DOI
https://doi.org/
10.18690/um.fov.2.2025.18

ISBN
978-961-286-963-2

Prediction of Student Performance in a Series of Related Courses of the Undergraduate Program of the Department of Computer, Informatics and Telecommunications Engineering – International Hellenic University

Panagis Eleftherios, Athanasios Angeioplastis, Alkiviadis Tsimpiris, Dimitrios Varsamis

International Hellenic University, Department of Computer, Informatics and Telecommunications Engineering, Serres, Greece
lefter.panag@gmail.com, aagiop@gmail.com, atsimpiris@ihu.gr, dvarsam@ihu.gr

It is a fact that educational institutions at all levels are now focusing their attention on analyzing the behavior and performance of their students. The main objective of this study is to examine whether it is possible to predict the grades of students in the Department of Computer, Informatics and Telecommunications Engineering (ICT) at the International Hellenic University (IHU), in a course, based on their performance in previous related courses within real error margins, as well as the optimization of the prediction error. Different models were used for prediction in order to evaluate the performance and impact of each model separately. The goal of this work is to provide significant results regarding the long-term performance of the students. The models used were capable of predicting the performance within the defined absolut error margin.

Keywords:
prediction,
student performance,
regression,
random forest,
decision trees,
neural networks

University of Maribor Press

1 Introduction

Undoubtedly, prediction models are powerful tools for understanding, examining, and interpreting complex phenomena and data for every aspect, as well as academic institutions. For all levels of education, it is a growing tension, to analyze and attempt to predict the students' performance [Devasia, Vinushree, & Hegde, 2016] [Verma, Srivastava, & Singh, 2021] and even inform students for the possible outcome [Dias, Hadjileontiadou, Diniz, J., & Hadjileontiadis, 2020].

In the context of this study, the aim is to explore trends, student profiles, academic outcomes, and to examine whether key prediction models can estimate student grades through a series of related academic courses following the previous statistical research for the bachelors' program [Angeioplastis, Panagis, Tsakiridis, Tsimpiris, Varsamis, 2024]. The main features and data related to the undergraduate program of the Department of Computer, Informatics and Telecommunications Engineering at the International Hellenic University were analyzed.

In this research, the grades of 6 related courses are used to predict the grade in the 7th course in terms of time. This study has been contacted by the open software *«Orange Data Mining»* [Demsar, Curk, Erjavec, Gorup, Hocevar, Milutinovic, Mozina, Polajnar, Toplak, Staric, Stajdohar, Umek, Zagar, Zbontar, Zitnik, Zupan, 2013] to apply the prediction models of Regression, Random Forest, Decision Trees and Neural Networks. Success is defined as the mean absolute error being smaller than one unit for the predictions of students' grades.

2 Dataset

The dataset used concerns the set of grades for the 6 selected and related courses: Programming I, Programming II, Object-Oriented Programming, Software Development Environments, Programming Methodology, and Software Technology, which are offered in the undergraduate program of the Department of Computer Engineering, Computers, and Telecommunications at the University of Applied Sciences of Serres. This is under the condition that students have passed the course exam on their first attempt. The data was processed anonymously, respecting the privacy of all students.

3 Orange Data Mining

The software chosen for the implementation of these models is the open-source software Orange Data Mining [Scarselli, Gori, Hagenbuchner, & Monfardini, 2008] [Kaur, Stoltzfus, & Yellapu]. Orange Data Mining is an open-source platform for visual data analysis, machine learning, and data mining. It is designed for data analysis and model creation with a user-friendly, visually oriented environment. One of the main features of Orange Data Mining is that it enables visual programming analysis, as it allows users to create data analysis workflows using a graphical interface. Widgets are used to perform various tasks, such as loading data, cleaning data, analysis, and visualization. Additionally, it is a suitable tool for data mining, as it includes a variety of machine learning and data mining algorithms for classification, regression, clustering, and association rule analysis.

4 Methodology

The methodology followed as:

1. **Data Extraction**: Data related to students' successful examination attempts on their first try for all courses in the faculty were extracted. The data collection was performed using SQL queries within the database environment, and the final file was converted into a CSV format for easier processing in the Orange Data Mining environment.
2. **Selection of Relevant Course Sequence**: The courses of interest were selected. These include the programming courses: Programming I, Programming II, Object-Oriented Programming, Software Development Environments, Programming Methodology, and Software Technology. The courses were chosen in increasing order of the semester of study.
3. **Importing the CSV File**: The CSV file containing the data was imported into the Open-Source software Orange Data Mining for processing.
4. **Application of Prediction Models**: Prediction models were applied within the Orange Data Mining environment. The prediction models chosen and applied were Multiple Linear Regression, Decision Tree, Random Forest, and Neural Network models.

5. **Evaluation of Results**: The results of the prediction models were evaluated.

5 Results

5.1 Multiple Linear Regression

The first attempt at prediction was made using Multiple Linear Regression [Montgomery, Peck & Vining, 2021] without normalization, employing the Cross Validation method with various Number of Folds to obtain better results.

Table 1: Results of Multiple Linear Regression for Each Number of Folds

Multiple Linear Regression – Cross validation				
Number of folds	MSE	RMSE	**MAE**	MAPE
2	1.481	1.217	**0.875**	0.163
3	1.424	1.203	**0.855**	0.162
5	1.543	1.242	**0.881**	0.166
10	1.582	1.258	**0.910**	0.171
20	1.601	1.275	**0.933**	0.175

It was observed that the model can generate predictions within the target of one unit that was set. The best prediction was with 3 Number of Folds, resulting in a Mean Absolute Error (MAE) of 0.855 and a Mean Squared Error (MSE) of 1.424, while the worst prediction was with 20 Number of Folds, with an MAE of 0.933 and an MSE of 1.601. It is worth noting that the error results remain consistent regardless of how many times the experiment is repeated.

The reason the best predictions come with smaller folds is that this is a relatively small dataset, and using more folds can lead to smaller training and test subsets. With 3 folds, each fold contains a larger percentage of data, which can improve model performance as the model is trained on more representative samples of the data. The dataset is split into three equal parts, and the model is trained on two of them and tested on the third. This process is repeated three times, with each fold being used once for testing.

Additionally, using fewer folds reduces the complexity of the training and testing process, potentially reducing the relative errors. In any case, the target of one unit was achieved for the Multiple Linear Regression model.

5.2 Neural Networks

A particularly interesting aspect that we wanted to examine is whether the individual Neural Networks [Agatonovic-Kustrin, 2000] [Rasamoelina, Adjailia & Sinčák, 2020] that we can develop through Orange Data Mining provide results within the target range. After experimenting with various combinations, we settled on the SGD – ReLU neural network model, with 20 neurons, 300 iterations, and 10 folds.

Table 2: Neural Network with SGD Solver and ReLU Activation Function

Neural Network SGD – ReLU 20 neurons, 300 iteration				
Number of folds	MSE	RMSE	**MAE**	MAPE
10	1.440	1.200	**0.954**	0.176

The forecasting experiments were conducted multiple times to ensure that the results were not due to random chance. The optimal Mean Absolute Error (MAE) value achieved was 0.954.

5.3 Decision Tree Algorithm

The next model used to predict the students' grades in the course Software Technology was the Decision Tree algorithm (Suthaharan, 2016) (Priyam, 2013). The choice to create a binary tree result in a simpler and clearer model, where each node has only two branches. This facilitates the interpretation and analysis of the decisions made by the model, as it helps avoid excessive overfitting compared to trees with multiple branches. By setting the condition that no subset should be smaller than 20 observations (Do Not Split Subset Smaller Than: 20), it ensures that the tree does not split into subsets that are too small to provide reliable predictions.

The maximum depth limit (Limit Max Tree Depth) ensures that the tree does not become overly complex for the given range of values.

Table 3: Decision Tree with Minimum Number of Observations in the Leaves 2 and 10 Folds

Decision Binary Tree				
Number of folds	MSE	RMSE	**MAE**	MAPE
10	1.383	1.176	**0.806**	0.159

As shown, the Mean Absolute Error (MAE) for the model is 0.806, making this decision tree the most reliable model, especially due to the use of 10 folds for cross-validation.

5.4 Random Forest

The last prediction algorithm evaluated is the Random Forest algorithm. For the evaluation of the results of a Random Forest model [Breiman, 2001], it is crucial to ensure that the results are stable and reliable. To achieve this, we applied 10-fold cross-validation, which is common in most machine learning applications. We ran each model with the respective settings 30 times, aiming to ensure that no extreme values outside the target range of one unit occurred.

Table 4: Largest Value of Mean Absolute Error (MAE) for 200 Trees

Random Forest				
Number of folds	MSE	RMSE	MAE	MAPE
10	1.477	1.215	0.905	0.175

After several trials, we settled on the choice of 200 trees and an average execution time of the algorithm of 5 seconds across the 30 runs. As shown, there was a significant reduction in the difference between the two extremes, with the Mean Absolute Error (MAE) ranging from 0.870 to 0.905. This indicates that the model is producing reliable and consistent results, with a minimal variance between the runs. The lower bound of the range is 0.870, while the upper bound is 0.905.

6 Model's Comparison

We investigated whether there are statistically significant differences between the predictive models using ANOVA Bonferroni's test [Judd, McClelland, & Ryan, 2017]. We chose to analyze the results of the algorithms by applying 10-fold cross-validation, which ensures that the model is evaluated on different subsets of the data, providing a reliable estimate of model performance while reducing the impact of randomness. For this reason, the models were selected for which there is evidence that they perform better with this number of folds.

Table 5:Combination of Models with 10-Folds and Statistical Results

ANOVA test between the models				
	Random Forest	Multiple Linear Regression	Neural Network	Decision Tree
Random Forest		0.356	0.402	0.575
Multiple Linear Regression	0.644		0.576	0.636
Neural Network	0.598	0.424		0.623
Decision Tree	0.425	0.364	0.377	

The results show that all models achieve the target of the Mean Absolute Error (MAE) being less than one unit. From the statistical analysis of the algorithms and the comparisons between them, p-values greater than 0.05 were obtained, which was set as the significance level of the test. This indicates that there are no statistically significant differences between the models in terms of predicting the grades for the course Software Technology.

7 Conclusion

For the sequence of 6 courses followed, the models of **Multiple Linear Regression**, **Decision Tree**, **Random Forest**, and **Neural Networks** were able to provide predictions within the target of one unit for the **Mean Absolute Error (MAE)**. The **Decision Tree algorithm** was the most capable, presenting the smallest error values, while the **Neural Network algorithm** showed the largest error values. After the statistical analysis, it was concluded that there are no statistically significant differences between the models in terms of predicting the grades for the course **Software Technology**. This scientific field remains open for future research, with the application of predictive models to other sequences of

courses. Additionally, the optimization of the model errors for more reliable predictions is left open for further exploration.

Acknowledgment

This research work was supported by the "Applied Informatics" Post Graduate Program of Computer, Informatics and Telecommunications Engineering Department, International Hellenic University, Greece

References

Agatonovic-Kustrin, S., & Beresford, R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis. 2000, Vol. 22, 5, pp. 717-727.

Angeioplastis, Athanasios & Panagis, Eleftherios & Tsakiridis, Sotirios & Tsimpiris, Alkiviadis & Varsamis, Dimitrios. (2024). Statistical Analysis of Demographic Data and Student Performance in the Courses of the bachelor's degree Program at the Department of Computer, Informatics and Telecommunications Engineering - International Hellenic University with ORACLE APEX Statistics. 13-24. 10.18690/um.fov.3.2024.2.

Breiman, L. Random forests. Springer. Machine Learning, 2001, pp. 5-32.

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B. Orange: data mining toolbox in Python. Journal of Machine Learning Research. 8 14, 2013, pp. 2349-2353.

Devasia, T., Vinushree, T. P., & Hegde, V. Prediction of students' performance using Educational Data Mining. In 2016 international conference on data mining and advanced computing (SAPIENCE). IEEE. 2016, pp. 91-95.

Dias, S. B., Hadjileontiadou, S. J., Diniz, J., & Hadjileontiadis, L. J. DeepLMS: a deep learning predictive model for supporting online learning in the Covid-19 era. Scientific reports - Nature.com. 10(1), 2020.

Judd, C. M., McClelland, G. H., & Ryan, C. S. Data analysis: A model comparison approach to regression, ANOVA, and beyond. s.l. : Routledge, 2017. 9781315744131.

Kaur, P., Stoltzfus, J., & Yellapu, V. Descriptive statistics. International Journal of Academic Medicine. Vol. 4, 1, pp. 60-63.

Montgomery, D. C., Peck, E. A., & Vining, G. G. Introduction to linear regression analysis. s.l. : John Wiley & Sons, 2021. 9781119578727.

Rasamoelina, A. D., Adjailia, F., & Sinčák, P. A review of activation function for artificial neural network. s.l. : IEEE, 2020. 18th World Symposium on Applied Machine Intelligence and Informatics (SAMI). pp. 281-286.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. The graph neural network model. IEEE transactions on neural networks. 2008, pp. 61-80.

Verma, B. K., Srivastava, D. N., & Singh, H. K. Prediction of Students' Performance in e-Learning Environment using Data Mining/Machine Learning Techniques. J. Univ. Shanghai Sci. Technol. 23(05), 2021, pp. 593-596.