Machine Learning and Diversity: Predicting Demographic Patterns Among Tuition and Non-Tuition Students

Abstract

This study used Machine Learning algorithms to indicate Tehran University students' demographic composition and diversity and differentiate tuition-paying from tuition-free students. This research area is highlighted due to growing tuition fee concerns worldwide. Furthermore, Machine Learning algorithms have become increasingly popular for predictive modeling, enhancing the research's precision. Tehran University's 14 faculties were examined in a comprehensive census from 2015 to 2020. The accuracy of prediction was maximized by harnessing supervised learning algorithms. It was found that the optimized neural network and random forest algorithms achieved the highest precision rate of 82%. Based on an F1 score of 0.89, the optimized neural network algorithm outperformed an optimized random forest algorithm. The machine learning approach differentiates student diversity and demographic composition based on tuition status using grade, age, department, and GPA factors. This study established a significant association between these variables using nonparametric correlation tests in SPSS. Therefore, machine learning algorithms help predict and understand the diversity and composition of the student population affected by tuition fees.

***Keywords*:** Access to Higher Education, Machine Learning in Higher Education, Diversity in Higher Education, Equity, Tuition.

Introduction

Based on Cattaneo et al. (2020), a nation's human capital is inarguably its most valuable resource and asset in today's highly competitive global landscape. As a result, the role of higher education in nurturing and developing skilled human capital is crucial to shaping the nation's future workforce. Therefore, it ensures a country's economy's long-term advancement and development throughout all socioeconomic facets. According to Trinh (2021), university graduates must be educated, and workforce competencies must be enhanced, including continuous professional development and training to achieve this transformation. The importance of higher education to students, families, institutions, organizations, and governments is clear from this perspective. Universities' critical role in advancing sustainable development and developing sustainable societies is further recognized. Based on Aparicio et al. (2023), they are crucial for providing countries with a competitive edge regionally and globally. Zumeta (2011), however, acknowledges that public funds for higher education have declined dramatically since the 2008 recession. Based on Bound et al. (2019), higher education institutions generally enjoy relatively favorable financial positions during economic prosperity but are often targeted for budget cuts during recessions.

It is also noticeable that university tuition fees have increased consistently in the past few decades. A study by Kelly & Shale (2004) showed that tuition costs in the United States soared 125% between 1980 and 1998, while household income increased by just 1%. There was a noticeable increase in tuition rates across various regions of the United States in 2000. The tuition at public universities has increased by 62%, and the tuition at public two-year colleges has risen by 40%. There was also a substantial increase in tuition fees at private universities, increasing by 42%. United States South and Southwest regions saw a particularly steep rise in tuition fees, according to Ma et al. (2016). The tuition increases in some geographical regions of the United States, notably

the state of Texas, reached such proportions that over six years from 2003 to 2009, there was a staggering 72% escalation in tuition fees, according to findings by Flores & Shepherd (2014). Bell (2020) also found that more than 70% of parents expressed apprehension about the cost of their children's college educations, confirming other studies. According to a separate survey performed by Johnston et al. (2009), 70% of university students admitted that financial constraints forced them to leave their studies, with 52% citing their inability to pay tuition.

There are both advantages and disadvantages to setting and raising tuition fees. Firstly, the negative consequences are discussed and then examine the positive ones. According to Williams (2016), tuition increases may adversely affect students from lower socio-economic levels. These increases may hinder their ability to achieve economic and social mobility, thus widening the societal class gap. Less privileged students often have access to financial aid and loans even when tuition fees rise. Although, this approach has not always proven successful, according to empirical evidence. According to Long (2006), many families and students are confused about their loan debts due to financial aid and loans. Student debt can adversely affect students' academic decisions during university enrollment and even after graduation, as debt-related anxieties may deter them from making life choices, including marriage, starting families, and homeownership. Based on Boatman et al. (2017), escalating university tuition fees results in declining access for low-income students. It is also possible for enrollment, retention, and graduation rates to decrease due to tuition hikes. The prevailing trend in studies suggests that these rates are falling despite conflicting evidence. Studies by Hemelt & Marcotte (2011), Cunningham & Santiago (2008), Perna (2006), and Paulsen & St. John (2002) indicated that higher tuition rates are associated with lower enrollments, retentions, and graduations. However, the establishment and growth of tuition fees have positive consequences. Students can view these fees as an investment that helps them choose

a field of study based on informed research. Universities' revenue streams can be augmented by them, according to policymakers. Financial constraints are a reality for universities all over the world. Tuition fees are, therefore, an effective method of addressing universities' financial deficits by bolstering their income.

A unique and diverse tuition framework led to the selection of the University of Tehran as the focus of this study. Tehran University has various program types, including E-learning programs, full-time programs with associated fees, and some campuses such as the Aras, Alborz, and Kish campuses. Student tuition fees will vary depending on which program they are enrolled in. Due to its reputation, the University of Tehran consistently attracts more applicants than other universities for these programs. It is notable that aside from the diversity of tuition amounts, the amounts themselves also exhibit annual continuous increases. Even though tuition fees and their escalation have been of great significance, little research has been conducted to examine what tuition fees have on Tehran University and, by extension, on higher education in Iran. In addition, no prior studies have examined the impact of tuition fees on student diversity and demographics. The multifaceted significance of predicting tuition-influenced diversity and makeup of student populations includes: (1) University administrators and educational policymakers can formulate more effective programs tailored to the needs of these groups if they know the student population's composition, which includes factors such as age, marital status, gender, number of children, geographical residence, and others; (2) Demographic diversity is crucial for social justice and equal access to tuition fees. It is a fundamental principle of higher education to ensure that all members of society have access to higher education without discrimination based on race, gender, economic status, etc. The aim of fostering diversity is to ensure that everyone has access to higher education; (3) Tuition policies within each university can be optimized by understanding the

diversity and composition of student populations. To enhance the efficiency of tuition-related policies, grants, and loans are allocated for educational expenses and student loans, respectively.

Research background

The authors of this research conducted an extensive search and found no prior studies related to the application of artificial intelligence algorithms in predicting demographic patterns among tuition and non-tuition students. Due to this lack of existing research, the authors have highlighted studies that have utilized artificial intelligence algorithms in assessing tuition-related activities.

The research titled "A Systematic Examination of Tertiary Level Student Tuition Fee Waiver Management During the Pandemic Using Machine Learning Approaches" was carried out by Shakir et al. (2022). The findings showed that Cross-validation (CV) yielded accuracy rates of 77.82 and 74.49%, respectively, for Random Forest Regression (RFR) and Decision Tree Regression (DTR) before applying cross-validation. As a result of the CV application, the DTR accuracy rate decreased to 71.41%, and the RFR accuracy rate dropped to 76.90%. According to Mubarak et al. (2021), a study titled "Prediction of Students' Early Dropout Based on their Interaction Logs in an Online Learning Environment" showed that their proposed models exceeded the baseline performance of Machine Learning Models by 84% in predicting students at risk of dropping out. For classifying new UKT students, Adli and Sahid (2021) applied the MKNN (K-Nearest Neighbor Modification) algorithm in their study titled "UKT (Single Tuition) Classification Prediction." The accuracy rate of their classification results was 71% when K-Fold Cross Validation was used. Another study titled "Predicting Student Dropout in a Self-Paced MOOC Course Using a Random Forest Model" was conducted by Dass et al. (2020). An analysis of their model demonstrated that it could predict dropouts and continuations in MOOC courses.

This was done with 94.5% AUC, 87.5% accuracy, 87.5% recall, 87.5% F1 score, and 88% precision.

The study was conducted by Agrusti et al. (2020) under the title "A Deep Learning Approach to Predicting University Dropout: A Case Study at Roma Tre University." Their results were compared to those from Bayesian networks based on deep learning models. The deep learning models varied in accuracy from 67.1% to 94.3% for first- and third-year students, respectively. Kemper et al. (2020), in their study titled "Predicting Student Dropout: A Machine Learning Approach," found that decision trees performed slightly better than logistic regressions. The two methods, however, achieved high classification accuracy, exceeding 83% after the first semester, and high prediction accuracy of 95% after three semesters. A research study performed by Aldino and Sulistiani (2020) titled "Decision Tree C4.5 Algorithms for Tuition Aid Grant Program Classification (Case Study: Department of Information Systems, Universitas Teknokrat Indonesia)." All classification components' accuracy, precision, and recall scores were 87% for the ten-fold cross-validation. It appears that the model is capable of being implemented into systems effectively. The study was conducted by Rohmayani (2020) under the title "Analysis of Student Tuition Fee Payment Delay Prediction Using the Naive Bayes Algorithm with Particle Swarm Optimization (Case Study: Politeknik TEDC Bandung)." According to the results of three classification models that used the Naive Bayes algorithm with Particle Swarm Optimization (PSO), accuracy, precision, recall, and AUC were 73.94%, 78.50%, and 0.771, respectively. Despite an additional 3 seconds in execution time, this was achieved. Basu et al. (2019) conducted a study titled "Predictive Models of Student College Commitment Decisions Using Machine Learning." In this study, the logistic regression classifier had an AUC score of 79.6% and outperformed other models in predicting acceptance of admission offers to students. The

significance of this research lies in its demonstration of how institutions can enhance the accuracy of their class size estimations using machine learning algorithms, which will lead to improved resource allocation and more control over net tuition income. Nagy and Molontay (2018) examined a study titled "Predicting Dropout in Higher Education Based on Secondary School Performance." In a 10-fold cross-validation analysis, Gradient Boosted Trees and Deep Learning, as the best models, produced AUC values of 0.808 and 0.811.

Methodology

In recent years, machine learning has become an essential field of research. A supervised and unsupervised learning method can be distinguished based on this learning method. We briefly present the methods we used to construct the predictive models in this study, which focused on supervised learning.

A) Algorithms used

A-1) Logistic regression

Logistic regression (Hosmer and Lemeshow 1989) can be viewed as arising from a Bernoulli model. Given a set of predictors, xn, we wish to determine the probability of a binary outcome yn. We define a probability model:

$$\mathrm{P}(Y_n = 1 \mid \mathbf{x}_n) \doteq \sigma(\mathbf{w} \cdot \mathbf{x}_n)$$

with corresponding likelihood function:

$$\begin{aligned} \mathrm{P}(\mathbf{y} \mid \mathbf{x}_n, n = 1 \ldots N) &= \prod_n \sigma(\mathbf{w} \cdot \mathbf{x}_n)^{y_n} \left(1 - \sigma(\mathbf{w} \cdot \mathbf{x}_n)\right)^{(1-y_n)} \\ &= \prod_n \sigma(\mathbf{w} \cdot \mathbf{x}_n)^{y_n} \sigma(-\mathbf{w} \cdot \mathbf{x}_n)^{(1-y_n)} \end{aligned}$$

where the logistic function

$$\sigma(\theta) = \frac{1}{1 + \exp\,[-\theta]}$$

is a continuous increasing function mapping any real valued θ into the interval (0, 1), and thus is suitable for representing the probability of a Bernoulli trial outcome. A useful variant for scientific and sociology experiments employs a binomial (Bickel and Doksum 2001) rather than Bernoulli formulation to facilitate repeated trials.

A-2) Neural network

Neural networks have gained widespread application across various real-world problems, particularly excelling in scenarios that are too intricate for conventional technologies. Their primary advantage lies in their capacity to tackle complex problems lacking algorithmic solutions or situations where algorithmic solutions are excessively convoluted. Neural networks are especially suited for tasks that humans excel at but challenge traditional computers. These tasks encompass pattern recognition and forecasting, requiring the discernment of data trends. The true strength of neural networks lies in their ability to represent both linear and non-linear relationships while autonomously learning these relationships from the provided data. Traditional linear models are ill-suited for modeling data featuring non-linear attributes. Among neural network models, the multi-layer perceptron (MLP) stands as one of the most common. This supervised network requires a desired output for learning purposes. Its objective is to create a model that accurately maps input to output using historical data, enabling the model to generate output when the desired result is unknown. The learning process of MLP and many other neural networks employs a technique known as backpropagation. During backpropagation, the neural network is repeatedly exposed to input data. After each exposure, the network's output is compared to the desired output, and an error is computed. This error is then fed back (backpropagated) to the neural network to adjust the

weights. This iterative process aims to decrease the error with each iteration, gradually aligning the neural model with the desired output. This iterative refinement process is referred to as "training" (Hyndman & Athanasopoulos, 2014).

A-3) Decision tree

A decision tree is a supervised learning method primarily employed for classification tasks, although it can also be adapted for regression purposes. The decision tree commences with a root node, representing the initial decision point for dividing the dataset. This root node consists of a single feature that effectively segregates the data into distinct classes. Each division is represented by an edge connecting either to another decision node, housing an additional feature for further data division into homogeneous groups, or to a terminal node, which provides predictions for the classes. This iterative process of partitioning the data into two binary segments is termed recursive partitioning (James et al., 2013).

A-4) Random Forest

A random forest is an ensemble method that extends the decision tree approach by generating multiple decision trees. Instead of employing all features for each decision tree, a subset of features is randomly selected to create individual decision trees within the random forest. Each tree makes predictions regarding class outcomes, and the final class prediction for the model is determined through a majority vote among these trees (Hastie et al., 2009).

A-5) Other algorithms for prediction

K-Nearest Neighbors, K-Nearest Neighbors, nearest neighbor methods particularly in the context of machine learning and decision theory. It begins with a foundational overview of machine learning and decision theory, emphasizing their role in classification and regression tasks. Nearest

neighbor methods operate by considering the labels of the K-nearest data points in the data space. These methods are particularly effective for large datasets and low-dimensional problems due to their local nature. Furthermore, they have variants suitable for multi-label classification, regression, and semi-supervised learning, making them applicable to a wide range of machine learning scenarios. Decision theory is discussed as it provides valuable insights into the outcomes of nearest neighbor learning (Kramer & Kramer, 2013).

Support Vector Classifier, Support vector machines (SVMs) are part of the maximum margin classifier family and operate on the principle of structural risk minimization (SRM). SRM seeks to choose a hypothesis function with low capacity from a nested set of functions, with the goal of minimizing both the true error rate (classification error on unseen data) and the empirical error rate (error on the training dataset). (Burges, 1998).

AdaBoost Classifier, AdaBoost is such an algorithm developed by Freund and Schapire (1996) at AT&T labs. The advantages of AdaBoost include less memory and computational requirements. Boosting is a method of combining performances of weak learners to build a strong classifier whose performance is better than any of the individual weak classifiers. A weak learner is a simple rule whose classification accuracy may be only slightly better than a random guess. Enhanced performance of the resulting combined classifier is due to added weights given to training examples which are difficult to classify.

Gradient Boosting is introduced as an ensemble method, specifically a boosting algorithm. Boosting algorithms are designed to enhance the predictive power of a base class of models that are relatively weak in terms of prediction, such as decision trees. They achieve this by transforming the base model into a more potent learning algorithm capable of producing stronger predictions. In classification tasks, this transformation involves creating a weighted majority vote over the base

models, while in regression tasks, it entails forming a linear combination of the base models. Gradient Boosting is highlighted as one of these boosting algorithms, emphasizing its ability to significantly improve prediction quality (Beygelzimer et al., 2015).

Linear Discriminant Analysis (LDA) is a technique aimed at finding an optimal linear transformation to reduce the dimensionality of the original data. The primary objective of LDA is to identify a linear transformation that maximizes the separation between different classes within the reduced-dimensional space. To achieve this, LDA formulates criteria for dimension reduction that focus on maximizing the variance between different classes while minimizing the variance within each class (Park & Park, 2007).

Quadratic Discriminant Analysis (QDA) is a common technique employed in supervised classification tasks. QDA models the likelihood of each class as a Gaussian distribution and utilizes posterior distributions to make predictions about the class for a given test point (Hastie et al., 2001). To determine the Gaussian parameters for each class, QDA typically employs Maximum Likelihood (ML) estimation based on the training data.

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

All students at Tehran University throughout the five-year period from 2015 to 2020 were included in the statistical population. Our data collection comprised a comprehensive survey of students from various faculties, including Management, Psychology, Educational Sciences, Economics, and several technical and engineering faculties. There were 11 faculties in this category. Over the last five years, these data have been collected and analyzed.

It is estimated that 24% of the total student population is tuition-paying, and 76% is tuition-free. Machine learning models are developed by dividing datasets into two subsets: training and testing. Most data (80%) is used to train the model, while a smaller portion (20%) is used to test itb

C) Features used in the dataset

Independent attributes 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 attribute: Type. The table below shows the attributes used in the dataset;

Table 1: Features used in the dataset

Features	References
degree	Dietrich & Gerner (2012)/ Dwenger, Storck & Wrohlich (2012)/ Hemelt & Marcotte (2016)/ Moulin et al. (2016)
Faculty	Callender & Jackson (2008)
age	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)
type of course	Vasigh & Hamzaee (2004)/ Dwenger, Storck & Wrohlich (2012)/ Havranek, Irsova, & Zeynalova, (2017)/ Hemelt & Marcotte. (2011)/Garrett & Greene (2018)/ Neill (2015)/ Kim, DesJardins & McCall (2009)
nationality	Moulin et al. (2016)
marital status	Callender & Jackson (2008)/ Callender & Jackson (2005)
number of children	Callender & Jackson (2008)/ Hemelt & Marcotte (2016)/ Neill (2015)/ Callender & Jackson (2005)
year	Dickson & Pender (2013)/ Dwenger, Storck & Wrohlich (2012)/ Havranek, Irsova, & Zeynalova,/ (2017)/ Hemelt & Marcotte. (2011)/ Moulin et al. (2016)/ Allen & Wolniak (2019)/ Arendt (2013)
financial aid	Vasigh & Hamzaee (2004)/ Havranek, Irsova, & Zeynalova, (2017)/ Neill (2015)/ Andrieu & John. (1993)/ Shin & Milton (2008)/ Dowd (2004)/ Kim, DesJardins & McCall (2009) Chen & DesJardins (2008)
gender	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)
Academic decisions (decision to transfer, decision to withdraw, decision to remove semester, decision to leave, decision to be a guest and decision to change major)	Hübner (2012)/ Moulin et al. (2016)/ Arendt (2013)/ Martinello (2015)
GPA	Hemelt & Marcotte (2016)/ Andrieu & John. (1993)/ Arendt (2013)/ Dowd (2004)/ Chen & DesJardins (2008)

For this purpose, we have employed a dummy variable, feature 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, 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 = female); Academic decisions (Transfer, drop out, Remove, Leave, Change filed and Guest) (1 = Yes, 0 = No).

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 first part) Application of Machine Learning algorithms in predicting

The findings are presented in several sections below;

1) Descriptive statistics indicators of data

In the first part of the findings, descriptive statistics are reported. Table (2) describes the results of descriptive statistics

Table 2: Descriptive statistics indicators[1]

Features	Count	Mean	Median	Std	Min	25%	50%	75%	Max
Department	13709.0	*	2.0	*	100	1.00	2.00	4.00	4.0
Age	13709.0	26.68	*	3.40	17.00	22.00	25.00	29.00	67.0
Grade	13709.0	*	1.0	*	0.00	0.00	1.00	1.00	2.0
Type	13709.0	*	1.0	*	0.00	1.00	1.00	1.00	1.0
GPA	13709.0	16.77	2.15	1.91	10.04	15.67	17.17	18.25	20.0
Nationality	13709.0	*	1.0	*	0.00	1.00	1.00	1.00	1.0
Marital status	13709.0	*	1.0	*	0.00	1.00	1.00	1.00	1.0
Children	13709.0	0.12	*	0.49	0.00	0.00	0.00	0.00	9.0
Year	13709.0	4.043	*	1.58	1.00	3.00	4.00	5.00	6.0
Financial Aid	13709.0	*	1.0	*	0.00	1.00	1.00	1.00	1.0
Gender	13709.0	*	1.0	*	0.00	0.00	1.00	1.00	1.0
Transfer	13709.0	*	0.0	*	0.00	0.00	0.00	0.00	1.0
Drop out	13709.0	*	0.0	*	0.00	0.00	0.00	0.00	1.0
Remove	13709.0	*	0.0	*	0.00	0.00	0.00	0.00	1.0
Leave	13709.0	*	0.0	*	0.00	0.00	0.00	0.00	1.0
Change filed	13709.0	*	0.0	*	0.00	0.00	0.00	0.00	1.0

[1] We have employed the mean for continuous variables and utilized the median for discrete variables.

Guest	13709.0	*	0.0	*	0.00	0.00	0.00	0.00	1.0

2) Common findings between models

Gaining insights into the relationships among input features allows for strategic feature engineering, leading to a more informed and refined approach to building machine learning models. The heatmap in Figure 1, represents the correlation matrix, where each cell's color and intensity indicates the strength and direction of the correlation between variables. The variables with the highest influence on the target (Type) are age, grade, department. According to tuition, 76% of students are tuition-free and 24% are tuition-paying students. A fundamental step in machine learning model development is division of a dataset into two subsets: one for training and one for testing. the larger portion of the data, 80%, is allocated for training the model, while the smaller portion 20% is reserved for testing the model's performance

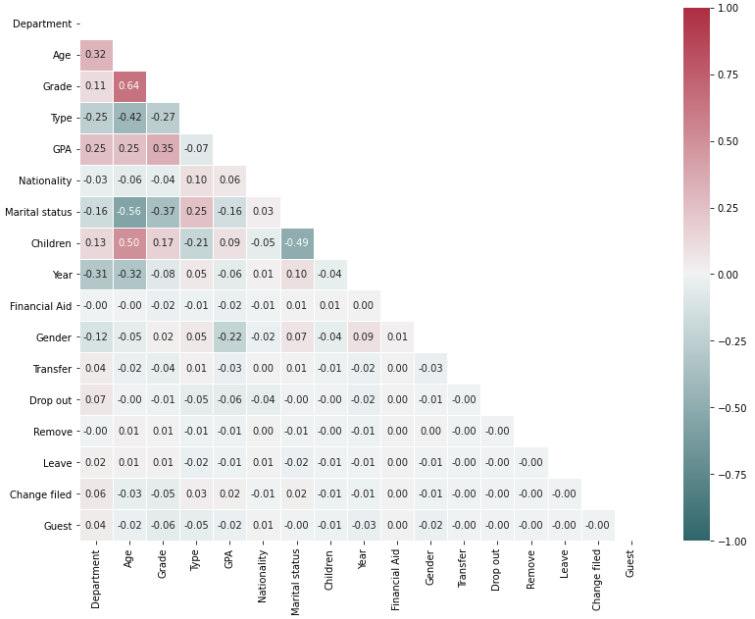

Figure 1: Correlation coefficients between variables

3) Analytical findings

3-1) l-Logistic regression

Figure 2 shows the coefficients assigned to each predictor variable based on the logistic regression model. The coefficients indicate how strongly and in what direction each predictor variable correlates with the log odds of each binary outcome (type variable). There is a high correlation between age and department features, which reflects how influential they are in predicting outcome probability.

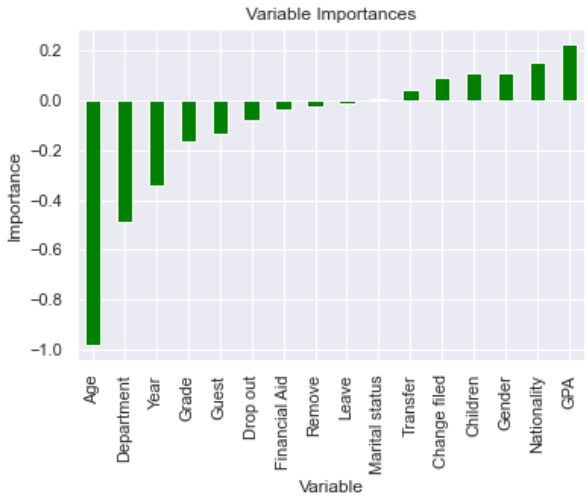

Figure 2: Importance of variables in logistic regression

Logistic regression results are shown in the following table. In class[2] 1, the model correctly identifies positive instances with an 81% precision rate, and class[3] 0 instances are accurately classified with a 66% precision rate. It also ranks among the top models for recalls with 95%. An appropriate balance between precision and recall is achieved by the harmonized F1 score of 87%.

Table 3: Results of applying logistic regression

	precision	recall	f1-score	accuracy

[2] Tuition-free students
[3] Tuition-paying students

0 1	0.66 0.81	0.30 0.95	0.41 0.87	0.79

3-2) Neural network

Two hidden layers of 20 and 15 neurons use the ReLU activation function to receive and transmit the initial data to the input layer. The output layer is ultimately responsible for generating the network's predictions. Weights and biases are accounted for in this structure, are trainable parameters. Loss function variations are illustrated in the graph below (determining the difference between actual and predicted values and optimizing to minimize error as the optimization target). In the training dataset, the loss function is represented in blue; in the testing dataset, it is expressed in orange. The disparity between actual and predicted values decreases with each training iteration (epoch), which indicates that the neural network model is becoming more accurate.

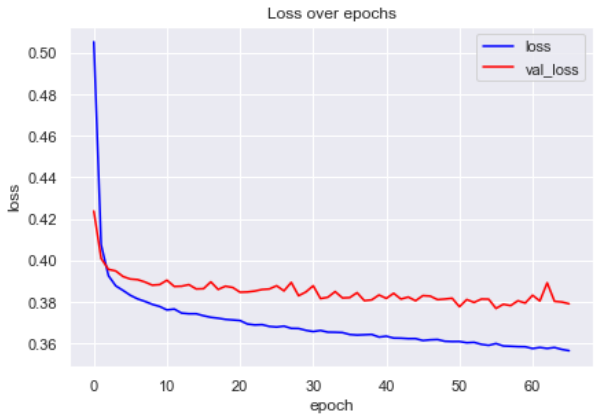

Figure 3: Loss function chart

Table 4: The results of neural network application

	precision	recall	f1-score	accuracy
0 1	0.66 0.86	0.54 0.91	0.59 0.88	0.81

The results of the neural network application are listed in the following table. The F1 score for class 0 and class 1 is 53 and 88%, respectively, higher than logistic regression.

Hyperparameter optimization of neural network model

The neural network's structure was optimized using Keras hyperparameter tuning techniques in this analysis to increase prediction accuracy. This procedure selects the most appropriate values for hyperparameters. These include the number of hidden layers, the activation function, the number of neurons within each layer, the strength of regularization, and the learning rate. These hyperparameters were fine-tuned using two methods: random search and grid search. These techniques entail systematically exploring the hyperparameter space to identify optimal settings to maximize neural networks' performance. The table below describes the structure of a neural network. As indicated in this table, there are two intermediate layers with 32 neurons with a 10% drop rate and a binary end layer for prediction in the neural network structure.

Table 5: Neural network 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, 32)	33

Total params: 1,633
Trainable params: 1,633
Non-trainable params: 0

In Figure 4, the left side shows the progress of the loss function, and the right indicates the model's accuracy in each period. The accuracy metric measures the percentage of correctly classified samples in the training or validation dataset. Accuracy typically improves over time, like loss function.

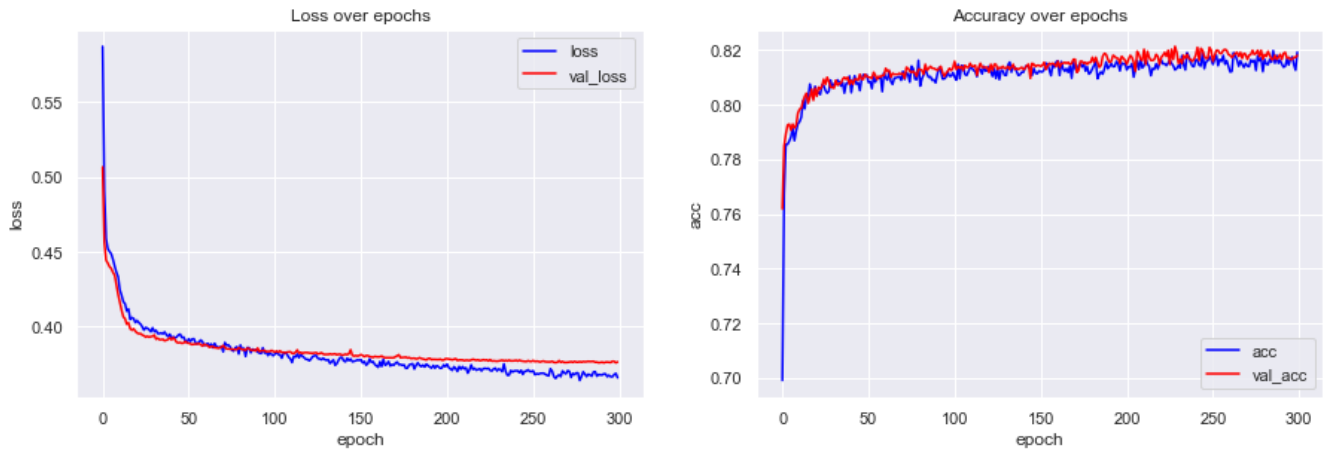

Figure 4: Loss function chart

An optimized neural network application is presented in the following table. After optimization, the F1 score for both classes is improved.

Table 6: The results of the optimized neural network application

	precision	recall	f1-score	accuracy
0	0.70	0.47	0.56	0.82
1	0.85	0.93	0.89	

3-3) Decision tree

In Figure 5, each feature is illustrated for its impact on the collective decision-making process within the ensemble of decision trees. Decision trees identify which features (such as GPA and grade) significantly influence the model's predictions based on importance scores assigned to each feature.

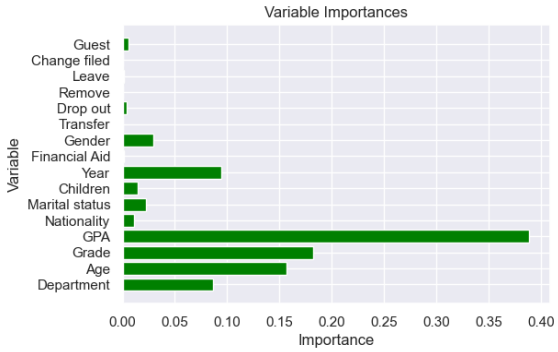

Figure 5: Importance of variables in decision tree algorithm

The decision tree algorithm results are presented in the following table. There is the highest precision (85%) in class 1 and the weakest recall (85%) among all models.

Table 7: The results of applying the decision tree algorithm

	precision	recall	f1-score	accuracy
0	0.54	0.54	0.54	0.77
1	0.85	0.85	0.85	

3-4) Random Forest

As a first step, the default hyperparameters are used during training to forecast random forest outcomes. Prediction accuracy decreases when a specific input feature is excluded from the analysis so we can assess the importance of individual input features in prediction. As shown in Figure 5, similar to the decision tree, GPA and Age, in order, make the most significant contributions to the model's decision-making process.

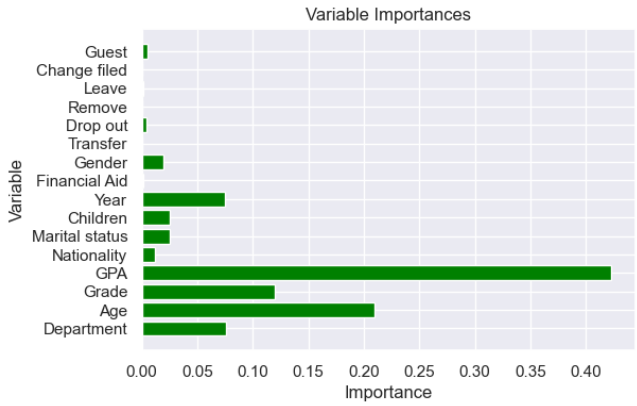

Figure 5: Importance of variables in random forest algorithm

Random forest hyperparameters are adjusted to improve the model's predictive accuracy or other relevant performance measures. The most common hyperparameters for random forests include the maximum depth of trees, the minimum sample size for splitting internal nodes, the number of trees, and the maximum number of features to consider per split. Random or grid search methods are used to optimize this process. In these methods, the model is trained and cross-validated on a dataset to test various combinations of hyperparameters. Increasing the number of trees in the random forest model improves predictions, as illustrated in Figure 6.

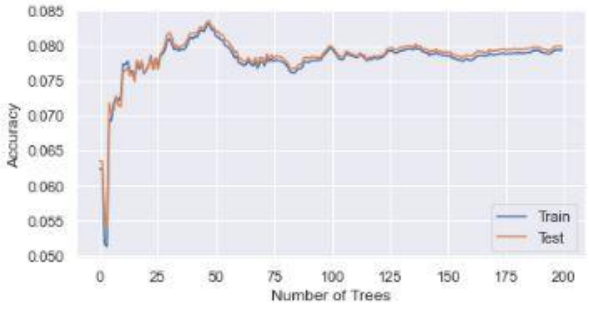

Figure 6: Prediction model of the optimized random forest

Based on both default and optimized hyperparameters, Figure 9 shows the performance of a random forest. As a result of training, recall and F1 scores for class 1 improved slightly.

Table 8: The results of applying the optimized random forest

algorithm		precision	recall	f1-score	accuracy
the optimized random forest	0 1	0.72 0.83	0.41 0.95	0.52 0.89	0.82
random forest	0 1	0.60 0.86	0.56 0.88	0.58 0.87	0.80

4) Comparison of the results of the used algorithms

Various well-known machine learning classifiers are presented in Figure 7 to provide a comprehensive overview of this analysis. Support Vector Classifier, K-Nearest Neighbors, Gradient Boosting, Quadratic Discriminant, Gaussian Naive Bayes, AdaBoost classifier, and Linear discriminant are some of the algorithms used in this study. Gradient Boosting and AdaBoost had a remarkable 82% accuracy among the models. In this study, all these classifiers fail to perform as well as the optimized neural network, which is the top-performing model.

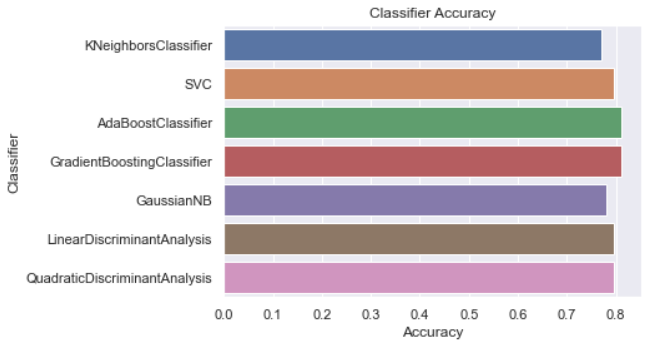

Figure 7: Other algorithms for prediction

The second section) Extracting the demographic composition and diversity of students

Students' diversity and demographic composition are analyzed through frequency and statistical measures (such as standard deviation, mean, and mode). Details are given in table (9).

Table 9: Demographic composition and diversity of students

Variables	Mean	Mode	Frequency	Standard Deviation

GPA[4]_0 > Type[5] GPA_1> Type	17.01 16.70	*	*	1.65 1.98
Age[6]_0> Type Age_1 > Type	31.46 25.17	*	*	7.52 5.15
Grade_0> Type Grade_1> Type	*	mode Grade[7]_0: 1 mode Grade_1: 0	Grade_0: 2832 Grade_1: 5203	*
Gender_0> Type Gender_1> Type	*	mode Gender[8]_0: 1 mode Gender_1: 1	Gender_0: 1812 Gender_1: 6343	*
Marital Status_0> Type Marital Status_1>Type	*	Marital status[9]_0: 1 Marital status_1: 1	Marital status_0: 2249 Marital status_1: 9314	*
Gender>Marital status> Type	*	mode Gender_0_0: 1 mode Gender_0_1: 1 mode Gender_1_0: 1 mode Gender_1_1: 1	Gender_0_0: 950 Gender_0_1: 3558 Gender_1_0: 1299 Gender_1_1: 5756	*
Age> GPA > Type	*	mode GPA_1_0: 2.0 mode GPA_1_1: 2.0 mode GPA_2_0: 2.0 mode GPA_2_1: 2.0 mode GPA_3_0: 2.0 mode GPA_3_1: 2.0 mode GPA_4_0: 2.0 mode GPA_4_1: 2.0	GPA_1_0: 473 GPA_1_1: 1889 GPA_2_0: 770 GPA_2_1: 2245 GPA_3_0: 368 GPA_3_1: 1174 GPA_4_0: 102 GPA_4_1: 660	*
GPA > Grade> Type	*	mode GPA_1_0: 1 mode GPA_1_1: 1 mode GPA_2_0: 1 mode GPA_2_1: 0 mode GPA_3_0: 1 mode GPA_3_1: 0 mode GPA_4_0: 1 mode GPA_4_1: 0	GPA_1_0: 900 GPA_1_1: 1525 GPA_2_0: 1280 GPA_2_1: 1735 GPA_3_0: 539 GPA_3_1: 1404 GPA_4_0: 113 GPA_4_1: 1049	*
GPA > Gender > Type	*	mode GPA_1_0: 1 mode GPA_1_1: 1 mode GPA_2_0: 1 mode GPA_2_1: 1 mode GPA_3_0: 1 mode GPA_3_1: 1 mode GPA_4_0: 0 mode GPA_4_1: 1	GPA_1_0: 19 GPA_1_1: 1727 GPA_2_0: 983 GPA_2_1: 3588 GPA_3_0: 562 GPA_3_1: 905 GPA_4_0: 253 GPA_4_1: 123	*
Department[10] > GPA > Type	DP_1_0: 16.30 DP_1_1: 16.04 DP_2_0: 17.87 DP_2_1: 18.23 DP_3_0: 16.80 DP_3_1: 16.81 DP_4_0: 17.02 DP_4_1: 17.40823263460748	*	*	DP_1_0: 1.61 DP_1_1: 2.00 DP_2_0: 1.29 DP_2_1: 1.17 DP_3_0: 1.68 DP_3_1: 1.84 DP_4_0: 1.64 DP_4_1: 1.59
Department > Grade> Type	*	mode Gender_1_0: 1 mode Gender_1_1: 0 mode Gender_2_0: 1	Gender_1_0: 644 Gender_1_1: 3234 Gender_2_0: 620	*

[4] The GPA grading system is structured as follows: scores between 90 and 100 receive an "A," scores between 80 and 90 receive a "B," scores between 70 and 80 receive a "C," and scores less than 70 a "D."

[5] type (1= tuition-free students, 0= tuition-paying students)

[6] The age categorization is delineated as follows: individuals aged 16 to 22 fall into category 1, those aged 22 to 30 belong to category 2, individuals aged 30 to 40 are classified under category 3, and those aged 40 and beyond are categorized as 4.

[7] Grade (0 = bachelor's degree, 1 = master's degree and number 2 = PhD)

[8] gender (1 = men, 0 = female)

[9] marital status (1 = single, 0 = married)

[10] Department (1 = technical and engineering faculties, 2 = psychology and educational sciences, 3 = economics and 4 = management faculty)

		mode Gender_2_1: 0 mode Gender_3_0: 1 mode Gender_3_1: 0 mode Gender_4_0: 1 mode Gender_4_1: 1	Gender_2_1: 657 Gender_3_0: 213 Gender_3_1: 357 Gender_4_0: 1355 Gender_4_1: 1216	
Age > Marital status> Type	*	mode Gender_1_0: 1 mode Gender_1_1: 1 mode Gender_2_0: 1 mode Gender_2_1: 1 mode Gender_3_0: 1 mode Gender_3_1: 1 mode Gender_4_0: 1 mode Gender_4_1: 1	Gender_1_0: 719 Gender_1_1: 5460 Gender_2_0: 376 Gender_2_1: 1108 Gender_3_0: 231 Gender_3_1: 669 Gender_4_0: 923 Gender_4_1: 2076	*
Age > Gender > Type	*	mode Grade_0_0: 2.0 mode Grade_0_1: 1.0 mode Grade_1_0: 2.0 mode Grade_1_1: 2.0	Grade_0_0: 1558 Grade_0_1: 0 Grade_1_0: 1558 Grade_1_1: 3131	*
Department > Marital Status > Type	*	mode Gender_1_0: 1 mode Gender_1_1: 1 mode Gender_2_0: 1 mode Gender_2_1: 1 mode Gender_3_0: 1 mode Gender_3_1: 1 mode Gender_4_0: 1 mode Gender_4_1: 1	Gender_1_0: 719 Gender_1_1: 5460 Gender_2_0: 376 Gender_2_1: 1108 Gender_3_0: 231 Gender_3_1: 669 Gender_4_0: 923 Gender_4_1: 2076	*
Department > Age > Type	*	mode DP_1_0: 2.0 mode DP_1_1: 2.0 mode DP_2_0: 2.0 mode DP_2_1: 2.0 mode DP_3_0: 2.0 mode DP_3_1: 2.0 mode DP_4_0: 2.0 mode DP_4_1: 2.0	DP_1_0: 625 DP_1_1: 3204 DP_2_0: 331 DP_2_1: 790 DP_3_0: 163 DP_3_1: 484 DP_4_0: 594 DP_4_1: 1490	*
GPA > Marital status> Type	*	mode DP_1_0: 2.0 mode DP_1_1: 2.0 mode DP_2_0: 2.0 mode DP_2_1: 2.0 mode DP_3_0: 2.0 mode DP_3_1: 2.0 mode DP_4_0: 2.0 mode DP_4_1: 2.0	DP_1_0: 625 DP_1_1: 3204 DP_2_0: 331 DP_2_1: 790 DP_3_0: 163 DP_3_1: 484 DP_4_0: 594 DP_4_1: 1490	*
Grade> Age > Type	*	mode Grade_0_0: 2.0 mode Grade_0_1: 1.0 mode Grade_1_0: 2.0 mode Grade_1_1: 2.0	Grade_0_0: 1558 Grade_0_1: 0 Grade_1_0: 1558 Grade_1_1: 3131	*

Based on GPA and type, tuition-paying students generally have slightly higher GPAs than tuition-free. In addition, tuition-free students have significantly higher GPA variability than tuition-paying students, as reflected in their higher standard deviations. According to Age and Type, tuition-paying students generally are older with more age diversity, as the higher standard deviation shows, precisely the converse for tuition-free students. Regarding degree and type, the data set shows that the "bachelor" degree is the most common degree with the highest frequency (2,832

times) of all degrees among tuition-free students. On the other hand, the "master's" degree has the highest frequency (5203 times) among tuition-paying students.

Most tuition-paying and tuition-free students in the GPA category of A were in the age group of 20-30. On the contrary, most tuition-free students were in the GPA category of D. There seems to be a positive correlation between tuition fees and students' motivation to achieve a better GPA. According to the GPA category A, tuition-paying master's male students represented the dominant portion, whereas tuition-free students tended to be undergraduates in the same GPA category. Based on gender distribution, most students in the GPA category D were tuition-free undergraduates. An examination of the GPA analysis revealed that there were more males in both tuition-paying and tuition-free categories within the GPA category of A. The proportion of female students was more significant among tuition-paying students, suggesting women prefer tuition-paying courses. Regarding marital status, the GPA category of A had more single students than tuition-free students. However, married women constituted a larger portion of tuition-paying students, suggesting that married women choose tuition-paying courses more often. The GPA category of B was higher in the Faculty of Management for students who paid tuition fees. By contrast, technical and engineering faculties have the most extensive tuition-free students. In addition, most tuition-paying students choose educational sciences, psychology, and management at the master's degree level. According to an analysis of student diversity across faculties, the Faculty of Management had the highest proportion of married students. In contrast, the Faculty of Engineering and Technology had the highest proportion of single students. There is a notable concentration of tuition-paying students in the Faculty of Management in the 20 to 30 age range, making up most of the student population. Among the tuition-paying students within the GPA category of A, it was found that the majority were in the 20 to 30 age group. Nevertheless, this age

group and GPA category had substantial tuition-paying students. In contrast, tuition-free students usually were ranged in age from the early 20s to the 30s. In analyzing the GPA category of A based on gender, a small percentage of female students were tuition-paying. However, the proportion of female tuition-paying was higher in the GPA category of C.

The third section) Combining the findings of Machine Learning algorithms with the demographic composition and diversity of students

Table 10 shows the non-parametric correlations between the variables deemed most significant by machine learning algorithms. The population's diversity and makeup were analyzed according to these variables. After that, pairwise correlations between these pivotal variables were examined.

Table 10: Correlation between variables with high importance in random forest algorithm

Variables	Metric	Phi	Cramer's V	Lambda	Chi-Square	Contingency Coefficient	Uncertainty Coefficient
Department > Type	Approximate Significance	.000	.000	.000	.000	.000	.000
	Value	.280	.280	.065	1077.583	.270	.047
Grade> Type	Approximate Significance	.000	.000	.000	.000	.000	.000
	Value	.431	.431	.130	2545.183	.396	.142
Age > Type	Approximate Significance	.000	.000	.000	.000	.000	.000
	Value	.109	.109	.032	161.753	.108	.007
GPA > Type	Approximate Significance	.000	.000	*	.000	.000	.000
	Value	.388	.388	*	2062.640	.362	.101

The table above provides tests for assessing the correlation between qualitative and discrete variables. Based on these tests, a probability value is computed called "Approximate Significance." This value represents the statistical significance or p-value of the statistical test. Statistical significance is determined by determining the p-value of a test. In all tests, significant correlations were observed between the variables. The highest relationship value between grade and type

variables was 0.431 in the Phi test. Also, Cramer's V test showed 0.431 as the highest value for the same variables. The two variables mentioned above also registered the highest value of 0.130 in the Lambda test. In addition, in the Uncertainty Coefficient and Contingency Coefficient tests, the variables obtained the highest values at 0.396 and 0.142, respectively. Finally, both variables got the highest values in the Chi-square tests.

Discussion and conclusion

This study used Machine Learning algorithms to indicate Tehran University students' demographic composition and diversity and differentiate tuition-paying from tuition-free students. Tuition fees significantly influence students' academic choices in higher education. As a result of reduced government funding (due to inflation) and rising spending in other sectors, universities face financial challenges. Students' tuition fees at higher education institutions significantly impact their academic choices. University financial constraints are mainly due to escalating service and welfare costs and heightened competition in higher education (Mughan et al, 2022). Increasing higher education tuition fees is one fundamental approach many countries use to resolve this dilemma. In response to reduced state funding and escalating expenditures, families and students are being forced to assume a more significant share of their educational expenses (Ison, 2022). Students' academic choices and encounters have been impacted by the escalation of tuition fees in different ways. There are still some uncertainties regarding tuition's impact, equating to an indeterminate black box. Based on review of existing literature, it has identified three perspectives regarding tuition's influence:

Based on the first perspective, tuition increases have not affected enrollment or continuing education decisions (Flores, 2010; Darolia and Potochnick, 2015; Kaushal, 2008; Cameron and

Heckman, 2001). Secondly, tuition fee increases have unfavorably affected academic decisions regarding enrollment and pursuit of higher education (Allen and Wolniak, 2019; Garrett and Greene, 2018; Garibaldi et al, 2007; Hemelt and Marcotte, 2011; Boatman et al, 2017; Cunningham and Santiago, 2008; Perna, 2006). Ultimately, it is the type of university and location of residence that impact the outcomes of academic decisions regarding enrollment and continued education, according to the third perspective (Do, 2004; Bozick & Miller, 2014; Dwenger et al, 2012; Frenette, 2005). There has been a lack of clarity regarding the role of tuition in students' academic decision-making due to the three identified trends regarding the impact of tuition. In consequence, empirical evidence has influenced student demographics and diversity. Thus, tuition has a significant effect on demographic structures.

The neural network algorithm and the optimized neural network algorithm produced the best results in our research. An optimized neural network algorithm achieved an accuracy of 82%. Compared to previous studies, these predictive values are higher. The random forest algorithm achieved 77% accuracy, according to Shakir et al. (2022). The MKNN algorithm was reported to be 71% accurate by Adli and Sahid (2021). The decision tree algorithm was 87% accurate in the Aldino and Sulistiani (2020) study. The logistic regression algorithm was found to have 79% accuracy by Basu et al. (2019). According to Rohmayani (2020), 77% accuracy was achieved by using the Naive Bayes algorithm. Finally, Nagy and Molontay (2018) used a decision tree algorithm to predict with 80% accuracy. Research on forecasting minority enrollment rates among students should be conducted using machine learning algorithms in the future. This study would help assess the demographics and inclusivity of students in tuition-paying higher education institutions. A comparative study should be conducted using data from other Tehran-based universities to analyze the impact of tuition fees on diversity and student composition. It is possible

to investigate the diversity and composition of selected multinational student groups using comparative research methods such as the George Brady model, which compares tuition policies and financial aid.

References

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

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

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.

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.

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.

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.

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

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

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.

Beygelzimer, A., Hazan, E., Kale, S., & Luo, H. (2015). Online gradient boosting. Advances in neural information processing systems, 28.

Bickel, P. J., & Doksum, K. A. (2001). Mathematical statistics (2nd ed., Vol. 1). Englewood Cliffs: Prentice Hall.

Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

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.

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.

Bozick, R., & Miller, T. (2014). In-state college tuition policies for undocumented immigrants: Implications for high school enrollment among non-citizen Mexican youth. *Population Research and Policy Review*, 33(1), 13-30

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167.

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.

Cameron, S. V., & Heckman, J. J. (2001). The dynamics of educational attainment for black, hispanic, and white males. *Journal of political Economy*, 109(3), 455-499.

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.

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.

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.

Darolia, R., & Potochnick, S. (2015). Educational "when," "where," and "how" implications of in-state resident tuition policies for Latino undocumented immigrants. *Review of Higher Education*, 38(4), 507-535.

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

Davidson, A. (2015). Is college tuition really too high. *The New York Times Magazine*, 25.

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.

Do, C. (2004). The effects of local colleges on the quality of college attended. *Economics of Education Review*, 23(3), 249–257.

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.

Dynarski, S., Libassi, C. J., Michelmore, K., & Owen, S. (2018). Closing the gap: The effect of a targeted, tuition-free promise on college choices of high-achieving, low-income students (No. w25349). *National Bureau of Economic Research.*

Flores, S. M. (2010). State dream acts: The effect of in-state resident tuition policies and undocumented Latino students. *Review of Higher Education*, 33(2), 239-283.

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.

Frenette, M. (2005). The impact of tuition fees on university access: Evidence from a large-scale price deregulation in professional programs (No. 2005263e). Statistics Canada, Analytical Studies Branch.

Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In icml (Vol. 96, pp. 148-156).

Garibaldi, P., Giavazzi, F., Ichino, A., & Rettore, E. (2007). College cost and time to obtain a degree: Evidence from tuition discontinuities. *NBER Working Paper*, 12863.

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

Hastie, T. (2001). Tibshirani R. Friedman J.: The elements of statistical learning. Friedman J.: The elements of statistical learning.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

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.

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.

Hosmer, D. E., & Lemeshow, S. (1989). Applied logistic regression. New York: Wiley.

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

Hyndman, R. J., & Athanasopoulos, G. (2014). Optimally reconciling forecasts in a hierarchy. Foresight: The International Journal of Applied Forecasting, (35), 42-48.

Ison, M. P. (2022). The viability of tuition-free community college. *Educational Policy*, 36(5), 1054-1077.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

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.

Kaushal, N. (2008). In-state tuition for the undocumented: Education effects on Mexican young adults. *Journal of Policy Analysis and Management*, 27(4), 771–792.

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

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

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.

Kramer, O., & Kramer, O. (2013). K-nearest neighbors. Dimensionality reduction with unsupervised nearest neighbors, 13-23.

Kreighbaum, A. (2019). Democratic contenders draw contrasts on free college, student debt. Inside Higher Ed. https://www.insidehighered.com/news/2019/06/28/democratic-contenders-draw-contrasts-free-college-student-debt#.XaN2yPaPGG0.link.

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.

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

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

Millon, J. (2021). Free Tuition in Higher Education. *Journal of the Student Personnel Association at Indiana University*.

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.

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.

Mughan, S., Sherrod Hale, J., & Woronkowicz, J. (2022). Build It and will They Come? The Effect of Investing in Cultural Consumption Amenities in Higher Education on Student-Level Outcomes. *Research in Higher Education*, 63(1), 60-91.

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.

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

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

Ontivero-Ortega, M., Lage-Castellanos, A., Valente, G., Goebel, R., & Valdes-Sosa, M. (2017). Fast Gaussian Naïve Bayes for searchlight classification analysis. Neuroimage, 163, 471-479.

Park, C. H., & Park, H. (2008). A comparison of generalized linear discriminant analysis algorithms. Pattern Recognition, 41(3), 1083-1097.

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.

Potochnick, S. (2014). How states can reduce the dropout rate for undocumented immigrant youth: The effects of in-state resident tuition policies. *Social science research*, 45, 18-32.

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.

Rosenberg, B. (2019). Free public college is a terrible idea. The Chronicle of Higher Education. https://www.chronicle.com/article/Free-Public-College-Is-a/247134?cid=wcontentgrid_40_2

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.

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.

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

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.

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.

Winograd, M., & Staisloff, R. (2016). Student debt. *CQ Researcher*, 26(41),965-988.

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.