Machine Learning Models to Predict Students' Study Path Selection

https://doi.org/10.3991/ijim.v16i01.20121

Amir Dirin[1](✉), Charlese Adriana Saballe[2]
[1]Digital Economy Department, Haaga-Helia University of Applied Science, Helsinki, Finland
[2]Zolando.Com
amir.dirin@haaga-helia.fi

Abstract—Selecting a proper study path in higher education is a difficult task for many students. They either have a lack of knowledge on the study path offered or are unsure of their interest in the various options. The current educational setups enable us to collect valid and reliable data on student success and learning behaviour. This study explores and solves the problem of what path to select by proposing possible study paths with the help of machine learning algorithms. Learning analytics (LA) and educational data mining (EDM) are technologies that aid in the analysis of educational data. In this quantitative study, we applied a questionnaire to collect data from students at the Business Information Technology Department (Bite) at the Haaga-Helia University of Applied Science. We managed to collect 101 samples from students during 2017–2018. We used various machine learning algorithms and prediction models to assess the best approach for study path selection. We applied three performance scores of accuracy, Cohen's Kappa, and ROC curve to measure the accuracy of the algorithm results. KNIME analytics was selected as a proper tool to pre-process, prepare, analyse, and model the data. The results indicate that Random Forest (94% accuracy) and Decision Tree (93% accuracy) are the best classification models for students' study path selection. The contribution of this study is for educational data mining research to assess the comparison of various algorithms. Furthermore, this is a novel approach to predict students' study path selection, which educational institutes should develop to assist students in their study path selection.

Keywords—educational data mining, decision trees, random forest, logistic regressions

1 Introduction

Technological advancements in recent years have enabled educational institutes to collect data about students' performance and behavior even at the early age, e.g., [1]. Machine learning (ML) algorithms and approaches assist in identifying patterns and predicting the student's future performance based on collected data. Machine learning data-driven analysis has become popular in the educational context for improving educational offerings. However, selecting the proper algorithms in machine learning is very important; hence, many algorithms have been developed and evaluated to assess the most accurate one for the given dataset.

Data mining in the educational context has become very popular in recent years due to the fact that data collection on students' performance and behaviour has become easier than ever before. Contemporary educational infrastructures such as online registration, online learning, and tracing students' activities and performance have enabled researchers to collect valid and reliable data. Educational data mining (EDM) aims to assist researchers in this field in carrying out various analyses and predictions. This initiative has resulted in educational institutes fulfilling students' needs more efficiently and effectively [2]. The Finnish Ministry of Education recently launched its Vision 2030 (FEC, 2017) for higher education to develop innovative digital learning platforms. [3] reviewed data mining from 1995 to 2005 in an education context. Their findings indicate that various data mining approaches have been applied since 1995. Data mining is done with either statistical, clustering, classification, or outlier detection using various mining techniques such as association rule, patterning, and text mining. All these efforts aim to ease the education offering and facilitate the learning and teaching process more efficiently and effectively.

After the first semester, students often choose their academic path. For some students, making a decision is difficult, especially when the provided study route is unfamiliar. [4] identified factors that impact the selection of study paths: the experience, the habit, the instructor's role, the ability, the curriculum and the university atmosphere and the study culture. Figure 1 presents the process for selecting the study path by students at the business information technology (Bite) at Haaga-Helia University of Applied Science (HHUAS).

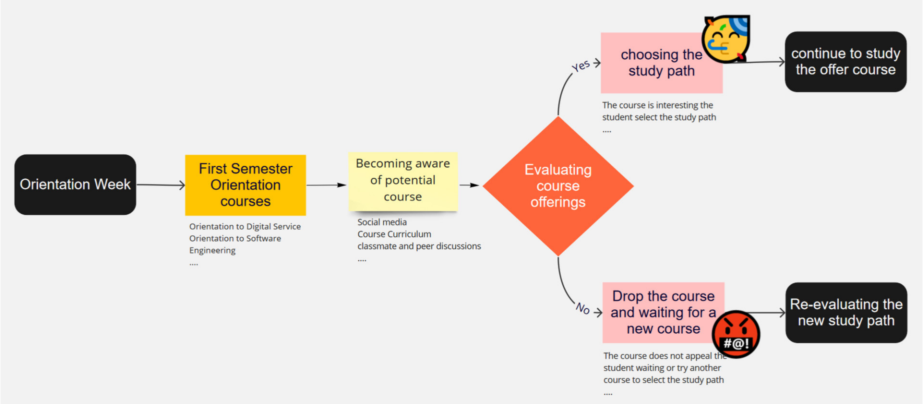

Orientation Week
First Semester Orientation courses
Orientation to Digital Service
Orientation to Software Engineering
....
Becoming aware of potential course
Social media
Course Curriculum
classmate and peer discussions
....
Evaluating course offerings
Yes
choosing the study path
The course is interesting the student select the study path
....
continue to study the offer course
No
Drop the course and waiting for a new course
#@!
The course does not appeal the student waiting or try another course to select the study path
....
Re-evaluating the new study path

Fig. 1. Study path selection process at BITe degree program at HH UAS

In line with this, [5] also recognized that the students' selection is mainly on the basis of the student's prior knowledge, student's judgement of their own competence, and student's own career preferences. [5] revealed that students often engaged with peers and educational institutes through social media to collect information on the university's offered study path or forthcoming courses. Nonetheless, deciding on a study path selection is stressful and difficult for students. Students' motivation suffers as a result of this circumstance, particularly in their first year of college. [6] have demonstrated that students' motivation drops due to the development of high pressure over time. [7] concluded that students need group counselling for study paths and major selections.

In this article, the study path refers to the study modules that universities provide to students as part of their B.Sc. degree curriculum. This research aimed to create a prediction model using machine learning algorithms and statistical analysis. We collected data from students in two different study paths from the (Bite) department at HHUAS. The selected case study paths were digital service design (DSD) and software development (SW), which students need to decide on in the second semesters. In this study, we examined various statistical techniques to determine the most accurate and efficient models to predict students' preference for study paths. The results of this study help students to select the right study paths. Additionally, it assists the degree programs in anticipating students' expectations regarding course implementation.

2 Related studies

2.1 Structure of UAS degree program and study path selection

Studyinfo.fi [8] provides in-depth details about the University of Applied Science (UAS) in Finland. For students with solid knowledge, UAS offers research opportunities within the selected study path. In addition, professional studies help students to develop in-depth competences in the selected major/s. On the other hand, students are obliged to pass elective studies offered by a home degree program or any other UAS. Furthermore, during the practical training period, students receive hands-on experience in industry. Finally, the student is required to complete a final thesis, which is often based on the student's own study path and major selection.

There are many reasons that students select certain study paths. For example, [9] identified that female students select the study path based on their aptitude for the subject, while male students focus on job opportunities and job status. In alignment with this study, [7] revealed that students' personal motivation and intrinsic and extrinsic work values affect their choices for study path selection. [9] developed a proof-of-concept computerized prototype for the academic advising process and the study path selection. [10] created a data mining model to analyse knowledge and help students in their major selection process.

2.2 Data mining

Never before has the data been generated in such high volumes as it is today. Data mining refers to extracting knowledge from datasets. The aim is to explore knowledge that is hidden in the data, for example, patterns among data. This knowledge helps in making reliable and accurate predictions and hence decisions for further actions. [11] defines data mining as identifying and discovering interesting, unexpected, and valuable structures in a large database. The general process of data mining is presented in Figure 2.

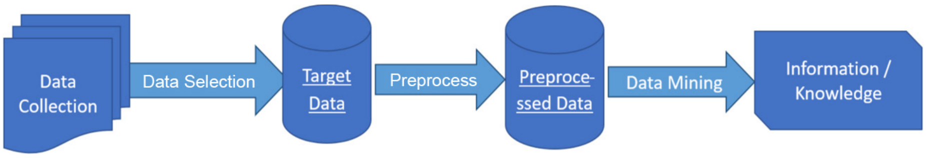


Fig. 2. The general process of data mining

The most essential step is to collect reliable and accurate data. Next is the data preparation, in which proper data are selected. The pre-processed phase makes the data clean, for example, by filling or ignoring the vacancy value of data and removing the irrelevant data. Finally, proper data mining methods are applied to extract information and hence knowledge for prediction or for making decisions. Many methods and techniques have been developed to ease data mining over time, for example, clustering [13] in which similar and related data are grouped and categorised. Similar to this, time-series data mining [14] seeks to extract meaningful knowledge from the data over time. Data mining has been applied in various domains, such as the medical, science, and educational contexts. Regression trees [15] are a static method for constructing a predicational model; this method is more suitable for dependent variables that take continuous or ordered discrete values. The regression model is an iterative process that divides the data into branches and iteratively makes a smaller group. There are many approaches for neural network data mining [16]; however, fuzzy neural networks and self-organization are the most used approaches.

2.3 Data mining in the education context

The world is overwhelmed with data—medical data, scientific data, financial data, marketing data, and educational data. These data are assets to the stakeholders' success if they are properly processed and analysed. Data mining means applying a set of methods to process complex data in a database. Data mining techniques have been widely investigated and developed in recent years [12]. According to [13], educational data mining (EDM) is a research field that performs computational analysis on data that originate in an educational context. A review of the literature on EDM indicates that many statistical and clustering approaches have been developed to help educational institutions [20]. The EDM aimed at helping students and educators to understand the essential needs for offering a proper learning environment and the development of competence. [14] have identified that EDM increases graduation rates, to making pedagogical decisions, and optimizing students' success. A prediction model is widely used in different fields of study, such as finance [15] and [16]. In the educational context, data mining in recent years has increased significantly due to the availability of various data mining and data analytics, such as [17], which applies big data analysis to study challenges in higher education. Likewise, [18] investigated the role of learning analytics and educational data mining in communication and collaboration. [19] believed that the evolution of big data affected learning in higher education. Table 1 presents a summary of the EDM adapted from [20] and extended by the authors. Machine learning is a sub-discipline of artificial intelligence, which implements an algorithm that learns from an input source and predicts the output accordingly. [28] considered machine learning as enabling computers to make successful predictions using past experiences. To make an accurate and reliable prediction, computers must acquire knowledge. This knowledge is constructed through continuous training and testing of the selected data. Training is the process that enables the system to learn. Its performance depends on the type of training, background knowledge, and learning algorithms. To achieve this, there are two main options: modeling and optimization. In modeling, we apply standard algorithms and techniques to predict based on the existing data set on business needs. The output from modeling is, therefore, a trained model, which can make inferences and predictions from a new data set, for example, live stream. Data-intensive machine learning methods have been adapted in science, technology, and commerce [21]. The success of the prediction also depends on the adapted algorithm and the datasets. The algorithms are divided into supervised and unsupervised learning. In supervised learning, the machine is trained to use data that are well labelled, while in unsupervised training, data are not labelled. In general, supervised learning is used for prediction, and unsupervised data are used for analysis.

2.4 Machine learning

Machine learning is a subset of AI that has become popular with the fast growth of data sets. Statistics is the foundation of machine leaning, upon which ML is built. The quality of the machine learning performance depends on the accuracy of the datasets, the processing power, and the amount of training data set [22]. The amount of data set impacts how accurately recurring patterns and the correlations among the data are identified. Machine learning is categorized into three main groups. First, "supervised learning," often called the 'task-driven approach', for which the objective is known. Classification and regressions are the main approaches for supervised learning. In classification, the data is attributed to a specific label in which the output is either nominal or discrete. On the other hand, regression is employed to predict unknown elements on the basis of the existing observations. The second group is unsupervised learning, or the 'data-driven approach'. We do not need to define goals; the patterns are identified through clustering or associations. This is an approach that allows us to identify hidden patterns. Finally, reinforcement learning is the third type of ML focus in training AI to perform complicated tasks through a reward system. The decision is made through trial and error that the algorithm optimizes itself, learning from the previous iterations. Deep learning is a subset of machine learning that is inspired by the structure of the human brain using a neural network (NN). The most common ML algorithms that are used especially in the educational context are predictions: prediction algorithms are often applied for supervised learning either in classification or regression. In this approach, we use the features and the interrelation among these data to define the key variable. Specifically, we use linear regression, decision tree, support vector machine (SVM), and k-nearest neighbour (KNN):

- Linear regression is the most basic prediction model and is based on dependent and independent variables. We use training data as a basis, and accordingly, a linear equation is created to fit the observation as closely as possible by applying the least square method. The aim is to minimize the sum of the errors, which are the difference between the estimations and their projections.
- Logistic regression, on the other hand, is applicable when the outcome variable is categorized, referred to as binomial logistic regression.
- The decision tree is also a classification model, which categorizes the data and presents the results in a tree-like structure. It consists of a root node and then divides into sub branches, which represent more nodes.
- Random forest is an algorithm in which it uses a tree learning method for classification and regression to construct a multitude decision tree.

2.5 Data mining tools

The challenges associated with analyzing data have led to the development of various statistical and data mining tools. The most well-known data mining tools are IBM SPSS [30] and R [23], which are statistical and graphical software for data mining. KNIME Analytics Platform [24] and Bayesian labs [25]. These tools are applied based on the nature of the data and the purposes of the analysis. These tools also contain robust educational data mining features and capabilities.

2.6 Motivation theory

Motivation refers to the individual's internal processes that help to sustain behaviour. The attempt to identify and develop models and theories of motivations has a long history. These theories are rooted in factors of psychological and physiological motivation. Some of the motivational theories are discussed below. [26] showed that perceived successes and failures share three common properties: locus, stability, and controllability. All these properties affect common emotional experiences such as anger, gratitude, guilt, pride, pity, and shame. [35] The theory of human motivation categorizes human needs as basic physiological requirements, safety, belonging, esteem, and self-actualization. Maslow asserted that these needs must be satisfied. Therefore, through motivational theory, we are able to justify the reasons behind people's behaviour. Cognitive evaluation theory investigates extrinsic or intrinsic motivation [27]. This theory endeavours to explain the external impact on individual internal motivations. This theory has been widely used in education to identify the intrinsic and extrinsic factors that affect students' motivations. Motivational factors in the context of education have also been investigated thoroughly by [28]. In accordance with the recommendation of a theoretical framework of students' motivations, performance, and behaviour, there have been extensive activities of machine learning algorithms and data mining modeling. For example, behavioural modeling [29], prediction of performance [30], predictions of dropout rates [31], and improvement of assessments [32] For contemporary students, the extrinsic impact has mainly been rooted in the impact of smart gadgets on students' motivation and performances. One example is a [42] study of Facebook absorption on student achievement. [29] identified methods for recognizing patterns in educational data, such as clustering, regression (logistic/multiple), and discovery models.

3 Research questions and methods

The aim of this study is to use various statistical modeling approaches to find the best prediction models for students' preferences for selecting their study paths modeling.

3.1 Research questions

This study aims to answer the following questions:

1. Are there any accurate data model techniques that would accurately predict students' study path selection?
2. Is there a model that classifies or clusters the students based on the study paths, mastery orientation, motivation, or demographic attributes?

The answers to this question help students to select proper study paths and hence develop competence for their future careers. Furthermore, the educational institutions may apply the results in planning and allocating resources for students.

3.2 Research methods and questionnaire

In this study, we utilized the CRISP-DM methodology [33]. The CRISP-DM has proven to be efficient and has been widely used for data mining workflow processes. The CRISP-DM consists of six steps: business understanding, data understanding, data preparation modeling, evaluation, and deployment. Furthermore, this quantitative research approach helps us to quantify opinions, beliefs and behaviours along with other defined variables. We designed and tailored a questionnaire for collecting data from students in four different courses in business information technology (BITe) at Haaga-Helia University of Applied Science. The data were collected through a pre-registration survey during February and March 2019. The questionnaire was customized from the motivation survey done at the University of Oulu [34] and the mastery intrinsic/mastery extrinsic orientation by Niemivirta [28]. The questionnaire consisted of two distinct parts. The first part of the questionnaire contained questions to collect participants' basic information, such as age, gender, geographical origin, semester level, and specialization path. The second part consisted of 26 tailored questions, which reflected students' affinity for either software development (SW) or digital service design (DSD) paths, motivation orientation, and career orientation. In the second set of questions, we used a 7-point Likert-like scale in which 1 means strongly disagree and 7 means strongly agree. The aim of this questionnaire was to identify three kinds of statements. The first type expressed a motivational goal (MG). Examples of this type are '*Career development and promotions are important for me' or 'Salary means a lot to me*'. The second type revealed mastery extrinsic orientation (MEO) (using statements like, '*It is important for me that I get good grades*', 'A*n important goal for me is* *to do well in my studies*', and '*My goal is to succeed in school*') and mastery intrinsic orientation (MIO) (using statements like, '*I study in order to learn new things*', '*An important goal for me is to learn as much as possible.*' and '*To acquire new knowledge is an important goal for me in school.*' from theories of motivation.

3.3 Participants

The data were collected as a hard copy from the following courses: Orientation to Software Engineering, Programming (Java), User Experience Design, and Digital Service Design. Respondents were given time in class to complete the survey. We decided to collect data from third-level students and above, since students select their study path after the second semester. We reached students through respective classes or a WhatsApp group using an online version of the questionnaire, which was created in Google Forms. Students participated voluntarily in the study, and the information collected was recorded anonymously. A total of 101 samples were collected, which represents one-quarter of the BITe student population in 2019. Figure 3a presents the gender distributions of the participation, whose age varies between 17–35, and Figure 3b presents the country origin of the respondents.

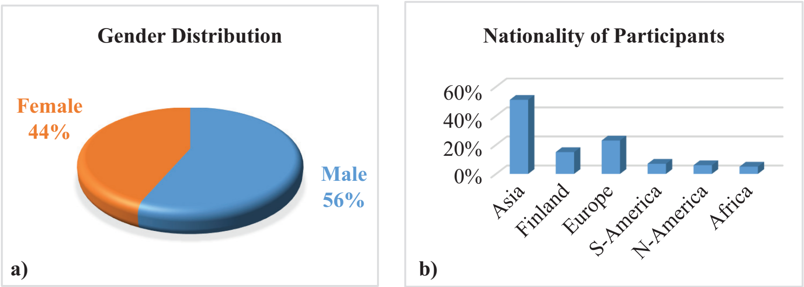
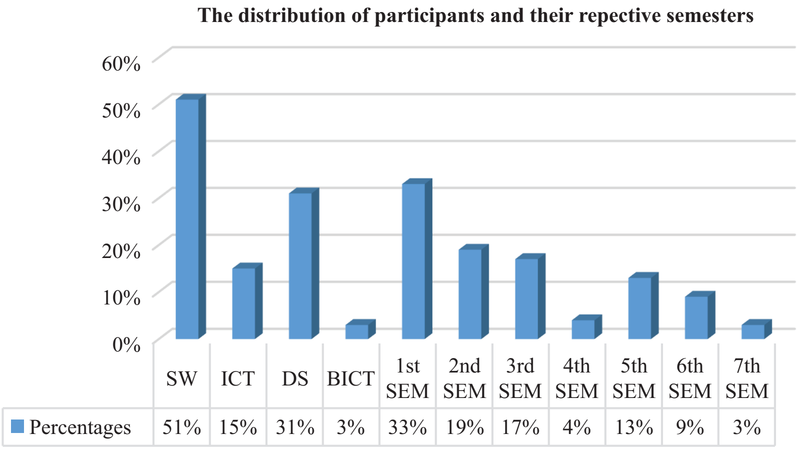


Fig. 3. a) Gender distribution of the participants; b) Nationality of the participants

Figure 4 presents the distribution of the populations in different courses, e.g., Software (SWD), ICT infrastructure (ICT), Digital Service Design (DS), and Business ICT (BICT) paths.



	SW	ICT	DS	BICT	1st SEM	2nd SEM	3rd SEM	4th SEM	5th SEM	6th SEM	7th SEM
■ Percentages	51%	15%	31%	3%	33%	19%	17%	4%	13%	9%	3%

Fig. 4. Study and semester distributions of the samples

3.4 Data preparation

As a first step, we transferred the data to Excel spreadsheets and then exported the data in csv format to KNIME. The analysis process began with a preliminary inspection of the gathered data. In this phase, we traced and identified the missing and incorrect values in the dataset. Therefore, we use the Most Frequent Value [35] to fix the missing values in two rows. Outliers are handled through the Numeric Outlier node in KNIME and replaced by the closest permitted values. Furthermore, data were filtered that showed only Software Development (SWD) and DSD study paths for binary classification, which is required for a logistic regression algorithm. Additionally, to improve the accuracy of the prediction in machine learning, we applied z-score normalization to rescale the data for standard normal distribution.

3.5 Data modeling

Dependent Variable: In this study, the students' study path selection is considered as a dependent variable. The variable is not numeric and based on two study paths: software development (SD) and digital service design (DSD).

Independent Variable: The first set of independent attributes is related (SWD, DSD factors), and the second set, motivational attributes, is based on [37], such as (MG, MEO, MIO factors), while the third set of attributes is student demographic variables (age, gender, and geographic area).

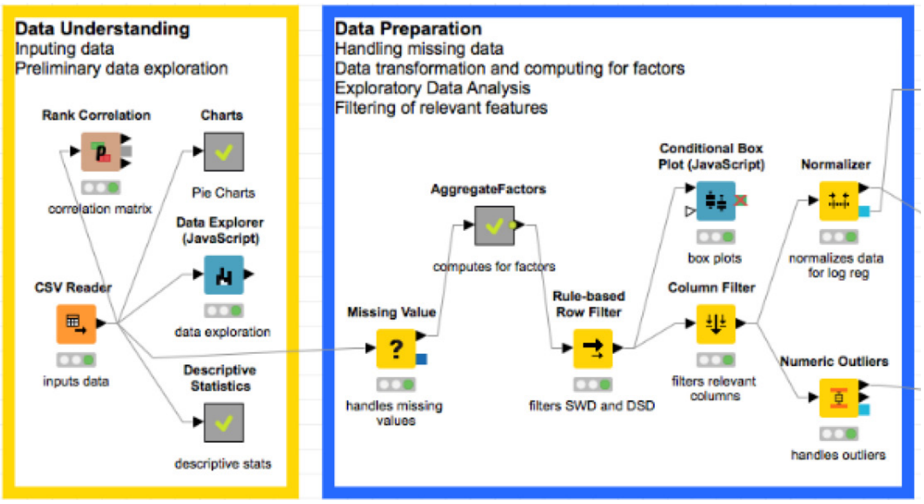
Approaches:

Linear regression is not a relevant approach for the prediction since the dependent variable (SD or DSD) is not numeric. Therefore, we used logistic regression for binary outcome (SWD or DSD), which also answers the research question.

Random Forest and decision trees were used to identify students' categorization of study path, mastery orientation, motivational goals, and demographic attribution, such as semester level, gender, age, and geographic area. Furthermore, as the numerical variables for the model building technique for the logistic regression, stepwise procedures are used for backward elimination. For logistic regression modeling, the data was bootstrapped to have n=2000 to approximate a larger sample size. In the decision trees and random forested modeling, the data were bootstrapped to have n=500 in order to have enough sample size but also avoid overfitting. The data were partitioned into 80% training dataset and 20% testing dataset, which meant for logistic regression that the training data was n=1600 and the testing data was n=400.

3.6 Data analysis

Average scores for the four factors were computed, and exploratory data analysis was performed using descriptive statistics, correlation matrix, and boxplot. The data are filtered to show only the SW and DSD study paths to ensure the binary classification of data, as required by the logistic regression algorithm. We applied z-score normalization to rescale the data to get the standard normal distribution with a mean of 0 and a standard deviation of 1, required for optimizing the logistic regression algorithm. Figure 5 shows the KNIME workflow for the different phases of the study.



Data Understanding
Inputing data
Preliminary data exploration

Data Preparation
Handling missing data
Data transformation and computing for factors
Exploratory Data Analysis
Filtering of relevant features

Fig. 5. Workflow of different phases of the data processing

3.7 Applied machine learning models

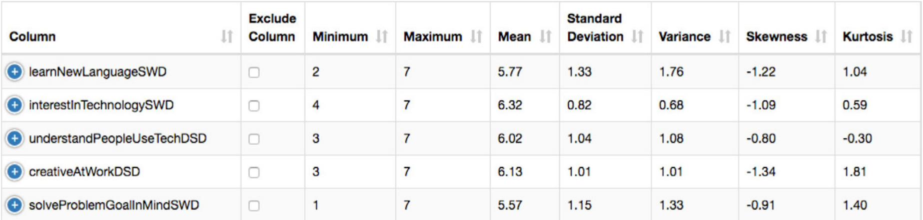
In this study, we applied supervised learning techniques along with various modeling experiments. We applied logistic regression to predict specialization paths to ensure a binary outcome of SW or DSD for the response variable. For classification, random forest and decision tree were used to categorize students' specialization paths. Furthermore, we used a bootstrapping method [36] to simulate a larger dataset. Finally, the results of multiple modeling efforts were evaluated based on three performance measures: accuracy, Cohen's kappa [47], and area under the receiver operating characteristic (ROC) curve [37].

4 Results

4.1 Selecting predictors for DSP and SW

The details of the analysis report are presented as B.Sc. thesis work at [38]. We classified the results into two different factors: comparisons and predictions. We defined the numerical variables after performing the essential pre-processing phase, such as the most frequent value technique. The next step we pursued was to identify the correct predictors for classification for the DSP and SW study path. A sample of the descriptive statistics of the numerical variables is presented in Table 1.

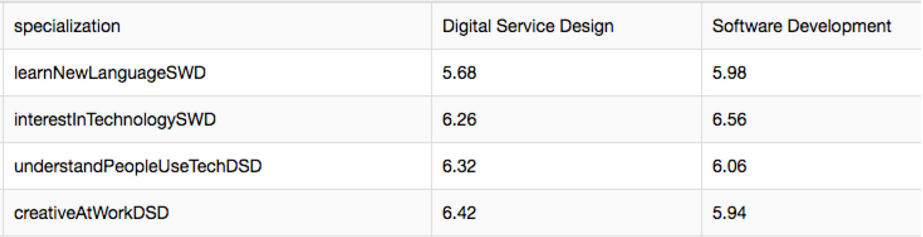
Table 1. Sample of descriptive statistics of numerical features



Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis
⊕ learnNewLanguageSWD	□	2	7	5.77	1.33	1.76	-1.22	1.04
⊕ interestInTechnologySWD	□	4	7	6.32	0.82	0.68	-1.09	0.59
⊕ understandPeopleUseTechDSD	□	3	7	6.02	1.04	1.08	-0.80	-0.30
⊕ creativeAtWorkDSD	□	3	7	6.13	1.01	1.01	-1.34	1.81
⊕ solveProblemGoalInMindSWD	□	1	7	5.57	1.15	1.33	-0.91	1.40

Table 1 presents the minimum, maximum, mean, standard deviation, variance, skewness, and kurtosis values. As indicated in the table, "To acquire new knowledge is an important goal for me in school" received the highest mean (6.39) and the lowest variation (0.77) among the factors. This indicates that participants from the specialization paths gave similar ratings for the statement. Therefore, this is not a good predictor for the selection of study path since all students were in strong agreement regardless of their study path. In line with this, the statement "I am interested in technology" had the second-highest mean (6.32) and the second-lowest standard deviation (0.82). This also demonstrates that there were high levels of agreement between students' ratings. The analysis of the mean values of the selections of the study path by digital service design and software development is presented in Table 2.

Table 2. Sample of mean values for the selection of study paths



specialization	Digital Service Design	Software Development
learnNewLanguageSWD	5.68	5.98
interestInTechnologySWD	6.26	6.56
understandPeopleUseTechDSD	6.32	6.06
creativeAtWorkDSD	6.42	5.94

The statement "*I enjoy working in an environment where there is always something new going on*" had a higher mean for DSD (5.77) than for SW (5.63), even though the question was targeted for SW groups. The same can be noted for two other factors for SW: "*I always keep myself up to date on information about new technological innovations*" had a higher mean for DS (5.77) than for SW (5.63), and "*I would like to invent and develop new devices and applications*" had a higher mean for DSD (5.87) than for SW (5.66). Therefore, these three statements are not good predictors or classifiers for SW.

4.2 Collinearity of the data using correlation matrix

We have applied a collinearity approach to identify the highly correlated independent variables. The two statements for the mastery extrinsic orientation, "*An important goal for me is to do well in my studies*" and "*My goal is to succeed in school*", were strongly correlated, with a measure of 0.69. This means that one of the statements should be taken out of the selection to avoid the possibility of skewing the results of the regression model due to the redundancy of the information. None of the other variables exhibited a high correlation with each other, so they would not disturb the results of the regression. Figure 6 shows the high collinearity of the independent variables.

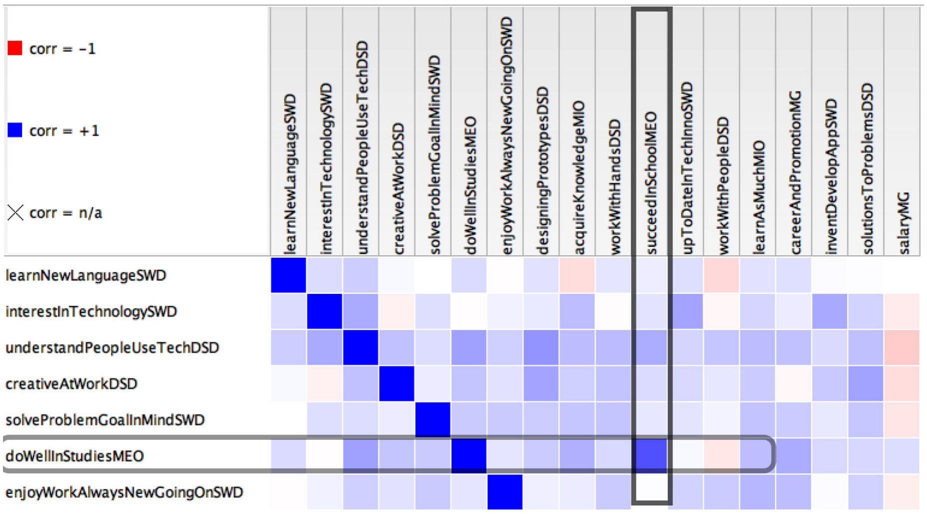


Fig. 6. Sample of correlation matrix between features

4.3 The distribution of variables using box plots

We use box plots to show the distribution of the numerical variables (DS, SWD, MEO, MIO, and MG). Figure 7a and b show the box plots sample of the untrimmed data on the distribution of the ratings given by DSD versus SDW students.

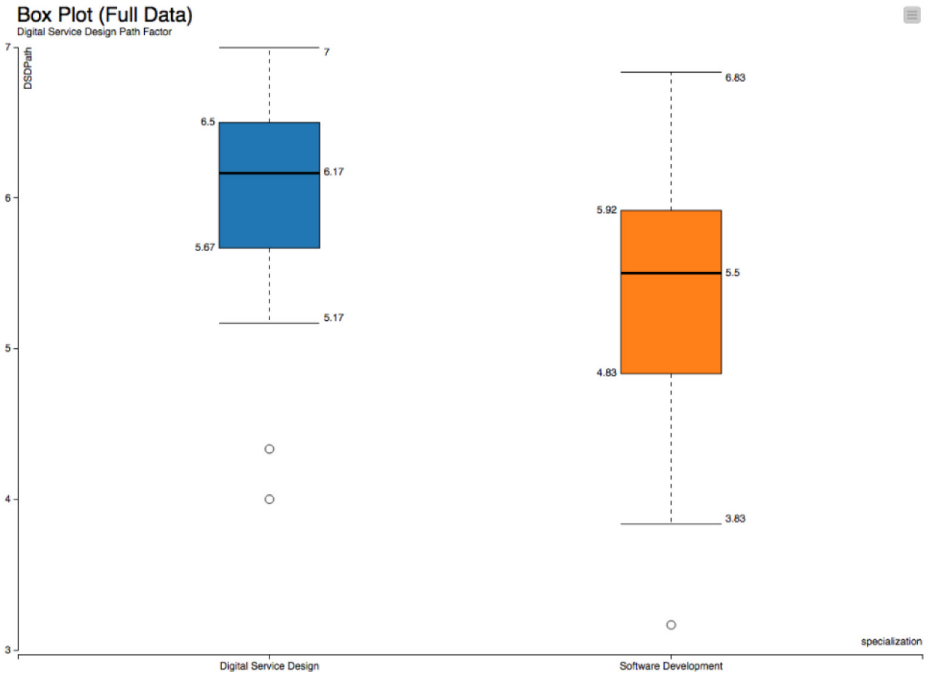


Fig. 7a. Distribution of ratings for DSD factors by study path

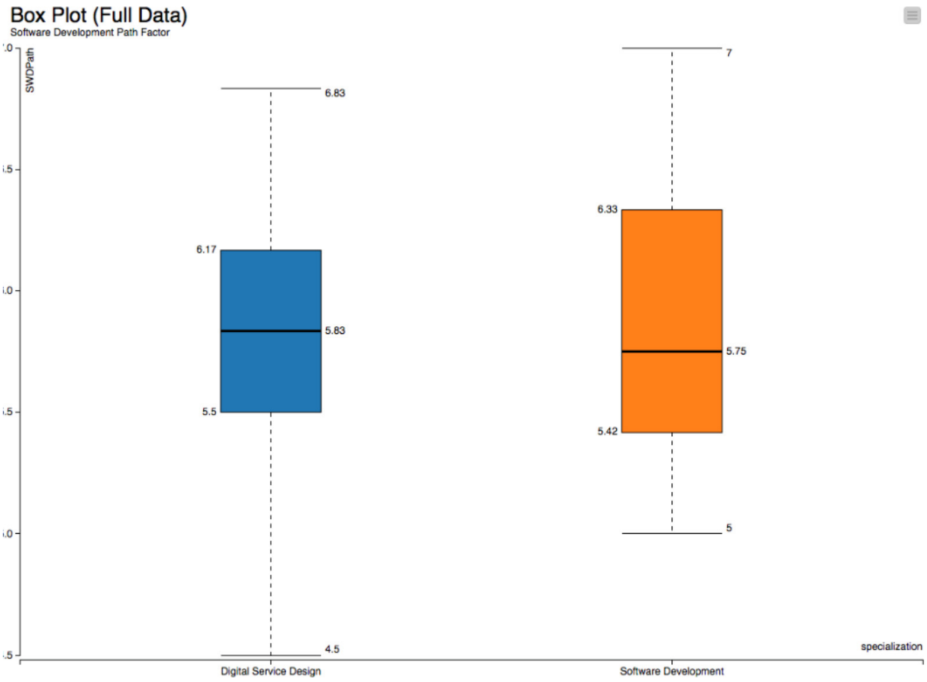


Fig. 7b. Distribution of rating for SWD factors by study path

The boxplots for DSD students are shorter than for SWD students. This means that DSD students have a higher level of agreement in terms of their ratings in comparison with SWD. The median response of DSD students (6.17) was also higher than the median response of SWD students (5.5) in terms of agreement with DSD-related statements. This suggests that the DSD factor is a viable predicator for the study path. The ratings for SWD factors by DSD versus SWD students for all the statements are shown in Figure 7b. Regarding the median for both groups, the plots do not exhibit a clear difference in the distribution between the two groups. This suggests that using the SWD factor with all six SWD statements may not be a good predictor.

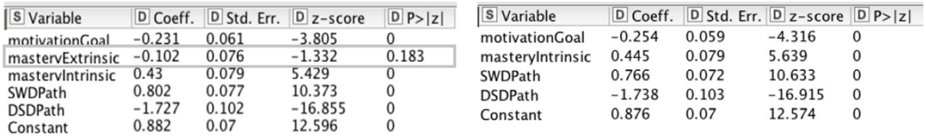
The exploratory data analysis indicated that at least six statements were candidates for being removed from the feature selection of the regression and classification models. Four were in SWD: "*I am interested in technology*", "*I enjoy working in an environment where there is always something new going on*", "*I always keep myself up to date with information on new technological innovations*", and "*I would like to invent and develop new devices and applications".* In addition, there was one each in Mastery Extrinsic Orientation ("*My goal is to succeed in school*") and Mastery Intrinsic Orientation ("*To acquire new knowledge is an important goal for me in school*"). Therefore, these statements were excluded from the factors, the averages were recomputed, and the boxplots were re-inspected to verify the aptness of the trimmed factors.

4.4 Prediction models

We have applied logistic regression models to predict the students' study path selection of DSD and SWD. We utilized bootstrapped samples of the training dataset (n=1600) for the development of logistic regression models. The modeling began with full numerical features, and then the next insignificant variable was trimmed until all that remained in the model were significant predictors. Afterwards, demographic factors were added and checked for their significance.

Table 3a demonstrates the coefficient of regression and relevant statistics. The regression model of all the numerical factors is based on the training dataset (n=1600). The Chi square test [39] (P > |z|) value for mastery extrinsic orientation is not less than 0.05. This suggests that there is not enough evidence to conclude that MEO is a significant predictor, so it was dropped from the regression model.

Table 3. a) Coefficient of regression and statistics. The logistic regression model b) Coefficient of regression and statics. Sample logistic regression model for DS, SWD, MIO, and MG factors. The logistic regression model 1



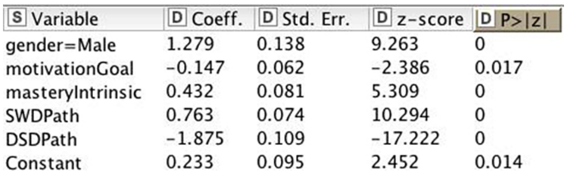
S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
motivationGoal	-0.231	0.061	-3.805	0
masteryExtrinsic	-0.102	0.076	-1.332	0.183
masteryIntrinsic	0.43	0.079	5.429	0
SWDPath	0.802	0.077	10.373	0
DSDPath	-1.727	0.102	-16.855	0
Constant	0.882	0.07	12.596	0

S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
motivationGoal	-0.254	0.059	-4.316	0
masteryIntrinsic	0.445	0.079	5.639	0
SWDPath	0.766	0.072	10.633	0
DSDPath	-1.738	0.103	-16.915	0
Constant	0.876	0.07	12.574	0

In Table 3b, the sample of logistic regression model for DS, SWD, MIO, and MG factors are presented. Note that the p-values are at or near zero for all the factors.

Table 4 presents the coefficient of regression and statistics logistic regression model 2. The gender was added 4. The p-value of gender is 0.017, which means that it was significant at 0.05.

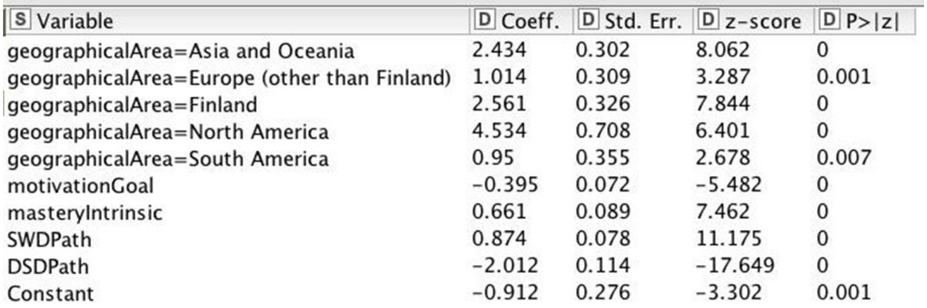
Table 4. The coefficient of regression and statistics logistic regression model 2



S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
gender=Male	1.279	0.138	9.263	0
motivationGoal	−0.147	0.062	−2.386	0.017
masteryIntrinsic	0.432	0.081	5.309	0
SWDPath	0.763	0.074	10.294	0
DSDPath	−1.875	0.109	−17.222	0
Constant	0.233	0.095	2.452	0.014

Table 5 presents the logistic regression and statistical regression model 3 of the geographical area. The geographical areas were all significant variables with p-values close to zero.

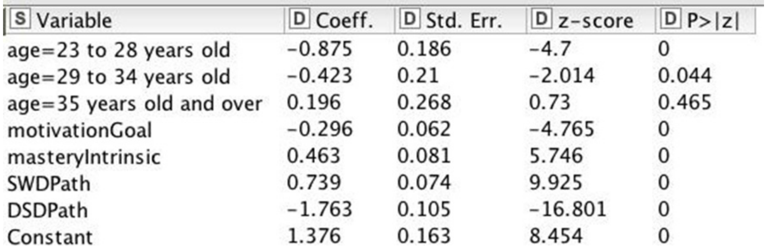
Table 5. The coefficient of regression and statistics logistic regression model 3



S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
geographicalArea=Asia and Oceania	2.434	0.302	8.062	0
geographicalArea=Europe (other than Finland)	1.014	0.309	3.287	0.001
geographicalArea=Finland	2.561	0.326	7.844	0
geographicalArea=North America	4.534	0.708	6.401	0
geographicalArea=South America	0.95	0.355	2.678	0.007
motivationGoal	−0.395	0.072	−5.482	0
masteryIntrinsic	0.661	0.089	7.462	0
SWDPath	0.874	0.078	11.175	0
DSDPath	−2.012	0.114	−17.649	0
Constant	−0.912	0.276	−3.302	0.001

As with the previous models in model 4, as shown in Table 6, the age variable is added. All the age categories were significant in the model except for age 35 and above, for which the p-value equals 0.465.

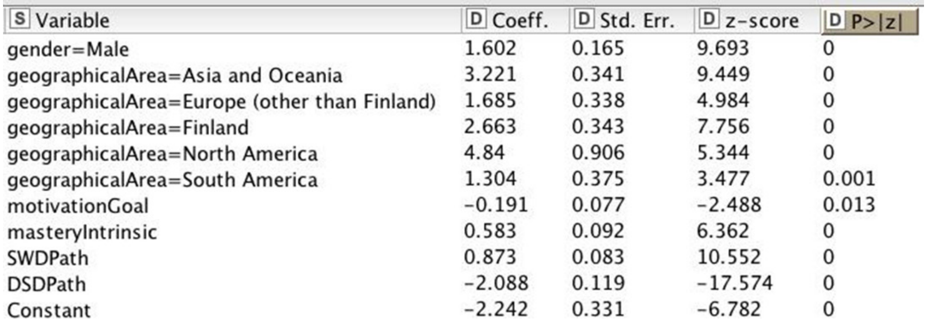
Table 6. The coefficient of regression and statistics logistic regression model 4



S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
age=23 to 28 years old	−0.875	0.186	−4.7	0
age=29 to 34 years old	−0.423	0.21	−2.014	0.044
age=35 years old and over	0.196	0.268	0.73	0.465
motivationGoal	−0.296	0.062	−4.765	0
masteryIntrinsic	0.463	0.081	5.746	0
SWDPath	0.739	0.074	9.925	0
DSDPath	−1.763	0.105	−16.801	0
Constant	1.376	0.163	8.454	0

Table 7 presents that logistic regression model 5 is the same as the logistic regression model except the addition of gender and geographical area. As the table shows, all the variables, including the two dummy variables, were significant predictors.

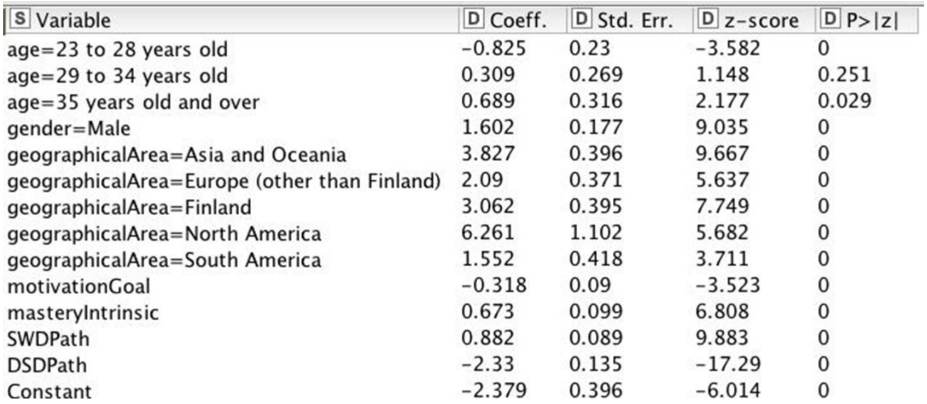
Table 7. The coefficient of regression and statistics logistic regression model 5



S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
gender=Male	1.602	0.165	9.693	0
geographicalArea=Asia and Oceania	3.221	0.341	9.449	0
geographicalArea=Europe (other than Finland)	1.685	0.338	4.984	0
geographicalArea=Finland	2.663	0.343	7.756	0
geographicalArea=North America	4.84	0.906	5.344	0
geographicalArea=South America	1.304	0.375	3.477	0.001
motivationGoal	-0.191	0.077	-2.488	0.013
masteryIntrinsic	0.583	0.092	6.362	0
SWDPath	0.873	0.083	10.552	0
DSDPath	-2.088	0.119	-17.574	0
Constant	-2.242	0.331	-6.782	0

Table 8 displays the resulting coefficients of regression and z-statistics. It shows that all variables were significant except the category age=29 to 34 years old, which had a p-value of .251.

Table 8. Sample of the coefficient of regression and statistics. Logistic regression model 6

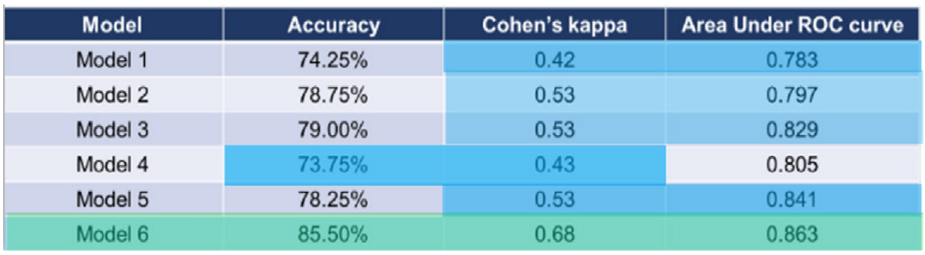


S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
age=23 to 28 years old	-0.825	0.23	-3.582	0
age=29 to 34 years old	0.309	0.269	1.148	0.251
age=35 years old and over	0.689	0.316	2.177	0.029
gender=Male	1.602	0.177	9.035	0
geographicalArea=Asia and Oceania	3.827	0.396	9.667	0
geographicalArea=Europe (other than Finland)	2.09	0.371	5.637	0
geographicalArea=Finland	3.062	0.395	7.749	0
geographicalArea=North America	6.261	1.102	5.682	0
geographicalArea=South America	1.552	0.418	3.711	0
motivationGoal	-0.318	0.09	-3.523	0
masteryIntrinsic	0.673	0.099	6.808	0
SWDPath	0.882	0.089	9.883	0
DSDPath	-2.33	0.135	-17.29	0
Constant	-2.379	0.396	-6.014	0

4.5 Performance measures

We applied the performance accuracy measurement of the logistic regression models through bootstrap testing with a data subset (n=400). Table 9 presents the performance accuracy, Cohen's kappa, and ROC scores for the logistic regression models.

Table 9. Sample of performance accuracy, Cohen's kappa, and ROC scores for the LR models

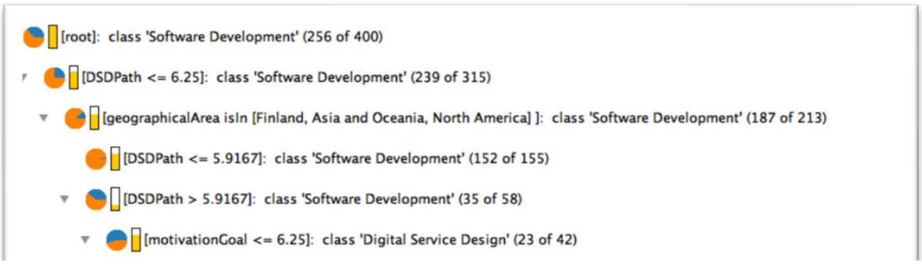


Model	Accuracy	Cohen's kappa	Area Under ROC curve
Model 1	74.25%	0.42	0.783
Model 2	78.75%	0.53	0.797
Model 3	79.00%	0.53	0.829
Model 4	73.75%	0.43	0.805
Model 5	78.25%	0.53	0.841
Model 6	85.50%	0.68	0.863

Logistic regression model 6 had the highest scores for all three measures: performance accuracy of 85.5%, Cohen's kappa of 0.68, and ROC probability of 0.863. Logistic regression Model 1 received the lowest performance values for ROC (0.78) and Cohen's kappa (0.42). Logistic regression model 4 achieved the lowest accuracy score (73.75%) and second-lowest kappa (0.43). Logistic regression Models 2, 3, and 5 had comparable performance scores with moderate kappa scores (0.53) and good ROC scores.

4.6 Classification modeling

We applied the Decision Tree and Random Forest algorithms for students' classification modeling. For the classification (n=400), a training dataset was used. The initially hypothesized classifiers were DSD, SWD, MEO, MIO, MG factors, semester level, gender, age, and geographical area. Several iterations and combinations of factors were modelled for each of the two classification algorithms. The resulting models were then validated using the testing data subset (n=100). The resulting scores for accuracy, Cohen's kappa, and probability of the area under the ROC curve were noted. To keep this report concise, only the best model from the two methods is presented. The best in terms of the set accuracy criteria was DTModel 26, which is presented in Figure 8. The following classifiers were used in this DT model: DS, SWD, MG, and geographical area of origin.

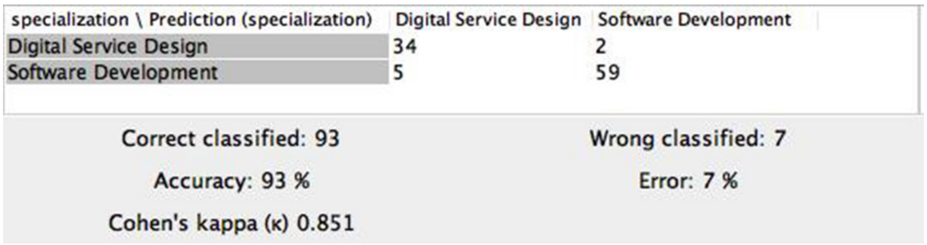


```
[root]: class 'Software Development' (256 of 400)
  ▾ [DSDPath <= 6.25]: class 'Software Development' (239 of 315)
    ▾ [geographicalArea isIn [Finland, Asia and Oceania, North America] ]: class 'Software Development' (187 of 213)
        [DSDPath <= 5.9167]: class 'Software Development' (152 of 155)
      ▾ [DSDPath > 5.9167]: class 'Software Development' (35 of 58)
        ▾ [motivationGoal <= 6.25]: class 'Digital Service Design' (23 of 42)
```

Fig. 8. Sample view of the decision tree model

An accuracy test was performed on the model using the testing data (n=100). As shown in Table 10, the correctly classifying (n=34) DSD and (n=59) SWD students had a combined accuracy rate of 93% and a Cohen's kappa of 0.851.

Table 10. Confusion matrix and accuracy scores of the decision tree model



specialization \ Prediction (specialization)	Digital Service Design	Software Development
Digital Service Design	34	2
Software Development	5	59

Correct classified: 93

Wrong classified: 7

Accuracy: 93 %

Error: 7 %

Cohen's kappa (κ) 0.851

Figure 9 presents the resulting plot of the ROC curve, which yielded an area under the curve with a .959 probability of correctly classifying the selection of study path using the factors as opposed to classifying randomly.

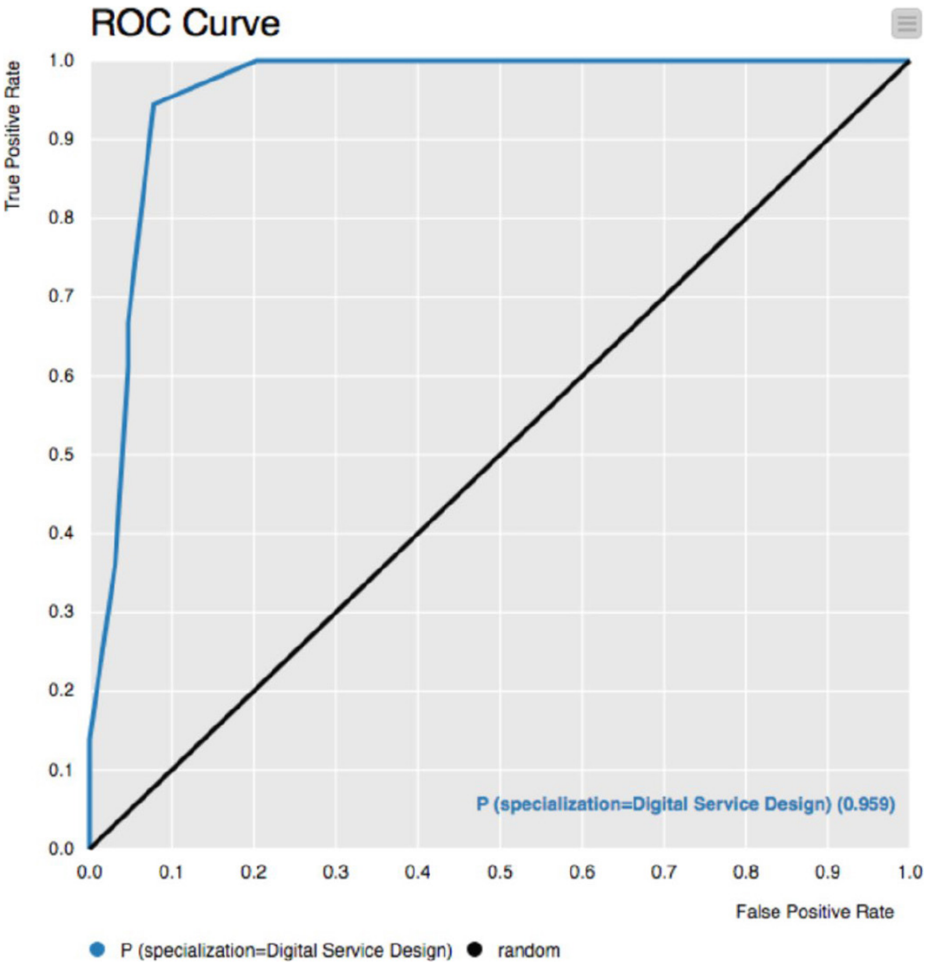
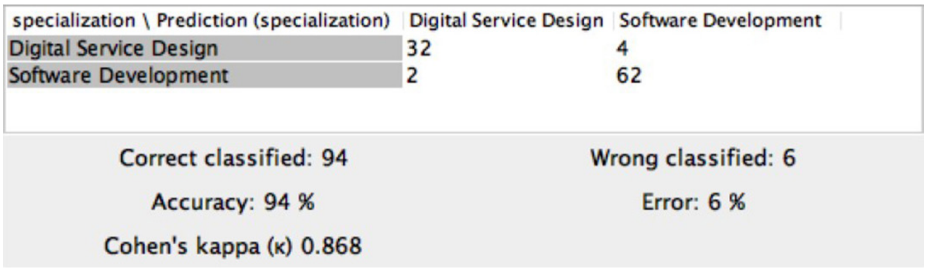


Fig. 9. Receiver operating characteristic curve – decision tree

4.7 Selecting the best random forest model

The best model formed by using the random forest algorithm was RFModel 22, which is presented in Table 11. The assessed model has DSD and SWD, motivational goal, age, and geographical area of origin as the classifiers. Fitting the model on the testing dataset (n=100) produced the confusion matrix shown in Table 11. The model performed extremely well by correctly classifying 32 students from DSD and 62 students from DWD. It had an overall accuracy score of 94% and a Cohen's kappa of 0.868.

Table 11. Confusion matrix and accuracy scores of the random forest model



specialization \ Prediction (specialization)	Digital Service Design	Software Development
Digital Service Design	32	4
Software Development	2	62

Correct classified: 94

Wrong classified: 6

Accuracy: 94 %

Error: 6 %

Cohen's kappa (κ) 0.868

Figure 10 shows that the area under the ROC curve generated a probability of 0.987 that the model is able to correctly categorize between the two study paths using the classifiers as opposed to categorizing at random.

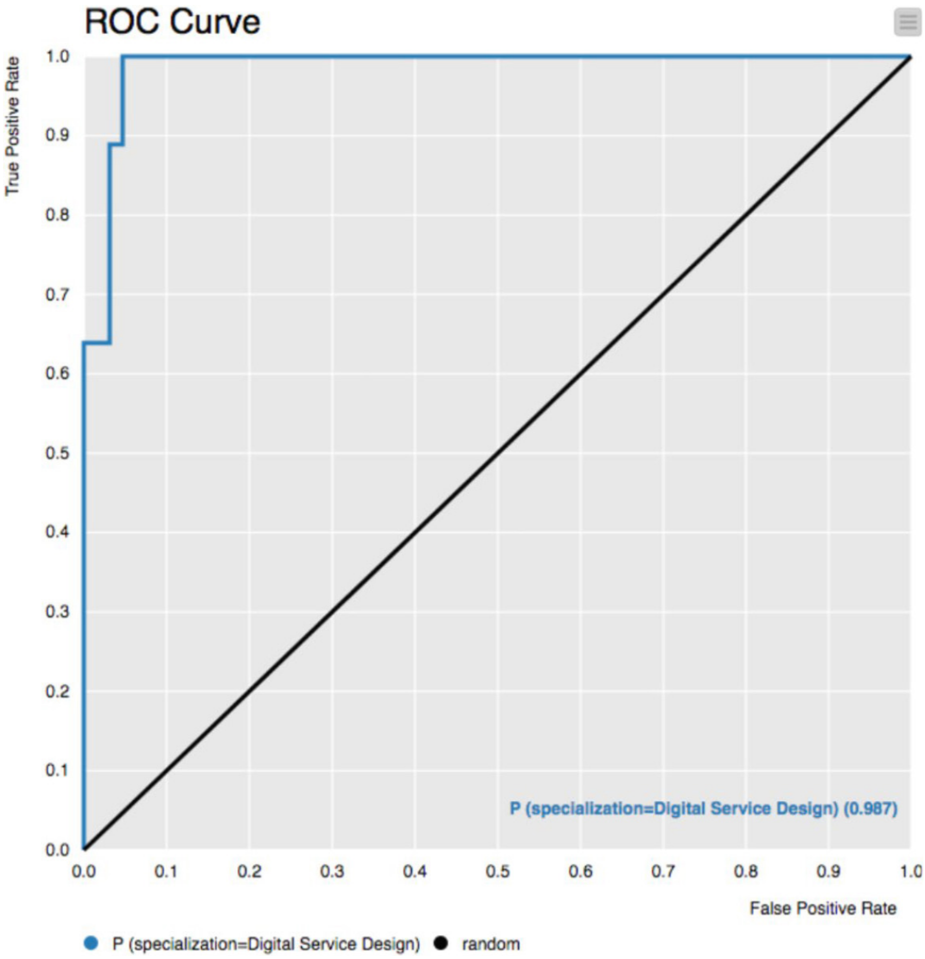


Fig. 10. The area under receiver operating characteristic cure – random forest

The variables measuring affinity to SWD (two statements), DS (six statements), motivational goal (two statements), mastery intrinsic orientation (one statement), and the demographic data of age, gender, and geographical origin were significant predictors of study paths. The results showed that the final logistic regression model could predict the selection of the study path of BITe students (either SWD or DSD) with 85.5% accuracy. There was also an 86.3% probability (area under the ROC curve) that the model could distinguish between the two study paths. A validation test of the model resulted in even higher accuracy (94.12%) in predicting the students' selection of software development or digital service design for their study path. Similarly, the classification models derived from both random forest and decision tree algorithms resulted in very high measures of accuracy: 94% for the random forest model and 93% for the decision tree model. Because there were only very slight differences in the performance measures of these models, both are recommended to be used for classification. The random forest algorithm would provide a slightly higher accuracy rate, but it is more challenging illustrating that model clearly. On the other hand, the decision tree model would be easier to interpret by extracting a classification rule from its tree view.

5 Discussion

Data mining through machine learning models is a common practice in different fields, for example, in marketing and sales. In the context of education, data mining is a relatively new approach to predicting students' behavior. Accurate prediction depends on many factors, such as the reliability and accuracy of the collected data. Data mining in the educational context has increased in recent years due to the high accessibility and availability of reliable educational data. For example, the prediction of students' dropout rates [50], students' performance predictions [51], [40], and students' social behaviour [52]. The prediction helps students and educational institutes to save money, time, and resources. All these types of predictions have mainly focused on the students' performance. But they have not predicted what students desire to learn and in which fields they pursue to develop their professional competencies. Therefore, this study aims to overcome the existing gaps in recommending an approach to predict students' study path selection. The current study complements the previous publications for students' study path selection through social media [5]. This study recommends approaches to predict and anticipate students' study path selections.

Predicting a student's preference in selecting a study path is vital for academic institutions and students. It would help students to save time and effort in identifying their passion and motivation for their future career. And it would help educational institutes in allocating resources and saving costs. Therefore, we applied several statistical approaches, for example, decision trees and random forests, for classifying motivations and preferences on the study path selection among various options that have been provided by the degree program. [41] shows that college students have to take many things when they want to choose a college major. [41] demonstrates that students' decision making for their career is influenced by their parents, coaches, religious figures, or any other role models and peers. This study, however, focuses mainly on the factors and the promotions at educational institutes, which articulate students' decision on study paths and hence their career selection.

5.1 Answer to the research questions

The following two research questions led us to conduct this research study:

1. Are there any accurate data model techniques that would accurately predict students' study path selection?
2. Is there a model that classifies or clusters the students based on the study paths, mastery orientation, motivation, or demographic attributes?

To answer these questions, we selected two distinct case studies of study paths, that is, digital service design and software development in the business information technology (Bite) department at the Haaga-Helia University of Applied Science. After pre-processing the data in Table 2 and Figure 2, we selected the mastery extrinsic orientation (MEO) and mastery intrinsic orientation (MIO) motivational goals (MG) as appropriate predictors for this study. These predictors have also been applied in previous studies (e.g., Hamedi & Dirin, 2018NIEMIVIRTA, 2002; Tuominen-Soini, Salmela-Aro, & Niemivirta, 2012; Winne & Baker, 2013) [43]. Furthermore, we have used demographic factors such as age, gender, and nationalities as variables for clustering purposes. We clustered the statements in the questionnaire as a selection factor; for example, as demonstrated in Table 1, we assumed that the statement 'I am interested in technology' would be rated more highly by software students than by design students. However, an inspection of the data showed that the statement was rated equally by both SWD and DS students. However, [53] demonstrated that the interest in technology correlates with the amount of time the pupils taught technology and the use of technology in the environment in which they interact, for example, at home or at school. Furthermore, as shown in Table 2, three statements ('*I enjoy working in an environment where there is always something new going on*', '*I always keep myself up to date with information on new technological innovations*', and '*I would like to invent and develop new devices and applications*') were hypothesized to be factors linked with SWD students. However, the result is contrary to the initial assumptions. The analysis indicated that DSD students have higher mean ratings on these statements than SWD students. Hallström [54] elaborates that technological determinism appears at all levels and is not limited to specific fields. Based on Figure 7a and b, we have identified that DSD factors are viable predictors for study path selections. These demonstrate that mastery extrinsic orientation (MEO) and mastery intrinsic orientation (MIO) are not good predictors or classifiers of study path selection. This is somehow aligned with the findings of [44] that mastery-oriented students' personal lives impact their educational performance and study effort. The logistic regression model of all numeric factors is run on the trained dataset (see Table 4). The coefficient of the regression model also demonstrates that the MEO is not a significant predictor. The final six models put forward for consideration received high accuracy ratings, moderate to substantial kappa scores, and good ROC accuracy measures (see Tables 6 and 7). The best among the models in terms of the set evaluation criteria is the logistic regression model, which achieved the highest scores for all three performance measures. The overall results of the proposed models and testing reveal that the logistic regression model has the highest accurate prediction probability for the selection of study paths. This is also proven to be a reliable prediction model in our case study. Pearce and Ferrier [55] evaluated the reliability of the probabilities of occurrence of the logistic regression model and the discrimination capacity of correctly identifying the dependent and independent variables. In line with this study, [45] compared decision tree and logistic regressions to analyse the learning curve. They found that logistic regressions were better for smaller data sets and tree induction was better for larger data sets. The accuracy of logistic regression is the reason that this model has been widely used in the medical field [46]. In the context of education, logistic regressions have been applied in various studies, for example, to determine the success factors of international student marketing [47] and to investigate the success factors for online teaching [48]. Our comparison of models also demonstrated that logical regression is the best model to use for students' predictions. Moreover, using the following significant classifiers, two classification models were generated by using the random forests and decision trees algorithms: DSD and SWD factors, motivational goal, age, and geographical area of origin. Testing of the models indicates that they also have very high accuracy for the prediction of the study path selection. Random forests are the most popular algorithms for prediction in machine learning [49]. The accuracy of the random forest and decision tree algorithms for prediction is also demonstrated by [56].

5.2 Reliability and validity

This study was conducted with real users and in students who chose to select a study path or had already selected the study path. Therefore, the data was collected from real users and in a real environment. Furthermore, the questionnaire, for example, questions about student's efficacy are based on the previously proven academic questionnaires. Hence, we believe that the results of this study are reliable in the context in which we conducted the study. Furthermore, to ensure the validity of the results, we examined the prediction through various algorithms. However, we believe that with more sample data, we may end up with more accurate results.

6 Conclusions

Advancement in technologies and data analysis enabled us to measure students' behaviour and performance at an individual level and in various courses. EDM empowered us to identify the dependent and independent factors that affect all students to some degree. Furthermore, machine learning algorithms will enable us to predict students' behaviour and performance in the future. This study sought to identify the right algorithms for predicting the students' study path selection. Therefore, we have conducted a case study of business information technology (BITe) at Haaga-Helia University of Applied Sciences (UAS). The main objective of the research was to help find an optimal algorithm for detecting and predicting students' study path selection. The results support the educational institute in improving educational offerings through the use of data-driven insights into students' study path preferences. We conducted exploratory research to apply machine learning techniques to develop: (i.) A prediction model of the two most common study paths in the program. (ii.) A classification model based on several factors such as students' affinity to Software Development (SWD) or Digital Service Design (DSD), motivational goals, mastery orientation, and other demographic variables. As a future plan, we aim to extend the research by engaging with more study paths and exploring more algorithms and their accuracy in predicting students' preferred study paths. This plan helps us to collect more samples to improve the accuracy of the predictions.

7 Acknowledgement

Special thanks to students in the DSD and SW study paths at BITe for providing reliable data.

8 References

[1] S. Papadakis, "Robots and Robotics Kits for Early Childhood and First School Age," Int. J. Interact. Mob. Technol., vol. 14, no. 18, 2020. https://doi.org/10.3991/ijim.v14i18.16631

[2] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," IEEE Access. 2017. https://doi.org/10.1109/ACCESS.2017.2654247

[3] FEC, "Vision for Higher Education and Research in 2030." [Online]. Available: https://minedu.fi/en/vision-2030. [Accessed: 11-Apr-2020].

[4] K. Määttä and S. Uusiautti, "How to Enhance the Smoothness of University Students' Study Paths?" Int. J. Res. Stud. Educ., 2011. https://doi.org/10.5861/ijrse.2012.v1i1.16

[5] A. Dirin, M. Nieminen, and A. Alamäki, "Social Media and Social Bonding in Students' Decision-Making Regarding their Study Path," Int. J. Inf. Commun. Technol. Educ., vol. 17, no. 1, pp. 88–104, 2020. https://doi.org/10.4018/IJICTE.2021010106

[6] K. Salmela-Aro and S. Read, "Study Engagement and Burnout Profiles among Finnish Higher Education Student," Burn. Res., 2017. https://doi.org/10.1016/j.burn.2017.11.001

[7] "A Study on the Effects of Paramedic Students' Major Selection Motivation and Occupational Values on Employment Preparation Behavior," J. Digit. Converg., vol. 18, no. 8, pp. 263–270, 2020.

[8] "University of Applied Science." [Online]. Available: https://studyinfo.fi/wp2/en/higher-education/polytechnics-universities-of-applied-sciences/polytechnic-uas-bachelors-degree/. [Accessed: 01-Jul-2020].

[9] Y. Bouaiachi, M. Khaldi, and A. Azmani, "A Prototype Expert System for Academic Orientation and Student Major Selection," Int. J. Sci. Eng. Res., vol. 5, no. 11, pp. 25–28, 2014.

[10] K. Kularbphettong and C. Tongsiri, "Mining Educational Data to Support Students' Major Selection," World Acad. Sci. Eng. Technol. Int. J. Educ. Pedagog. Sci., vol. 8, no. 1, 2014.

[11] D. J. Hand, "Principles of Data Mining," in Drug Safety, 2007. https://doi.org/10.2165/00002018-200730070-00010

[12] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 2012.

[13] C. Romero and S. Ventura, "Educational Data Mining: A Review of the State of the Art," IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews. 2010. https://doi.org/10.1109/TSMCC.2010.2053532

[14] G. Javidi, L. Rajabion, and E. Sheybani, "Educational Data Mining and Learning Analytics: Overview of Benefits and Challenges," in Proceedings – 2017 International Conference on Computational Science and Computational Intelligence, CSCI 2017, 2018. https://doi.org/10.1109/CSCI.2017.360

[15] A. Blanco-Oliver, A. Irimia-Dieguez, and N. Reguera-Alvarado, "Prediction-Oriented PLS Path Modeling in Microfinance Research," J. Bus. Res., 2016. https://doi.org/10.1016/j.jbusres.2016.03.054

[16] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," MIS Q. Manag. Inf. Syst., 2012. https://doi.org/10.2307/41703503

[17] D. Leung and R. Law, "Information Technology, Tourism," Encyclopedia of Tourism, vol. 29, no. 3, 2015, pp. 885–886. https://doi.org/10.1016/S0160-7383(02)00009-9

[18] G. Siemens and R. S. J. D. Baker, "Learning Analytics and Educational Data Mining: Towards Communication and Collaboration," in ACM International Conference Proceeding Series, 2012. https://doi.org/10.1145/2330601.2330661

[19] P. Baepler and C. Murdoch, "Academic Analytics and Data Mining in Higher Education," Int. J. Scholarsh. Teach. Learn., 2010. https://doi.org/10.20429/ijsotl.2010.040217

[20] R. Stegmann, "Student Performance in an Online Learning Platform – Predicting Grades given Student Features and Behaviour.," Aalto University, 2016.

[21] M. I. Jordan and T. M. Mitchell, "Machine Learning: Trends, Perspectives, and Prospects," Science, 2015. https://doi.org/10.1126/science.aaa8415

[22] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine Learning on Big Data: Opportunities and Challenges," Neurocomputing, vol. 237, 2017. https://doi.org/10.1016/j.neucom.2017.01.026

[23] Y. Zhao, R and Data Mining: Examples and Case Studies, 2012.

[24] KNIME.COM AG, "KNIME Analytics Platform," KNIME Analytics Platform product sheet, 2016.

[25] S. Conrady and L. Jouffe, Bayesian Networks and Bayesia Lab, vol. 1, no. 1, 2007.

[26] B. Weiner, "An Attributional Theory of Achievement Motivation and Emotion," Psychol. Rev., 1985. https://doi.org/10.1007/978-1-4612-4948-1

[27] M. Gagné and E. L. Deci, "Self-Determination Theory and Work Motivation," J. Organ. Behav., 2005. https://doi.org/10.1002/job.322

[28] M. Niemivirta, "Motivation and Performance in Context: The Influence of Goal Orientations and Instructional Setting on Situational Appraisals and Task Performance," Psychol. – An Int. J. Psychol. Orient, vol. 45, no. 4, pp. 250–270, 2002. https://doi.org/10.2117/psysoc.2002.250

[29] Z. Papamitsiou and A. A. Economides, "Temporal Learning Analytics for Adaptive Assessment," J. Learn. Anal., 2014. https://doi.org/10.1145/2567574.2567609

[30] T. Mishra, D. Kumar, and S. Gupta, "Mining Students' Data for Prediction Performance," in International Conference on Advanced Computing and Communication Technologies, ACCT, 2014. https://doi.org/10.1109/ACCT.2014.105

[31] A. Hamedi and A. Dirin, "A Bayesian Approach in Students' Performance Analysis," in EDULEARN18 Proceedings, 2018. https://doi.org/10.21125/edulearn.2018.2498

[32] D. Schultz, S. Duffield, S. C. Rasmussen, and J. Wageman, "Effects of the Flipped Classroom Model on Student Performance for Advanced Placement High School Chemistry Students," J. Chem. Educ., 2014. https://doi.org/10.1021/ed400868x

[33] C. Pete et al., "Crisp-Dm 1.0," Cris. Consort., 2000.

[34] U. of Oulu, "Study choice selections," 2019. [Online]. Available: http://cases.zef.fi/unioulu/. [Accessed: 04-Aug-2020].

[35] J. Zhang, "Most Frequent Value Statistics and the Hubble Constant," Publ. Astron. Soc. Pacific, 2018. https://doi.org/10.1088/1538-3873/aac767

[36] D. A. Freedman, "Bootstrapping Regression Models," Ann. Stat., 1981. https://doi.org/10.1214/aos/1176345638

[37] P. Sedgwick, "Receiver Operating Characteristic Curves," BMJ (Online). 2013. https://doi.org/10.1136/bmj.f2493

[38] C. Saballe, "Using Machine Learning Models to Predict the Study Path Selection of Business Information Technology Students," Haaga-Helia University University of Applied Science, 2019.

[39] N. Pandis, "The Chi-Square Test," American Journal of Orthodontics and Dentofacial Orthopedics, 2016. https://doi.org/10.1016/j.ajodo.2016.08.009

[40] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised Data Mining Approach for Predicting Student Performance," Indones. J. Electr. Eng. Comput. Sci., 2019. https://doi.org/10.11591/ijeecs.v16.i3.pp1584-1592

[41] D. Fizer, "Factors Affecting Career Choices of College Students Enrolled in Agriculture," 2013.

[42] H. Tuominen-Soini, K. Salmela-Aro, and M. Niemivirta, "Achievement Goal Orientations and Academic Well-Being Across the Transition to Upper Secondary Education," Learn. Individ. Differ., 2012. https://doi.org/10.1016/j.lindif.2012.01.002

[43] P. H. Winne and R. S. J. d. Baker, "The Potentials of Educational Data Mining for Researching Metacognition, Motivation and Self-Regulated Learning," J. Educ. Data Min., 2013.

[44] C. Senko and K. M. Miles, "Pursuing their Own Learning Agenda: How Mastery-Oriented Students Jeopardize their Class Performance," Contemp. Educ. Psychol., 2008. https://doi.org/10.1016/j.cedpsych.2007.12.001

[45] C. Perlich, F. Provost, and J. S. Simonoff, "Tree induction vs. Logistic regression: A learning-curve analysis," J. Mach. Learn. Res., 2004.

[46] D. W. Kononen, C. A. C. Flannagan, and S. C. Wang, "Identification and Validation of a Logistic Regression Model for Predicting Serious Injuries Associated with Motor Vehicle Crashes," Accid. Anal. Prev., 2011. https://doi.org/10.1016/j.aap.2010.07.018

[47] T. Mazzarol, "Critical success Factors for International Education Marketing," Int. J. Educ. Manag., 1998. https://doi.org/10.1108/09513549810220623

[48] T. Volery and D. Lord, "Critical Success Factors in Online Education," Int. J. Educ. Manag., 2000. https://doi.org/10.1108/09513540010344731

[49] B. F. F. Huang and P. C. Boutros, "The Parameter Sensitivity of Random Forests," BMC Bioinformatics, 2016. https://doi.org/10.1186/s12859-016-1228-x

[50] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting Students Drop out: A Case Study," in EDM'09 – Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, 2009.

[51] A. Daud, M. D. Lytras, N. R. Aljohani, F. Abbas, R. A. Abbasi, and J. S. Alowibdi, "Predicting Student Performance Using Advanced Learning Analytics," in 26th International World Wide Web Conference 2017, WWW 2017 Companion, 2019. https://doi.org/10.1145/3041021.3054164

[52] J. Bayer, H. Bydžovská, J. Géryk, T. Obšívač, and L. Popelínský, "Predicting Drop-out from Social Behaviour of Students," in Proceedings of the 5th International Conference on Educational Data Mining, EDM 2012, 2012.

[53] J. Ardies, S. De Maeyer, D. Gijbels, and H. van Keulen, "Students Attitudes Towards Technology," Int. J. Technol. Des. Educ., 2014. https://doi.org/10.1007/s10798-014-9268-x

[54] J. Hallström, "Embodying the Past, Designing the Future: Technological Determinism Reconsidered in Technology Education," Int. J. Technol. Des. Educ., 2020. https://doi.org/10.1007/s10798-020-09600-2

[55] J. Pearce and S. Ferrier, "Evaluating the Predictive Performance of Habitat Models Developed Using Logistic Regression," Ecol. Modell., 2000. https://doi.org/10.1016/S0304-3800(00)00322-7

[56] R. Accorsi, R. Manzini, P. Pascarella, M. Patella, and S. Sassi, "Data Mining and Machine Learning for Condition-Based Maintenance," Procedia Manuf., 2017. https://doi.org/10.1016/j.promfg.2017.07.239

9 Authors

Dr. Amir Dirin is an adjunct professor in educational technology at the University of Helsinki and lecturer at Haaga-Helia university of applied science. His main research area is in immersive experience, machine learning, and user experience design and development.

Charlese Adriana Saballe holds a bachelor degree in Business Information Technology (Bite) and currently working as a data mining researcher at Zolando.com.

Article submitted 2020-11-29. Resubmitted 2021-05-02. Final acceptance 2021-05-08. Final version published as submitted by the authors.