**Predicting the Academic Performance of Tehran University Students with Supervised Algorithms and Comparing it with the Results of Previous Studies**

**Abstract**

The aim of this research is to predict the academic performance of Tehran University students with supervised algorithms and compare it with previous studies' results. Therefore, the data of Tehran University students in the period (2016-2021) was collected by the whole census method in 14 faculties. Supervised learning algorithms used in the research included: linear regression, Neural network regression (Neural Network optimized), Random forest (Random Forest optimized) and Decision Tree. Among the models, the best value RMSE (1.53) related to the algorithm Random Forest optimized.

**Keywords: Artificial Intelligence, Academic Performance, GPA, Supervised Learning, Higher Education**

**Introduction**

Undoubtedly, students are the most critical asset of universities and higher education institutions. Students are recognized as the most important human capital for higher education institutions and society in contributing to the comprehensive development of a country. Especially in the present era, with the transition from a production-based economy to a knowledge-based economy, this issue has become important, especially in developing countries (Rehman et al, 2021). The knowledge economy is an economy that uses knowledge as the main engine of economic growth. This economy is the main factor of growth, wealth creation, and employment in all fields (Phale et al, 2021). One critical element for realizing a knowledge-based economy is human capital, especially university capital (students). Many studies have investigated the effect of knowledge-based human capital on economic growth, which proves the role of knowledge-based capital on economic growth (Ngepah et al, 2021). A country's economic and social development has a direct relationship with students' academic performance. Students' academic performance plays a vital role in creating the best quality graduates who will become the leaders and human resources of a particular country and thus be responsible for that country's social and economic development (Ali et al, 2009). For this reason, the issue of student's academic performance has always been given considerable attention in research in different countries. Students' performance is influenced by many factors, such as psychological, economic, social, individual, and environmental factors. Although these factors strongly influence student performance, they vary from country to country and from individual to individual (Singh et al, 2016). In this research, we have equated students' GPAs with academic performance.

GPA is an essential indicator of showing students' academic ability. We group students into good and bad categories based on GPA. Also, GPA is a crucial indicator in students' subsequent decisions, such as deciding to continue their studies, applying for studies in other countries, etc. In the labor market, GPA is a vital sign of the student's educational quality in the eyes of employers. Many factors affect the grade. Among these factors: are study motivation, regular study throughout the semester, attendance at lectures, use of counseling during studies, watching specific educational videos on YouTube, first-year university average, high school average, frequency of illegal drug use, amount of social media use, the number of hours of sleep during the day, the duration of the study, the provisional GPA of the first semester, the temporary GPA of the second semester, the student's employment status, the results of the level determination test and IELTS score, gender, the type of scholarship awarded, previous academic record, the type of admission, talent Education, the province of high school education, the family background of students, which has been investigated in various studies of predicting students' GPA using artificial intelligence algorithms (Falát & Piscová, 2021; Khan et al, 2021; Maulana & Defriani, 2020; ALLAH, 2019; Putpuek et al, 2018; Ahmad et al, 2015).

Therefore, students' GPAs are a function of various factors, which are different in each country's higher education system. Because every country has its own economic, social, cultural, and political conditions, which affect the higher education system of that country at different levels and, accordingly, the student's GPA (academic performance). In this study, we predict the GPA of Tehran University students. Tehran University is a public university and one of Iran's largest higher education centers. This university is known as the "Mother University" and the "Symbol of Higher Education" of Iran. Also, artificial intelligence algorithms have been used for forecasting due to the importance and rapid growth of artificial intelligence achievements in the fields of study and its high accuracy in predicting compared to common statistical analysis. Therefore, this research has been done to predict the academic performance of Tehran University students with supervised algorithms and compare it with previous studies results. The innovation and contribution of this research in the existing literature is that a study using artificial intelligence algorithms has been conducted in Iran. In addition, the results of this study will be compared with international studies that predict students' GPAs using artificial intelligence algorithms. It is worth noting that the GPA variable in Iran's higher education system is continuous (from 0 to 20). While in most of the world's higher education systems, the GPA is a discrete variable (A+ to F).

**Research literature**

Further, in Table (1), the researchers that used learning algorithms to predict GPA are mentioned.

Table 1: Studies on learning algorithms and GPA prediction

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Results | hyper parameter optimization | overfit | metrics | Algorithms | year | author | Row |
| 84.6% | has not been stated | has not been stated | Accuracy | neural network | 2015 | Abu-Naser etal | 1 |
| 6Only the importance of the variables is mentioned | has not been stated | has not been stated | Accuracy | J48 decision tree | 2016 | Al-Barrak, & Al-Razgan | 2 |
| Naïve Bayes classifier overall prediction accuracy of 86% | has not been stated | has not been stated | Accuracy Precision  Recall  Specificity | Naïve Bayes, Neural Network, and Decision  Tree | 2016 | Mueen etal | 3 |
| Naïve Bayes  (68.7%) | has not been stated | has not been stated | TP  FP  Precision  Recall  F-Measure  ROC  Class | Naïve Bayes, LibSVM, J48, Random Forest, and JRip | 2016 | Shamsi, M & Lakshmi | 4 |
| The best algorithm was naïve Bayes classification for the first data set, with 98 percent accuracy, and decision trees for the second data set, with 78 percent accuracy | has not been stated | has not been stated | Accuracy  Precision  Recall  F | Linear regressionDecision trees  Naïve Bayes classifier | 2017 | Pojon | 5 |
| Tree and Na¨ıve Bayes classifiers which are having 63.7 % and 72.6 % respectively | has not been stated | has not been stated | Accuracy  Sensitivity  Specificity | The hybrid of Decision Tree and Naïve Bayes algorithms | 2017 | Chuan etal | 6 |
| J48  Precision  0.616 | has not been stated | has not been stated | Precision  Recall | J48  RepTree  Random Tree | 2018 | Hamoud etal | 7 |
| Decision Tree  66.9%. | has not been stated | has not been stated | Accuracy  Precision  Recall  FMeasure | Decision Tree and Random Forest | 2018 | Kaunang & Rotikan | 8 |
| Random Forest 99% | has not been stated | has not been stated | J48 Classifier  BayesNet  Random Forest  PART Classifier | Sensitivity (Recall or True positive rate)  Precision  F-score  Accuracy (MAE, RMSE, RAE, RRSE) | 2018 | Hussain etal | 9 |
| Support Vector Machine  69.15%. | has not been stated | has not been stated | Accuracy | Naïve Bayes  Support Vector Machine (SVM) | 2018 | Kesumawati & Utari | 10 |
| The experimental results show that accuracy algorithm  (AC) of 78.57% with true positive rate (TP) of 76.72% by using quality training  data of 90% have best performance accuracy value | has not been stated | has not been stated | Accuracy | Decision Tree C4.5 | 2018 | Budiman etal | 11 |
| The ANN (fully connected feed forward multilayer ANN) model achieved the best performance that is equal to 0.807 and achieved the best classification accuracy that is equal to 77.04% | has not been stated | has not been stated | Precision  Recall  F- Measure  AccuracyClassification Error  ROC index | Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression | 2019 | Altabrawee etal | 12 |
| J48 و Random Forest  ?? | has not been stated | has not been stated | Mean Absolute Error (MAE)  Root Mean Squared Error (RMSE):  Relative Absolute Error (RAE):  Root Relative Squared Error (RRSE) | Naive Bayes (NB)  Logistic Classifier (LC)  J48 Classifier (J48)  Support Vector Machine (SVM)  Random Forest (RF)  Logistic Regression (LR) | 2019 | Canagareddy etal | 13 |
| MLP 97% | has not been stated | has not been stated | ACCURACY | Decision Tree (J48) and Artificial Neural Network  (ANN) | 2019 | Alsalman etal | 14 |
| Decision Tree 98.86% | has not been stated | has not been stated | ACCURACY | Decision Tree  Neural Network algorithm  Naive Bayes  Support Vector Machine  K-Nearest Neighbor | 2019 | Kumar & Salal | 15 |
| Deep Neural Network model 85% | Is mentioned | has not been stated | Accuracy, Precision, Recall and F-Score | Naïve Bayes  Decision Tree  Random Forest  Support Vector Machine Deep Neural Network | 2019 | Vijayalakshmi & Venkatachalapathy | 16 |
| K-Nearest-Neighbor 54% | has not been stated | has not been stated | Accuracy | K-Nearest-Neighbor  Decision Tree  Naïve Bayes | 2019 | Mohammadi etal | 17 |
| accuracy of up to 100% | has not been stated | has not been stated | CHI SQUARE | Linear Regression | 2019 | Bum etal | 18 |
| Multilayer Perceptron – MLP and Naïve Bayes 89.19% | has not been stated | has not been stated | TP rate  FP rate  Precision  Recall | OneR algorithm  PART  J48 Decision tree  Random Tree (RT)  Random Forest (RF)  Naïve Bayes  Support Vector Machine  Multilayer Perceptron - MLP | 2020 | Ha etal | 19 |
| SVM regression method with Kernel RBF which is equal to 0.1505. | has not been stated | has not been stated | MAE  RMSE | Support vector regression  Linear Regression  Simple Linear Regression | 2020 | Dewi & Widiastuti | 20 |
| Catboost  91% | has not been stated | has not been stated | Accuracy  ROC-AUC  Recall  (RMSE)  F1 | Logistic Regression  Random Forest  XGBoost  Catboost | 2020 | Oreshin etal | 21 |
| The support vector  regression linear algorithm  R2 score 0.83442 | has not been stated | has not been stated | MAE  MSE  RMSE  R2 score | Linear regression  Support vector regression | 2021 | Dabhade etal | 22 |

In the following, the variables selected in this research are examined with the variables of the studies that predicted the grade point average using artificial intelligence algorithms.

***Gender*** (Abu-Naser et al., 2015; Shamsi & Lakshmi, 2016; Pojon, 2017; Chuan et al., 2017; Hamoud et al., 2018; Kaunang & Rotikan, 2018; Hussain et al., 2018; Kesumawati & Utari, 2021; Budiman et al., 2018; Altabrawee et al., 2019; Al-Salman et al., 2019; Vijayalakshmi and Venkatachalapathi, 2019: Falát & Piscová, 2022; Gipson, 2018); ***Nationality*** (Al-Barrak & Al-Razgan, 2016: Vijayalakshmi and Venkatachalapathi, 2019), ***Department or Faculty*** (Hamoud et al., 2018; Kesumawati & Utari, 2021; Altabrawee et al., 2019; Al-Salman et al., 2019); ***Age*** (Mueen et al., 2016; Pojon, 2017; Hamoud et al., 2018; Budiman et al., 2018; Al-Salman et al., 2019; Vijayalakshmi and Venkatachalapathi, 2019; : Falát & Piscová, 2022), ***Type of Course*** (Shamsi & Lakshmi, 2016); ***Financial Aid*** (Shamsi & Lakshmi, 2016: Hamoud et al., 2018; Gipson, 2018); ***Family Size[[1]](#footnote-1)*** (Chuan et al., 2017; Hussain et al., 2018; Al-Salman et al., 2019; Falát & Piscová, 2022), ***Marital Status*** (Hussain et al., 2018; Altabrawee et al., 2019; Al-Salman et al., 2019); **Year** (Shamsi & Lakshmi, 2016: Kaunang & Rotikan, 2018; Falát & Piscová, 2022), ***Degree*** (Cheewaprakobkit, 2015).

**Research Method**

This section mentions data collection, variables, used algorithms, validation criteria, and implementation of algorithms.

1. **Data collection**

This research's statistical population consists of all Tehran University students in the last five years (2016-2021). from using the method of all census data of students of faculties (all the members of the population are enumerated); Psychology and educational sciences, management, economics, and All engineering faculties (11 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 students in all faculties of Tehran University.

1. **Variables**

Independent variables in this research include; Faculty, age, degree, type of course, nationality, marital status, number of children, year, financial aid, and gender. Dependent variable: GPA was. In the table below, its description is given;

Table 2: Independent and dependent variables

|  |  |  |
| --- | --- | --- |
| Variable type | Independent variables | Description |
| discrete variable | degree | This variable has three educational levels (bachelor's, master's, and Ph.D.) |
| discrete variable | Faculty | The studied faculties include four faculties (management, economics, psychology, educational sciences, and engineering). |
| Continuous variable | age | \* |
| discrete variable | type of course | The division of courses is based on tuition payment. In our category, students who paid an amount to the university as tuition fees every semester were labeled as tuition-paying students, and students who did not pay any amount as tuition fee each semester were marked as non-tuition-paying students. In Iran, based on the national exam (university entrance), higher ranks do not pay for education, and lower classes pay an amount as tuition for each semester. |
| discrete variable | nationality | Regarding nationality, students are divided into two groups (Iranian students and international students). |
| discrete variable | marital status | In terms of marital status, students are classified into two categories (single and married). |
| Continuous variable | number of children | \* |
| discrete variable | year | The years consist of 6 years (2017-2022). |
| discrete variable | financial aid | Financial aid is given by the government to low-income students. Based on this, students are divided into two categories (recipients of financial assistance and non-recipients of financial aid) |
| discrete variable | gender | In terms of gender, there are two groups (women and men). |
| Continuous variable | GPA | \* |

**C) Evaluation Metrics**

We used three metrics: Mean Square Error (MSE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) to evaluate the performance of a regression models. They are measures of the average difference between the predicted values and the actual values. MSE is the average of the squared differences between the predicted values and the actual values. It is given by:

where y\_pred is the predicted value, y\_actual is the actual value, and n is the number of data points. MAE is the average of the absolute differences between the predicted values and the actual values. It is given by:

where |x| represents the absolute value of x.

RMSE is the square root of the MSE. It is given by:

The RMSE is often used as a preferred metric because it has the same unit as the dependent variable and is more interpretable than the MSE. In general, a lower value of these metrics indicates better performance of the model. However, the choice of metric depends on the specific problem and the goals of the analysis.

**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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | mean (for continuous variables) and median (for discrete variables) | standard deviation | minimal | 25% | 50% | 75% | maximum |
| Faculty | 2.0 | \* | 1.00 | 1.00 | 2.00 | 4.00 | 4.0 |
| Age | 26.688818 | 6.402831 | 2.00 | 22.00 | 25.00 | 29.00 | 67.0 |
| degree | 1.0 | \* | 0.00 | 0.00 | 1.00 | 1.00 | 2.0 |
| type of course | 1.0 | \* | 0.00 | 1.00 | 1.00 | 1.00 | 1.0 |
| GPA | 16.777675 | 1.918553 | 10.04 | 15.67 | 17.17 | 18.25 | 20.0 |
| nationality | 1.0 | \* | 0.00 | 1.00 | 1.00 | 1.00 | 1.0 |
| marital status | 1.0 | \* | 0.00 | 1.00 | 1.00 | 1.00 | 1.0 |
| number of children | 0.125830 | 0.495526 | 0.00 | 0.00 | 0.00 | 0.00 | 9.0 |
| year | 4.0 | \* | 1.00 | 3.00 | 4.00 | 5.00 | 6.0 |
| financial aid | 1.0 | \* | 0.00 | 1.00 | 1.00 | 1.00 | 1.0 |
| GPA | 1.0 | \* | 0.00 | 0.00 | 1.00 | 1.00 | 1.0 |

Table (3) shows the descriptive statistics of 11 variables investigated in the research. Age, grade point average, and number of children were continuous among the variables. The average values of these variables were 26.688818, 16.777675, and 0.125830, respectively. Eight other variables were discrete, for which the median value was calculated.

***b) Common findings between models***

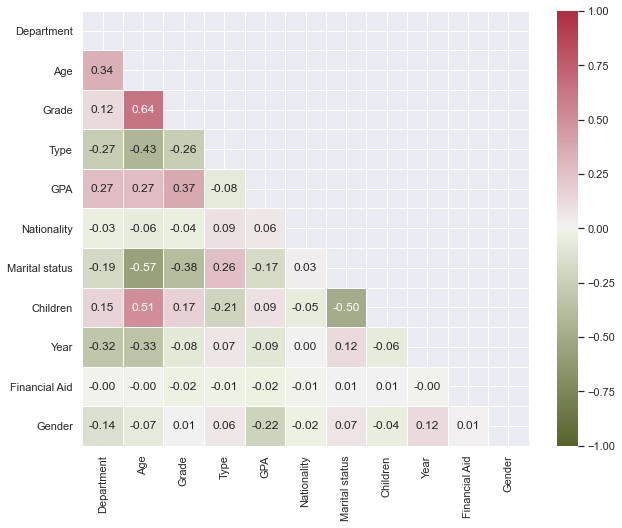


Figure 1: Correlation coefficients between variables

***c) Analytical findings***

*C-1) linear regression*

Linear regression is a commonly used statistical technique for modeling the relationship between a dependent variable and one or more independent variables. It is used to predict the value of the dependent variable based on the values of the independent variables. In linear regression, a linear relationship is assumed between the dependent variable and the independent variable(s), and the aim is to find the best-fitting line that minimizes the sum of the squared differences between the predicted values and the actual values. Linear regression can be used for both simple and multiple regression models. Simple linear regression involves one independent variable, while multiple linear regression involves two or more independent variables

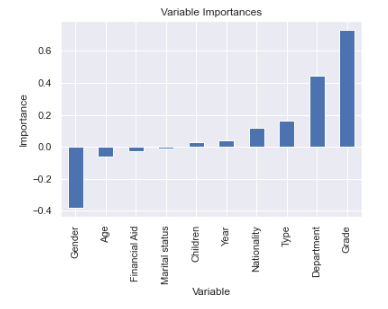


Figure 2: Importance of variables in multiple linear regression

In the following table, the results of applying multiple linear regression are stated.

Table 4: Results of applying multiple linear regression

|  |  |
| --- | --- |
| Value | Metrics |
| 1.290476995584436 | MAE |
| 2.694159658032994 | MSE |
| 1.6413895509698464 | RMSE |

The mean square error of an estimator measures the squared difference between the estimated values and what is to be estimated whereas, root-mean-square error implies frequently used a measure of differences between values predicted by the model and the values observed. Mean absolute error can be defined as a measure of a difference between two continuous variables. According to the table above, MAE means the average difference between the predicted value between the model and the actual value, which was found to be 0.21 here. MSE means the average squared difference between the predicted value of the model and the actual value, which was obtained here as 0.07. RMSE means the square root of the average difference between the predicted value of the model and the actual value, which was found to be 0.27 here. The closer the values of these indicators are to zero, the better the fit.

Variable importance of all input variables are calculated and shown in fig. 2. Each independent variable had a coefficient which indicates the direction and magnitude of the relationship between the variable and the outcome. Larger coefficient values (both positive and negative) imply a stronger association between the variable and the log-odds of the outcome.

In logistic regression, each independent variable has a coefficient associated with it. These coefficients indicate the direction and magnitude of the relationship between the independent variable and the log-odds of the dependent variable. A positive coefficient suggests that the variable increases the log-odds of the outcome, while a negative coefficient suggests the opposite. The magnitude of the coefficient represents the strength of the association.

*C2) Neural network regression*

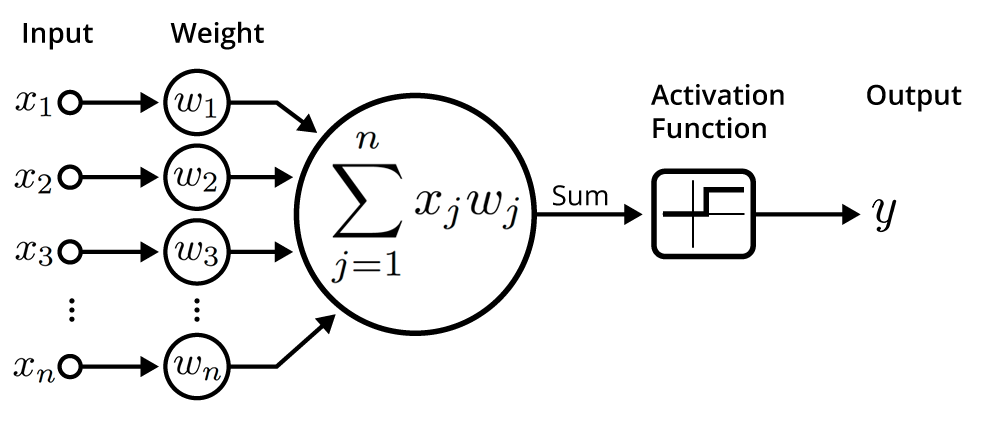


Figure 3: Neural regression network

Neural networks (Kaur etal, 2016; Guo etal, 2015) are a class of machine learning algorithms that are inspired by the structure and function of the human brain. They consist of interconnected nodes, called neurons, that work together to perform complex computations. Neural networks are capable of learning patterns and relationships in data, making them suitable for a wide range of applications such as image and speech recognition, natural language processing, and autonomous control. The ability to learn from data makes neural networks useful in situations where traditional rule-based systems would be impractical or impossible to implement.

Neural networks can be trained using various algorithms, including backpropagation and stochastic gradient descent. The training process involves adjusting the weights and biases of the neurons in the network to minimize the error between the predicted output and the actual output. A reference for further reading on neural networks can be found in the book "Deep Learning" by Ian Good fellow. Neural network structure is optimized in this analysis via hyper parameter tuning to have more precise prediction. Hyper parameter tuning is a critical step in building neural network models that involves selecting the optimal values for the hyper parameters. Hyper parameters are parameters that are set before the training process begins, and they affect the behavior and performance of the neural network. Common hyper parameters that require tuning include the learning rate, number of hidden layers, number of neurons in each layer, activation function, and regularization strength. The optimal values for these hyper parameters depend on the specific problem, dataset, and architecture of the neural network. Hyper parameter tuning can be performed using various methods, including grid search, random search, Bayesian optimization, and genetic algorithms. These methods involve searching the hyper parameter space to find the optimal combination of hyper parameters that maximize the performance of the neural network on a validation set.

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

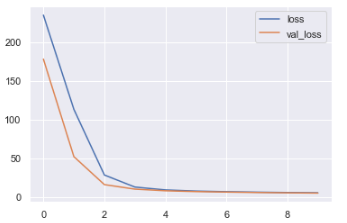


Figure 4: LOSS FUNCTION chart

Table 5: Neural network regression structure

|  |  |  |
| --- | --- | --- |
| Layer (type) | Output Shape | Param # |
| dense\_128 (Dense) | (None, 16) | 176 |
| dense\_129 (Dense) | (None, 16) | 272 |
| dense\_130 (Dense) | (None, 1) | 17 |

Total params: 465

Trainable params: 465

Non-trainable params: 0

In the following table, the results of the application of neural network regression are stated.

Table 6: The results of neural network regression application

|  |  |
| --- | --- |
| Value | Metrics |
| 1.6618198995049809 | MAE |
| 4.384664729322111 | MSE |
| 2.093959104023312 | RMSE |

The mean square error of an estimator measures the squared difference between the estimated values and what is to be estimated whereas, root-mean-square error implies frequently used a measure of differences between values predicted by the model and the values observed. Mean absolute error can be defined as a measure of a difference between two continuous variables. According to the table above, MAE means the average difference between the predicted value between the model and the actual value, which was found to be 0.22 here. MSE means the average squared difference between the predicted value of the model and the actual value, which was obtained here as 0.08. RMSE means the square root of the average difference between the predicted value of the model and the actual value, which was found to be 0.28 here. The closer the values of these indicators are to zero, the better the fit.

***Hyper parameter optimization of neural network regression***

The table below shows the regression structure of the neural network.

Table 7: Neural network regression structure

|  |  |  |
| --- | --- | --- |
| Layer (type) | Output Shape | Param # |
| dense\_131 (Dense) | (None, 128) | 1408 |
| dropout\_9 (Dropout) | (None, 128) | 0 |
| dense\_132 (Dense) | (None, 128) | 16512 |
| dropout\_10 (Dropout) | (None, 128) | 0 |
| dense\_133 (Dense) | (None, 128) | 16512 |
| dropout\_11 (Dropout) | (None, 128) | 0 |
| dense\_134 (Dense) | (None, 1) | 129 |

Total params: 34,561

Trainable params: 34,561

Non-trainable params: 0

Based on this table, the neural network structure used together has two intermediate layers with a 10% drop rate and a binary end layer for prediction.

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 values and the actual values reflects the increase in the accuracy of the neural network model in each training period of the model.

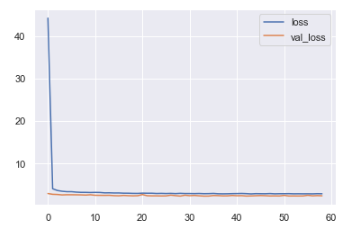


Figure 5: LOSS FUNCTION chart

In the following table, the results of the optimized neural network regression application are stated.

Table 8: Results of optimal neural network regression application

|  |  |
| --- | --- |
| Value | Metrics |
| 1.1981220878759695 | MAE |
| 2.4148747618820536 | MSE |
| 1.5539867315656377 | RMSE |

The mean square error of an estimator measures the squared difference between the estimated values and what is to be estimated whereas, root-mean-square error implies frequently used a measure of differences between values predicted by the model and the values observed. Mean absolute error can be defined as a measure of a difference between two continuous variables. According to the table above, MAE means the average difference between the predicted value between the model and the actual value, which was found to be 1.19. MSE means the average squared difference between the predicted value of the model and the actual value, which is 2.41 here. RMSE means the square root of the average difference between the predicted value of the model and the actual value, which was found to be 1.53 here. The closer the values of these indicators are to zero, the better the fit.

*C-3) Random forest algorithm*

Random forests (Sorour & Mine, 2018) are a popular machine learning algorithm that is used for classification and regression tasks. They are an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the model. In a random forest, multiple decision trees are trained on random subsets of the data and random subsets of the features. This randomness helps to reduce overfitting and improve the generalization ability of the model. The final prediction is made by averaging the predictions of all the trees in the forest. Random forests are widely used in various fields such as finance, healthcare, and image analysis, where they have shown to be effective in handling high-dimensional data and dealing with missing or noisy data. A reference for further reading on random forests can be found in the paper "Random Forests" by Leo Breiman. To improve the preformance of the model, some key hyperparameters are tuned such as: number of decision trees to be included in the random forest, the maximum depth of each decision tree, the minimum number of samples required to be at a leaf node, controling the boostrap sampling in training for individual trees.

In the figure below, the importance of variables in the random forest algorithm is expressed. According to this figure, the most important variables in learning the model with high correlation included the level of education and the faculty.

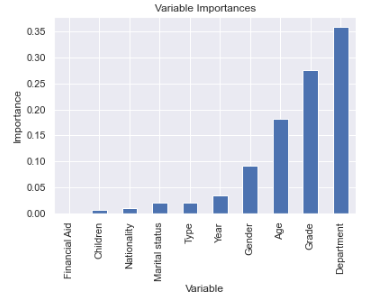


Figure 6: Importance of variables in random forest algorithm

The following table states the results of applying the randomized forest algorithm.

Table 9: The results of applying the random forest algorithm

|  |  |
| --- | --- |
| Value | Metrics |
| 1.1837001855958573 | MAE |
| 2.346364834468825 | MSE |
| 1.5317848525392934 | RMSE |

The mean square error of an estimator measures the squared difference between the estimated values and what is to be estimated whereas, root-mean-square error implies frequently used a measure of differences between values predicted by the model and the values observed. Mean absolute error can be defined as a measure of a difference between two continuous variables. According to the table above, MAE means the average difference between the predicted value between the model and the actual value, which was found to be 1.18. MSE means the mean squared difference between the predicted value of the model and the actual value, which was obtained here as 2.34. RMSE means the square root of the average difference between the predicted value of the model and the actual value, which was found to be 1.53 here. The closer the values of these indicators are to zero, the better the fit.

C-4) Decision tree algorithm

Decision trees (Pandey & Sharma, 2013) are a popular machine learning algorithm that is used for classification and regression tasks. They are a type of supervised learning algorithm that works by recursively partitioning the data into subsets based on the values of the input features, until a stopping criterion is met. In a decision tree, each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a class label or a numerical value. The structure of the tree is learned from the training data using various algorithms, such as the ID3, C4.5, or CART algorithm. Decision trees are widely used in various fields such as finance, healthcare, and image analysis, where they have shown to be effective in handling high-dimensional data and dealing with missing or noisy data.

The graph below shows the difference between the predicted values and the actual values for the grade point variable. The accumulation of data close to zero indicates the high accuracy of the model.

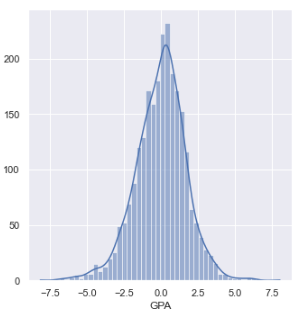


Figure 7: The graph of the difference between the predicted values and the actual values in the decision tree algorithm

The following figure shows the importance of variables in the decision tree algorithm. According to this figure, the most important variables in learning the model with high correlation included the level of education and the faculty.

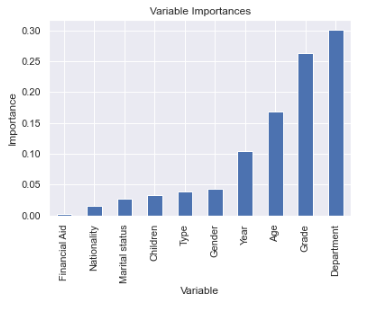


Figure 8: Importance of variables in decision tree algorithm

In the following table, the results of using the decision tree algorithm are stated.

Table 10: The results of applying the decision tree algorithm

|  |  |
| --- | --- |
| Value | Metrics |
| 1.305568574702099 | MAE |
| 2.922975618633861 | MSE |
| 1.7096712019080924 | RMSE |

The mean square error of an estimator measures the squared difference between the estimated values and what is to be estimated whereas, root-mean-square error implies frequently used a measure of differences between values predicted by the model and the values observed. Mean absolute error can be defined as a measure of a difference between two continuous variables. According to the table above, MAE means the average difference between the predicted value between the model and the actual value, which was found to be 1.30. MSE means the average square of the difference between the predicted value of the model and the actual value, which was found to be 2.92 here. RMSE means the square root of the average difference between the predicted value of the model and the actual value, which was obtained here as 1.79. The closer the values of these indicators are to zero, the better the fit.

***D) Comparing the results of multiple linear regression model, neural network regression, random forest and decision tree***

The comparative comparison table states the MAE, MSE, and RMSE values for all implemented algorithms.

Table 11: Comparative comparison of supervised learning algorithms for predicting students' academic performance

|  |  |  |  |
| --- | --- | --- | --- |
| Algorithms | MAE | MSE | RMSE |
| Linear Regression | 1.290477 | 2.694160 | 1.641390 |
| Linear Regression regularized | 1.290571 | 2.694440 | 1.641475 |
| Neural Network | 1.661820 | 4.384665 | 2.093959 |
| Neural Network optimized | 1.198122 | 2.414875 | 1.553987 |
| Random Forest optimized | 1.183700 | 2.346365 | 1.531785 |
| Random Forest | 1.222495 | 2.455876 | 1.567124 |
| Decision Tree | 1.305569 | 2.922976 | 1.709671 |

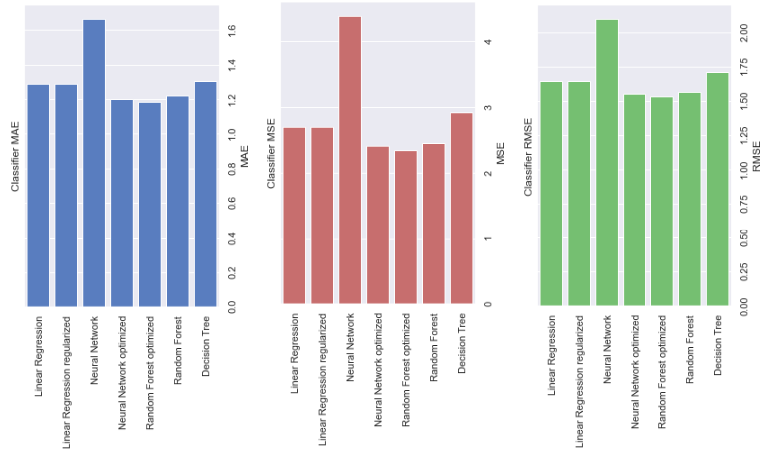


Figure 9: Comparison of implemented algorithms

**Discussion**

This research discussed the prediction of the GPA of Tehran University students based on artificial intelligence algorithms. Based on the implemented models, the optimized random forest model had the best RMSE with a value of 1.53.

در مطالعه Dabhade etal (2022) بهترین مقدار RMSE مربوط به Support Vector (kernel = linear) با مقدار 0.47 بود.

در مطالعه Çakt & Dağdeviren (2022) بهترین مقدار RMSE مربوط به Extreme gradient boosting با مقدار 3.01 بود.

در مطالعه Beckham etal (2023) بهترین مقدار RMSE مربوط به Multi-Layer Perceptron (12 Neurons) با مقدار 0.216 بود.

در مطالعه (2020) Dewi& Widiastuti بهترین مقدار RMSE مربوط به SVR RBF با مقدار 0.18 بود.

**References**

Abu-Naser, S. S., Zaqout, I. S., Abu Ghosh, M., Atallah, R. R., & Alajrami, E. (2015). Predicting student performance using artificial neural network: In the faculty of engineering and information technology.

Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students’ academic performance using classification data mining techniques. Applied mathematical sciences, 9(129), 6415-6426.

Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting students final GPA using decision trees: a case study. International journal of information and education technology, 6(7), 528.

Ali, N., Jusof, K., Ali, S., Mokhtar, N., & Salamat, A.S.A. (2009). THE FACTORS INFLUENCING STUDENTS’ PERFORMANCE AT UNIVERSITI TEKNOLOGI MARA KEDAH, MALAYSIA. Management Science and Engineering, 3(4), P 81-90

ALLAH, A. Q. G. F. (2019). Using machine learning to support students’ academic decisions (Doctoral dissertation, The British University in Dubai (BUiD)).

Alsalman, Y. S., Halemah, N. K. A., AlNagi, E. S., & Salameh, W. (2019, June). Using decision tree and artificial neural network to predict students academic performance. In 2019 10th international conference on information and communication systems (ICICS) (pp. 104-109). IEEE.

Altabrawee, H., Ali, O. A. J., & Ajmi, S. Q. (2019). Predicting students’ performance using machine learning techniques. JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences, 27(1), 194-205.

Aluko, R. O., Adenuga, O. A., Kukoyi, P. O., Soyingbe, A. A., & Oyedeji, J. O. (2016). Predicting the academic success of architecture students by pre-enrolment requirement: Using machine-learning techniques. Construction Economics and Building, 16(4), 86-98.

Beckham, N. R., Akeh, L. J., Mitaart, G. N. P., & Moniaga, J. V. (2023). Determining factors that affect student performance using various machine learning methods. Procedia Computer Science, 216, 597-603.

Budiman, E., Haviluddin, Kridalaksana, A. H., Wati, M., & Purnawansyah. (2018). Performance of decision tree C4. 5 algorithm in student academic evaluation. In Computational Science and Technology: 4th ICCST 2017, Kuala Lumpur, Malaysia, 29–30 November, 2017 (pp. 380-389). Springer Singapore.

Bum, S., Iorliam, I. B., Okube, E. O., & Iorliam, A. (2019). Prediction of Student’s Academic Performance Using Linear Regression. NIGERIAN ANNALS OF PURE AND APPLIED SCIENCES, 2, 259-264.

Çakıt, E., & Dağdeviren, M. (2022). Predicting the percentage of student placement: A comparative study of machine learning algorithms. Education and Information Technologies, 27(1), 997-1022.

Canagareddy, D., Subarayadu, K., & Hurbungs, V. (2019). A machine learning model to predict the performance of university students. In Smart and Sustainable Engineering for Next Generation Applications: Proceeding of the Second International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM 2018), November 28–30, 2018, Mauritius 2 (pp. 313-322). Springer International Publishing.

Cheewaprakobkit, P. (2015). Predicting student academic achievement by using the decision tree and neural network techniques. Human Behavior, Development And Society, 12(2), 34-43.

Chuan, Y. Y., Husain, W., & Shahiri, A. M. (2017). An exploratory study on students’ performance classification using hybrid of decision tree and naïve Bayes approaches. In Advances in Information and Communication Technology: Proceedings of the International Conference, ICTA 2016 (pp. 142-152). Springer International Publishing.

Dabhade, P., Agarwal, R., Alameen, K. P., Fathima, A. T., Sridharan, R., & Gopakumar, G. (2021). Educational data mining for predicting students’ academic performance using machine learning algorithms. Materials Today: Proceedings, 47, 5260-5267.

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Computer Science, 57, 500-508.

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015, July). Predicting students performance in educational data mining. In 2015 international symposium on educational technology (ISET) (pp. 125-128). IEEE.

Sorour, S. E., & Mine, T. (2016, July). Building an interpretable model of predicting student performance using comment data mining. In 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI) (pp. 285-291). IEEE.

Pandey, M., & Sharma, V. K. (2013). A decision tree algorithm pertaining to the student performance analysis and prediction. International Journal of Computer Applications, 61(13), 1-5.

Dewi, K. E., & Widiastuti, N. I. (2020, July). Support vector regression for GPA prediction. In IOP Conference Series: Materials Science and Engineering (Vol. 879, No. 1, p. 012112). IOP Publishing.

Falát, L., & Piscová, T. (2022). Predicting GPA of University Students with Supervised Regression Machine Learning Models. Applied Sciences, 12(17), 8403.

Gipson, J. A. (2018). Predicting Graduation and College GPA: A Multilevel Analysis Investigating the Contextual Effect of College Major (Doctoral dissertation, Purdue University).

Ha, D. T., Loan, P. T. T., Giap, C. N., & Huong, N. T. L. (2020). An empirical study for student academic performance prediction using machine learning techniques. International Journal of Computer Science and Information Security (IJCSIS), 18(3), 21-28.

Hamoud, A., Hashim, A. S., & Awadh, W. A. (2018). Predicting student performance in higher education institutions using decision tree analysis. International Journal of Interactive Multimedia and Artificial Intelligence, 5, 26-31.

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students’ academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447-459.

Iqbal, Z., Qadir, J., Mian, A. N., & Kamiran, F. (2017). Machine learning based student grade prediction: A case study. arXiv preprint arXiv:1708.08744.

Kaunang, F. J., & Rotikan, R. (2018, October). Students' academic performance prediction using data mining. In 2018 Third International Conference on Informatics and Computing (ICIC) (pp. 1-5). IEEE.

Kesumawati, A., & Utari, D. T. (2018, October). Predicting patterns of student graduation rates using Naïve bayes classifier and support vector machine. In AIP conference proceedings (Vol. 2021, No. 1, p. 060005). AIP Publishing LLC.

Khan, F., Weiss, G. M., & Leeds, D. D. (2021). Predicting the Academic Performance of Undergraduate Computer Science Students Using Data Mining. In Advances in Software Engineering, Education, and e-Learning: Proceedings from FECS'20, FCS'20, SERP'20, and EEE'20 (pp. 303-317). Springer International Publishing.

Kumar, M., & Salal, Y. K. (2019). Systematic review of predicting student's performance in academics. Int. J. of Engineering and Advanced Technology, 8(3), 54-61.

Maulana, M. F., & Defriani, M. (2020). Logistic model tree and decision tree J48 algorithms for predicting the length of study period. PIKSEL: Penelitian Ilmu Komputer Sistem Embedded and Logic, 8(1), 39-48.

Mohammadi, M., Dawodi, M., Tomohisa, W., & Ahmadi, N. (2019, February). Comparative study of supervised learning algorithms for student performance prediction. In 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) (pp. 124-127). IEEE.

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. International Journal of Modern Education and Computer Science, 8(11), 36.

Ngepah, N., Saba, C. S., & Mabindisa, N. G. (2021). Human capital and economic growth in South Africa: A cross-municipality panel data analysis. South African Journal of Economic and Management Sciences, 24(1), 11.

Oreshin, S., Filchenkov, A., Petrusha, P., Krasheninnikov, E., Panfilov, A., Glukhov, I., ... & Kozlova, D. (2020, October). Implementing a Machine Learning Approach to Predicting Students’ Academic Outcomes. In Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System (pp. 78-83).

Phale, K., Fanglin, L., Adjei Mensah, I., Omari-Sasu, A. Y., & Musah, M. (2021). Knowledge-Based Economy Capacity Building for Developing Countries: A Panel Analysis in Southern African Development Community. Sustainability, 13(5), 2890.

Pojon, M. (2017). Using machine learning to predict student performance (Master's thesis).

Putpuek, N., Rojanaprasert, N., Atchariyachanvanich, K., & Thamrongthanyawong, T. (2018, June). Comparative study of prediction models for final GPA score: a case study of Rajabhat Rajanagarindra University. In 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS) (pp. 92-97). IEEE.

Rehman, W., Degirmen, S., & Waseem, F. (2021). Propensity for and Quality of Intellectual Capital Divulgence Across the BRICS Banking Sector: A Knowledge-Based Perspective from Emerging Economies. Journal of the Knowledge Economy, 1-28.

Shamsi, M. S., & Lakshmi, J. (2016). A Comparative Analysis of classification data mining techniques: Deriving key factors useful for predicting students performance. arXiv preprint arXiv:1606.05735.

Singh, S. P., Malik, S., & Singh, P. (2016). Research paper factors affecting academic performance of students. Indian Journal of Research, 5(4), 176-178.

Sugiharti, E., Firmansyah, S., & Devi, F. R. (2017). Predictive evaluation of performance of computer science students of unnes using data mining based on naÏve bayes classifier (NBC) algorithm. Journal of Theoretical and Applied Information Technology, 95(4), 902.

Vijayalakshmi, V., & Venkatachalapathy, K. (2019). Comparison of predicting student’s performance using machine learning algorithms. International Journal of Intelligent Systems and Applications, 11(12), 34.

1. In this research, this variable is the number of children [↑](#footnote-ref-1)