Analysis of the situation of tuition paying students by applying artificial intelligence algorithms

Abstract

Fill this part plz!

Keywords: tuition, artificial intelligence in higher education, supervised learning, financial resources of higher education

Introduction

In today's competitive world, the most important source and asset of any country is its human capital (Cattaneo et al, 2020). Based on this, higher education plays an essential role in creating and developing capable human capital and providing any country's future workforce. Therefore, in the long term, it gives the basis for countries' economic growth and development in all dimensions of their biological life. This is done through the education of university graduates and the development of the capabilities of the workforce (for example, in-service training). For this reason, governments, organizations and institutions, families and students have a great desire to invest in higher education (Trinh, 2021). In addition, universities are known as the driving force of sustainable development and moving towards a sustainable society, which plays a strategic and decisive role in creating a competitive advantage in countries at the regional and international levels (Aparicio et al, 2023). However, Zumeta (2011) argued that the public funding of higher education had decreased significantly since the economic recession in 2008. Higher education institutions have a relatively favorable financial position in good economic conditions. But they are the first institutions to face budget cuts in bad economic times (Bound et al, 2019).

Examining the historical trend also shows that the amount of university tuition is continuously increasing. For example, between 1980 and 1998, tuition fees in the United States increased by 125%, while household income grew by only 1% (Kelly & Shale, 2004). In 2000, there was a significant increase in tuition rates in different geographic regions of the United States. In the public higher education sector, tuition in public universities grew by 62% and in public two-year colleges by 40%. In the private higher education sector, tuition fees in private universities also increased by 42%. This tuition increase has occurred most in the South and Southwest regions of the United States (Ma et al, 2016). The wave of increase in the amount of tuition in some geographical regions of America, especially the state of Texas, has been so high that during a six-year period from 2003 to 2009, there was a 72% increase in the amount of tuition (Flores & Shepherd, 2014). Based on the findings of other studies, more than 70% of parents have also expressed concern about how to finance their children's college education (Bell, 2020). In another experimental sample, 70% of students in a university admitted that they dropped out of the university because of their financial support, and 52% of students said that they dropped out of the university because of their inability to pay tuition fees (Johnston et al, 2009).

Setting and increasing the amount of tuition has many positive and negative consequences. First, the negative consequences and then the positive consequences are discussed. One of the negative consequences that an increase in tuition can have is the reduction of access to students from lower social and economic classes (Williams, 2016). Because by increasing the amount of tuition, they cannot experience social and economic mobility, and by improving their own social and economic base, they reduce the existing class gaps in society. Even with the increase in tuition fees, financial aid and loans were considered for the participation of students with lower social and economic classes. But empirical evidence shows that this has also failed to be effective. According to Long (2006), there are many concerns about financial aid and loans because students and their families do not have a clear understanding of loan debt. This can negatively affect students' academic decisions during university registration and even after graduation. For example, the fear of debt prevents students from making subsequent decisions such as buying a house, getting married, and having children. There is a lot of empirical evidence about the decrease in the access of low-income students due to the increase in tuition fees in universities (Boatman et al, 2017). Another negative consequence of the increase in the tuition fee is the decrease in students' enrollment, retention and graduation rates. Although there is conflicting evidence in this regard, the main direction of the studies is

reducing the mentioned rates. For example, we can refer to studies (Hemelt & Marcotte, 2011; Cunningham & Santiago, 2008; Paulsen & St. John, 2002; Perna, 2006). They concluded that the increase in the tuition rate leads to a decrease in the enrollment, retention and graduation rates. However, the establishment and growth in tuition also have positive consequences. Among these positive consequences, from students' point of view, accuracy in choosing a field is a kind of investment. From the point of view of policymakers, it is the development of the income stream of universities. Universities' lack of financial resources is an undeniable fact everywhere in the world. At the same time, setting and increasing tuition fees is a guaranteed policy to increase universities' income.

مورد پژوهی پژوهش حاضر دانشگاه تهران است. علت انتخاب این دانشگاه، برخورداری از ساختار شهریه‌ای متفاوت و چندگانه می‌باشد. در دانشگاه تهران، علاوه بر دوره‌های شبانه، دوره‌های الکترونیکی و پردیس‌های گوناگونی نظیر: پردیس کیش، البرز و ارس وجود دارند. هر کدام از این دوره‌ها شهریه‌های متفاوتی را از دانشجویان دریافت می‌کنند. با توجه به برند دانشگاه تهران نسبت به دیگر دانشگاه‌ها متقاضی بیشتری برای این دوره‌ها در این دانشگاه وجود دارد. علاوه بر تفاوت شهریه‌ای، میزان این شهریه نیز پیوسته و هر ساله در حال افزایش است. در ادامه بر اساس اطلاعات موجود در سایت دانشگاه تهران در بازه زمانی سه ساله میزان شهریه دانشگاه تهران در دوره شبانه، مجازی و پردیس‌ها ارائه شده است. در جدول (1) روند افزایش میزان شهریه در مقاطع مختلف دانشگاه تهران در دوره شبانه آمده است.

جدول 1: روند افزایش میزان شهریه در مقاطع مختلف دانشگاه تهران در دوره شبانه

مقاطع تحصیلی	رشته تحصیلی	سال 1399	سال 1398	سال 1397
کارشناسی	علوم انسانی و رفتاری	8607785	7173154	6237525
	هنر	10877398	9064498	7882172
	فنی و کشاورزی	12303603	10253003	8915655
کارشناسی ارشد	علوم انسانی و رفتاری	9726533	9726533	8457855
	سایر گروه‌ها	10729635	10792635	9384900
دکتری	علوم انسانی و رفتاری	101339819	89582161	77897531
	سایر گروه‌ها	123859792	109489307	85208093
نرخ شهریه ثابت فقط لحاظ شده است اعداد به صورت میلیون ریال و شهریه ترمی می‌باشد				

منبع:

سایت دانشگاه تهران (1400). اطلاعات آماری مربوط به شهریه:https://academics.ut.ac.ir/fa/page/1057 /

بر اساس اطلاعات موجود در جدول (1)، شاهد افزایش میزان شهریه در کلیه مقاطع به ویژه در مقطع دکتری هستیم. از نظر ماهیت رشته‌ای، نیز رشته فنی و کشاورزی افزایش بیشتری نسبت به رشته‌های علوم انسانی و رفتاری و هنر داشته است.

بنابراین، از یک سو، میزان تخصیص منابع مالی دولتی به دانشگاه‌ها رو به کاهش است. لیکن، سیاستگذاران نیازمند به باز اندیشی در ساختار و سیاست‌های مالی آموزش‌عالی هستند. چراکه با توجه به نقش کلیدی دانشگاه‌ها در توسعه همه‌جانبه زیست‌بوم‌های گوناگون در صورت نبود منابع‌مالی مکفی کارایی نظام دانشگاهی نیز کاهش می‌یابد.

بر این اساس، هدف از این پژوهش، پیش بینی تفاوت بین دانشجویان شهریه پرداز و بدون شهریه در دانشگاه تهران با استفاده از الگوریتم های یادگیری نظارت شده می باشد.

Research background

Shakir et al. (2022), in research titled a systematic study on tertiary level student tuition fee waiver management during the pandemic using machine learning approaches, concluded that results appear that the accuracy of DTR and RFR are 74.49% and 77.82%, respectively, for before applied CV; and, 71.41%, and 76.90% respectively, for after applied CV. Mubarak et al. (2021), in research titled prediction of students' early dropout based on their interaction logs in online learning environment, results showed that the proposed models achieved an accuracy of 84% compared to the baseline of Machine Learning Models for prediction of the students at-risk of dropping out. Adli and Sahid (2021), in research titled UKT (single tuition) classification prediction, use MKNN (K-Nearest Neighbor Modification) algorithm in conducting classifications to establish UKT new students. Using the MKNN method and supported by the K-Fold Cross Validation validation method, the classification results get an accuracy value of 71%. Muqorobin et al. (2020), in research titled estimation system for late payment of school tuition fees the final results of this, study indicate that the Naïve Bayes Method + Information Gain Method produces the highest accuracy, namely 95%, compared to the Naïve Bayes method alone, namely 85% and the K-NN method, namely 81%. Dass et al. (2020), in research titled predicting student dropout in self-paced MOOC course using random forest model, the model developed can predict the dropout or continuation of students on any given day in the MOOC course with an accuracy of 87.5%, AUC of 94.5%, precision of 88%, recall of 87.5%, and F1-score of 87.5%, respectively.

Agrusti et al. (2020), in research titled deep learning approach for predicting university dropout: A case study at Roma Tre University, the results obtained using deep learning models to the ones using Bayesian networks. The accuracy of the obtained deep learning models ranged from 67.1% for the first-year students up to 94.3% for the third-year students. Kemper et al. (2020), in research titled predicting student dropout: A machine learning approach, find decision trees to produce slightly better results than logistic regressions. However, both methods yield high prediction accuracies of up to 95% after three semesters. A more than 83% accuracy classification is already possible after the first semester. Aldino and Sulistiani (2020), in research titled decision tree C4. 5 algorithms for tuition aid grant program classification (Case Study: Department of Information System, Universitas Teknokrat Indonesia), the results of the classification are validated with ten-fold cross-validation with accuracy, precision, and recall with the score of 87 % for all part. It means the model performs quite well to be implemented into the system. Rohmayani (2020), in research titled, analysis of student tuition fee pays delay prediction using Naive Bayes algorithm with particle swarm optimization (Case Study: Politeknik TEDC Bandung), the testing result of those three classification models using the highest Naive Bayes accuracy algorithm are testing using the Naive Bayes algorithm based on Particle Swarm Optimization (PSO), with an accuracy of 73.94%, precision 78.50%, 69% recall, and AUC 0.771, even though the execution time is 3 seconds longer.

Basu et al. (2019), in research titled, predictive models of student college commitment decisions using machine learning, the results from this study indicate that the logistic regression classifier performed best in modeling the student college commitment decision problem, i.e., predicting whether a student will accept an admission offer, with an AUC score of 79.6%. The significance of this research is that it demonstrates that many institutions could use machine learning algorithms to improve the accuracy of their estimates of entering class sizes, thus allowing more optimal allocation of resources and better control over net tuition revenue. Kurniadi et al. (2018), in research titled the prediction of scholarship recipients in higher education using the k-Nearest neighbor algorithm, the results of the simulation of the prediction model show that the determining factor of training data from both the number and the variation of different values can improve the performance of the k-NN algorithm with the best accuracy rate of 95.83 percent in predicting students who have the greatest chance of getting the scholarship. Nagy and Molontay (2018), in research titled, predicting dropout in higher education based on secondary school performance, the methods were tested using 10-fold cross-validation, and the AUC of the best models, Gradient Boosted Trees, and Deep Learning, were 0.808 and 0.811 respectively. Solis et al (2018), in research titled, perspectives to predict dropout in university students with machine learning, using a classification variable that defines the non-dropout as the graduated student. In a first validation sample, this approach correctly predicted 91% of dropouts, with a sensitivity of 87%.

Methodology

Machine learning has emerged as a research field in artificial intelligence in recent years. Based on this learning method, we distinguish between supervised and unsupervised learning. As we focused on supervised learning in this research, we briefly present the supervision methods we later used to build the predictive models.

A) Algorithms used

A-1) Logistic regression

Logistic regression models are defined as “statistical models which describe the relationship between a qualitative dependent variable (that is, one which can take only certain discrete values, such as the presence or absence of a disease) and an independent variable. Synonymous terms are logistic regression, logistic models, and logit models. Logistic regression models are used to study effects of predictor variables on categorical outcomes. Normally, the outcome is binary, such as presence or absence of disease (e.g., non–Hodgkin lymphoma), in which case the model is called a binary logistic model. When there is only one predictor variable in a logistic regression model, the model is referred to as a simple logistic regression. When there are multiple predictors (e.g., risk factors and treatments), including categorical and continuous variables as predictors, the model is referred to as a multiple or multivariable logistic regression (Nick & Campbell, 2007).

A-2) Artificial neural network

Artificial neural network (ANN) technology is a compelling alternative modeling and prediction tool. Unlike traditional physical-based numerical models, ANNs do not require explicit

characterization and quantification of physical properties and conditions, nor accurate representation of the governing physical laws. Rather, ANNs ‘‘learn’’ system behavior of interest from representative data. By reducing common modeling data uncertainties and eliminating the need for imposing a rigidly defined physical and mathematical framework, superior prediction performance, enhanced hydrogeologic understanding, and improved management decisions may be achieved with the ANN approach (Coppola et al, 2005).

A-3) Decision tree

The Decision Tree algorithm is a member of the supervised learning algorithm family used in statistics, data mining, and machine learning. The decision tree approach, unlike other supervised learning lgorithms, may also be utilized to solve regression and classification issues. We have selected the decision tree algorithm for many reasons, the result of the classification tree is easier to understand and interpret, and it supports multiple data types such as numeric, nominal, categorical, etc. The objective of employing a Decision Tree is to build a training model that can predict the class or value of the target variable by learning basic decision rules from past data (training data). (El Massari et al, 2022).

A-4) Random Forest

A Random Forest (RF) is, as the name suggests, an ensemble of stochastically built decision trees used for classification, or indeed for regression. Random Forest is widely considered relatively immune to overfitting. Each tree is grown by stochastic recursive partitioning, and the individual trees carry independent information because of the substantial random element in their construction. Every decision tree in the forest is firstly randomized using a bootstrap sample of Q instances from the training data, chosen by sampling with replacement from the Q objects in the training set. Thus, each object may be selected one or possibly more times for a given tree’s dataset, while about 37% of the instances remain unchosen in any particular bootstrap sample and constitute the so-called out-of bag (OOB) data. The OOB data may serve as an internal validation set for the given tree. The combined performance on the separate OOB datasets of each tree can be aggregated, and constitutes a fair test of the overall predictive performance of the Random Forest (Su et al, 2021).

B) Society, sample (sample size) and sampling method

This research's statistical population consists of all Tehran University students in the last five years (2015-2020). from using the method of all census data of students of faculties; Psychology and educational sciences, management, economics and the collection of technical and engineering faculties (11 technical and engineering faculties) were collected and analyzed in the last five years. One of the main limitations of the research was the need for more access to the complete data set of graduate students in all faculties of Tehran University.

C) Variables and how to code

Independent variables in this research include; Department, Age, GPA, Grade, Type, Nationality, Marital Status, Children, Year, Financial Aid, Gender, Transfer, drop out, Remove, Leave, Change filed, Guest and dependent variable: Type. The table below shows the types of variables and their coding methods;

جدول 1: انتخاب متغیرهای منتخب پژوهش

متغییرهای مستقل	متغیرها
مقطع تحصیلی	Dietrich & Gerner (2012)/ Dwenger, Storck & Wrohlich (2012)/ Hemelt & Marcotte (2016)/ Moulin et al. (2016)
دانشکده	Callender & Jackson (2008)
سن	Callender & Jackson (2008)/ Dwenger, Storck & Wrohlich (2012)/ Neill (2015)/ Moulin et al. (2016)/ Allen & Wolniak (2019)/ Andrieu & John. (1993)/ Arendt (2013)/ Chen & DesJardins (2008)/ Callender & Jackson (2005)/ Martinello (2015)
نوع دوره[1]	Vasigh & Hamzaee (2004)/ Dwenger, Storck & Wrohlich (2012)/ Havranek, Irsova, & Zeynalova, (2017)/ Hemelt & Marcotte. (2011)/Garrett & Greene (2018)/ Neill (2015)/ Kim, DesJardins & McCall (2009)
ملیت	Moulin et al. (2016)
وضعیت تاهل	Callender & Jackson (2008)/ Callender & Jackson (2005)
تعداد فرزندان[2]	Callender & Jackson (2008)/ Hemelt & Marcotte (2016)/ Neill (2015)/ Callender & Jackson (2005)
سال	Dickson & Pender (2013)/ Dwenger, Storck & Wrohlich (2012)/ Havranek, Irsova, & Zeynalova,/ (2017)/ Hemelt & Marcotte. (2011)/ Moulin et al. (2016)/ Allen & Wolniak (2019)/ Arendt (2013)
کمک مالی	Vasigh & Hamzaee (2004)/ Havranek, Irsova, & Zeynalova, (2017)/ Neill (2015)/ Andrieu & John. (1993)/ Shin & Milton (2008)/ Dowd (2004)/ Kim, DesJardins & McCall (2009) Chen & DesJardins (2008)
جنسیت	Callender & Jackson (2008)/ Dickson & Pender (2013)/ Dwenger, Storck & Wrohlich (2012)/ Havranek, Irsova, & Zeynalova, (2017)/ Garrett & Greene (2018)/ Acton (2018)/ Hübner (2012)/ Neill (2015)/ Moulin et al. (2016)/ Andrieu & John. (1993)/ Arendt (2013)/ Dowd (2004)/ Chen & DesJardins (2008)/ Callender & Jackson (2005)
تصمیمات تحصیلی (تصمیم به انتقال تحصیلی، تصمیم به انصراف، تصمیم به حذف ترم، تصمیم به مرخصی، تصمیم به مهمان شدن و تصمیم به تغییر رشته)	Hübner (2012)/ Moulin et al. (2016)/ Arendt (2013)/ Martinello (2015)
معدل	Hemelt & Marcotte (2016)/ Andrieu & John. (1993)/ Arendt (2013)/ Dowd (2004)/ Chen & DesJardins (2008)

variable of Grade (0 = bachelor's degree, 1 = master's degree and number 2 = PhD); Department (1 = technical and engineering faculties, 2 = psychology and educational sciences, 3 = economics and 4 = management faculty); age (continuously); type (1= tuition-free students, 0= tuition-paying students); nationality (1= Iranian students, 0= international students); marital status (1 = single, number 0 = married); children (continuously); year (1= 2015, 2= 2016, 3= 2017, 4= 2018, 5= 2019, number 6= 2020); GPA (continuously); financial aid (1 = not receiving financial aid, 0 = receiving financial aid); gender (1 = men, 0 = women); Academic decisions (Transfer, drop out, Remove, Leave, Change filed and Guest) (1 = Yes, 0 = No).

[1] در برخی از مطالعات میزان شهریه در نظر گرفته شده است و در برخی از مطالعات بین دانشجویان شهریه پرداز و غیر شهریه ای مقایسه انجام شده است

[2] در مطالعات تعداد اعضای خانواده مد نظر بوده است در این پژوهش تعداد فرزندان

D) Evaluation Metrics

Precision, recall, and F1 score are common metrics used to evaluate the performance of classification models. Precision measures the proportion of true positive predictions out of all positive predictions. It is calculated as:

precision = true positives / (true positives + false positives)

Recall measures the proportion of true positive predictions out of all actual positives. It is calculated as:

recall = true positives / (true positives + false negatives)

The F1 score is the harmonic mean of precision and recall and provides a single metric that balances both measures. It is calculated as:

F1 score = 2 * (precision * recall) / (precision + recall)

These metrics are useful for evaluating the performance of classification models in different scenarios. For example, high precision is important when the cost of false positives is high, while high recall is important when the cost of false negatives is high

Findings

The findings are presented in several sections below;

A) Descriptive statistics indicators of data

In the first part of the findings, descriptive statistics are reported. Table (3) describes the results of descriptive statistics

Table 3: Descriptive statistics indicators

Features	Count	Mean	Std	Min	25%	50%	75%	Max
Department	13709.0	2.154059	1.296688	100	1.00	2.00	4.00	4.0
Age	13709.0	26.688818	6.402831	2.00	22.00	25.00	29.00	67.0
Grade	13709.0	0.728208	0.661330	0.00	0.00	1.00	1.00	2.0
Type	13709.0	0.759720	0427269	0.00	1.00	1.00	1.00	1.0
GPA	13709.0	16.777675	1.918553	10.04	15.67	17.17	1825	20.0
Nationality	13709.0	0.987162	0.112581	0.00	1.00	1.00	1.00	1.0
Marital status	13709.0	0.843461	0.363379	0.00	1.00	1.00	1.00	1.0
Children	13709.0	0.125830	0495526	0.00	0.00	0.00	0.00	9.0
Year	13709.0	4.043402	1.587015	1.00	3.00	4.00	5.00	6.0

Financial Aid	13709.0	0.997155	0.053263	0.00	1.00	1.00	1.00	1.0
Gender	13709.0	0.594865	0490936	0.00	0.00	1.00	1.00	1.0
Transfer	13709.0	0.001605	0.040029	0.00	0.00	0.00	0.00	1.0
Drop out	13709.0	0.006346	0.079413	0.00	0.00	0.00	0.00	1.0
Remove	13709.0	0.001021	0.031941	0.00	0.00	0.00	0.00	1.0
Leave	13709.0	0.002261	0.047501	0.00	0.00	0.00	0.00	1.0
Change filed	13709.0	0.002626	0.051179	0.00	0.00	0.00	0.00	1.0
Guest	13709.0	0.002553	0.050465	0.00	0.00	0.00	0.00	1.0

b) Common findings between models

The heatmap in Figure 1, represents the correlation matrix, where each cell's color and intensity indicates the strenght and direction of the correlation between variables. The variables with the highest influence on the target (Type)

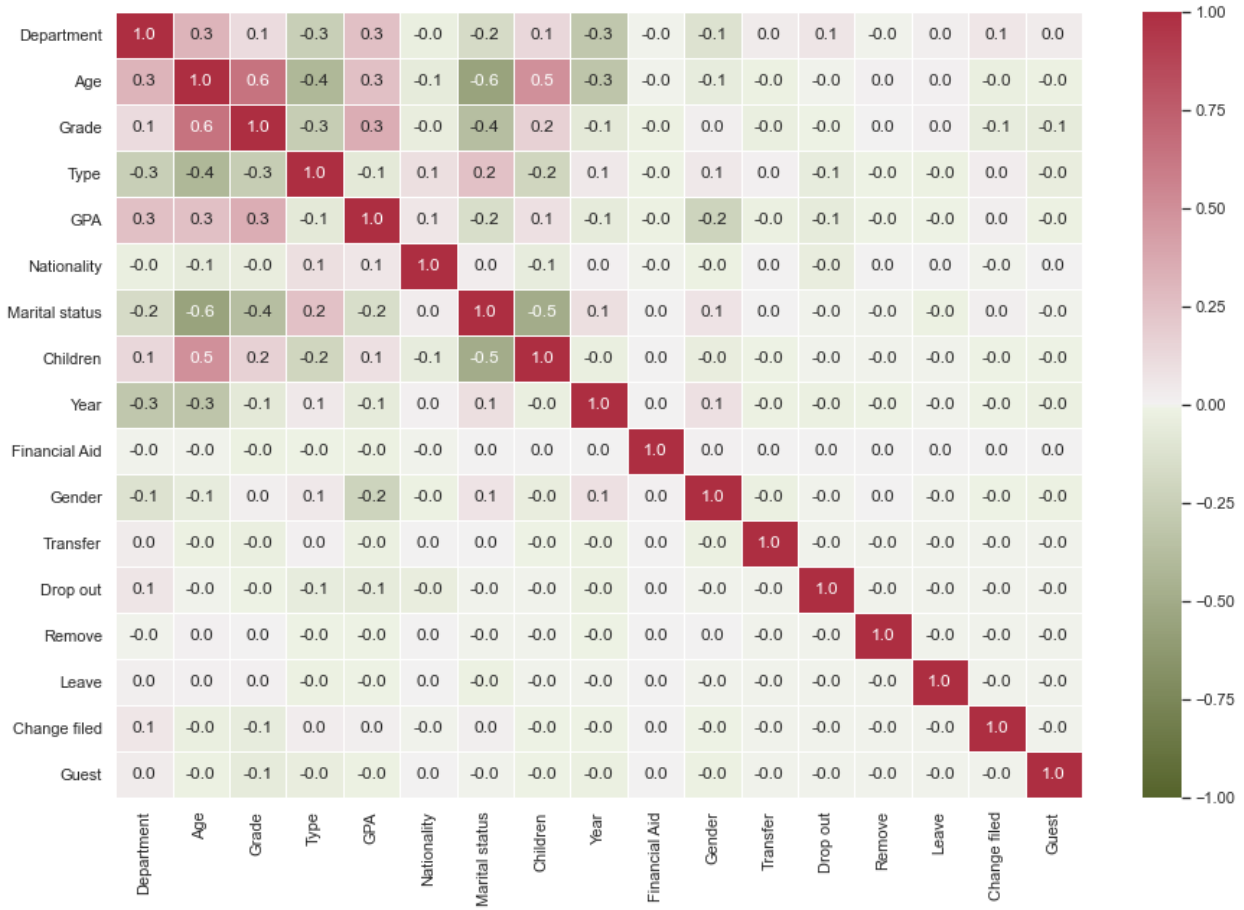


Figure 1: Correlation coefficients between variables

Next, in figure (2), the percentage chart of students is drawn based on the dependent variable (Type)

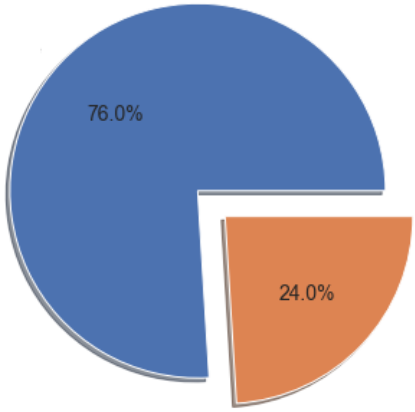


Figure 2: Percentage chart of students based on the dependent variable

According to this chart, 76% of students are tuition-free and 26% are tuition-paying students.

c) Analytical findings

C-1) logistic regression

در شکل زیر اهمیت متغیرهای در رگرسیون لاجستیک بازنمایی شده است. بر اساس این شکل، مهمترین متغیرهای در یادگیری مدل با همبستگی مثبت شامل، معدل، ملیت و جنسیت با مقادیر 0.22 و 0.15 و 0.11 بود. علاوه بر این، متغیر سن با مقدار 0.98 – مهمترین متغیری بود که همبستگی منفی با متغیر معدل داشت.

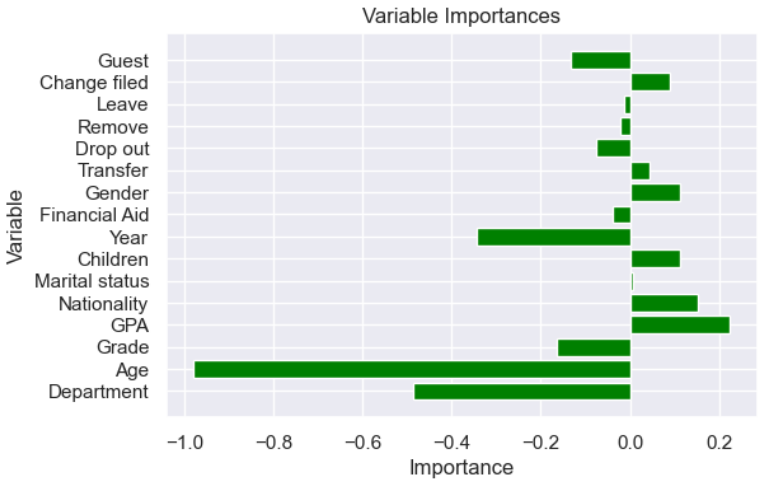


Figure: Importance of variables in logistic regression

In the following table, the result of applying logistic regression is stated. The precision of the model in classifying positive cases is 81% for class 1 and 66% for class 0. Additionally, the recall

score of 95% is one of the highest among all the models. The harmonized F1 score of 87% indicates a balanced trade-off between precision and recall.

Table: Results of applying logistic regression

	precision	recall	f1-score
0	0.66	0.30	0.41
1	0.81	0.95	0.87

C2) Neural network

The figure below shows the amount of Loss function changes. The Loss function is in blue for the training data and in orange for the test data. The decreasing trend in the difference between the predicted and the actual values reflects the increase in the accuracy of the neural network model in each training period of the model (epoch).

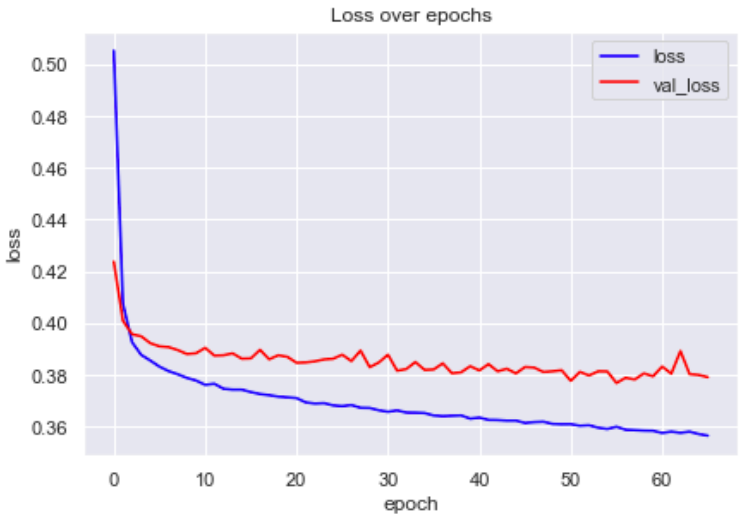


Figure: Loss function chart

In the following table, the results of the application of neural network are listed. The F1 score of 53% for class 0 and 88% for class 1 are higher compared to logistic regression.

Table: The results of neural network regression application

	precision	recall	f1-score
0	0.64	0.45	0.53
1	0.84	0.92	0.88

Hyper parameter optimization of neural network model

In this analysis, the neural network's structure is fine-tuned using Keras hyperparameter optimization techniques to enhance prediction precision. This process involves selecting the most suitable values for as hyperparameters such as the learning rate, the number of hidden layers, the quantity of neurons within each layer, the activation function, and the strength of regularization. Two methods, including grid search, and random search, are employed to adjust these hyperparameters. These methods involve a systematic exploration of the hyperparameter space to identify the ideal combination of settings that maximizes the neural network's performance on a validation dataset. The table below shows the structure of the neural network.

Table: Neural network regression structure

Layer (type)	Output Shape	Param #
dense_3 (Dense)	(None, 32)	544
dropout (Dropout)	(None, 32)	0
dense_4 (Dense)	(None, 32)	1056
dropout_1 (Dropout)	(None, 32)	0
dense_5 (Dense)	(None, 1)	33

Total params: 1,633

Trainable params: 1,633

Non-trainable params: 0

بر اساس این جدول، ساختار شبکه عصبی استفاده شده روی هم رفته دارای دو لایه میانی با 32 نورون میزان افت 10 درصد و یک لایه انتهایی باینری برای پیش بینی وجود دارد.

The figure () demonstrates the evolutaion of loss funcrion (left) and accuracy (right) of the model in each epoch. The accuracy metric measures the proportion of correctly classified samples in the training or validation dataset. Similar to the loss function, accuracy generally improves during training.

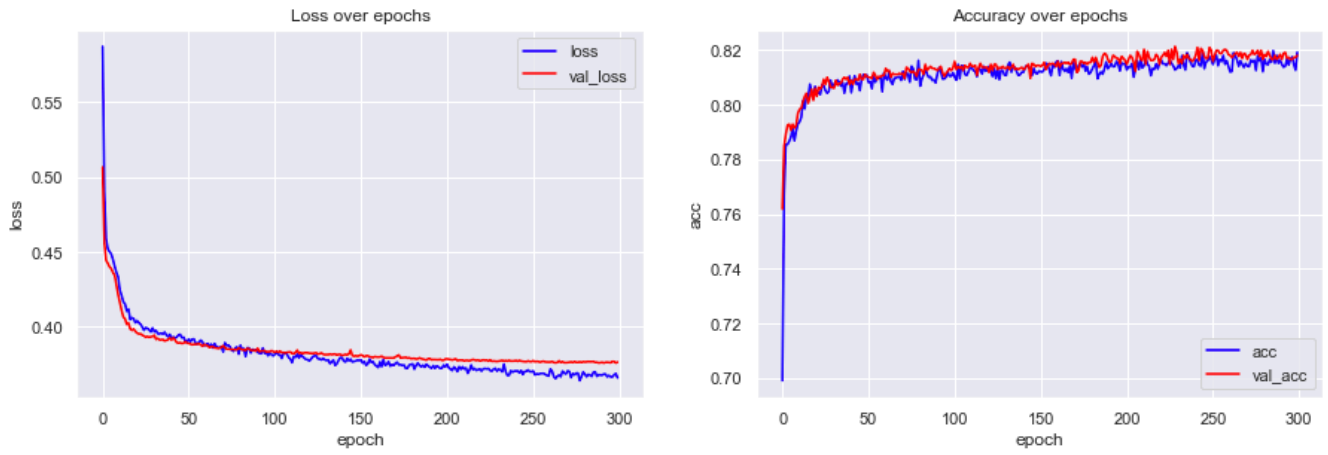


Figure: LOSS FUNCTION chart

In the following table, the results of the optimized neural network application are stated. The F1 score for both classes are enhanced after optimization.

Table: Results of optimized neural network regression application

	precision	recall	f1-score
0	0.72	0.45	0.56
1	0.84	0.94	0.89

C3) decision tree

Figure () quantifies how much each feature contributes to the overall decision-making process within the decision tree ensemble. By assigning importance scores to features, decision tree identifies which factors (GPA and grade in this context) have the most influence on the model's predictions.

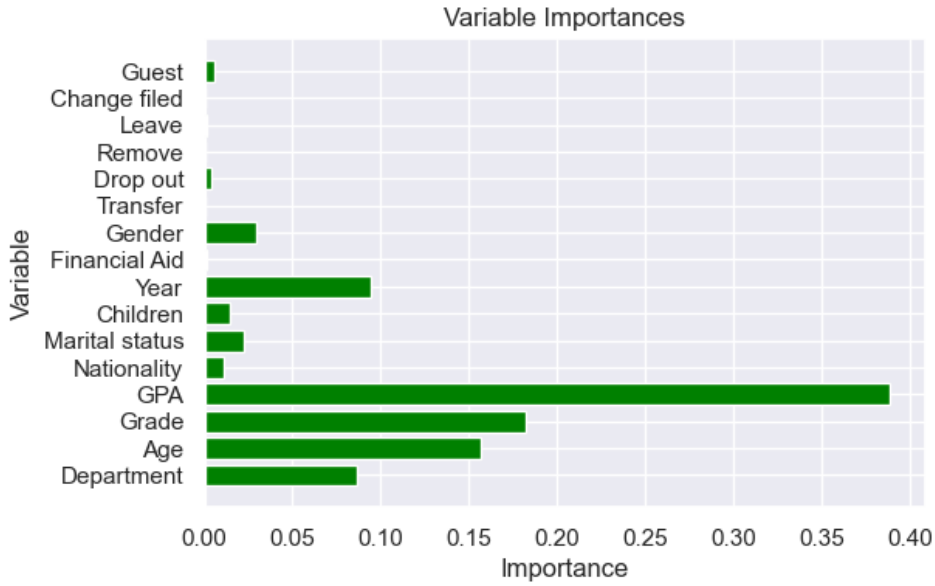


Figure: Importance of variables in decision tree algorithm

In the following table, the results of using the decision tree algorithm are stated. The precision of 85% for class 1 is the highest precision among all the models while the recall of 85% is weak.

Table: The results of applying the decision tree algorithm

	precision	recall	f1-score
0	0.48	0.53	0.51
1	0.85	0.85	0.83

C4) Random forest

In the first step the ranodm forest with the default hyperparameters are trained to predict the outcome. By examining the decrease in prediction accuracy when the feature is removed from the analysis, we are able to assess the significance of individual input features in predicting. Figure shows that like decision tree, the GPA and then the Age features have the largest contributions to the model's decision-making.

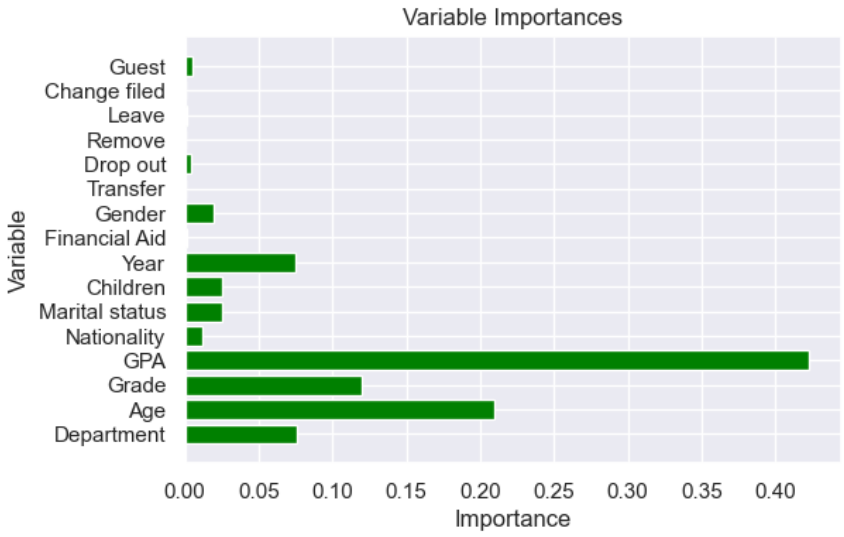
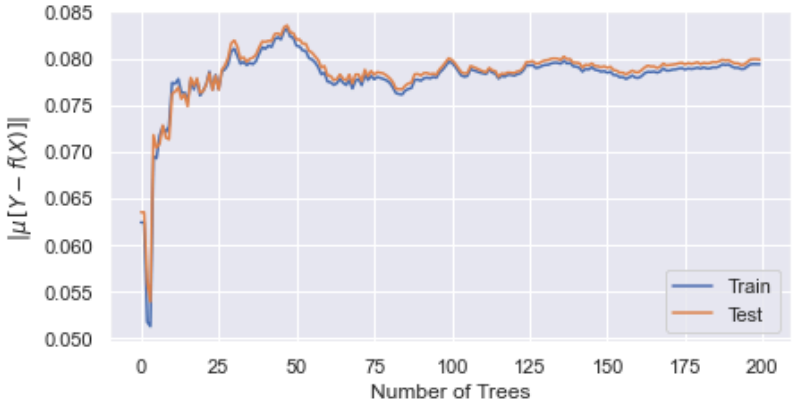


Figure: Importance of variables in random forest algorithm

In order to maximize the model's predictive accuracy or other relevant performance metrics, hyperparameters of random forest are tuned. Common hyperparameters for Random Forest include the number of trees, the maximum depth of each tree, the minimum number of samples required to split an internal node, and the maximum number of features considered for each split. This is achieved through techniques like grid search or random search, where different combinations of hyperparameters are evaluated by training and cross-validating the model on a dataset. By increasing the number of trees, the random forest model can make better prediction which is shown in Figure ().



The performance of random forest with default and optimized hyperparameters are shown in table . The recall and F1 score for class 1 increased by few percents after tunning.

Table: The results of applying the optimized random forest algorithm

	precision	recall	f1-score
0	0.53	0.50	0.52
1	0.85	0.86	0.85
	precision	recall	f1-score
0	0.72	0.41	0.52
1	0.83	0.95	0.89

C5) Comparison of the results of the used algorithms

To assess classification performance for different models, ROC (Receiver Operating Characteristic) is shown in Figure (). ROC curves are graphical representations of a model's ability to distinguish between classes at various decision thresholds. By comparing these curves, we can gauge the trade-off between a model's true positive rate and false positive rate across different threshold settings. A model with a curve that reaches closer to the top-left corner represents better discriminatory power and superior classification performance. The best ROC rate is obtained for neural network and optimized neural network models.

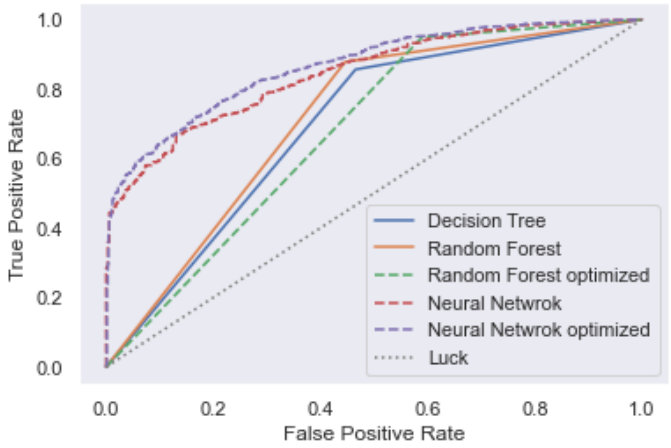
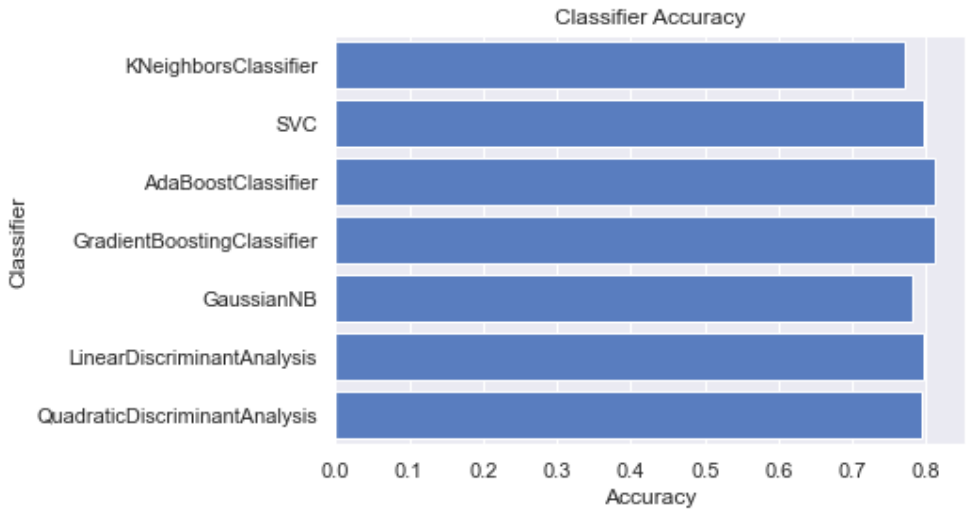


Figure: ROC comparison chart for all implemented algorithms

For the completeness in this analysis, the accuracy for other known machine learning classifiers are shown in Figure . Ada boost [ref] and Gradient boostt [ref] models achieved the highest accuracy of 82%. The performance of all these classifiers are weaker than the optimized neural network as the best model in this paper.



بحث و نتیجه گیری

Fill this part plz!

شکیر[3] و همکاران (2022) با الگوریتم جنگل تصادفی دقت 77 درصد

عدلی و سهید[4] (2021) با الگوریتم MKNN با دقت 71 درصد

مقروبین[5] و همکاران (2020) با الگوریتم Naïve Bayes با دقت 95 درصد

آلدینو و سولیستیانی[6] (2020) با الگوریتم درخت تصمیم با دقت 87 درصد

کورنیادی[7] و همکاران (2018) با الگوریتم k-NN با دقت 95 درصد

باسه[8] و همکاران (2019) با الگوریتم رگرسیون لجستیک با دقت 79 درصد

روهمایانی[9] (2020) با الگوریتم Naïve Bayes با دقت 77 درصد

ناگی و مولانتای[10] (2018) با الگوریتم درخت تصمیم با دقت 80 درصد

[3] Shakir
[4] Adli & Sahid
[5] Muqorobin
[6] Aldino & Sulistiani
[7] Kurniadi
[8] Basu
[9] Rohmayani
[10] Nagy & Molontay

فهرست منابع

Cattaneo, M., Civera, A., Meoli, M., & Paleari, S. (2020). Analysing policies to increase graduate population: do tuition fees matter? European Journal of Higher Education, 10(1), 10-27.

Trinh, N. T. H. (2021). Factors Affecting the Tuition Fee Policy of Public Higher Education. Business Excellence and Management, 11(3), 22-42.

Shakir, A. K., Sutradhar, S., Sakib, A. H., Akram, W., Saleh, M. A., & Abedin, M. Z. (2022). A Systematic Study on Tertiary Level Student Tuition Fee Waiver Management During Pandemic Using Machine Learning Approaches. In Advances in Information, Communication and Cybersecurity: Proceedings of ICI2C'21 (pp. 259-273). Springer International Publishing.

El Massari, H., Gherabi, N., Mhammedi, S., Sabouri, Z., & Ghandi, H. (2022). Ontology-based decision tree model for prediction of cardiovascular disease. Indian J. Comput. Sci. Eng, 13(3), 851-859.

Basu, K., Basu, T., Buckmire, R., & Lal, N. (2019). Predictive models of student college commitment decisions using machine learning. Data, 4(2), 65.

Adli, D., & Sahid, D. S. S. (2021). UKT (Single Tuition) Classification Prediction uses MKNN (K-Nearest Neighbor Modification) algorithm. International ABEC, 81-84.

Muqorobin, M., Kusrini, K., Rokhmah, S., & Muslihah, I. (2020). Estimation System For Late Payment Of School Tuition Fees. International Journal of Computer and Information System (IJCIS), 1(1), 1-6.

Aldino, A. A., & Sulistiani, H. (2020). Decision Tree C4. 5 Algorithm For Tuition Aid Grant Program Classification (Case Study: Department Of Information System, Universitas Teknokrat Indonesia). Jurnal Ilmiah Edutic: Pendidikan dan Informatika, 7(1), 40-50.

Rohmayani, D. (2020). Analysis of Student Tuition Fee Pay Delay Prediction Using Naive Bayes Algorithm With Particle Swarm Optimization (Case Study: Politeknik TEDC Bandung). Jurnal Teknologi Informasi dan Pendidikan, 13(2), 1-8.

Kurniadi, D., Abdurachman, E., Warnars, H. L. H. S., & Suparta, W. (2018, November). The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm. In IOP conference series: materials science and engineering (Vol. 434, No. 1, p. 012039). IOP Publishing.

Aparicio, G., Iturralde, T., & Rodríguez, A. V. (2023). Developments in the knowledge-based economy research field: A bibliometric literature review. Management Review Quarterly, 73(1), 317-352.

Nagy, M., & Molontay, R. (2018, June). Predicting dropout in higher education based on secondary school performance. In 2018 IEEE 22nd international conference on intelligent engineering systems (INES) (pp. 000389-000394). IEEE.

Zumeta, W. (2010). The great recession: Implications for higher education. In M. F. Smith (Ed.), NEA 2010 Almanac of Higher Education (pp. 29-42). Washington, DC: National Education Association.

Bound, J., Braga, B., Khanna, G., & Turner, S. (2019). Public universities: The supply side of building a skilled workforce. RSF: The Russell Sage Foundation Journal of the Social Sciences, 5(5), 43-66.

Kelly, W., & Shale, D. (2004). Does the Rising Cost of Tuition Affect the Socio-Economic Status of Students Entering University? Online Submission.

Ma, J., Baum, S., Pender, M., & Welch, M. (2016). Trends in College Pricing 2016. The College Board. https://trends.collegeboard.org/sites/default/fles/2016-trends-college-pricing-web_0.pdf

Flores, S. M., & Shepherd, J. C. (2014). Pricing out the disadvantaged? The effect of tuition deregulation in Texas public four-year institutions. The ANNALS of the American Academy of Political and Social Science, 655(1), 99-122.

Bell, E. (2020). The politics of designing tuition-free college: How socially constructed target populations influence policy support. The Journal of Higher Education, 91(6), 888-926.

Johnson, J., Rochkind, J., Ott, A. N., & DuPont, S. (2009). With their whole lives ahead of them. Public Agenda. 1-48. Retrieved from: http://www.publicagenda.org/files/theirwholelivesaheadofthem.

Williams, J. C. (2016). "It's always with you, that you're different": Undocumented students and social exclusion. Journal of Poverty, 20(2), 168-193.

Long, B. T. (2006). College tuition pricing and federal financial aid: Is there a connection. Testimony before the US Senate Committee on Finance, Hearing: Report card on tax exemptions and incentives for higher education: Pass, fail, or need improvement.

Boatman, A., Evans, B. J., & Soliz, A. (2017). Understanding loan aversion in education: Evidence from high school seniors, community college students, and adults. AERA Open, 3(1), 1–16.

Hemelt, S. W., & Marcotte, D. E. (2011). The impact of tuition increases on enrollment at public colleges and universities. Educational Evaluation and Policy Analysis, 33(4), 435-457.

Cunningham, A. F., & Santiago, D. (2008). Student aversion to borrowing: Who borrows and who doesn't. Institute for Higher Education Policy and Excelencia in Education. December.

Su, Y., Weng, K., Lin, C., & Zheng, Z. (2021). An improved random forest model for the prediction of dam displacement. IEEE Access, 9, 9142-9153.

Paulsen, M. B., & St. John, E. P. (2002). Social class and college costs: Examining the financial nexus between college choice and persistence. The Journal of Higher Education, 73(2), 189–236.

Perna, L. W. (2006). Understanding the relationship between information about college costs and financial aid and students' college-related behaviors. American Behavioral Scientist, 49, 1620–1635.

Mubarak, A. A., Cao, H., & Zhang, W. (2022). Prediction of students' early dropout based on their interaction logs in online learning environment. Interactive Learning Environments, 30(8), 1414-1433.

Dass, S., Gary, K., & Cunningham, J. (2021). Predicting student dropout in self-paced MOOC course using random forest model. Information, 12(11), 476.

Agrusti, F., Mezzini, M., & Bonavolontà, G. (2020). Deep learning approach for predicting university dropout: A case study at Roma Tre University. Journal of e-Learning and Knowledge Society, 16(1), 44-54.

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. European Journal of Higher Education, 10(1), 28-47.

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018, July). Perspectives to predict dropout in university students with machine learning. In 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI) (pp. 1-6). IEEE.

Acton, R. (2018). The impact of public tuition subsidies on college enrollment decisions: Evidence from michigan.

Allen, D., & Wolniak, G. C. (2019). Exploring the effects of tuition increases on racial/ethnic diversity at public colleges and universities. Research in Higher Education, 60(1), 18-43.

Andrieu, S. C., & John, E. P. S. (1993). The influence of prices on graduate student persistence. Research in Higher Education, 34(4), 399-425.

Arendt, J. N. (2013). The effect of public financial aid on dropout from and completion of university education: evidence from a student grant reform. Empirical Economics, 44, 1545-1562

Callender, C., & Jackson, J. (2005). Does the fear of debt deter students from higher education?. Journal of social policy, 34(4), 509-540.

Callender, C., & Jackson, J. (2008). Does the fear of debt constrain choice of university and subject of study?. Studies in higher education, 33(4), 405-429.

Chen, R., & DesJardins, S. L. (2008). Exploring the effects of financial aid on the gap in student dropout risks by income level. Research in Higher education, 49, 1-18.

Dickson, L., & Pender, M. (2013). Do in-state tuition benefits affect the enrollment of non-citizens? Evidence from universities in Texas. Economics of Education Review, 37, 126-137.

Dietrich, H., & Gerner, H. D. (2012). The effects of tuition fees on the decision for higher education: evidence from a German policy experiment. Economics Bulletin, 32(3), 2407-2413.

Dowd, A. C. (2004). Income and financial aid effects on persistence and degree attainment in public colleges. education policy analysis archives, 12, 21.

Dwenger, N., Storck, J., & Wrohlich, K. (2012). Do tuition fees affect the mobility of university applicants? Evidence from a natural experiment. Economics of Education Review, 31(1), 155-167.

Garrett, H., & Greene, A. (2018). Tuition and Fees for Public In-State Four-Year Institutions and the White/Black Education Gap (2006-2016).

Havranek, T., Irsova, Z., & Zeynalova, O. (2017). Tuition Reduces Enrollment Less Than Commonly Thought.

Hemelt, S. W., & Marcotte, D. E. (2011). The impact of tuition increases on enrollment at public colleges and universities. Educational Evaluation and Policy Analysis, 33(4), 435-457.

Su, Y., Weng, K., Lin, C., & Zheng, Z. (2021). An improved random forest model for the prediction of dam displacement. IEEE Access, 9, 9142-9153.

Hemelt, S. W., & Marcotte, D. E. (2016). The changing landscape of tuition and enrollment in American public higher education. RSF: The Russell Sage Foundation Journal of the Social Sciences, 2(1), 42-68.

Hübner, M. (2012). Do tuition fees affect enrollment behavior? Evidence from a 'natural experiment'in Germany. Economics of Education Review, 31(6), 949-960.

Kim, J., DesJardins, S. L., & McCall, B. P. (2009). Exploring the effects of student expectations about financial aid on postsecondary choice: A focus on income and racial/ethnic differences. Research in Higher Education, 50(8), 741-774.

Coppola Jr, E. A., Rana, A. J., Poulton, M. M., Szidarovszky, F., & Uhl, V. W. (2005). A neural network model for predicting aquifer water level elevations. Groundwater, 43(2), 231-241.

Martinello, F. (2015). Course withdrawal dates, tuition refunds, and student persistence in university programs (No. 1501).

Moulin, L., Flacher, D., & Harari-Kermadec, H. (2016). Tuition fees and social segregation: lessons from a natural experiment at the University of Paris 9-Dauphine. Applied Economics, 48(40), 3861-3876.

Neill, C. (2015). Rising student employment: The role of tuition fees. Education Economics, 23(1), 101-121.

Shin, J. C., & Milton, S. (2008). Student response to tuition increase by academic majors: Empirical grounds for a cost-related tuition policy. Higher Education, 55(6), 719–734.

Vasigh, B., & Hamzaee, R. G. (2004). Testing sensitivity of student enrollment with respect to tuition at an institution of higher education. International Advances in Economic Research, 10(2), 133-149.

Nick, T. G., & Campbell, K. M. (2007). Logistic regression. Topics in biostatistics, 273-301.

El Massari, H., Gherabi, N., Mhammedi, S., Sabouri, Z., & Ghandi, H. (2022). Ontology-based decision tree model for prediction of cardiovascular disease. Indian J. Comput. Sci. Eng, 13(3), 851-859.