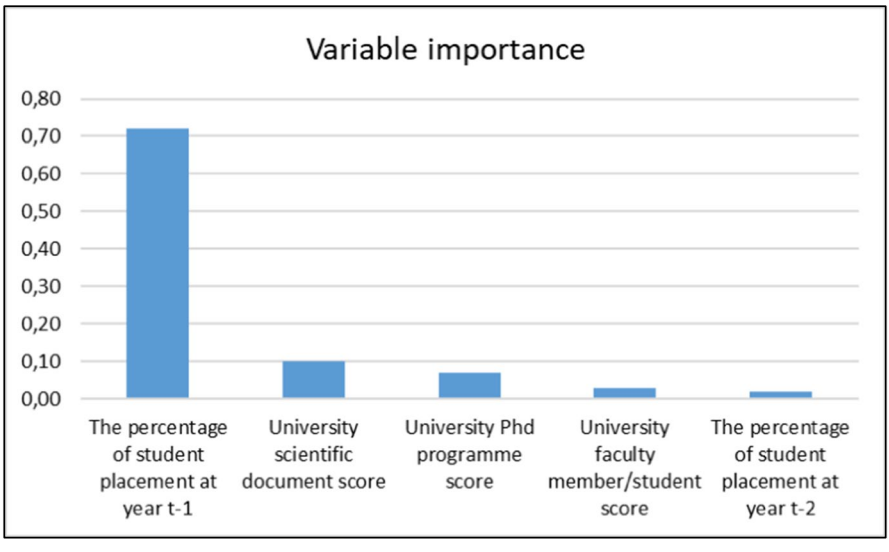


Fig. 17 Sensitivity of input variables versus output variable

5 Conclusions

The main aim of this study was to develop machine learning algorithms for predicting the percentage of student placement based on the data related to the university's academic reputation, opportunities of the city where the university is located, facilities and cultural opportunities of the university. Thirteen machine learning algorithms including linear models and non-linear models were employed to achieve the best results in terms of RMSE, MSE, MAE and R^2 values. The estimated output values generated by extreme gradient boosting were found to be the most satisfactory when the performance values were calculated. The three most sensitive factors that contributed over 85% of the sensitive weights together represented the percentage of student placement in year *t-1*, the university scientific document score, and university Phd programme score. Prediction results obtained in this study can be used in many different ways, such as determining the quotas for universities, allocating resources, regulating university enrollments, and making new regulations.

In order to make the universities more effective and productive, The Council of Higher Education in Turkey regulates university enrollments. Thus, the results obtained from this study helps The Council of Higher Education to find out the required quotas that the universities need. The reliability of machine learning algorithms depends mainly on its ability to learn from the previous data. Further studies should include additional experimentation with machine learning algorithms,obtain more training records by including other engineering departments or majors to generalize the model, and boost the performance of machine learning algorithms.

References

Absher, K., & Crawford, G. (1996). Marketing the community college starts with understanding students' perspectives. *Community College Review, 23*(4), 59–68.

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Anılan, H., Çemrek, F., & Anagün, Ş. S. (2008). Ortaöğretim öğrencilerinin meslek seçimi ve üniversite tercihlerine ilişkin görüşleri (Eskişehir örneği). *E-journal of World Sciences Academy, 3*(2), 238–249.

Azizi, V., & Hu, G. (2019). Machine Learning Methods for Revenue Prediction in Google Merchandise Store. In INFORMS International Conference on Service Science (pp. 65–75). Springer.

Beaulac, C., & Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education, 60*(7), 1048–1064.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Briggs, S. (2006). An exploratory study of the factors influencing undergraduate student choice: The case of higher education in Scotland. *Studies in Higher Education, 31*(6), 705–722.

Bringula, R., & Basa, R. (2011). Institutional image indicators of three universities: Basis for attracting prospective entrants. *Educational Research for Policy and Practice, 10*, 53–72.

Brown, C., Varley, P., & Pal, J. (2009). University course selection and services marketing. *Marketing Intelligence & Planning, 27*(3), 310–325. https://doi.org/10.1108/02634500910955227.

Çakıt, E., & Karwowski, W. (2017a). Predicting the occurrence of adverse events using an adaptive neuro-fuzzy inference system (ANFIS) approach with the help of ANFIS input selection. *Artificial Intelligence Review, 48*(2), 139–155.

Çakıt, E., & Karwowski, W. (2017b). Understanding the social and economic factors affecting adverse events in an active theater of war: a neural network approach. In International Conference on Applied Human Factors and Ergonomics (pp. 215–223). Springer.

Çakıt, E., Karwowski, W., Bozkurt, H., Ahram, T., Thompson, W., Mikusinski, P., & Lee, G. (2014). Investigating the relationship between adverse events and infrastructure development in an active war theater using soft computing techniques. *Applied Soft Computing, 25*, 204–214.

Çakit, E., Durgun, B., & Cetik, O. (2015). A neural network approach for assessing the relationship between grip strength and hand anthropometry. *Neural Network World, 25*(6), 603.

Çakıt, E., Karwowski, W., & Servi, L. (2020). Application of soft computing techniques for estimating emotional states expressed in twitter® time series data. *Neural Computing and Applications, 32*(8), 3535–3548.

Capraro, A. J., Patrick, M. L., & Wilson, M. (2004). Attracting college candidates: The impact of perceived social life. *Journal of Marketing for Higher Education, 14*(1), 93–106.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794).

Chen, T., Xu, J., Ying, H., Chen, X., Feng, R., Fang, X., ... & Wu, J. (2019). Prediction of Extubation Failure for Intensive Care Unit Patients Using Light Gradient Boosting Machine. *IEEE Access,* 7, 150960–150968.

Cover, T., & Hart., P. (2006). Nearest neighbor pattern classification. *IEEE Trans Inf Theor, 13*(1), 21–27.

Dao, M. T. N., & Thorpe, A. (2015). What factors influence Vietnamese students' choice of university? *International Journal of Educational Management, 29*(5), 666–681. https://doi.org/10.1108/IJEM-08-2014-0110.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1996). Support vector regression machines. *Advances in Neural Information Processing Systems, 9*, 155–161.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics, 4*, 325–327.

Esen, S. K. (2019). Understanding university choice decisions of Turkish students. In Handbook of research on contemporary approaches in management and organizational strategy (pp. 508–537). IGI Global.

Fausett, L. V. (2006). Fundamentals of neural networks: Architectures, algorithms and applications. Pearson Education India.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research, 15*(1), 3133–3181.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Genuer, R., Poggi, J. M., Tuleau-Malot, C., & Villa-Vialaneix, N. (2017). Random forests for big data. *Big Data Research, 9*, 28–46.

Guleria, P., & Sood, M. (2015). Predicting student placements using Bayesian classification. In 2015 Third International Conference on Image Information Processing (ICIIP) (pp. 109–112). IEEE.

Gurney, K. (1997). *An introduction to neural networks*. CRC press.

Hao, J., & Ho, T. K. (2019). Machine learning made easy: A review of scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics, 44*(3), 348–361.

Haykin, S. (2007). *Neural networks: A comprehensive foundation*. Prentice-Hall.

Iatrellis, O., Savvas, I. K., Fitsilis, P., & Gerogiannis, V. C. (2021). A two-phase machine learning approach for predicting student outcomes. *Education and Information Technologies, 26*(1), 69–88.

Kalkan, S. B., Başar, Ö. D., & Özden, Ü. H. (2015). Üniversite tercihlerinde Urap sıralamasında kullanılan değişkenlerin etkilerinin genelleştirilmiş tahmin denklemleri ile incelenmesi. Marmara Üniversitesi İ.İ.B. *Dergisi, 37*(2), 95–110.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146–3154).

Lagrari, F. E., Ziyati, H., & El Kettani, Y. (2018). An Efficient model of text categorization based on feature selection and random forests: Case for business documents. In International conference on advanced intelligent systems for sustainable development (pp. 465–476). Springer.

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education, 53*(3), 950–965.

Manvitha, P., & Swaroopa, N. (2019). Campus placement prediction using supervised machine learning techniques. *International Journal of Applied Engineering Research, 14*(9), 2188–2191.

Marsland, S. (2015). *Machine learning: An algorithmic perspective*. CRC press.

Mason, C., Twomey, J., Wright, D., & Whitman, L. (2018). Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression. *Research in Higher Education, 59*(3), 382–400.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Moogan, Y. J., Baron, S., & Harris, K. (1999). Decision-making behaviour of potential higher education students. *Higher Education Quarterly, 53*(3), 211–228. https://doi.org/10.1111/1468-2273.00127.

Mostafa, L., & Beshir, S. (2021, June). University Selection Model Using Machine Learning Techniques. In The International Conference on Artificial Intelligence and Computer Vision (pp. 680–688). Springer.

Olcme, Secme ve Yerleştirme Merkezi (OSYM). (2020). Available online at: https://www.osym.gov.tr/ (Accessed 20 Dec 2020).

Pal, A. K., & Pal, S. (2013). Classification model of prediction for placement of students. *International Journal of Modern Education and Computer Science, 5*(11), 49.

Pampaloni, A. M. (2010). The influence of organizational image on college selection: What students seek in institutions of higher education. *Journal of Marketing for Higher Education, 20*(1), 19–48.

Park, T., & Kim, C. (2020). Predicting the variables that Determine University (re-) entrance as a career development using support vector machines with recursive feature elimination: The case of South Korea. *Sustainability, 12*(18), 7365.

Pliakos, K., Joo, S. H., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers & Education, 137*, 91–103.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In Advances in neural information processing systems (pp. 6638–6648).

Qin, Z., Myers, D. B., Ransom, C. J., Kitchen, N. R., Liang, S. Z., Camberato, J. J., ... & Shanahan, J. F. (2018). Application of machine learning methodologies for predicting corn economic optimal nitrogen rate. *Agronomy Journal,* 110(6), 2596–2607.

Ramasubramanian, S., Gyure, J. F., & Mursi, N. M. (2002). Impact of internet images: Impression-formation effects of university web site images. *Journal of Marketing for Higher Education, 12*(2), 49–68.

Richardson, A., van Florenstein Mulder, T., & Vehbi, T. (2021). Nowcasting GDP using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting, 37*(2), 941–948.

Robinson, D. (2017). The incredible growth of Python. Retrieved from https://stackoverflow.blog/2017/09/06/incredible-growth-python/. Accessed 26 Mar 2021.

Scikit-Learn: Machine Learning in Python (2021). Available online at: http://scikit-learn.org/stable/ (Accessed 10 Jan 2021).

Sheela KG, Deepa SN (2013) Performance analysis of modeling framework for prediction in wind farms employing artificial neural networks. Soft Computing, 1–9.

Simões, C., & Soares, A. M. (2010). Applying to higher education: Information sources and choice factors. *Studies in Higher Education, 35*(4), 371–389. https://doi.org/10.1080/03075070903096490.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199–222.

Soutar, G. N., & Turner, J. P. (2002). Students' preferences for university: A conjoint analysis. *International Journal of Educational Management, 16*(1), 40–45. https://doi.org/10.1108/09513540210415523.

Swamynathan, M. (2019). *Mastering machine learning with python in six steps: A practical implementation guide to predictive data analytics using python*. Apress.

Tatar, E., & Oktay, M. (2006). Search, choice and persistence for higher education: A case study in Turkey. *Eurasia Journal of Mathematics, Science and Technology Education, 2*(2), 115–129.

The University Ranking and Academic Performance (URAP). (2020) Available online at: https://www.urapcenter.org/ (Accessed 2 Dec 2020).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288.

Ustuner, M., & Balik Sanli, F. (2019). Polarimetric target decompositions and light gradient boosting machine for crop classification: A comparative evaluation. *ISPRS International Journal of Geo-Information, 8*(2), 97.

Veloutsou, C., Lewis, J. W., & Paton, R. A. (2004). University selection: Information requirements and importance. *International Journal of Educational Management, 18*(3), 160–171. https://doi.org/10.1108/09513540410527158.

Vrontos, S. D., Galakis, J., & Vrontos, I. D. (2021). Modeling and predicting US recessions using machine learning techniques. *International Journal of Forecasting, 37*(2), 647–671.

Walsh, C., Moorhouse, J., Dunnett, A., & Barry, C. (2015). University choice: Which attributes matter when you are paying the full Price? *International Journal of Consumer Studies, 39*(6), 670–681. https://doi.org/10.1111/ijcs.12178.

Wang, G., & Liu, Z. (2020). Android malware detection model based on lightgbm. In Recent Trends in Intelligent Computing, Communication and Devices (pp. 237–243). Springer.

Witten, I. H., & Frank, E. (2002). Data mining: Practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record, 31*(1), 76–77.

Yousafzai, B. K., Hayat, M., & Afzal, S. (2020). Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student. *Education and Information Technologies, 25*(6), 4677–4697.

Yuksek Ogretim Kurulu (YOK). (2020). Available online at: https://www.yok.gov.tr/ (Accessed 20 Nov 2020).

Zhang, S., Hu, Q., Xie, Z., & Mi, J. (2015). Kernel ridge regression for general noise model with its application. *Neurocomputing, 149*, 836–846.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.

Zurada, J. M. (1992). *Introduction to artificial neural systems* (Vol. 8). West.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.