*Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi*
*Sayı 7, S. 1-4, 2, 2023*
*© Telif hakkı IJANSER'e aittir*
***Araştırma Makalesi***

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

*International Journal of Advanced Natural Sciences and Engineering Researches*
*Volume 7, pp. 1-4, 2, 2023*
*Copyright © 2023 IJANSER*
***Research Article***

# Predicting the University Placement Status of University Students Using Artificial Intelligence

Seyhun ÇELEBİOĞLU*[1] and Selim SÜRÜCÜ [2]

[1] *Department of Computer Engineering, Çankırı Karatekin University, Çankırı, Turkey. ORCID iD: 0000-0002-0668-6328*
[2] *Department of Computer Engineering, Çankırı Karatekin University, Çankırı, Turkey. ORCID iD: 0000-0002-8754-3846*

*(seyhun721@gmail.com) *Email of the corresponding author*

*(Received: 1 March 2023, Accepted: 10 March 2023)*

(2nd International Conference on Scientific and Academic Research ICSAR 2023, March 14-16, 2023)

**ATIF/REFERENCE:** Çelebioğlu, S. & Sürücü, S. (2023). Predicting the University Placement Status of University Students Using Artificial Intelligence. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(2), 1-4.

***Abstract*** – A university, also known as a higher education institution, is an institution where the highest level of education, research and knowledge is produced. Universities, which are divided into various disciplines, generally consist of units that provide higher education, undergraduate and graduate education. Students who want to attend university after high school are placed in universities according to many criteria such as high school average score, aptitude exams, general exams, language exams, etc. Since there are paid/unpaid institutions in the American university system, it is seen that the student's family status (family income, housing status, etc.) is also an important factor in studying at university. In this study, it is aimed to predict whether the student will be able to go to university or not by using 4 machine learning models (Decision Tree, Random Forest, K-Nearest Neighbours, Logistic Regression) and an artificial neural network (Multi Layer Perceptron - MLP) methods using the "Go to College" dataset, which is a synthetic and open source 1000 student data. In the training phase, 5-fold cross validation was used to obtain more accurate results. For a two-state classification problem, 92% accuracy was obtained after training the artificial neural network for 2000 iterations. This value appears to be the best result.

*Keywords – Machine Learning, Artificial Intelligence, Student Prediction, MLP, Go To College*

## I. INTRODUCTION

The American university system is an education model that attracts attention with its different structure from other university systems in the world and is frequently preferred by international students. In addition, the American university system stands out with its wide range of programmes, course options that provide students with a wide range of freedom and flexibility, the opportunity to transfer between different disciplines, research-oriented education approach, student-centred teaching approach and various support programmes for international students.

To be eligible to study at American universities, you must first have a high school diploma. Then, a specific exam or exams (such as SAT or ACT) and/or a certain grade point average are required to meet the admission requirements of the universities. In addition, the application process should be completed by preparing all of the documents (transcripts, reference letters, personal statements, etc.) requested by the universities during the application process. The application process usually takes a few months and at the end of the application process, the university makes an evaluation to determine whether the student is accepted or not.

The American university system offers two options: free and paid. Some states may have free tuition programmes or scholarships for free education. For example, public universities in New York state offer the Excelsior Scholarship programme, which offers free tuition to students living in the state. There are also scholarship and tuition assistance programmes provided by the federal and state governments, depending on the financial situation of the students. However, free education opportunities are limited, and for most students, a fee-based education model applies. Therefore, students are encouraged to seek scholarships, student loans and other financial aid options to cover the cost of their education.

The American university system generally has a fee-based education model, with the exception of public universities that offer free tuition. However, because public universities are funded from the state's tax revenues, they often offer a more favourable fee to students living in the state. It can also vary between universities and according to the programme or courses the student takes. Private universities and top-tier public universities generally have higher fees, while public universities have lower fees. In the 2021-2022 academic year, annual tuition fees at private universities usually range from $30,000 to $50,000, while at public universities these fees are usually between $10,000 and $30,000.

## II. LITERATURE REVIEW

Arqawi et al. trained 1100 student data with 20 machine learning models and 1 deep learning model. They compared the performance of these models according to metrics such as accuracy and f1 score. While the deep learning model gave an accuracy of 93%, the classical machine learning models gave an accuracy of approximately 91% [1]. In another study using many different data sets, 1000 data were trained with decision tree, decision support machine and artificial neural network. In the test process, 73% accuracy value was obtained as a result of 2000 iterations. By increasing the number of iterations up to 5000, 93% accuracy value was obtained [2].

Mia et al. used enrolment data from a private university in Bangladesh to estimate the number of new enrolments. In their study where they used machine learning such as SVM, Naive Bayes, they obtained the best accuracy result with SVM with 85.76%. In another study using student data, it was attempted to predict the completion of courses according to certain characteristics of students. In this study with Naive Bayes classifier, the accuracy value was calculated as 84% [4].

## III. MATERIALS AND METHOD

In this study, a synthetic student dataset called "Go to College" dataset is created to analyse whether students can enter a college or university in the United States. There are 1000 records of data in total. Among the features in the dataset, there are different factors such as the quality of colleges and universities (A or B), the student's grade point average (100 system), the school status of the parents, the income status of the parents, the status of the house they live in.  The "Go to College" dataset is an open source dataset and is available to users free of charge [5]. Many data science platforms or data source websites can be used to access the dataset.

In the training phase of the dataset, 5 machine learning methods were applied using 5-fold validation. 5-fold cross-validation is a commonly used validation technique for measuring the performance of a machine learning model (See Fig.1) . This method works by dividing the data into five equal parts [6]. One of these parts is reserved as test data, while the remaining four parts are used as training data. Then, the model is trained in four different combinations and tested each time using the fifth split. This process is repeated five times, each time using a different part as test data. As a

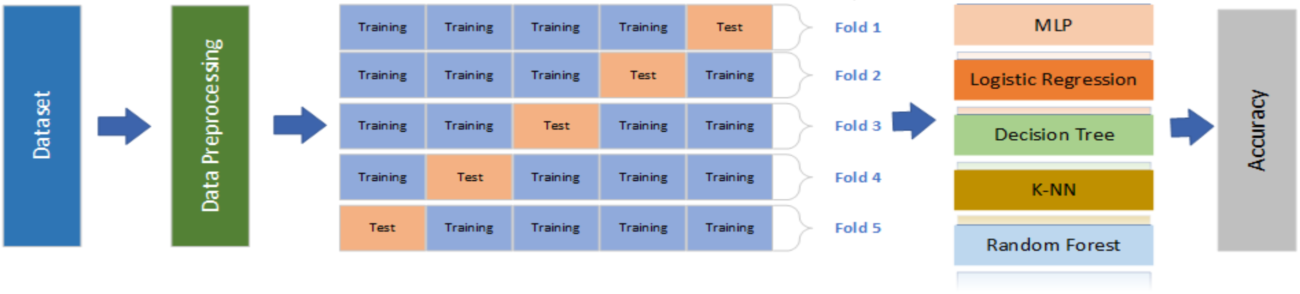

Figure 1 Flowchart of the study

result, five different test results are obtained and the performance of the model is calculated by averaging these results. 5-fold cross-validation is also an effective method to reduce the overfitting problem of the machine learning model. Thanks to this method, a better generalization is provided for the model to obtain more successful results in real life.

When the dataset is analysed, it is seen that there is a two-class problem according to whether a student is placed in a university or not. In this study, 5 machine learnings were used. The first of these methods is Logistic regression, which is a classification model. The logistic regression model is based on the assumption that data can be described by a linear equation. However, the output variable belongs to one of two categorical classes. Therefore, this model attempts to determine the probability differences between the classes using a probabilistic approximation of the output variable. Logistic regression is widely used in solving classification problems, especially in medical research, marketing research and social sciences [7].

Decision Tree was used as the second classifier. Decision tree is a machine learning model in which the features in the data set are used as independent variables and used to classify the target variable [8]. Decision tree creates a tree structure using a specific set of rules for classifying the features in the data set. It works particularly effectively in classification problems and is often used to obtain simple and straightforward results.

Another classifier used is k-NN (k-Nearest Neighbours). This algorithm is a classification technique that assigns a data sample to a class or value defined by its nearest neighbours. k-NN algorithm is classified by the k nearest neighbours of a data point for classification [9]. These neighbours are calculated using a distance measure such as Euclidean or Manhattan distance. For regression analysis, the k-NN algorithm makes predictions by calculating the average of the k nearest neighbours of a data point. The k-NN algorithm is particularly useful for classification problems and is often used in recognition systems, medical diagnostics, prediction models and robotics.

The random forest model, which is used for a better classification by combining more than one decision tree, is used as the fourth model in our study. In this model, the importance of the features is determined and decision trees are created accordingly [10]. By combining these decision trees, more accurate classification results are obtained. In addition, the Random Forest algorithm trains each decision tree on a different subset of data, thus minimizing the overfitting problem.

MLP (Multi-Layer Perceptron) is an artificial neural network model used for classification [11]. MLP consists of at least 3 layers: an input layer, at least one intermediate layer and an output layer. MLP consists of layers where many neurons are connected to each other and these neurons process the input data and produce results. MLP performs many iterations in the learning process and updates the weights using back propagation algorithm. This algorithm is used to minimise the error in the learning process. MLP learns based on training data and then classification is used on new data. In this study, MLP is used as a very simple model consisting of two hidden layers.

In this study, the accuracy metric is used as a measure of success. this metric expresses the accuracy rate of a classification model and is calculated by dividing it by the total number of correctly classified instances [12]. This metric is often used to measure the performance of a classification model. While calculating the overall accuracy value, since 5-fold validation was used, the average of the accuracy values in 5 folds was found.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

## IV. Results

When Table 1 is analysed, it is seen that the best result is obtained with MLP. According to the dataset used in the study, it is seen that artificial neural networks give better results than classical machine learning models. After 2000 iterations of training in MLP, the highest result of 92% was obtained.

Looking at the results of machine learning models, it was seen that accuracy values between 84% and 87% were obtained. It is noteworthy that there is not a big difference between the results of the models.

Table 1. Mean Accuracy Values

| Model | Accuracy |
|---|---|
| Logistic Regression | 84% |
| Decision Tree | 85% |
| K-NN | 85% |
| **MLP** | **92%** |
| Random Forest | 87% |

.

## V. DISCUSSION

Table 2. Example of a table

| Previous Work | Machine Learning Model | Accuracy | dataset status |
|---|---|---|---|
| [1] | NuSVC | 91% | Same |
| [2] | ANN | 71%-93% | Different |
| [3] | SVM | 85.76% | Different |
| [4] | Naïve Bayes | 84% | Different |

In the studies conducted on this subject in the literature, it is seen that student characteristics are used for different purposes. When Table 2 is analysed, it is seen that there is a study that exceeds our results. According to another study using the same dataset, our results are much better.

## REFERENCES

[1] Arqawi, S. M., Eman, A. Z., Anees, H. R., Abunasser, B. S., and Abu-Naser, S. “Predicting university student retention using artificial intelligence,” *International Journal of Advanced Computer Science and Applications*, 13(9), 2022. - doi:https://doi.org/10.14569/IJACSA.2022.0130937

[2] Kumbhar, C. and Sridhar, S.S.“Trend analysis of university placement by using machine learning algorithms.” in International Journal of Engineering & Technology. 2018. Doi: 7. 178. 10.14419/ijet.v7i2.4.13034.

[3] Mia, MD. J., Biswas, A. A., Sattar, A. and Habib MD. T. “Registration Status Prediction of Students Using Machine Learning in the Context of Private University of Bangladesh.” In *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 9(1), Nov. 2019.

[4] Perez, J. G., and Eugene S. P. "Predicting Student Program Completion Using Naïve Bayes Classification Algorithm." *International Journal of Modern Education & Computer Science*. 13.3, 2021.

[5] (2023) Go To College Dataset. [Online]. Available: https://www.kaggle.com/datasets/saddamazyazy/go-to-college-dataset.

[6] Jung, Yoonsuh, and Jianhua Hu. “A K-fold Averaging Cross-validation Procedure.” *Journal of nonparametric statistics*. vol. 27,2, pp.167-179, 2015 doi:10.1080/10485252.2015.1010532

[7] LaValley, M. P. "Logistic regression." *Circulation*. Vol. 117.18, 2395-2399, 2008.

[8] Kotsiantis, S. B. "Decision trees: a recent overview." *Artificial Intelligence Review*. vol. 39, pp.261-283, 2013.

[9] Guo, G., "KNN model-based approach in classification." On The Move to Meaningful Internet Systems 2003: In *CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003*, Catania, Sicily, Italy, Proceedings. Springer Berlin Heidelberg. Nov. 2003.

[10] Probst, P., Marvin N. W. and Boulesteix A. "Hyperparameters and tuning strategies for random forest." *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*. vol. 9.3. 2019.

[11] Pinkus, A. "Approximation theory of the MLP model in neural networks." *Acta numerica*. Vol. 8, pp.143-195, 1999.

[12] SÜRÜCÜ, S. and ECEMİŞ, İ.N. "Garbage Classification Using Pre-Trained Models." *Avrupa Bilim ve Teknoloji Dergisi* 36: 73-77.