JOURNAL OF E-LEARNING AND KNOWLEDGE SOCIETY
Vol. 20, No. 2 (2024), pp. 19-27

Predicting student specializations: a Machine Learning Approach based on Academic Performance

Athanasios Angeioplastis, Nikolaos Papaioannou, Alkiviadis Tsimpiris[1], Angeliki Kamilali, Dimitrios Varsamis

International Hellenic University, Dept. of Computer, Informatics and Telecommunications Engineering – Serres (Greece)

(submitted: 22/3/2024; accepted: 19/8/2024; published: 28/8/2024)

Abstract

Education is a cornerstone of societal progress, equipping people with essential skills and knowledge. In today's dynamic global society, personalized learning experiences are crucial. Data-driven methodologies, especially Educational Data Mining (EDM), play pivotal roles. This study employs machine learning algorithms to predict specializations for Greek high school students based on their previous grades. The aim is to provide a practical tool for educators and parents, aiding in the optimal selection of specializations. The paper outlines the methodology, presents comparative study results, and concludes with insights into the potential impact on educational decision-making. This research advances the integration of data-driven approaches in education, enhancing students' learning experiences and prospects.

KEYWORDS: Machine Learning Algorithms, Educational Data Mining, Prediction, High School.

DOI
https://doi.org/10.20368/1971-8829/1135904

CITE AS
Angeioplastis, A., Papaioannou, N., Tsimpiris, A., Kamilali, A., & Varsamis, D. (2024). Predicting student specializations: a Machine Learning Approach based on Academic Performance. *Journal of e-Learning and Knowledge Society*, *20*(2), 19-27. https://doi.org/10.20368/1971-8829/1135904

1. Introduction

One of the primary foundations of society development is education, which gives people the knowledge and abilities they need to function in a constantly changing environment. It is universally acknowledged as the essential component of economic progress (Chang, Chen, & Xiong, 2018; Alani, Yawe & Mutenyo, 2022), societal improvement and personal development (Zheng, 2023), making it a fundamental human right. It is more than just a process of acquiring information, but mostly a transformative journey that enables people to think critically, to confront complex problems and generally, to make significant contributions to their communities (Kurnia, 2021). In today's interconnected global society, the role of education has become even more pivotal, as it equips people with the tools they need to navigate a complicated and rapidly evolving environment (Schleicher, 2018). Owing to these conditions, teachers are required to modify the curriculum and address the particular requirements and learning preferences of a broad spectrum of students (Kilag, Comighud, Amontos, Damos & Abendan, 2023). As a result, dynamic, customized learning experiences replace conventional, one-size-fits-all methods. This shift is supported by the integration of data-driven methodologies, which provide valuable insights into how students learn, what interests them, and where they may need additional support, which lead educators and institutions to increasingly using innovative technologies and approaches in their quest to optimize learning experiences and outcomes (Siemens & Long, 2011). Data mining is one such effective technique (Romero & Ventura, 2007).

[1] corresponding author - email: alkisser@gmail.com

Data mining is the process of discovering valuable patterns, dependencies, insights, and knowledge from datasets that contain large amount of data (Chen, Abtahi, Carrero, Fernandez-Llatas & Seoane, 2023). More precisely, it involves employing a variety of computer tools, statistical algorithms, and machine learning approaches that facilitate the extraction of information, hidden relationships and correlations from raw data, that at first glance may not be immediately apparent (Mittal, Shuja & Zaman, 2016). Data mining encompasses a wide range of techniques to extract valuable insights from large datasets such as classification (Dol Aher & Jawandhiya, 2023;Tsimpiris & Kugiumtzis, 2012a; Kaur, Singh & Josan, 2015), association (Antonello et al., 2021), decision trees (Jin, Li, Ma & Wang, 2022), clustering analysis (Romanazzi, Scocciolini, Savoia & Buratti, 2023; Papaioannou et al., 2023b; Hartigan & Wong, 1979; Correa-Morris, Urra Yglesias & Puente, 2023), neural networks (Papaioannou et al., 2023a; Rutkowska et al., 2023), random forest (Schnitzler, Ross & Gloaguen, 2019), k-nearest neighbors (Tsimpiris, Vlachos, & Kugiumtzis, 2012b) etc. In general, this process is essential in diverse fields such as business (Wang, Omar, Alotaibi, Daradkeh & Althubiti, 2022), healthcare (Jothi, Abdul Rashid & Husain, 2015), finance (Jin & Hu, 2022), and education,(Altabrawee, Ali, & Qaisar, 2019; Strikas, et al., 2023; Amelia, Gafar Abdullah, Mulyadi & Ijost, 2019; Ordoñez-Avila, Reyes, Meza, & Ventura, 2023; Aldowah, Al-Samarraie, & Fauzy, 2019; Rodrigues, Zárate, & Isotani, 2018), as it enables informed decision-making, prediction (Sultana, Rani, & Farquad, 2019), and optimization. The part that pertains to education is known as educational data mining.

Educational Data Mining (EDM) refers to the application of data mining techniques in the field of education (Mohamad & Tasir, 2013). EDM aims to extract, evaluate, and comprehend knowledge from massive datasets related to the teaching and learning process (Baker & Yacef, 2009). Information about student performance, teaching methods, educational materials, and other elements that influence the learning process may be included in this. By using data analysis techniques such as predictive models and clustering algorithms, EDM can provide valuable insights into how the teaching and learning process can be improved (Peña-Ayala, 2014). Furthermore, it may anticipate the needs of the students (Shaik et al., 2022), recommend customized strategies, and assist in decision-making to improve the learning environment (Chalaris, Gritzalis, Maragoudakis, Sgouropoulou, & Tsolakidis, 2014).

Educational data mining offers a range of impactful applications in the field of education such as personalized learning paths by analyzing students' learning patterns and preferences (Gobert, Kim, Pedro, Kennedy & Betts, 2015), predicting student's performance that enabling the educators to provide targeted support and resources (Amrieh, Hamtini & Aljarah, 2016; Nabil, Seyam & Abou-Elfetouh, 2021; Sandra, Lumbangaol & Matsuo, 2021), and finally feedback and assessment improvement by examining how students respond to one another and their interactions (Gushchina & Ochepovsky, 2020).

In this paper we will use machine learning algorithms to predict the specialization that Greek high school students will follow. In Greece, in the end of the first year of high school, students have the opportunity to select one of the available offered specializations. The specializations provide students with the ability to delve deeper into specific fields of knowledge and prepare them for the national examinations based on the subjects related to their chosen specialization. Each specialization includes different courses and leads to different career options. Here, we will focus on theoretical and practical specializations. The theoretical and practical specializations are two different directions within the educational system, during high school, that offer different courses and prepare students for different educational and professional paths.

The theoretical specialization focuses on theoretical knowledge and analysis. It includes subjects like Literature, History, Philosophy, Foreign Languages, Ancient Greek and it is suitable for students that are interested in humanities and social sciences and philosophy, as well as for those planning to pursue professional paths that require a strong understanding and analysis of theoretical principles.

The practical specialization emphasizes practical applications, mathematics and physical sciences. It includes subjects like Mathematics, Physics, Chemistry, Biology, Computer Science, and Technology. It is suitable for students interested in sciences and technology and who aim to pursue paths that require practical applications and data analysis.

The problem is that students often struggle with selecting the most suitable specialization, leading to choices that do not align with their strengths and interests. This misalignment can result in poor academic performance and decreased motivation. Research indicate that students typically struggle with this decision-making process, highlighting the necessity of a more supervised approach (Kallio, 1995).

To address this issue, machine learning is employed to develop a scalable and reliable system that can effectively generalize to new data and offer students tailored recommendations based on their academic performance. A variety of machine learning algorithms are utilized, including Random Forest, Naive Bayes, Support Vector Machines (SVM), Neural Networks, Logistic Regression, k-Nearest Neighbors (kNN), and CN2 Rule Induction. These algorithms will analyze students' previous grades, obtained when they all attended the same courses, in order to predict their future specializations. In summary, the primary aim of this article is to explore the potential of becoming a straightforward and valuable tool for educators and parents, that suggests the optimal choice of specialization for students, leveraging their performance in various courses from previous years. To the best of

our knowledge, there is no existing literature specifically addressing this issue within the Greek educational system.

The rest of the paper is structured as follows: Section 2 introduces the fundamental elements of the theory and methodology employed. Section 3 presents the results of the comparative study and finally, Section 4 offers the conclusions.

2. Methods

The primary objective of this study is to develop a model for predicting the specialization that Greek high school students should pursue. This involves leveraging historical data from nine distinct courses and employing machine learning algorithms to identify the most effective approach. To achieve this, a variety of supervised machine learning algorithms including Random Forest, Naive Bayes, Support Vector Machines (SVM), Neural Networks, Logistic Regression, k-Nearest Neighbors (kNN) and CN2 Rule Induction, are utilized. The evaluation of the methods will be conducted using confusion matrices, accuracy, and additional metrics provided by Orange (described in detail below). The entire procedure is executed using the Orange machine learning software.

Orange is a platform of open-source machine learning and data mining tools (Demšar et al., 2013). Predictive modeling, data preprocessing, visualization and other data analysis tasks are all made possible by its comprehensive toolkit and user-friendly interface. Orange is made to be user-friendly for both beginner and experienced data scientists, enabling users to create machine learning models and deal with data efficiently, without requiring a deep understanding of programming. Additionally, Orange offers a visual programming interface that allows users to create data workflows and perform complex analyses with ease. A workflow in Orange is a sequence of interconnected data processing and analysis components, that are performed in a particular order on a dataset. These elements – also referred to as widgets – may comprise tools for evaluation, modeling, preprocessing, data loading, and visualization.

This process included projecting their distributions, identifying missing values, and calculating key metrics such as mean, median, dispersion, minimum and maximum values. Then, the data sampler split the dataset into training and testing subsets. This was crucial for evaluating the performance of the models on unseen data, ensuring that the models generalized well.

Subsequently, several supervised learning algorithms were employed to predict the students' future specializations based on their grades. Specifically, Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), Neural Networks (NN), Logistic Regression (LOGR), k-Nearest Neighbors (kNN), and CN2 Rule Induction (CN2) were employed (see Section 2.2 for a detailed description of the parameters used).

Finally, the models were evaluated using the 'Test and Score' widget. This step involved training each model on the training subset and testing it on the testing subset to assess its performance. The actual specialization chosen by each student, served as the target variable for the supervised learning models. This means that the models were trained to predict this specific outcome based on the input features, which were the grades from nine courses (Modern Greek Literature, Modern Greek Language, Ancient Greek Language, Algebra, Geometry, Physics, Chemistry, Biology, and History), serving as the independent variables. After training, the models could forecast the most probable specialization for new students, based on their grades in the same set

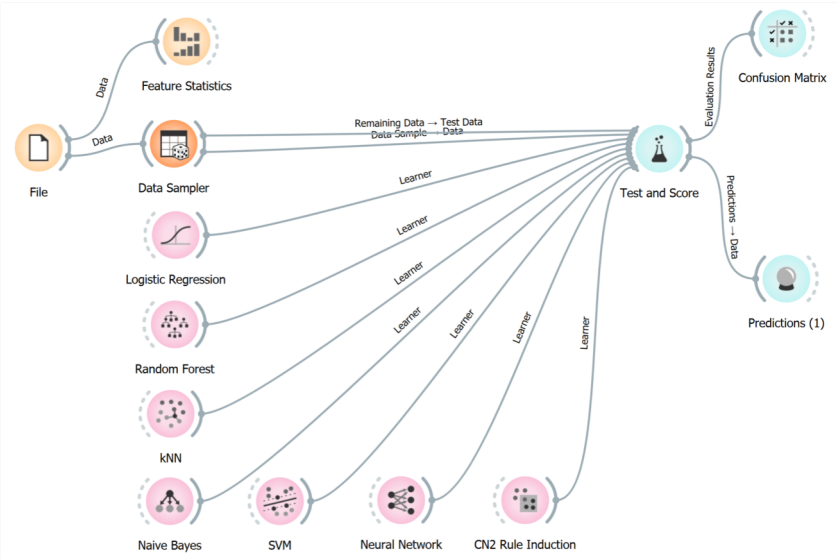

Figure 1 - Workflow in Orange.

of courses. To evaluate the performance of the models, metrics such as confusion matrix, area under ROC curve (AUC), classification accuracy (CA), F1, Precision (Prec), Recall and Matthews correlation coefficient (MCC) were employed.

2.1 Dataset

The Dataset in this paper consists of 530 records of 11 features each. Each record refers to a student. Specifically, they pertain to data from Greek students in the first year of high school from a High school in Serres, during the academic years 2013/2014-2021/2022. Each record has 11 features, such as the ID, the chosen specialization, and grades from nine courses, specifically Modern Greek Literature, Modern Greek Language, Ancient Greek Language, Algebra, Geometry, Physics, Chemistry, Biology, and History. These courses were selected because they are the subjects taken by students in Greece during the Panhellenic exams. The grades in these subjects ultimately determine their admission to university. While the focus was on the grades of these subjects, other potential features, such as attendance records, participation in extracurricular activities, and socio-economic background, were considered. However, these were either not available or not consistent across all records, leading to their exclusion from the current dataset. For each subject, the possible marks range from 0 to 20. The dataset was divided into training set and test set. The training set, which was used to train the different machine learning algorithms, consists of 530 instances randomly selected.

2.2 Machine Learning Algorithms

The machine learning algorithms employed, are discussed in this section. Each algorithm provides unique benefits and faces specific challenges, making them ideal for different aspects of prediction and analysis. The algorithms and the selected parameters and metrics for each algorithm are described below.

1) *Naive Byes*: in machine learning, Naive Bayes is a robust and popular classification algorithm (Zhang, 2004). Based on the Bayes theorem, it makes the assumption that the attributes/features utilized for classification are independent to one another. Naive Bayes classifiers are computationally efficient, quickly and able to achieve impressive results, especially when working with large datasets. However, circumstances in which the independence assumption might not hold true, can impact the accuracy of the model.

2) *Random Forest*: Random Forest is an ensemble learning method used for classification, regression and other tasks (Breiman, 2001). Decision trees are constructed using Random Forest. Each tree is developed from a bootstrap sample from the training data. The term "Random" refers to the arbitrary subset of characteristics that are pulled when creating individual trees, from which the optimal attribute for the split is chosen. The majority vote from each independently formed tree in the forest forms the basis of the final model. In this paper, the parameters are set as follows: the number of trees is set to ten, the minimum subset size for splits is five, and the number of attributes considered at each split is five.

3) *SVM*: Support Vector Machines (SVMs) are supervised learning algorithms used for both classification and regression tasks (Cortes & Vapnik, 1995). SVMs operate by finding the optimal hyperplane in a high-dimensional feature space that maximally separates the different classes of data points. This hyperplane is determined by selecting support vectors, which are the data points closest to the decision boundary. SVMs are unique in that they can handle data that is not linearly separable by using methods such as kernel functions, which convert the data into a higher-dimensional space where separation is feasible. Consequently, this makes SVMs adaptable and efficient for a wide range of applications. In this paper, the parameters for SVM are set as follows: the cost (C) is set to 1, regression loss epsilon to 0.1, numerical tolerance to 0.001, iteration limit to 100 and the kernel type is radial basis function (RBF).

4) *Neural Networks*: Neural networks are computational models that consist of interconnected nodes, or neurons, that process input data to make predictions and to help on decisions (Goodfellow, Bengio & Courville, 2016). A weight is assigned to each connection, and it changes as the connection is trained in order to take advantage of the data. Neural networks are organized in layers, including an input layer, hidden layers for complex pattern recognition, and an output layer for final predictions. In this paper, the parameters for the neural network are set as follows: the number of neurons in the hidden layer is 100, the selected solver is adam and the maximal number of iterations is 200.

5) *Logistic Regression*: Logistic regression is a statistical technique that predicts the probability of an event occurring by considering one or more independent variables (Hosmer, Lemeshow & Sturdivant, 2013). It employs the logistic function to constrain predictions between 0 and 1. In logistic regression, each independent variable's impact on the probability of the event is represented by its coefficient. In this paper, the parameters for logistic regression are set as follows: the regularization type is ridge (L2) and the regularization strength (C) is 1.

6) *k-Nearest Neighbors*: the k-Nearest Neighbors (kNN) algorithm is a versatile and intuitive machine learning method (Cover & Hart, 1967). It functions according to the similarity principle, in which a new data point is categorized in the feature space according to the majority class of its k nearest neighbors. The value of k is a crucial parameter that determines the number of neighbors that will be considered. When decision boundaries are complex or hard to specify mathematically, k-Nearest Neighbors (kNN) is particularly useful. In this paper, the parameters for kNN are set as follows: the number of neighbors is 5 and the distance metric is Euclidean.

7) *CN2 Rule Induction*: CN2 Rule Induction is a machine learning algorithm that is used for classification tasks (Clark & Niblett, 1989). It is also particularly well-suited for generating rule-based models from data. In order to forecast the target variable based on the values of its attributes, CN2 builds rules iteratively. It adds conditions to the rule that maximize information gain, starting with the most influential attribute. Subsequently, by iteratively taking consideration of new attributes, the algorithm improves the rule. This process is carried out by CN2 until no further improvements are possible. In this paper, the parameters for CN2 Rule Induction are set as follows: rule ordering is ordered, the covering algorithm is exclusive, the evaluation measure is entropy, the beam width is 5, the minimum rule coverage is 1, and the maximum rule length is 5.

3. Results

In this section the results are presented. As mentioned above, the primary focus of this research is to evaluate various machine learning algorithms for assessing the selection of specialization of Greek high school students. To accomplish this, a variety of machine learning algorithms including Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), Neural Networks (NN), Logistic Regression (LOGR), k-Nearest Neighbors (kNN), and CN2 Rule Induction (CN2) were employed. Tenfold cross validation was used to evaluate the prediction accuracy. The dataset consists of records (greek high school students) of 11 features each. The performance of the model was measured from different metrics, using tenfold cross validation. In a 10-fold cross-validation with 530 records, each fold will have approximately 53 records (since 530 divided by 10 is 53). During each fold, one-tenth of the data will be used for testing, which means 53 records will be used as test set for each fold, while the remaining nine folds will be used as training set. This process is repeated for each of the folds. Metrics such as confusion matrix, area under ROC curve (AUC), classification accuracy (CA), F1, Precision (Prec), Recall and Matthews correlation coefficient (MCC) were employed to evaluate the performance of the model. During the development of the model, the grades of the courses (Algebra, Biology, etc.) were determined as independent variables, while the selected specialization was determined as the dependent one.

In Table 1, the predicted values of the examined models and the actual values are presented for five randomly selected students. The results for this sample of 5 students indicate that Neural Networks (NN) and Logistic Regression (LOGR) were the most successful, as they did not make any mistakes on their predictions for the selected instances, whereas all the other examined methods made a few mistakes.

Specifically, for students 1, 3, and 4, all methods performed admirably, accurately predicting their chosen specializations. Likewise, for students 2 and 5, Neural Networks (NN) and Logistic Regression (LOGR) excelled, while the other methods (Naive Bayes, k-Nearest Neighbors, Random Forest, Support Vector Machines, and CN2 Rule Induction) encountered challenges in accurately predicting the actual selected specialization of these students. Naive Bayes (NB) and k-Nearest Neighbors (kNN) notably underperformed, as they were unable to accurately predict the actual selected specialization for students 2 and 5. Random Forest (RF) inaccurately estimated the specialization for student 5, as while the student chose theoretical, the estimation of RF was practical. Similarly, Support Vector Machines (SVM) and CN2 Rule Induction (CN2) incorrectly predicted the choices for student 2, since they predicted theoretical while the student actually chose practical.

These findings are also supported by Table 2, which summarizes the success rates achieved by each algorithm, assessed through a range of performance measures employed in this study. Specifically, Neural Networks and Logistic Regression outperformed other machine learning methods across all metrics considered.

Table 1 - Predicted and Actual Specializations for a sample of 5 students. T refers to Theoretical, while P refers to Practical.

Student	ID	LOGR	RF	kNN	NB	SVM	NN	CN2	Actual
1	2058	T	T	T	T	T	T	T	T
2	2059	P	P	T	T	T	P	T	P
3	2061	P	P	P	P	P	P	P	P
4	2062	P	P	P	P	P	P	P	P
5	2066	T	T	P	P	T	T	T	T

Table 2 - Performance metrics for the examined machine learning algorithms.

Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.75	0.70	0.70	0.72	0.70	0.37
RF	0.76	0.72	0.72	0.71	0.72	0.36
NN	0.83	0.76	0.76	0.76	0.76	0.46
NB	0.70	0.67	0.68	0.69	0.67	0.31
LOGR	0.83	0.76	0.75	0.75	0.76	0.45
kNN	0.73	0.70	0.70	0.69	0.70	0.32
CN2	0.69	0.65	0.65	0.65	0.65	0.24

They performed equally well in terms of Area Under the ROC Curve (AUC), Classification Accuracy (CA), and Recall, achieving impressive scores of 0.83, 0.76, and 0.76, respectively. When it comes to F1, Precision (Prec), and Matthews Correlation Coefficient (MCC), Neural Networks exhibited a slight advantage over Logistic Regression, boasting a 0.01 improvement. Random Forest emerged as the third-best performer in terms of AUC, CA, F1 and Recall achieving 0.76 for AUC and consistently achieving 0.72 for the rest performance measures. Support Vector Machine (SVM) claimed the third spot in terms of Precision and Matthews Correlation Coefficient (MCC), achieving scores of 0.72 and 0.37 respectively, an incremental improvement of 0.01 over Random Forest. Additionally,

it's worth noting that k-Nearest Neighbors (kNN) demonstrated a performance that closely aligned with Support Vector Machine (SVM) and Random Forest. However, it consistently lagged behind both SVM and Random Forest. Similarly, Naive Bayes showed performance closely aligned with kNN and both of them performed better compared to CN2, which performed less optimally for this specific task, registering the lowest scores across all performance measures employed in this study. Notably, CN2 consistently underperformed in this specific task, demonstrating scores that were 0.10 to 0.2 lower than the counterparts of Neural Networks and Logistic Regression in the examined measures.

From the Confusion matrix presented in Table 3, it is observed that Logistic Regression algorithm classifies correctly 401 from a total of 530 instances (76%).

Specifically, it accurately identifies 307 out of 351 students who have opted for practical specialization, demonstrating a strong accuracy rate of 87%. On the other hand, it exhibits a noticeably lower accuracy of 52% (94 out of 179) in correctly classifying students who have chosen theoretical specialization.

This suggests that the algorithm excels in predicting students inclined towards practical specialization, while facing relatively more challenge in accurately predicting those leaning towards theoretical specialization.

A similar trend is observed with the Neural Network algorithm (Table 4), as reflected in its confusion matrix, which closely resembles that of logistic regression. Notably, the Neural Network successfully classifies one additional student who has opted for theoretical specialization. A similar pattern is evident in the case of Random Forest, kNN, and CN2 rule inducer algorithms, as depicted in Tables 5, 6 and 7 respectively. These algorithms exhibit a comparable performance pattern to that of Logistic Regression and Neural Network, further emphasizing their effectiveness in predicting student that have chosen practical specialization.

This pattern experiences a subtle shift when considering the SVM and Naive Bayes algorithms, particularly in their accuracy in predicting students who have chosen theoretical specialization. Specifically, in Table 8, it is observed that SVM accurately identifies 253 out of 351 students who have opted for practical specialization, demonstrating an accuracy rate of 72%, while it exhibits a slightly lower accuracy of 61% (109 out of 179) in correctly classifying students who have chosen theoretical specialization. However, this percentage of accurate classification for students with theoretical specialization is comparatively higher than that achieved by other machine learning algorithms.

A similar trend is observed with the Naive Bayes algorithm (Table 9), as reflected in its confusion matrix, which closely resembles that of SVM.

Table 3 - Confusion matrix of the Logistic Regression algorithm.

Logistic Regression algorithm	**Practical**	**Theoretical**
Practical	307	44
Theoretical	85	94

Table 4 - Confusion matrix of the Neural Network algorithm.

Neural Network algorithm	**Practical**	**Theoretical**
Practical	307	44
Theoretical	84	95

Table 5 - Confusion matrix of the Random Forest algorithm.

Random Forest algorithm	**Practical**	**Theoretical**
Practical	303	48
Theoretical	89	90

Table 6 - Confusion matrix of the kNN algorithm.

kNN algorithm	**Practical**	**Theoretical**
Practical	295	56
Theoretical	94	85

Table 7 - Confusion matrix of the CN2 Rule Inducer algorithm.

CN2 Rule Inducer algorithm	**Practical**	**Theoretical**
Practical	280	71
Theoretical	84	95

Table 8 - Confusion matrix of the SVM algorithm.

SVM algorithm	**Practical**	**Theoretical**
Practical	253	98
Theoretical	70	109

Table 9 - Confusion matrix of the Naïve Bayes algorithm.

Naïve Bayes algorithm	**Practical**	**Theoretical**
Practical	252	99
Theoretical	70	109

4. Discussion and Conclusions

This study compared seven machine learning algorithms to investigate their accuracy in assessing the choice of specialization of Greek students in the end of the first year of high school. The data set used consists of 530 students that described by 11 features, such as id, chosen specialization and their final grades in nine core subjects in the first year of high school. Metrics such as confusion matrix, area under ROC curve (AUC), classification accuracy (CA), F1, Precision (Prec), Recall and Matthews correlation coefficient (MCC) were employed to evaluate the performance of the model. As for the results, on testing data, Neural Networks outperformed other machine learning methods across all metrics considered, followed by Logistic regression which was slightly worse when it comes to F1, Precision and Matthews Correlation Coefficient (MCC). In general, all the methods examined showed decent classification accuracy, as even CN2 rule inducer which was the worst compared to the other machine learning algorithms, achieved an accuracy of 65%. Neural Network which was the best overall achieved 76% accuracy.

Confusion matrices confirm that the class (Practical specialization) with larger sample size had improved classification accuracy, contrary to the class with fewer records (Theoretical specialization) for which the algorithms performed poorer. In summary, the results suggest that although challenging, automatic and accurate prediction of the specialization that students will select is feasible. Nevertheless, it could be further improved by using a larger and more diverse dataset, which could include additional relevant features such as attendance records, participation in extracurricular activities, and socio-economic background.

Additionally, examining other machine learning algorithms or even ensemble methods that combine multiple models could improve the prediction accuracy.

Acknowledgements

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

References

Alani, J., Yawe, B., & Mutenyo, J. (2022). Role of Higher Education Growth in Enhancing Economic Growth, Innovation Advancement and Technological Progress in Uganda (1970–2014). *The uganda Higher Education Review, 10*, pp. 1-18. doi:10.58653/nche.v10i1.01

Aldowah, H., Al-Samarraie, H., & Fauzy, W. (2019). Educational Data Mining and Learning Analytics for 21st century higher education: A Review and Synthesis. *Telematics and Informatics*. doi:10.1016/j.tele.2019.01.007

Altabrawee, H., Ali, O., & Qaisar, S. (2019). Predicting Students' Performance Using Machine Learning Techniques. *JOURNAL OF UNIVERSITY OF BABYLON for pure and applied sciences, 27*, pp. 194-205. doi:10.29196/jubpas.v27i1.2108

Amelia, N., Gafar Abdullah, A., Mulyadi, Y., & Ijost, I. (2019). Meta-analysis of Student Performance Assessment Using Fuzzy Logic. *Indonesian Journal of Science and Technology, 4*, pp. 74-88. doi:10.17509/ijost.v4i1.15804

Amrieh, E., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. *International Journal of Database Theory and Application, 9*, pp. 119-136. doi:10.14257/ijdta.2016.9.8.13

Antonello, F., Baraldi, P., Abdelaleem, A., Zio, E., Gentile, U., & Serio, L. (2021). A Novel Association Rule Mining Method for the Identification of Rare Functional Dependencies in Complex Technical Infrastructures from Alarm Data. *Expert Systems with Applications*. doi:10.1016/j.eswa.2021.114560

Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, pp. 3-17. doi:10.5281/zenodo.3554657

Breiman, L. (2001). Random Forests. *Machine Learning*, pp. 5-32. doi:10.1023/A:1010950718922

Chalaris, M., Gritzalis, S., Maragoudakis, M., Sgouropoulou, C., & Tsolakidis, A. (2014). Improving Quality of Educational Processes Providing New Knowledge Using Data Mining Techniques. *Procedia - Social and Behavioral Sciences, 147*. doi:10.1016/j.sbspro.2014.07.117

Chang, V., Chen, Y., & Xiong, C. (2018). Dynamic Interaction between Higher Education and Economic Progress: A Comparative Analysis of BRICS Countries. *Information Discovery and Delivery, 46*. doi:10.1108/IDD-07-2018-0023

Chen, K., Abtahi, F., Carrero, J.-J., Fernandez-Llatas, C., & Seoane, F. (2023). Process mining and data mining applications in the domain of chronic diseases: A systematic review. *Artificial Intelligence in Medicine, 144*, p. 102645. doi:10.1016/j.artmed.2023.102645

Clark, P., & Niblett, T. (1989). The CN2 Induction Algorithm. *Machine Learning*, pp. 261-283. doi:10.1023/A:1022641700528

Correa-Morris, J., Urra Yglesias, A., & Puente, O. (2023). Hybrids of K-means and linkage

algorithms., (pp. 10-17). doi:10.1109/ICAMCS59110.2023.00010

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, pp. 273-297. doi:10.1007/BF00994018

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, pp. 21-27. doi:10.1109/TIT.1967.1053964

Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., . . . Zupan, B. (2013). Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, pp. 2349-2353.

Dol Aher, S., & Jawandhiya, P. (2023). fication Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining — A survey. *Engineering Applications of Artificial Intelligence, 122*, p. 106071. doi:10.1016/j.engappai.2023.106071

Gobert, J., Kim, Y. J., Pedro, M., Kennedy, M., & Betts, C. (2015). Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. *Thinking Skills and Creativity, 18*. doi:10.1016/j.tsc.2015.04.008

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

Gushchina, O., & Ochepovsky, A. (2020). Data mining of students' behavior in E-learning system. *Journal of Physics: Conference Series, 1553*, p. 012027. doi:10.1088/1742-6596/1553/1/012027

Hartigan, J., & Wong, M. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics, 28*, pp. 100--108.

Hosmer, J. D., Lemeshow, S., & Sturdivant, R. (2013). *Applied Logistic Regression.* John Wiley & Sons.

Jin, C., Li, F., Ma, S., & Wang, Y. (2022). Sampling scheme-based classification rule mining method using decision tree in big data environment. *Knowledge-Based Systems, 244*, p. 108522. doi:https://doi.org/10.1016/j.knosys.2022.108522

Jin, X., & Hu, H. (2022). Research and implementation of smart energy investment and financing system design based on energy mega data mining. *Energy Reports, 8*, pp. 1226-1235. doi:10.1016/j.egyr.2022.02.044

Jothi, N., Abdul Rashid, N., & Husain, W. (2015). Data Mining in Healthcare – A Review. *Procedia Computer Science, 72*, pp. 306-313. doi:10.1016/j.procs.2015.12.145

Kallio, R. (1995). Factors influencing the college choice decisions of graduate students. *Research in Higher Education*, pp. 109-124. doi:10.1007/BF02207769

Kaur, P., Singh, M., & Josan, G. (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. *Procedia Computer Science, 57*, pp. 500-508. doi:10.1016/j.procs.2015.07.372

Kilag, O. K., Comighud, E., Amontos, C., Damos, M., & Abendan, C. F. (2023). Empowering Teachers: Integrating Technology into Livelihood Education for a Digital Future. *Journal of Education, Excellencia*.

Kurnia, R. (2021). A Case for Mezirow's Transformative Learning. *Diligentia: Journal of Theology and Christian Education*, p. 73. doi:10.19166/dil.v3i1.2945

Mittal, S., Shuja, M., & Zaman, M. (2016). A Review of Data Mining Literature. *IJCSIS, 14*, pp. 437-442.

Mohamad, S. K., & Tasir, Z. (2013). Educational Data Mining: A Review. *Procedia - Social and Behavioral Sciences, 97*, pp. 320-324. doi:10.1016/j.sbspro.2013.10.240

Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks. *IEEE Access*, pp. 1-1. doi:10.1109/ACCESS.2021.3119596

Ordoñez-Avila, R., Reyes, N., Meza, J., & Ventura, S. (2023). Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review. *9*, p. e13939. doi:10.1016/j.heliyon.2023.e13939

Papaioannou, N., Tsimpiris, A., Kounani, A., Stamatopoulos, I., Angeioplastis, A., & Varsamis, D. (2023a). Comparative analysis of convolutional neural networks for early diagnosis of plant diseases and pest in a multiclass dataset. *International Journal of Computing and Optimization, 10*, pp. 41-53. doi:10.12988/ijco.2023.9968

Papaioannou, N., Tsimpiris, A., Talagozis, C., Fragidis, L., Angeioplastis, A., Tsakiridis, S., & Varsamis, D. (2023b). Parallel Feature Subset Selection Wrappers Using k-means Classifier. *WSEAS TRANSACTIONS ON INFORMATION SCIENCE AND APPLICATIONS, 20*, pp. 76-86. doi:10.37394/23209.2023.20.10

Peña-Ayala, A. (2014). Review: Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications: An International Journal*, pp. 1432-1462. doi:10.1016/j.eswa.2013.08.042

Rodrigues, M., Zárate, L., & Isotani, S. (2018). Educational Data Mining: A review of evaluation

process in the e-learning. *Telematics and Informatics, 35*. doi:10.1016/j.tele.2018.04.015

Romanazzi, A., Scocciolini, D., Savoia, M., & Buratti, N. (2023). Iterative hierarchical clustering algorithm for automated operational modal analysis. *Automation in Construction, 156*, p. 105137. doi:10.1016/j.autcon.2023.105137

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, pp. 135-146. doi:10.1016/j.eswa.2006.04.005

Rutkowska, D., Duda, P., Cao, J., Rutkowski, L., Byrski, A., Jaworski, M., & Tao, D. (2023). The L2 convergence of stream data mining algorithms based on probabilistic neural networks. *Information Sciences, 631*. doi:10.1016/j.ins.2023.02.074

Sandra, L., Lumbangaol, F., & Matsuo, T. (2021). Machine Learning Algorithm to Predict Student's Performance: A Systematic Literature Review. *TEM Journal, 10*, pp. 1919-1927. doi:10.18421/TEM104-56

Schleicher, A. (2018). *World Class: How to Build a 21st-Century School System.* Paris: OECD Publishing. doi:10.1787/9789264300002-en

Schnitzler, N., Ross, P.-S., & Gloaguen, E. (2019). Using machine learning to estimate a key missing geochemical variable in mining exploration: Application of the Random Forest algorithm to multi-sensor core logging data. *Journal of Geochemical Exploration, 205*, p. 106344. doi:https://doi.org/10.1016/j.gexplo.2019.106344

Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2022). Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal, 2*, p. 100003. doi:10.1016/j.nlp.2022.100003

Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE Review*, pp. 30-32. doi:10.17471/2499-4324/195

Strikas, K., Papaioannou, N., Stamatopoulos, I., Angeioplastis, A., Tsimpiris, A., Varsamis, D., & Giagazoglou, P. (2023). State-of-the-art CNN Architectures for Assessing Fine Motor Skills: a Comparative Study. *WSEAS TRANSACTIONS ON ADVANCES in ENGINEERING EDUCATION, 20*, pp. 44-51. doi:10.37394/232010.2023.20.7

Sultana, J., Rani, M. U., & Farquad, H. (2019). Student's Performance Prediction using Deep Learning and Data Mining methods.

Tsimpiris, A., & Kugiumtzis, D. (2012a). Feature Selection for Classification of Oscillating Time Series. *Expert Systems, 29*, pp. 456 - 477. doi:10.1111/j.1468-0394.2011.00605.x

Tsimpiris, A., Vlachos, I., & Kugiumtzis, D. (2012b). Nearest neighbor estimate of conditional mutual information in feature selection. *Expert Systems with Applications, 39*, pp. 12697–12708. doi:10.1016/j.eswa.2012.05.014

Wang, J., Omar, A., Alotaibi, F., Daradkeh, Y., & Althubiti, S. (2022). Business intelligence ability to enhance organizational performance and performance evaluation capabilities by improving data mining systems for competitive advantage. *Information Processing & Management, 59*, p. 103075. doi:10.1016/j.ipm.2022.103075

Zhang, H. (2004). The Optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, pp. 562-567.

Zheng, Y. (2023). Promoting the Personal Development of Children Through Art Education. *Journal of Contemporary Educational Research, 7*, pp. 97-102. doi:10.26689/jcer.v7i4.4857