Haaga-Helia
University of Applied Sciences

Using Machine Learning Models to Predict the Study Path Selection of Business Information Technology Students

Charlese Adriana Saballe

Bachelor's Thesis
Degree Programme in Business Information Technology
2019

Haaga-Helia
University of Applied Sciences

Abstract

10 May 2019

Author(s)
Charlese Adriana Saballe

Degree programme
Degree Programme in Business Information Technology (BITe)

Report/thesis title
Using Machine Learning Models to Predict the Study Path Selection of Business Information Technology Students

Number of pages and appendix pages
59 + 13

Educational data mining (EDM) is an emerging field of research that puts into effective use advanced machine learning concepts in analysing numerous data from educational systems to improve the methods and tools in learning and teaching in educational institutions.

In line with the Finnish government's Vision 2030 for higher education which hopes to develop more pre-emptive and anticipation-led digital learning services, this thesis aims to explore the use of machine learning techniques to find accurate prediction and classification models of the two most common study path selection of students in the Degree Programme in Business IT of Haaga-Helia UAS using features such as student related characteristics and affinity to Software Development (SWD) and Digital Service Design (DSD), motivational goal, mastery orientation, and other demographic factors.

This quantitative research utilized questionnaire data gathered from 101 students in the BITe programme and followed the CRISP-DM framework as a guide to move forward to the various phases of the research. KNIME Analytics Platform, an open-source data mining tool, was used to pre-process, prepare, analyse and model the data.

Exploratory data analysis was undertaken to discover initial insights about the data and to trim the chosen factors. Bootstrap method was used as a sample-growing technique and data were partitioned into training (80%) and testing (20%) subsets. Three machine learning algorithms were used to model the data, and performance scores (accuracy, Cohen's kappa and ROC curve) were set as criteria to evaluate the models.

Results of the study revealed that the research was successful in establishing a predictive model using logistic regression. Using the significant predictors DSD & SWD factors, mastery intrinsic orientation, motivational goal, gender, geographical area and age, the model was able to predict study path selection with 85.5% accuracy. The validation test done on the model even achieved a higher accuracy score of 94%.

The research was also able to forward two highly accurate Random Forest (94% accuracy) and Decision Tree (93% accuracy) classification models. Due to only very slight differences between the performance measures of these models, both are recommended to be used for student classification. The Random Forest would result in a slightly higher accuracy rate while the Decision Tree model would be easier to interpret by extracting a classification rule from its tree view.

Model deployment was simulated in KNIME and the final models were exported in PMML format, thus opening the possibility for the models to be used in future researches or for the deployment in a study path recommender application for incoming students.

Keywords
Machine Learning, Educational Data Mining, Logistic Regression, Decision Trees, Random Forest, CRISP-DM, KNIME

Table of contents

List of Figures

List of Tables

Acknowledgements

My sincerest gratitude to my thesis advisor Amir Dirin for all the input, ideas, support and guidance he gave me during the whole research process, without which this thesis would not be completed. I very much humbly appreciate that he sets a high expectation from my work and that he pushes me to give the best output possible.

To Kari Silpiö and Kasper Valtakari for giving me a portion of their class times so I could gather the data from their students. To my academic advisor Riitta Blomster for always being there ready to help, for being accommodating whenever I need to talk, and for proofreading this report. To all the students, tutors, and teachers in BITe for the academic and social support all through my stay in Haaga-Helia. And of course, special mention to all the students who participated in the survey.

To Giang for sending me the link to the University of Oulu study choice test which sparked the inspiration for my questionnaire. To Carissa and Regina for their valuable comments in the questionnaire pre-testing. To Dominique and Hang for the constant keeping in touch and for the shared channel to vent when the pressure from our theses needs to be released.

To all of my friends both here and elsewhere for the encouragement and cheer-me-ups, virtual or otherwise.

To my family in the Philippines for the continuous source of inspiration and motivation. To baby Jrue for the joy that your videos bring täti. And most especially to Jaakko for the patience, love and support, and for giving me my space throughout my thesis-writing-bubble.

A warm and heartfelt thank you to all of you!

Charlese Saballe
Helsinki, May 2019

Terms and Abbreviations

BITe	Business Information Technology
Bootstrapping	A sample size growing approach by doing many random resampling with replacement from the original samples
CRISP-DM	Cross Industry Standard Process for Data Mining
DSD	Digital Service Design
EDM	Educational Data Mining
IEDM	International Educational Data Mining Society
JEDM	Journal of Educational Data Mining
KNIME	An open-source platform with a graphical user interface for data mining and machine learning
MEO	Mastery extrinsic orientation
MIO	Mastery intrinsic orientation
ML	Machine Learning
MG	Motivational goal
PMML	Predictive Model Markup Language
UAS	University of Applied Sciences
ROC	Receiver Operating Characteristic, a measure of accuracy of the model with the area under the curve equal to 1 is excellent and 0.5 is inaccurate
SWD	Software Development
Stratified sampling	A sampling technique where the population is divided into exclusive groups called strata and a sample of the population is collected within each stratum.
Z-score normalization	Re-scaling the data so that they follow a standard normal distribution with mean of 0 and standard deviation of 1.

1 Introduction

Machine Learning (ML) has lately been a hot topic and it comes to no surprise because ML is being utilized in various systems from a wide range of use cases and industries: from transportation to finance, from business to governments, from medicine to fashion retail, the list goes on. Most of these applications find hidden patterns in data in order to uncover meanings and relationships between factors or make predictions on future values of the data. (Reddi 2017, 1; Song 2018, 1; Ding 2018, 1-2.)

Although the analyses of patterns and the algorithms used in ML have been around for years with earlier concepts dating back 6th century BC in China, recent advances in data collection methods and increased computing machinery have fuelled the development of such technology and propelled these methods to be considered megatrends. (Park 2015, 46; Nisbet, Miner & Elder 2009, 5; Witten & Frank 2005.)

An area of interest that is emerging is in the education sector. Educational Data Mining (EDM) is a field of research that uses statistics, machine learning and data mining to analyse different types of data from various educational systems with the aim to further the methods and tools used to learn, teach and research in the field of education (Romero & Ventura 2010, 601; ElAtia, Ipperciel, & Zaiane 2016, xxiii).

The Finnish Ministry of Education launched its Vision 2030 for higher education where the government aims to increase the graduation rate of its citizen and have at least fifty percent of young adults aged 25 to 34 years old completing a higher education degree. The Ministry also hopes to develop more digital learning services that is pre-emptive and anticipation-led, has a student-oriented approach and is based on individual learning need (Ministry of Education and Culture 2019). Educational Data Mining, using Machine Learning in particular, could play a vital role in the Ministry's digitalisation efforts in reforming higher education and diversifying the Finnish learning environments.

The target group of this thesis is students in the Degree Programme in Business Information Technology (BITe) at Haaga-Helia University of Applied Science. The topic is connected to the researcher's previous projects with Dr. Amir Dirin under the theme of future learning research. It is also heavily linked to the author's background in Statistics and Data Science, as well as an interest in the promotion of student interests kindled by her work in Haaga-Helia student union and organizations.

1.1 Research Objectives and Research Questions

The overarching objective of the research is to contribute to the improvement of education by providing data-driven insights on student profiles. The specific goal of the thesis is to explore machine learning techniques and apply these to predict the study path selection of Business Information Technology students at Haaga-Helia University of Applied Sciences (UAS). The study also aims to find out if the students' mastery intrinsic-extrinsic orientation, motivation and demographic attributes could be used as features to model student classification.

To be more precise, the questions the research wishes to answer are:

1.) Are we able to find a suitable data model that would accurately predict if students' study path selection is Software Development (SWD) or Digital Service Design (DSD), based on their answers in a questionnaire?

2.) Can we find a model that could classify or cluster the students using features such as study paths, mastery orientation, motivation and demographic attributes such as age, country and gender?

1.2 Significance of the Study

Long, Ferrier & Heagney (2006, 170) identified that improving the match of students with their university courses could substantially help in decreasing discontinuation among younger students which would also open up spots for other potential applicants and decrease the inefficient use of public funds for higher education.

Finding a prediction model of students' specialization path could be useful in developing a system that would recommend study paths to potential students who are often unsure of which courses to specialize in. Dirin (2018) also stressed that knowing these factors that might affect student performance would help with lessening their confusion and anxiety and could lead to better course completion and graduation rates in the future.

Having an awareness of the estimated number of incoming students per specialization paths could also be helpful to BITe's programme administration. With this insight, they would be able to allocate teaching resources per specialization and plan course schedules better.

Knowing the orientation and motivation of the students could help the programme's faculty in designing course syllabus that would be best suitable to the learning needs of their students under their profiles and might help with the improvement efforts of the degree programme's teaching pedagogy.

1.3 Scope of the Thesis

The thesis is a research-oriented study that focuses on using machine learning methods to determine (i.) a prediction model of the two most common study paths of students in the Degree Programme in Business Information Technology: Software Development and Digital Service Design, and (ii.) a classification or clustering model with features such as motivational goal, mastery orientation, and other demographic factors.

The emphasis of this research is on designing the pipeline of the data science processes and extracting appropriate models from the data gathered. It does not cover the actual deployment of the data model in another system nor does it cover designing and developing an actual application that would give out study path recommendations to students.

Furthermore, this study does not dive deep in the mathematical theories of machine learning. Rather, it puts more emphasis on the practical and empirical application of its algorithms.

1.4 Thesis Structure

The thesis is presented as follows: Chapter 2 frames this study in the map of education research based on focus, method and topic of interest. It also presents the motivation theory in learning, and introduces the Degree Programme in BITe. It then summarizes the body of research done in Educational Data Mining.

Chapter 3 examines CRISP-DM as a data science methodology, describes some common concepts in data mining and machine learning, and looks at KNIME as a data mining tool. In Chapter 4, the methodology, concepts and tools discussed previously are assessed for their suitability in the research and the various phases and methods undertaken are presented. Chapter 5 shows the research results and Chapter 6 discusses the results uncovered. Finally, Chapter 7 focuses on the conclusions and recommendations for future research.

2 Conceptual Framework on Educational Data Mining Research

This chapter provides the context of the study in relation to education research. An overview of the field is discussed by looking at studies in Educational Data Mining. It presents the common methods, topics and focus of researches in an educational context. Next, a discussion on motivation theory is mentioned. Following that, details are presented about the Degree Programme in Business Information Technology whose students are the subject of this study. Lastly, a number of related studies undertaken in this area is also reviewed.

2.1 Educational data mining

The International Educational Data Mining Society (IEDM 2019) defines EDM as a flourishing field of study that aims to improve the methods of investigating large and various amount of data from educational systems to better understand learners and to provide the context in which students learn.

EDM seeks to obtain valuable insights for specific stakeholders (*focus*) through the analysis of datasets using data science techniques (*methods*) and use the results in order to improve the different areas of the learning process (*topics of interest*). Figure 1 illustrates the application of data mining to educational systems.

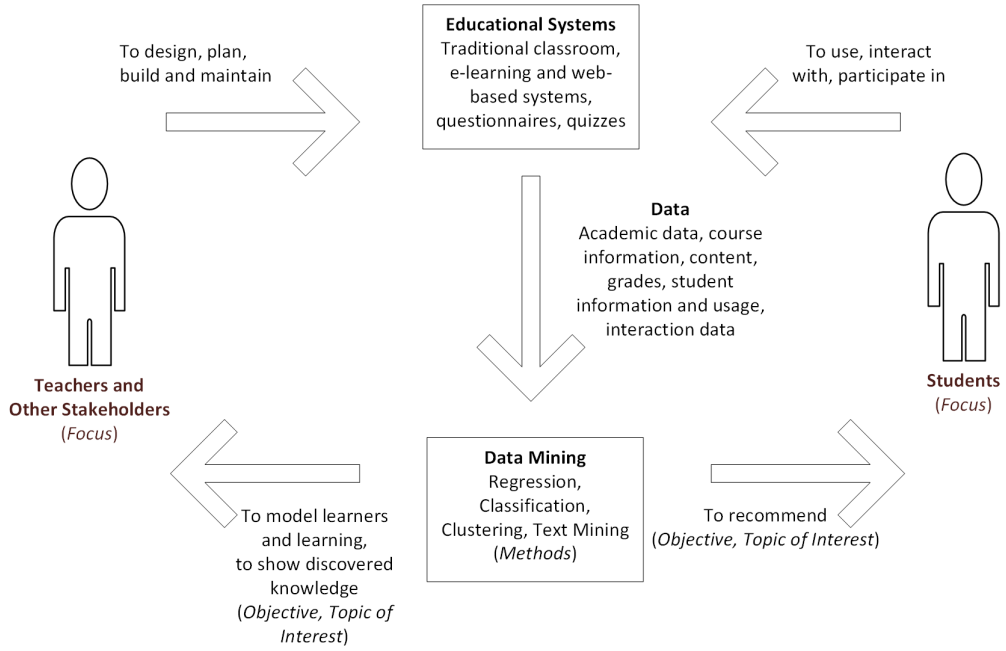


Figure 1. Focus, Methods and Topic of Interest in EDM (Bansal, Mishra & Singh 2017)

2.1.1 Focus

One area of EDM research can be specified towards the various actors in educational institutions. A question that could be asked is which among these stakeholders could benefit from the innovations made through EDM. Two obvious focus could be directed towards students and teachers.

Other actors that could also be considered are parents, tutors, course/programme administrators, academic researchers and system designers. The focus of educational data mining research could emphasize on increasing the variety of stakeholders who would benefit from the research and on discovering the channels in which information is received. (Chatti, Dyckhoff, Schroeder & Thus 2012, 8; Romero, Ventura, Pechnizkiy & Baker 2010, 3.) Stakeholders have different perspectives, goals and expectations, which is important to know when designing innovations for learning.

2.1.2 Methods

In order to discover the hidden underlying patterns in educational data, researchers employ a number of techniques. Papamitsiou & Economides (2014, 53) reviewed several case studies related to EDM and stated that the most popular data mining methods are classification, clustering, regression (logistic/multiple) and discovery with models. Stegmann (2016, 21) referenced this study to give an overview of the different key study methods used in EDM as presented in Table 1.

Table 1. Overview of Methods Often Used in EDM

Method	Description
Regression	Predict using regression; analyse which attributes correlate
Classification	Classify students
Clustering	Cluster students (non-supervised methods)
Text Mining	Perform text mining on assignments, posts, feedback
Association Rule Mining	Extract meaningful associations to aid teachers
Social Network Analysis	Analyse the network in Virtual Learning Environments with a social component
Discovery with models	Formulate models to aid either prediction or understanding
Visualisation	Provide visualisations to students or teachers
Statistics	Provide statistics to students or teachers

2.1.3 Topic of Interest

Researches in this area can also be classified according to their objectives. Papamitsiou & Economides (2014) also reveal that the majority of studies investigate issues related to student/student behaviour modeling, prediction of performance, increase reflection and awareness and prediction of dropout and retention. Table 2 lists the common research objectives in education research.

Table 2. Classification of EDM Research Based on Objectives

Objective	Description
Student/Student Behaviour Modelling	Detection, identification and modelling the behaviours of students
Prediction of performance	Predict indicators of performance such as grades, engagement in activities, enrolment, and student's moods.
Increase self-reflection and self-awareness	Provide information to the students to aid their reflection and awareness.
Prediction of drop out and retention	Predict which students will drop-out of the course.
Improve assessment and feedback services	Analyse the data to improve the test results
Recommendation of resources	Recommend new courses or additional explanations to students to aid learning.

Another dimension in EDM studies pertains to topic of interest. According to Romero & Ventura (2010, in Stegmann 2016), the most common topics in EDM research deal with developing tools, frameworks and methods; mining data from educational systems and research; process-related mining, data-driven adaptation and personalization of educational environment and systems; and the improvement of educational software as itemized in Table 3. A complete list of EDM topics of interest can be viewed in Appendix 1.

Table 3. Common Topics of Interest Used in EDM

Topic	Description
Generic frameworks and methods	To develop tool, frameworks, methods, algorithms, approaches, and so forth, specifically oriented to educational data mining research.
Mining educational data	Mining assessment data, mining browsing or interaction data, mining the results of educational research.
Educational process mining	To extract process-related knowledge from event logs recorded by educational systems.
Data-driven adaptation and personalization	To apply data mining methods for improving adaptation and personalization in educational environments and systems.
Improving educational software	Many large educational data sets are generated by computer software. Can we use our discoveries to improve the software's effectiveness?

2.2 Motivation Theory

Motivation is another conceptual framework that the study aims to include as part of the features in the classifying or clustering of the students. Winne & Baker (2013) stated that motivation theories attempt to explain the reasons why people behave in a certain way and why they continue or modify these behaviours.

Researchers assert that there are two types of motivation: intrinsic and extrinsic. Ryan & Desi (2000) offered definitions for both: “Intrinsic motivation is defined as the doing of an activity for its inherent satisfactions rather than for some separable consequence. Extrinsic motivation is a construct that pertains whenever an activity is done in order to attain some separable outcome”.

In other words, intrinsically motivated persons engage in an action based on personal enjoyment of the action itself and are moved to act for fun or for the challenge of it. Extrinsically motivated persons, on the other hand, perform an action because of tangible rewards such as grades, money, prize or recognition. (Smith 2011, 6-7; Ryan & Desi 2000, 56-60.)

Previous theses made by students from the BITe programme, for instance Make’s (2018) thesis, have cited Niemivirta’s (2002) study on motivation and goal orientation in relation to student performance and used his questionnaire to analyse university students’ performance.

Tuominen-Soini (2012) also referred to Niemivirta’s (2002) study in her research about motivation and achievement goal profiles of students. The research defined mastery orientation as the goal of obtaining new knowledge and coined it with the intrinsic and extrinsic motivation mentioned above to operationalize a conceptual definition of mastery-intrinsic orientation (goal of obtaining mastery based on inherent personal satisfaction of understanding) and mastery-extrinsic orientation (goal of gaining knowledge grounded on external rewards).

2.3 The Degree Programme in Business IT at Haaga-Helia UAS

The programme in Business Information Technology is a 210 ECTS credit degree offered in English at Haaga-Helia University of Applied Science in Pasila, Helsinki. It provides students with practical skills and theoretical knowledge about business and information technology (IT). It offers four study specialization paths: Software Development (SWD), Digital

Services (DSD), ICT Infrastructures (Infra) and Business and ICT (BICT). The programme takes in an average of 45 new students per semester. (Haaga-Helia 2019a.)

Software Development students are equipped with programming and software engineering courses that prepare them well for implementing software projects. Digital Service students are provided with user experience and design courses to ensure that they are able to design digital solutions that are user-centered. ICT Infrastructure students are trained in cloud and data security. Students from the Business ICT path are prepared to be experts in business processes and systems (e.g. business intelligence and customer relationship management). (Haaga-Helia 2019b)

The data about student demographics, graduation and dropout rates that are specific to the degree programme are unfortunately not publicly available. However, some data from a public statistics website were collected to show some estimate figures on number of students, gender breakdown and dropout rate.

Extrapolated data from Vipunen (2018), the Finnish education administration reporting portal, show that there are more than 362 students in the programme in 2018. Of these, about 67% are males and 33% are females. About 22% of the students are 22 years old and below, 45% are 23-28 years old, 21% are 29-34 years old and 11% are 35 years old and above. The latest data from Official Statistics Finland (2019) reveal that discontinued studies in the ICT field account to 12.8% in Universities of Applied Sciences for the academic year 2016-2017, the second highest dropout rate in any field of studies in Finland next to Social Sciences.

2.4 Review of Research Work in EDM

The past decade saw an expansive body of research done in relation to Educational Data Mining which means that it is a research area that is maturing. As a result, The International Conference on Educational Data Mining has been held annually since 2008. The Journal of Educational Data Mining has also been published since 2009 and is already on its tenth volume. (Romero, et al. 2010, 1; IEDM 2019; JEDM 2019.) This section presents some of those studies undertaken in the field.

Stegmann (2016) from Aalto University performed machine learning and data mining analytics to data collected from an online learning platform to predict the grades of the students using the platform. He compared several types of data transformations in the prediction model of the students' behaviour.

A framework of collaboration between human and machine to solve the problem in teaching programming in architectural design was proposed by Park (2015) from the Massachusetts Institute of Technology. He developed a tutoring system using machine learning and computer vision to improve the learning performance of students by behaviour modeling and personalization of the system. Rebolledo-Mendez, Boulay, Luckin & Benitez-Guerrero (2013) also proposed a motivationally-aware tutoring system that detects and reacts to the learner's motivational feedback.

Herold (2013) explored a novel approach in EDM data collection by digitizing students' handwritten coursework by using pens that can write on ordinary paper but can also turn pen strokes into digitized copy. He then performed various data mining and machine learning techniques in order to discover ways the students learn and uncover the most important features to the students' success in the course. Meanwhile, Spoon & al. (2016) used a Random Forest model to evaluate factors for student success and developed a criterion that identifies which students are encouraged to take intervention initiatives like joining a supplemental class.

Sarra, Fontanella & Zio (2018) gathered data from an online questionnaire and used the Bayesian Profile Regression method to identify students who are at risk of dropping out of the university. Using this framework, they were able to profile a student group that is more likely to face academic failure. A similar study was conducted by Hamedi & Dirin (2018) where they used Bayesian Network model to analyse BITe students' motivation factors to anticipate the influences that lead to academic dropouts.

The Decision Tree algorithm was used by Kai, Almeda, Baker, Heffernan & Heffernan (2018) to model students productive and unproductive persistence in learning. By identifying students who persist unproductively, the research provided insight as to when students are struggling and how to make their grit yield fruitful results.

To summarize, examining relevant researches indicate that data from various sources, even those gathered from questionnaires, can be used in EDM. Additionally, various data mining methods and machine learning algorithms such as regression analysis, decision trees, random forest and other classification techniques can be used to predict, classify and cluster students and student behaviour.

3 Data Mining Framework, Methods and Tools

This chapter discusses one data science methodology called Cross Industry Standard Process for Data Mining (CRISP-DM) and describes the data mining pipeline in each of the research phases. It would detail the basic concept of the common methods or algorithms usually used in EDM as mentioned in Chapter 2. It would also look at a data mining tool called KNIME to evaluate and map out its fit for the framework that will be set.

3.1 CRISP-DM Framework

Researches such as that of Palacios (2017), Nadali, Kakhky & Nosratabadi (2011) and Azevedo & Santos (2008) have evaluated and favoured CRISP-DM against other frameworks such as SEMMA (Sample, Explore, Modify, Model and Assess) and KDD (Knowledge Discovery in Databases).

Cross Industry Standard Process for Data Mining is a standard process model that describes an overview of the distinct phases, tasks and output needed in the implementation of data mining initiatives. The methodology structures the life cycle of a project in six stages, as shown in Figure 2. The process is flexible, and the sequence of the phases are not rigid with the arrows only signifying the most common and crucial dependencies between them. (Clark 2018; Palacios 2017; Shearer 2000.)

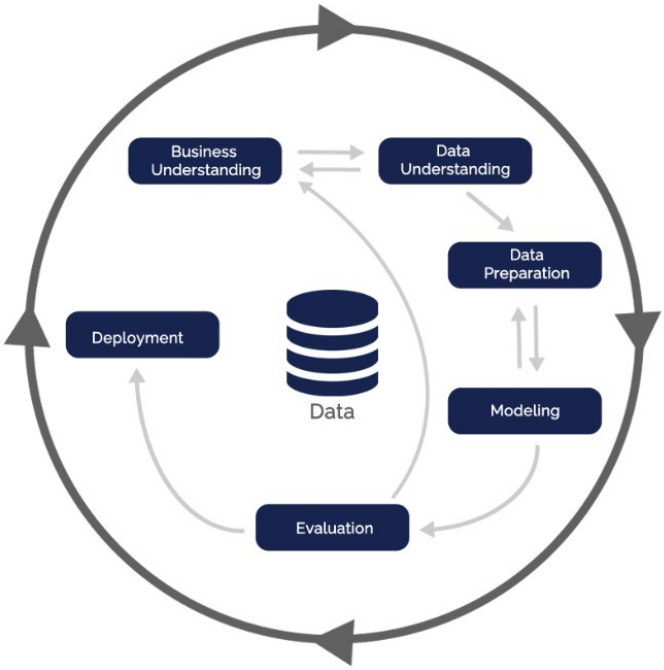


Figure 2. Phases of CRISP-DM Methodology (Otaris 2018)

These stages have been comprehensively detailed by Wirth & Hipp (2000) and Moreira, Carvalho & Horváth (2018) and a summary is described below:

1. *Business Understanding*. The initial phase aims at understanding the business area, defining the goal from a business perspective and transforming the goal into a data mining problem. The tasks for this stage include determining business objectives, assessing the situation, determining data mining goals and producing the project plan.

2. *Data Understanding*. This phase begins with collecting the needed data and inspecting the data to gain initial insights and identifying data issues such as missing data, outliers, errors and the likes. Tasks include collecting initial data, describing data, exploring data and verifying data quality.

3. *Data Preparation*. This phase comprises different data pre-processing operations to convert raw data into a suitable form ready for modeling in the next phase. The tasks included here are selecting data, cleaning data, constructing data, integrating data and formatting data.

4. *Modeling*. In this phase, appropriate modeling techniques and algorithms are selected and performed on the data. Parameters are also calibrated to optimal values. The tasks in this stage include selecting modeling technique, generating test design, building model and assessing models from a data analytics view.

5. *Evaluation*. This phase entails reviewing and evaluating the models to ensure that they are useful and meaningful from a business perspective. The decision on whether to deploy the results or not is based on the evaluation results. The tasks comprise of evaluating results, reviewing the process and determining next steps.

6. *Deployment*. The aim of this phase is to integrate the outcome of the model into the business process, whether in the tool or in a report or elsewhere. This phase includes tasks such as planning the deployment, planning the monitoring and maintenance, producing the final report and reviewing project.

Zhu (2017) argued that CRISP-DM is susceptible to the frequent changes in customer requirement, however, the framework is structured, organized, and extremely documented. It is also highly utilized by industry players, as evidenced by a survey conducted by KDnuggets (2014), a leading data mining website. Based on the results of the poll, 43% of

the respondents said that they used the framework for their data mining projects. Moreover, researchers like Oreski, Pihir & Konecki (2017) and Almahadeen, Akkaya & Sari (2017) have successfully utilized the framework in educational data mining in the area of prediction and classification.

3.2 Data Mining Methods

Nisbet & al. (2009) offered one definition of data mining as utilizing machine learning algorithms to find structural and meaningful patterns between data in a dataset. Data mining hopes to find solutions to problems by taking actionable steps to gain some sort of benefit or improvement.

Machine learning, on the other hand, can be broadly defined as data-driven methods that use available data - usually in the form of electronic information - and analyse these data to come up with accurate predictions that improve over time with additional knowledge and experience. It combines concepts in computer science, statistics and optimization as learning techniques. (Mohri, Rostamizadeh & Talwalkar 2012.)

Four machine learning approaches i.e. supervised learning, semi-supervised learning, reinforcement learning and unsupervised learning are described in Table 4. Bhilegaonkar (2016) asserts that the most common ML use cases are somewhat forms of supervised learning projects which have labelled inputs and outputs in the training data to come up with a model and then apply this model to classify new instances of the data.

Table 4. Common Machine Learning Approaches

Approach	Description
Supervised Learning	Inferring useful concepts or structure from data whose outcomes of interest are known to the learner (labelled data) in the form of training data set, which contains both observations and outcome.
Semi-supervised learning	Inferring useful structure or concepts from data whose outcomes of interest are known to the learner for partial observations. Could be thought of as a sub-category of supervised learning where a small fraction of training data is labelled and the remaining data is unlabelled.
Reinforcement learning	Correct observations, outcome pairs or laballed data pairs are not presented to the learner. However, based on learner's current state and action, the learning context changes its state and provides reward signal as feedback. Learner tries to maximize reward.
Unsupervised learning	Inferring useful structure or concepts from observations whose outcomes of interest are not known to the learner (unlabelled data). Since outcomes are not known, there is no error as in supervised learning, or a reward as in reinforcement learning.

3.2.1 Prediction and Classification Modeling

Prediction and classification modeling are types of supervised learning approach, and they are among the most common techniques used in educational data mining. In prediction, the aim is to construct a model that could infer an outcome or target variable from a mix of some other features of the data. In classification, responses can be grouped according to certain rule-based categories. The most commonly used predictive and classification modeling techniques are liner regression, logistic regression, decision trees and random forests. (Baker & Inventado 2014; Leventhal 2010.)

Linear regression is modeling the relationship of the predicted variable (the response) and the explanatory factors (a set of features) and is used when the outcome is numeric. The idea is to express the response as a linear equation of the features of pre-determined weights. These weights are calculated from training data. (Witten & Frank 2005.)

Logistic regression is a nonlinear regression technique that is used when the predicted value is binary. This algorithm models the probability that the response will occur based on the values of the features. This allows for the response values to represent belongingness to a class. (Roiger 2017.)

Decision tree is a classification technique that categorizes the data and represents the results in a tree-like structure. It starts with a variable called the root node which is then split into two or many branches that represent separate classes or ranges of the node. Recursive partitioning is continued where each branch is further split using some type of measure until a stopping rule is satisfied. (Nisbet & al. 2009, 241.)

Random forest is an algorithm proposed by Breiman in 2001. It is a tree-growing method that sets up tree predictors using bootstrapped data with the same distribution for all the trees in the forest. First, a random subset is sampled and further samples are done with replacement. Next, a subset of the variables is chosen, and the best split is determined from this subset. A testing data set is created for a third of the cases and the average error rate is calculated from all trees built. The predicted classes are counted based on some measures of variable importance and the average effect is calculated from all the trees resulting to a variable importance value. (Breiman 2001.)

3.2.2 Structure Discovery

Structure discovery is an unsupervised learning technique where data formations are found without any prior idea on what the outcome would be. The most common structure discovery algorithms in educational data mining include clustering and factor analysis. (Baker & Inventado 2014; Leventhal 2010.)

Cluster analysis is an algorithm that aims to find data instances that naturally group together and the output is shown as a graph that displays these clusters. If a set of clusters is optimal, each data point in a cluster will generally be more similar to the other data points in that cluster than the data points in other clusters. (Witten & Frank 2005.)

Factor analysis is an algorithm that discovers variables that group together inherently, as opposed to data points in clustering. It splits the data into a collection of features that are formed by some set of hidden factors. In EDM, factor analysis is used to reduce the number of variables. (Baker & Inventado 2014.)

3.3 KNIME Analytics Tool

KNIME Analytics Platform, a Java- and Eclipse-based open-sourced environment, is a scalable, powerful and modular software development platform that can perform a range of data mining tasks such as data loading, transformation, analysis, modeling and visualization. The users of the platform come from diverse fields of enterprises in life sciences, governments, research, finance, and retail, among others. (KNIME 2019.)

Muenchen (2019) reports that KNIME is one of the fastest growing packages in terms of scholarly articles using the software which means that academic researchers are beginning to use the platform in their studies. KNIME has also been lauded by IT advisory firms Gartner and Forrester as an industry leader for Machine Learning Platforms due to KNIME's strong presence and significant mind share in the market. (Muenchen 2019; KNIME 2019.)

There are several advantages in using KNIME as a platform of choice. It is free and it features an easy-to-use graphical user interface which allows users to build workflow-based modules by dragging and dropping a number of pre-built algorithms. This means that users do not have to write a single line of code to perform powerful data mining and machine learning techniques. KNIME also allows for some flexibility by integrating Python and R

into the workflow. For some ML models, KNIME lets users easily view their fully trained models. (Amarillo 2018.)

Rangra & Bansal (2014) compared six data mining tools and concluded that KNIME is recommended for both novice and expert users because of its very robust built-in functionalities as well as third-party integrations.

4 Research Methodology

This educational data mining study would defer to the body of research that highlights the suitability of CRISP-DM as a methodology for data mining projects and will utilize the process as basis for the thesis' framework. However, because the study is more of exploratory and research-focused, and the scope does not cover designing or developing an actual application, minor modifications will be applied.

The methodology will only use one cycle of the original CRISP-DM framework, *business understanding* phase will be framed as *research understanding* and deployment will only mean mostly reporting the outcome in this paper and simulating the model inside the tool of choice. Figure 3 depicts the modified framework to be used in the research and this chapter would detail the methods undertaken in each phase.

Figure 3. Modified CRISP-DM Framework to be Used in the Thesis

4.1 Research understanding

Initial topics of interest were suggested during the first meeting between the researcher and the thesis supervisor. It was agreed that the topic would be in relation to machine learning in education based on their previous collaboration in future learning research. From the discussions, the idea of recommending specialization paths to incoming BITe students was proposed. A preliminary review of related studies was conducted by the thesis author.

Subsequent meetings resulted in setting the research problem, objectives and assumptions as well as the data mining goals. The decision of using machine learning techniques to predict the study path selection of BITe students as the final topic for the thesis was thus made. It was also agreed to include motivation goals and mastery orientation be added as a set of possible features, supplementing other demographic variables such as age, gender and geographical area. Motivation theory from previous research proves its suitability in modeling the student behaviour/characteristic.

The underlying hypothesis is that the following factors would be good predictors and classifiers of study path selection: statements on affinity to Software Development or Digital

Service Design paths, motivation goals, mastery extrinsic orientation, mastery intrinsic orientation, age, gender, and geographic origin.

Soon after that, the project proposal and the project plan were written. Based on the conceptual framework described in Chapter 2, the study focus, objectives and methods were mapped against the body of other EDM researches and was classified accordingly, as Table 5 below reflects.

Table 5. Research Classification of the Thesis

	Dimension
Focus	Students
Objective and Topics of Interest	Data-driven insights and adaptation by increasing student's (self) awareness and recommendation of study paths and courses; modelling student classifications/clusters.
Target Methods	Prediction, Classifying, Clustering
Data	Questionnaire results

The suitability of KNIME as a tool was also examined. KNIME allows building flexible workflows and this could be suited to fit the CRISP-DM framework. KNIME will be used as the data mining platform for the research.

4.2 Data understanding

After setting the research objectives, framework, plan and tools, the following phase involved the collection of the necessary data for the research. This subchapter discusses the various tasks undertaken in understanding and collecting data.

The end goal for the use of the model derived from the study is to provide a baseline prediction and study path recommendation to prospective students, in alignment with the Ministry of Education's vision of an anticipation-led, individual learning need-based education. However, since the target is potential students, the data from the current learning systems at Haaga-Helia would not be relevant for this particular goal and is not used for this research. Instead, the features to be modelled should come from the data that could be gathered from incoming students.

Another challenge is determining what type of features can be considered relevant to the prediction and classification of these students. Most of the research done in EDM, as discussed in the review of related literature covers modeling behaviour, predicting study success and dropout, and increasing self-awareness of current, not potential, students.

The first problem can be solved by collecting data through a questionnaire. In theory, this could be implemented as a pre-registration survey to incoming students or could be integrated as part of the orientation week enrolment process.

The second challenge can be answered by employing similar concepts as the study choice tests that universities provide in their websites to help potential applicants determine relevant degree programmes.

4.2.1 Questionnaire Design

To obtain a proof-of-concept for this proposed solution, a survey instrument was designed to obtain the quantitative data needed in the research. The questionnaire was loosely based on the study choice test for master's degree programme by the University of Oulu (2018) and the mastery-intrinsic/extrinsic orientation questions from the Niemivirta (2002) study.

Initial Questionnaire

The first version of the questionnaire had two parts. The first part dealt with demographic questions about age, gender, geographic origin, semester level and specialization path. The second part initially had 26 questions which were tailored by the researcher and the thesis supervisor to reflect students' affinity to either Software Development or Digital Service Design paths as well as the motivational goal and mastery orientations based from the motivation theories explained in Chapter 2.2.

Software Development (SWD)

- I want to work with numbers.
- I'm interested in technology.
- I solve problems specifically with the end goal in mind.
- I enjoy working in an environment where there is always something new going on.
- I would want to manage people and things in my work.
- I always keep myself with up-to-date information on new technological innovations.
- I'd like to invent and develop new devices and applications.
- I dream about founding my own business.
- I want to sell things to people.

Digital Service Design (DSD)

- I like to learn new languages.
- I enjoy working together with others in a team.
- I want to understand how people use technology.
- It's important that I can be creative in my work.
- I am interested in designing archetypes/prototypes.
- I want to work with my hands.
- I would rather work with people than to work with machines.
- I'm interested in teaching and guiding others.
- I enjoy coming up with new solutions to problems.

Motivational Goal (MG)

- Career development and promotions are important for me.
- Salary means a lot to me.

Mastery Extrinsic Orientation (MEO)

- It is important for me that I get good grades.
- An important goal for me is to do well in my studies.
- My goal is to succeed in school.

Mastery Intrinsic Orientation (MIO)

- I study in order to learn new things.
- An important goal for me is to learn as much as possible.
- To acquire new knowledge is an important goal for me in school.

The questions were to be rated using a 7-point likert-type scale:

□1- Strongly disagree □ 2-Disagree □ 3-Slightly Disagree □ 4-Neutral □ 5-Slightly Agree □ 6-Agree □ 7-Strongly Agree

Pre-testing and Revised Questionnaire

The questionnaire was pre-tested to three people. A statistician checked for overall quality of the survey instrument, a market researcher checked the questionnaire structure and the clarity of the language, and a student checked for the ease of understanding of the statements.

Comments from the pre-testing phase were considered and the questionnaire was revised. Unclear statements were reworded to make sure that the participants would understand them. The number of items were decreased and only the most relevant statements were retained so as not to overwhelm the survey participants. The statement "I like to learn new languages" was reassigned to SWD. The two easy items about the study paths were placed in the beginning to set a feeling of ease in answering, then followed by the

section about rating statements which required more effort. The demographic part was moved to the end in order not to deplete the participants' energies so that they would have more thoughts in rating the statements in the previous section. The item placement was also randomized.

Eventually, the statements were trimmed down from 26 to18 and the survey instrument was approved for use by the thesis supervisor. A full copy of the final questionnaire can be found in Appendix 2.

4.2.2 Data Collection

As soon as the questionnaire was finalized, a permission was sought from the Manager of Research, Development and Innovation Services at Haaga-Helia. Once the approval was received, the data collection phase commenced and ran from February 12 to March 8, 2019.

Participants

The participants were students in the Business Information Technology programme at Haaga-Helia UAS. Convenience sampling method was used in data collection. To maximize the number of data points to be collected, the researcher decided to use a pen-and-paper approach with first year and second year students. The requests to distribute questionnaires were made to the professors of four BITe course: Orientation to Software Engineering, Programming (Java), User Experience, and Digital Service Design. Respondents were given time in class to accomplish the survey. To reach the students on their third year and above, the participants were sampled from the respective intake's WhatsApp groups using an online version of the questionnaire created in Google Forms.

The students were advised that the survey was confidential and the responses cannot be traced back to them. They were also told that the participation to the study was voluntary. A total of 101 participants, about a quarter of the total BITe student population, participated in the study.

Encoding and Importing data to KNIME

Since there was already an online version of the questionnaire created, the Google Form was used to encode the data gathered from the pen-and-paper questionnaires and the survey was linked to a google sheet which was eventually exported in comma-separated

values (csv) format. This allowed for easier importing of the data. The csv file from the previous step was fed into KNIME using csv reader and preliminary inspection of the data was done.

4.3 Data Preparation

Steps were made to ensure that the data is ready for use in the algorithms in the modeling phase. From the preliminary inspection, it was noted that two rows had missing data, these were completed by using “Most Frequent Value” in missing value handling. Outliers were handled through the Numeric Outlier node in KNIME and were replaced by the “Closest permitted value” strategy.

Average scores for the four factors were computed. Exploratory data analysis was done by inspecting the descriptive statistics, correlation matrix and box plot. Some insights were derived from the data which lead to the trimming of some statements from the list of factors. Data transformations and filtering were performed in relevant analysis. For instance, data were filtered to show only for SWD and DSD study paths to ensure the binary classification of data as required in the definition of the logistic regression algorithm referenced in Chapter 3.2.1.

Standardization using z-score normalization was also implemented. Z-score normalization is a data transformation technique with re-scales the data so that they follow a standard normal distribution with mean of 0 and standard deviation of 1, a requirement for optimizing the logistic regression algorithm. Adeyemo, Wimmer & Powell (2018) conclude that normalization techniques improve the accuracy of the predictions in machine learning algorithms.

4.4 Modeling

To try to answer the research questions, various modeling experiments were set-up and performed. Since the research data were labelled, supervised learning techniques were used.

Dependent variable

The label or dependent variable for this thesis was the students’ study path selections.

It is a non-numeric and categorical variable based on two study paths available for BITe students (either Software Development or Digital Service Design).

Independent variables

The features considered for this study were grounded on drivers used by previous researchers: the first set was attributes on characteristics related to the student profiles (SWD, DSD factors), the next set is about motivational attributes as based on Niemivirta's researches (MG, MEO, MIO factors) and the last was relating to student demographic variables (age, gender, geographic area of origin).

Approaches

For the prediction of students' study paths, linear regression would not be relevant because the target dependent variable (specialization paths) is not numeric. Instead, logistic regression was used since only binary outcome (SWD or DSD) are sought in the response variable. This approach addresses the first research question.

For the classification of students, random forest and decision tree were used to find student categorization based on study path, mastery orientation, motivational goals and demographic attribution (semester level, gender, age and geographic area of origin). This approach handles the second research question.

Stepwise selection using backward elimination for the numerical variables was used as the model building technique for the logistic regression. Menard (2013) noted that stepwise procedures are useful and applicable in purely predictive research where the concern is only with identifying a model that accurately predicts a variable, and exploratory research where the concern is developing prediction models of new use cases. Demographic variables were subsequently added individually and piecewise.

To simulate a larger dataset, bootstrapping method using a stratified sampling technique was also used. Bootstrap is an approach where sample size is grown by doing many random resampling with replacement from the original samples and Finch (2018) concludes that bootstrapping is useful for studies with small sample size and have complex models that are difficult to estimate. Lemm (2012) defines stratified sampling as a technique where the population is divided into exclusive groups called strata and a sample of the population is collected within each stratum which ensures that the distribution of the data

in each sample group is the same as the distribution of the population. For logistic regression modeling, the data was bootstrapped to have n= 2000 to approximate a large sample size. For decision trees and random forests modeling, the data was bootstrapped to have n=500, having enough sample size but also avoiding overfitting of the models.

Data were partitioned into 80% training dataset and 20% testing dataset. This meant that for logistic regression, the training data was n=1600 and the testing data was n=400. For the two classification models, the training data was n=400 and the test data was n=100.

Random seeds were used to ensure that the results of the modeling experiments are reproducible which is important in establishing the trustworthiness of the results. Goodman, Fanelli & Ioannidis (2016) refers to results reproducibility as obtaining the same results by repeating the experiment with procedures closely matching the original study.

4.5 Evaluation

Results from the multiple modeling experiments were evaluated based on three performance measures: accuracy, Cohen's kappa and Area under the Receiver Operating Characteristic (ROC) Curve.

Accuracy is the proportion of the correct number of predictions in the model. Cohen's kappa, as defined by Kampakis (2016), is a very helpful but under-used statistic that measures inter-rating agreement for categorical items and tells how much better the model is compared to random guesses. Landis and Koch (1977), as quoted by Kampakis (2016), provided a way to interpret Cohen's kappa as: less than 0 indicating no agreement, 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement.

Hsieh (2012) describes ROC curve as a two-dimensional plot of probabilities that measure the efficacy of the models. The area under the curve between 0.96-1 is considered as excellent accuracy, 0.90-0.96 as very good, 0.80-0.90 as good, 0.70-0.80 as fair, 0.60-0.70 as poor, and 0.50-0.60 as useless.

4.6 Deployment/Reporting

In the case of this study, deployment mainly meant as reporting of the research results and will constitute the succeeding chapters of the thesis. The model deployment was also

simulated inside KNIME. The final models were exported as Predictive Model Markup Language (PMML), a standard format used in data mining to make it easy to deploy machine learning models to different data mining tools (Guazzelli, 2010).

The PMML models were re-imported in KNIME and new datapoints were fed in the models and the results were verified if the prediction/classification were accurate.

4.7 Workflow Overview

The research initially started with the aim of contributing to the overall improvement of education by providing data-driven insights on students. The topic of interest was then narrowed down and a potential data mining problem was defined, scoped and the objectives were set.

The research followed a modified CRISP-DM methodology and used KNIME as its data mining tool. A research instrument was designed in the form of a questionnaire and data was collected from the target population – students of the Business IT programme in Haaga-Helia UAS. The rest of the research methodology can be illustrated by the workflow in Figure 4 below.

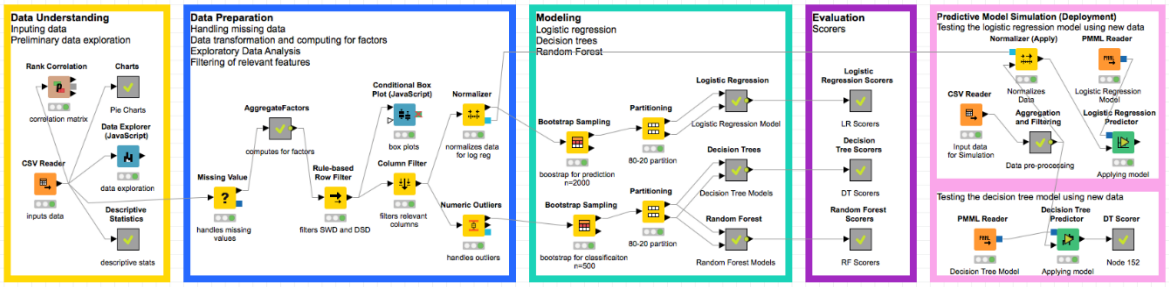


Figure 4. A KNIME Workflow of the Different Phases of the Research

Data from the questionnaire was inputted in KNIME. After that, a preliminary data exploration was done and basic descriptive statistics, correlation, and charts were inspected.

Missing values were handled through the Most Frequent Value method. Outliers were handled through the Closest Permitted Value technique. Factors were aggregated by computing for the averages of the scores for each statements belonging to each of the hypothesized factors. Features were trimmed and irrelevant variables were filtered. Data were accordingly transformed based on corresponding modeling techniques used.

Bootstrap sampling was done as a sample size growing method. The data points were partitioned in training and testing data sets. Three machine learning techniques were used to model the data: logistic regression, decision trees and random forests. Results were evaluated based on the performance measures: accuracy, Cohen’s kappa and ROC curve. Finally, simulation of the deployment of the final prediction and classification models were done inside KNIME.

5 Results

This chapter details the empirical outcome obtained from applying the CRISP-DM methodology to the research initiative. The results are based mainly from the data gathered from the survey instrument.

5.1 Profile of the Respondents

The population of interest is students of Business Information Technology at Haaga-Helia UAS. The basic demographic data about the population were shown in Chapter 2.

Fifty-seven of the 101 participants were male and forty-four were female. More than half of those who answered (51%) were taking Software Development path, 31% were pursuing Digital Service Design, 15% are from Business ICT and 3% are in ICT Infrastructure.

About a third (33%) of those sampled were 1st semester students, 19% were in 2nd semester, 17% were in 3rd semester, 4% in 4th semester, 13% in 5th semester, 9% in 6th semester, 3% in 7th semester and 2% in 8th semester or over.

In terms of age group, 23% were 17 to 22 years old, 44% were 23 to 28 years old, 25% were 29-34 years old and 8% were 35 years old and over.

Around 44% were from Asia and Oceania, 15% were from Finland, 23% were from elsewhere in Europe, 7% were South Americans, 6% were North Americans and 5% were from Africa. Details of the demographic profiles can be seen as graphs in Appendices 3a to 3e.

5.2 Exploratory Data Analysis

Initial data exploration was done in order to check any data quality issues and to uncover first insights from the data. From the preliminary inspection, potential features were dropped from selection and further analysis was done.

Because the data used in the research came from a questionnaire, data quality is assured. Two missing data points were already noted during the encoding stage. The missing values were handled by completion through Most Frequent Value technique.

5.2.1 Discovering Preliminary Insights Using Descriptive Statistics

The descriptive statistics of the numerical variables were checked to see any other irregularities in the data. Table 6 shows the information about the minimum (lowest rating received), maximum (highest rating received), mean (average rating), standard deviation and variance (measure of spread of the ratings), and skewness and kurtosis (measure of the shape of distribution of the ratings).

Table 6. 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
doWellInStudiesMEO	☐	1	7	5.91	1.17	1.36	-1.40	2.91
enjoyWorkAlwaysNewGoingOnSWD	☐	2	7	5.75	1.20	1.43	-1.19	1.19
designingPrototypesDSD	☐	2	7	5.40	1.28	1.64	-0.43	-0.55
acquireKnowledgeMIO	☐	4	7	6.39	0.77	0.60	-0.93	-0.24
workWithHandsDSD	☐	1	7	4.91	1.57	2.46	-0.59	-0.49
succeedInSchoolMEO	☐	1	7	5.54	1.33	1.77	-1.01	0.95
upToDateInTechInnoSWD	☐	2	7	5.58	0.98	0.97	-0.63	0.85
workWithPeopleDSD	☐	1	7	4.84	1.67	2.77	-0.54	-0.33
learnAsMuchMIO	☐	2	7	6.03	1.03	1.07	-1.50	3.43
careerAndPromotionMG	☐	1	7	5.79	1.27	1.61	-1.64	3.60
inventDevelopAppSWD	☐	3	7	5.57	1.07	1.15	-0.32	-0.58
solutionsToProblemsDSD	☐	1	7	6.11	1.01	1.02	-1.77	5.57
salaryMG	☐	1	7	5.54	1.28	1.63	-1.12	2.01

The statement "*To acquire new knowledge is an important goal for me in school*" received the highest mean (6.39) and lowest variation (standard deviation of 0.77) among the factors. This conveys that the participants from the specialization paths gave similar ratings for this statement which suggests that it might not be a good predictor for the study path selection since most students, regardless of study path, highly agree with it.

Similarly, the statement "*I am interested in technology*" had the second highest mean (6.32) and second lowest standard deviation (0.82), which suggest that the participants had high level of agreement with each other's ratings. This in turn means that the statement might not be a good predictor or classifier. The mean ratings of the features per study paths, shown in Table 7, were also inspected.

Table 7. Mean Rating of Features per Study Path

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
solveProblemGoalInMindSWD	5.29	5.69
doWellInStudiesMEO	5.9	5.94
enjoyWorkAlwaysNewGoingOnSWD	5.77	5.63
designingPrototypesDSD	6.29	5.08
acquireKnowledgeMIO	6.35	6.46
workWithHandsDSD	5.45	4.65
succeedInSchoolMEO	5.58	5.62
upToDateInTechInnoSWD	5.68	5.62
workWithPeopleDSD	5.26	4.54
learnAsMuchMIO	5.94	6.1
careerAndPromotionMG	5.84	5.79
inventDevelopAppSWD	5.87	5.65
solutionsToProblemsDSD	6.39	5.98
salaryMG	5.42	5.44

The statement "*I enjoy working in an environment where there is always something new going on*" was meant to be a factor for SWD, however the mean rating was slightly higher for DSD students (5.77) than SWD students (5.63). The same can also be noted for two other factors for SWD: "*I always keep myself with up-to-date information on new technological innovations*" had the mean for DSD (5.77) higher than the mean for SWD (5.63), and "*I would like to invent and develop new devices and applications*" with mean of 5.87 for DSD higher than the mean of 5.65 for SWD. This indicates that the three statements might not be good predictors or classifiers for SWD.

5.2.2 Checking Collinearity Using Correlation Matrix

The correlation matrix (Figure 5) was inspected to check the collinearity of the independent variables. The two statements for the mastery extrinsic orientation (*"An important goal for me is to do well in my studies." & "My goal is to succeed in school."*) were strongly correlated with a measure of 0.69 which 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 redundancy in information.

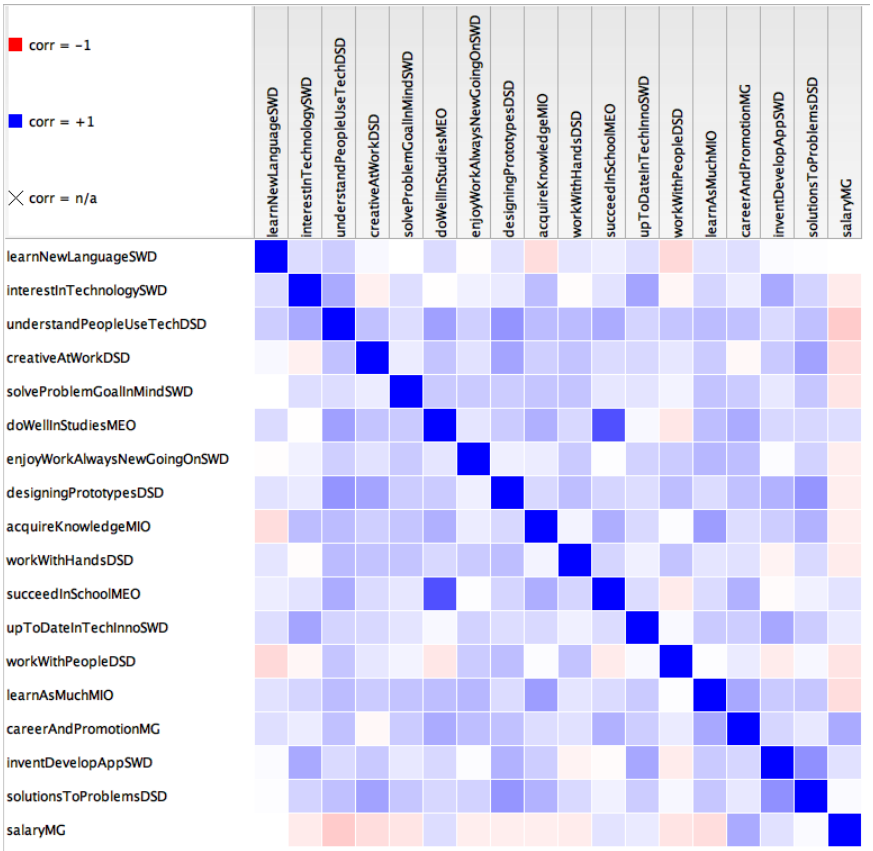


Figure 5. Correlation Matrix Between the Features

All the other variables did not exhibit high correlation with each other and would not pose disturbance in the results of the regression.

5.2.3 Examining Distribution of Variables Using Box plots

The box plots of the five numerical variables (DSD, SWD, MEO, MIO and MG) using untrimmed data were examined to see the distribution of the ratings given by Digital Services Design students versus Software Development students. The corresponding plots are attached as appendices.

Figure 6 shows the comparison of ratings for DSD-related factors by DSD versus SWD students. The plot for Digital Software Design students was relatively short which means that overall they have a higher level of agreement in terms of their ratings for DSD-related questions compared to Software Design students. The median response of DSD students

(6.17) was also higher than the median response of SWD students (5.5) in terms of agreement to DSD-related statements. These suggest that DSD factor is a good viable predictor for study path.

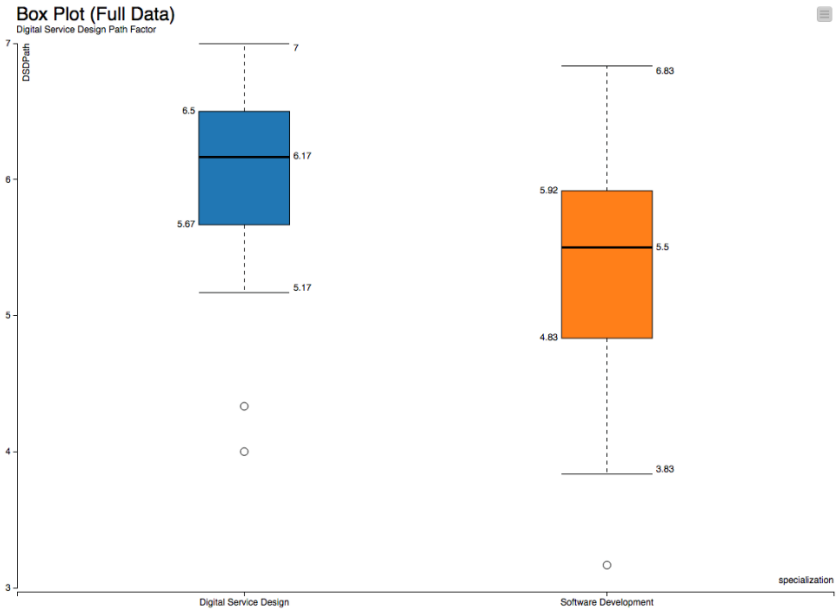


Figure 6. Distribution of Ratings for DSD Factor by Study Path

On the other hand, the ratings for SWD factors by DSD versus SWD students using all the statements (Figure 7) showed almost similar median for both groups. The plots did not exhibit a clear difference in the distribution between the two groups, suggesting that using SWD factor with all the six SWD-statements might not be a good predictor.

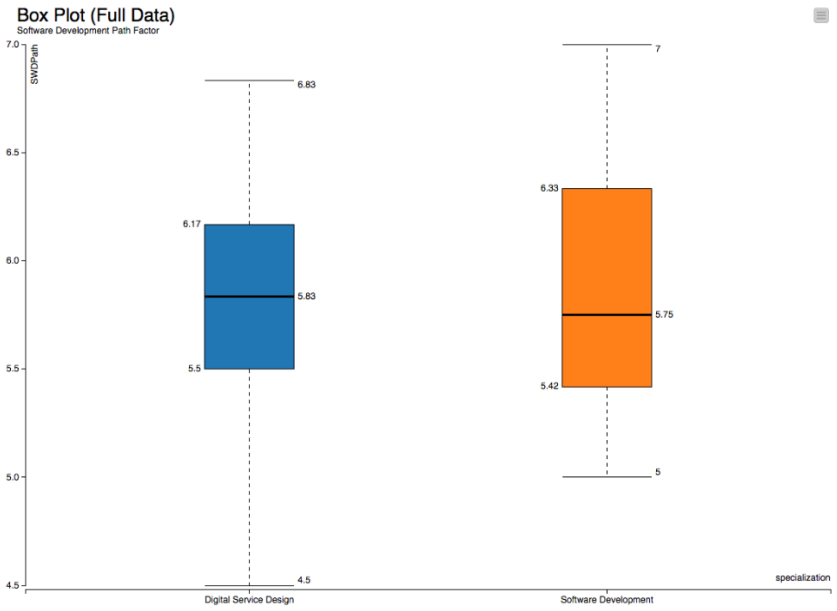


Figure 7. Distribution of Ratings for SWD Factor by Study Path

The boxplots of the ratings using the full sets of features for mastery extrinsic orientation (Appendix 5a) and mastery intrinsic orientation (Appendix 6a) also showed similar distributions between DSD and SWD students. This indicates that the two factors are possibly not good predictors or classifiers of study path selection.

5.2.4 Trimming of Factors Based from the Exploratory Data Analysis

Results from the exploratory data analysis reveal that at least six statements are candidates for dropping from the feature selection of the regression and classification models:

Software Development

- I am interested in technology
- I enjoy working in an environment where there is always something new going on
- I always keep myself with up-to-date information on new technological innovations
- I would like to invent and develop new devices and applications

Mastery Extrinsic Orientation

- My goal is to succeed in school

Mastery Intrinsic Orientation

- To acquire new knowledge is an important goal for me in school

These statements were eventually excluded from the factors. The averages were recomputed and the boxplots were re-inspected to verify the aptness of the trimmed factors.

Appendix 4b confirms that trimming the four statements from the SWD factors resulted in a shorter plot and a higher median was noted for SWD students than DSD students. This means that the adjusted SWD factor is now a viable predictor of study path compared to the full set of SWD statements.

Appendix 5b still shows that the distribution of ratings of MEO for both DSD and SWD study paths are relatively the same. This suggests that there is probably no distinct difference in the mastery extrinsic orientation between the two groups.

On the other hand, Appendix 6b illustrates that trimming the mastery intrinsic orientation factor somewhat improved the viability of MIO as a predictor of study path selection. It can now be noted that SWD students are in favorable agreement with their high ratings of MIO and their median rating is higher than DSD students.

5.3 Prediction Models

In order to predict the students' study path selection of either Digital Service Design (DSD) or Software Development (SWD), logistic regression models were developed from the bootstrapped samples of the training dataset (n=1600). Modeling started with the full numerical features and then trimming the next insignificant variable until all that were left in the model were significant predictors. Afterwards, demographic factors were added and checked for their significance.

5.3.1 Logistic regression models

The regression model of all the numerical factors was run on the training dataset. The coefficient of regression and relevant statistics are shown in the table below.

Table 8. Coefficient of Regression and Statistics – All Numerical Factors Model

Variable	Coeff.	Std. Err.	z-score	P>|z|
motivationGoal	-0.231	0.061	-3.805	0
mastervExtrinsic	-0.102	0.076	-1.332	0.183
mastervIntrinsic	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

Table 8 shows that the P > |z| value (p-value) of the mastery extrinsic orientation was not less than 0.05, suggesting that there is not enough evidence to conclude that MEO is a significant predictor and was dropped from the regression model. Table 9 below presents the coefficients and statistics when the regression model was run with DSD, SWD, MIO and MG factors.

Table 9. Coefficient of Regression and Statistics – LRModel 1

Variable	Coeff.	Std. Err.	z-score	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

Since the result displays that the p-values were all nearly zeros, all four factors – DSD, SWD, MIO and MG - were retained and this logistic regression model is referred to as LRModel 1. The demographic variable gender was added and Table 10 illustrates that the

p-value of gender = male is 0.017. This means that it was significant at 0.05 level. The model is denoted as LRModel 2.

Table 10. Coefficient of Regression and Statistics – LRModel 2

Variable	Coeff.	Std. Err.	z-score	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

In LRModel 3, geographical area was added. All of the geographical area categories were found out to be significant variables with their corresponding p-values at almost zeros, as shown in the table below.

Table 11. Coefficient of Regression and Statistics – LRModel 3

Variable	Coeff.	Std. Err.	z-score	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

LRModel 4 was the result of adding variable age to LRModel 1. Table 12 shows that all age categories were significant in the model except for age=35 years old and above with the p-value equal to 0.465.

Table 12. Coefficient of Regression and Statistics – LRModel 4

Variable	Coeff.	Std. Err.	z-score	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

Two more models were developed as a result of combining the variables by adding them piecewise to the first model. LRModel 5 had the factors for LRModel 1 with the addition of

gender and geographical area. The table below reveals that all the variables and all categories of the two dummy variables were significant predictors.

Table 13. Coefficient of Regression and Statistics – LRModel 5

Variable	Coeff.	Std. Err.	z-score	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

In LRModel 6, the variable age was added to the previous model. Table 14 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 with a p-value of 0.251.

Table 14. Coefficient of Regression and Statistics – LRModel 6

Variable	Coeff.	Std. Err.	z-score	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

5.3.2 Performance measures

The performance accuracy measures of the logistic regression models from the previous subchapter were generated by fitting the models with the bootstrapped testing data subset (n=400). The results are presented in the table below.

Table 15. Accuracy, Cohen's kappa and ROC Scores of the Logistic Regression Model

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

LRModel 6 had the highest scores for all three measures with an accuracy of 85.5%, kappa of 0.68 and ROC probability of 0.863. On the other hand, LRModel 1 got the lowest performance for both ROC (0.78) and kappa scores (0.42). LRModel 4 achieved the lowest accuracy score (73.75%) and second lowest kappa (0.43). LRModels 2, 3 and 5 had comparable performance scores with moderate kappa scores (0.53) and good ROC scores.

5.4 Classification Models

Student classification modeling using Decision Tree and Random Forest algorithms were done for the classification training dataset (n=400). The initial hypothesized classifiers were: DSD, SWD, MEO, MIO, MG factors, semester level, gender, age, and geographical area of origin. 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) and the resulting accuracy, Cohen's kappa and area under the ROC curve probability scores were noted. For conciseness of the report, only the best model from the two methods are presented.

5.4.1 Decision Trees

Of the more than twenty models developed using the Decision Tree algorithm, the best in terms of the set accuracy criteria was DTModel 26 with the following classifiers: DSD, SWD, MG and geographical area of origin. The simple view of the Decision Tree model is displayed in Figure 8 and the full view of the model can be checked in Appendix 9.

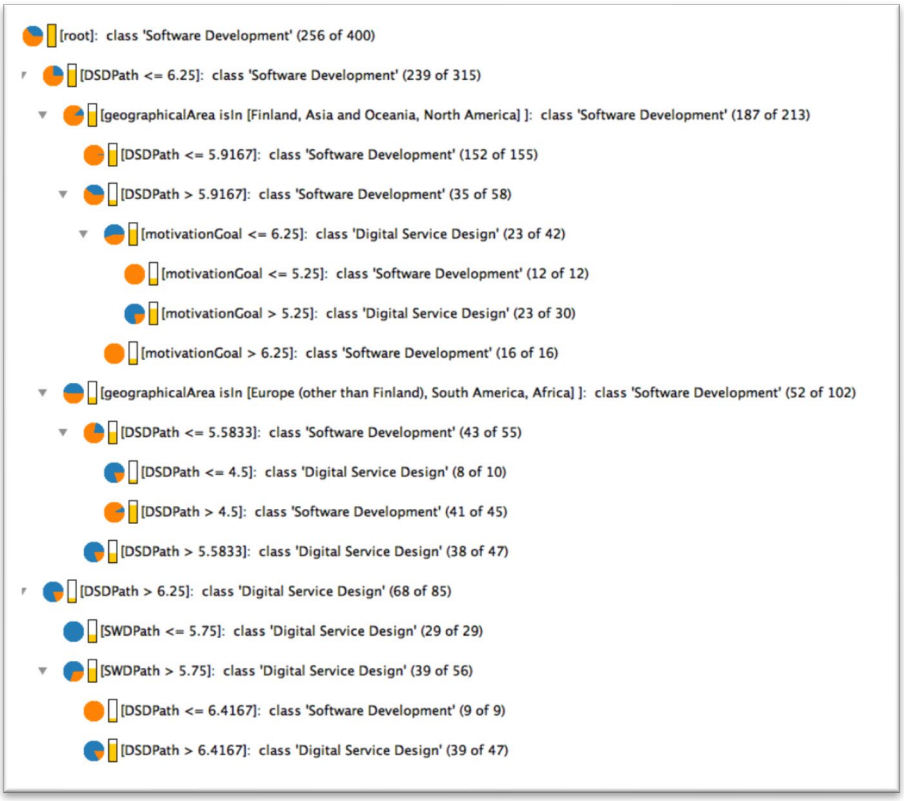


Figure 8. Simple View of the Decision Tree Model

Performing accuracy test on the model using the testing data (n=100) resulted in correctly classifying 34 Digital Service Design and 59 Software Development students with a combined accuracy rate of 93% and a Cohen's kappa of 0.851 as reflect in the figure below.

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. Confusion Matrix and Accuracy Scores of the Decision Tree Model

The resulting plot of the ROC curve (Figure 10) yielded an area under the curve with a 0.959 probability of correctly classifying the study path selection using the factors as opposed to classifying randomly by chance.

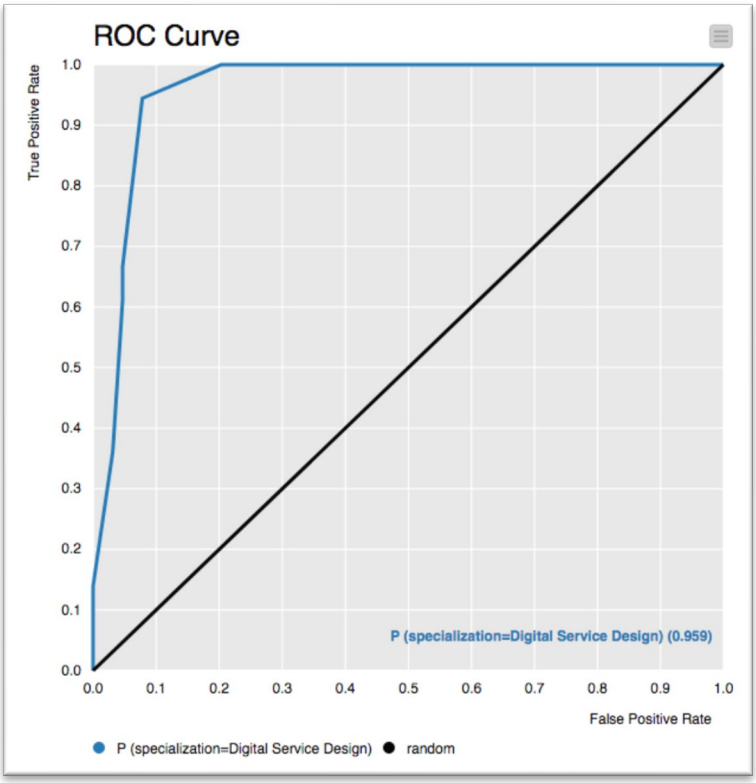


Figure 10. Area under the Receiver Operating Characteristic Curve - Decision Tree

5.4.2 Random Forest

The best model formed using the Random Forest algorithm was RFModel 22 which has DSD, SWD, motivational goal, age and geographical area of origin as classifiers. A sample tree view from the Random Forest is attached as Appendix 10. Fitting the model on the testing data set (n=100) produced the confusion matrix shown in Figure 11. The model performed extremely well in correctly classifying 32 students from Digital Service Design and 62 students from Software Development, with an overall accuracy score of 94% and a Cohen's kappa of 0.868.

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 11. Confusion Matrix and Accuracy Scores of the Random Forest Model

The area under the ROC curve (Figure 12) 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 by random.

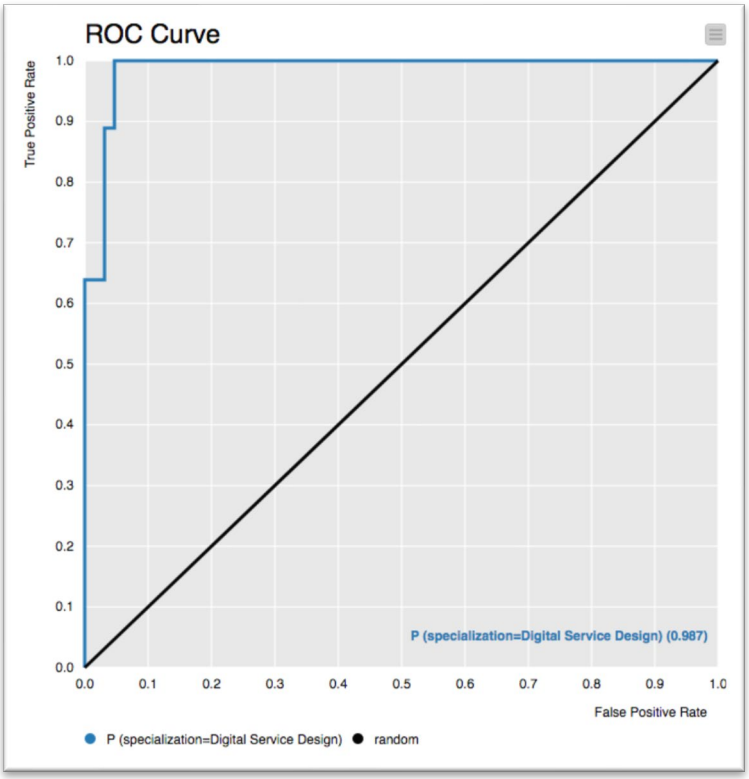


Figure 12. Area under the Receiver Operating Characteristic Curve - Random Forest

6 Discussion

The research developed several logistic regression models and evaluated the 6 best models to accurately predict the study path selection of Business IT students. Furthermore, several models using advanced machine learning techniques, in particular Decision Trees and Random Forests, were generated to classify the students from the degree programme.

The main factors of concern were from students' ratings on statement regarding their affinity to two study paths in the BITe programme (DSD factor and SWD factor). Additionally, mastery orientation (MEO and MIO factors) and motivational goals (MG factor) were also used as predictors and classifiers based on previous researches such that of Niemivirta (2002), Tuominen-Soini (2012) and Winne & Baker (2013). Finally, demographic factors (age, gender, geographical area of origin) were also taken as variables for consideration.

The following questions guided this research study:

1. Are we able to find a suitable data model that would accurately predict if students' study path selection is Software Development (SWD) or Digital Service Design (DSD), based on their answers in a questionnaire?

2. Can we find a model that could classify or cluster the students using features such as study paths, mastery orientation, motivation and demographic attributes such as age, country and gender?

6.1 Factor Selection

The statement "I am interested in technology" was intended to be a factor for Software Development students. It was assumed to be rated highly by software students than design students. However, inspection of the data revealed that the statement did not get considerable difference in ratings from the two groups. This suggests that Business IT students, whether from SWD or DSD study paths, were equally interested in technology.

More surprisingly, three statements ("I enjoy working in an environment where there is always something new going on", "I always keep myself with up-to-date information on new technological innovations", "I would like to invent and develop new devices and applications") which were hypothesized to be factors highly linked with Software Development

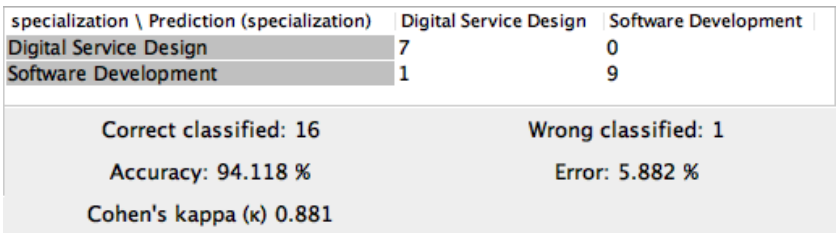
students, turned out contrary to initial assumptions. Exploratory analysis of the data revealed that Digital Service Design students had higher mean ratings to these statements than Software Development students.

6.2 Study Path Prediction

The first research question was answered by analyzing the data from the survey instrument and performing logistic regression analysis to the data. The final six models put forward for consideration received high accuracy ratings, moderate to substantial kappa scores and good ROC accuracy measures (see Table 15).

The best among the models in terms of the set evaluation criteria was LRModel 6 which achieved the highest scores for all three performance measures. The significant predictors for the model were DSD factor, SWD factor, mastery intrinsic orientation, motivational goal, gender, geographical area and age, and the coefficients of regression for each of these factors, their z-scores and p-values were presented in Table 14.

The model's accuracy was re-validated by fitting the unadjusted, non-bootstrapped testing data subset (n= 17) and the resulting confusion matrix is reflected in Figure 13. The chosen prediction model was able to correctly predict 7 Digital Service Design and 9 Software Development study path selections. Overall, the model forecasted the study path selection with a very high accuracy score of 94.12% and a kappa score of 0.881 indicating an almost perfect inter-rating agreement in correctly classifying the two study path categories compared to guessing the path selection randomly.



specialization \ Prediction (specialization)	Digital Service Design	Software Development
Digital Service Design	7	0
Software Development	1	9

Correct classified: 16 Wrong classified: 1

Accuracy: 94.118 % Error: 5.882 %

Cohen's kappa (κ) 0.881

Figure 13. Confusion Matrix and Accuracy Scores of the Final Predictive Model

The ROC curve of the model, which measures the overall efficacy of the prediction, resulted in a predictor with 100% probability that the model is able to assign an almost perfect separation of the values between the two study paths. The figure below plots the probability area under the ROC curve.

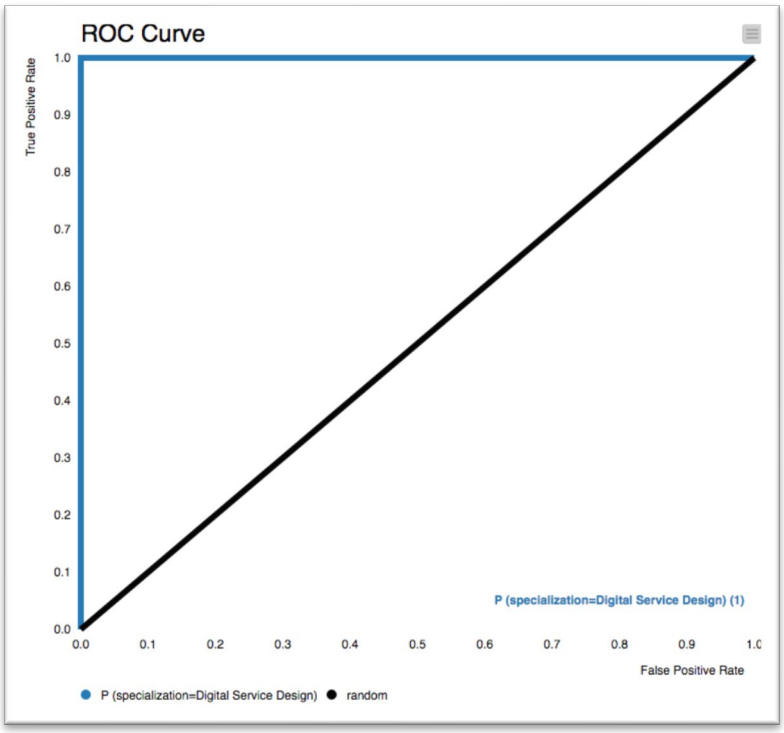


Figure 14. Area under the Receiver Operating Characteristic Curve - Predictive Model

Based from the results mentioned above, it can be asserted that the research was able to find a suitable model that could accurately predict student study path selection based on ratings gathered from a questionnaire. Therefore, the first research question was satisfied.

6.3 Student Classification

To address the second research question, Decision Trees and Random Forests were used and more than 50 classification models were developed from the various combination of the factors. The models were then tested using the testing data subset and were evaluated against the three performance criteria. The best model from each of the two methods were assessed for consideration as the final classification model.

The best Decision Tree model had DSD, SWD, motivational goal and geographical area of origin as classifiers. Meanwhile, the Random Forest model had DSD, SWD, motivational goal, age and geographical area of origin as significant factors.

Examining the figures from fitting the models with the testing dataset reveals that both models had comparable excellent results. The Random Forest model achieved an accuracy score of 94%, kappa of 0.868 and ROC curve probability of 0.987 while the scores

for the Decision Tree model were 93% accuracy, kappa score of 0.851 and 0.959 area under the ROC curve.

Although the Random Forest model slightly outperformed the Decision Tree model, the differences between the performance measure of the two were almost negligible. Therefore, both models can be recommended as final classification models for the study. The Random Forest model should be used for the direct fitting of data for student classification due of its higher accuracy scores.

However, an IF-THEN classification rule can be extracted for the Decision Tree model which could provide an easier way of interpreting the classes, in this case the study path selections. For instance, referring to the simple view of the Decision Tree (see Figure 8), it can be inferred that a survey respondent with an average DSD factor score greater than 6.25 AND an average SWD factor score less than or equal to 5.75 could be classified as a probable Digital Service Design student.

The research was able to successfully obtain models that can classify students using features such as study path factors, motivational goal, age and geographical area of origin, as affirmed by results of the Decision Trees and Random Forest modeling. Thus, it can be asserted that the second research question was answered.

6.4 Simulation of Model Deployment

KNIME Analytics Platform implemented a way for the Logistic regression and the Decision Tree models to be exported in PMML format which would allow the models to be easily used and deployed in other data mining tools. The tool did not provide the same functionality for Random Forest.

To simulate the deployment of the final models for study path prediction and student classification, a workflow (illustrated in Figure 15) was created in KNIME to replicate the inputting and pre-processing of new data, fitting the data to the model and checking the prediction outcomes.

Two survey responses simulating profiles assumed to correspond to a Digital Service Design student (female, with age range 23-28 years old, from Asia and Oceania and in 5th semester) and a Software Development student (male, 17-22 years old, from Europe (other than Finland) and in 3rd semester) were fitted in both the Logistic Regression and Decision Tree models.

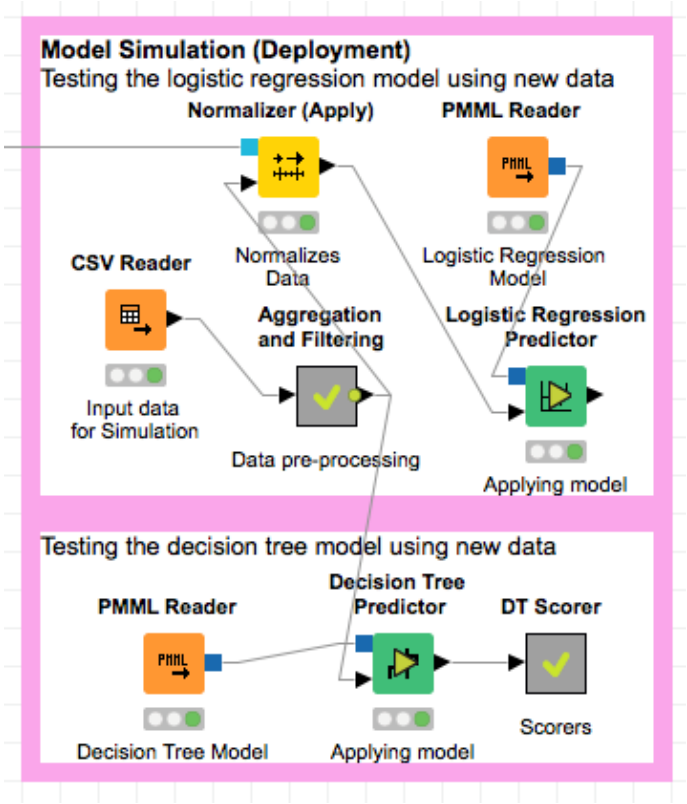


Figure 15. KNIME Workflow Simulating the Deployment of the Models

The logistic model correctly predicted both the study path selections, as evidenced by the outcome displayed in the table below:

Table 16. Outcome of the Simulated Deployment of the Logistic Regression Model

specialization	DSDPath	SWDPath	masteryIntrinsic	motivationGoal	gender	age	geographicalArea	Prediction (specialization)
Digital Service Design	1.769	-0.232	0.896	0.325	Female	23 to 28 years old	Asia and Oceania	Digital Service Design
Software Development	-1.214	1.466	-1.893	-0.531	Male	17 to 22 years old	Europe (other than Finland)	Software Development

Similarly, the Decision Tree model was able to classify the study path selections correctly, as illustrated by the prediction outcome of the study path selection and associated probabilities below:

Table 17. Outcome of the Simulated Deployment of the Decision Tree Model

specialization	P (specialization=Digital Service Design)	P (specialization=Software Development)	Prediction (specialization)
Digital Service Design	1	0	Digital Service Design
Software Development	0.089	0.911	Software Development

Although the deployment of the resulting models is not a scope of the thesis, this subsection illustrated the possibility of deploying the models in other systems. It has also validated the accuracy of the models in predicting and classifying study paths based on new data.

6.5 Summary and Other Findings

From the results of the model building using machine learning and the outcome of testing the final models, the research was able to successfully develop a logistic regression prediction model that could predict the study path selection (either SWD or DSD) with high accuracy and good probability that the prediction is correct. The data used were from a questionnaire with the following factors: *Software Development* ("I like to learn new languages."; "I solve problems specifically with the end goal in mind."), *Digital Service Design (*"I want to understand how people use technology."; "It's important that I can be creative in my work."; "I am interested in designing archetypes/prototypes."; "I want to work with my hands."; "I would rather work with people than to work with machines."; "I enjoy coming up with new solutions to problems."), *motivational goal (*"Career development and promotions are important for me."; "Salary means a lot to me."), *mastery intrinsic orientation* ("An important goal for me is to learn as much as possible."), *age*, *gender* and *geographical area origin*.

Moreover, using the following significant classifiers: *Digital Service Design* and *Software Development* factors, *motivational goal*, *age*, and *geographical area of origin*, two classification models were likewise generated by utilizing Random Forests and Decision Trees algorithms. Testing the models yielded very high accuracy results.

Educational Research Mining proved to be an interesting field of study with a huge amount of potential research problems to be explored. And as the Finnish Ministry of Education recently launched its Vision 2030 for higher education, now is currently an exciting time for researchers to take opportunities to develop innovative learning services that would offer help to the government's digitalisation efforts in education.

The CRISP-DM methodology was confirmed to be an effective framework that has helped guide this thesis forward through each of the research phases. KNIME Analytics platform had a gentle learning curve. Node and workflow setup and its adoption for use can be accomplished with a few hours of tutorials. Because the system had a convenient graphical user interface, advanced machine learning algorithms were utilized rapidly and without writing a single line of code. This has helped in developing the numerous models used in

the research quickly and efficiently. However, KNIME was not able to provide a highly customizable way of presenting the output from the analysis such as nicely formatted tables and graphs. It also was not able to provide PMML exporting for the Random Forest node.

The research also fortuitously found out that there is a need to start having effective data collection of basic information about the students and to make these data more accessible to researchers. For instance, the study path selections of BITe students are currently not officially tracked. Demographic and other relevant information (e.g. dropout and graduation rates) are likewise not readily and publicly available.

7 Conclusions and Recommendations

The study was an Educational Data Mining research focused on students of the Degree Programme in Business Information Technology at Haaga-Helia University of Applied Sciences. The general objective of the research was to help with improving education using data-driven insights on students. Specifically, the goal was to conduct an exploratory research to apply machine learning techniques in order to come up with: (i.) a prediction model of the two most common study paths in the programme, and (ii.) a classification model based on several factors such as students' affinity to Software Development or Digital Service Design, motivational goals, mastery orientation and other demographic variables.

The thesis was able to contribute to data mining research in education by using advanced machine learning algorithms to develop a novel way of predicting the study path selection of students in the Business IT programme. It was also able to construct classification rules that categorize the students based on specified factors. Moreover, the study was also successful in showing that the proposed method of gathering the data from a questionnaire resulted to having viable factors that could be used for machine learning.

The variables measuring affinity to Software Development (two statements), Digital Service Design (six statements), motivational goal (two statements), mastery intrinsic orientation (one statement) and the demographic data age, gender and geographical origin were significant predictors of study paths. The results were able to demonstrate that the final logistic regression model was able to predict the study path selection of BITe students (either SWD or DSD) with 85.5% accuracy and 86.3% probability (area under the ROC curve) that the model was able to distinguish between the two study paths. A validation test of the model was done and resulted to an even higher accuracy (94.12%) in correctly predicting the students' study path selection of Software Development or Digital Service Design.

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. Due to only very slight differences between the performance measures of these models, both were recommended to be used for classification. The Random Forest would result in a slightly higher accuracy rate but the model is more challenging to illustrate plainly. On the other hand, the Decision Tree would be easier to interpret by extracting a classification rule from its tree view.

Examining the results of the final models, it was confirmed that both Digital Service Design (DSD factor) and Motivational Goal (MG) performed significantly as predictors and classifiers of student study path selection. However, the research was not able to establish the significance of Mastery Extrinsic Orientation (MEO) in predicting and classifying the study paths. It can also be deduced that Digital Service Design and Software Development students are equally likely interested in technology as evidenced by the factor “I am interested in technology” being trimmed off of the prediction and classification features.

Grounding the thesis with theories of notable researchers in the field of educational data mining and examining the previous papers, theses and dissertations of academics from respected educational institutions provided credibility to the concepts used as viewpoints of the study. Utilizing CRISP-DM, a well-established framework in data mining, ensured that the study was guided by a scientific and trustworthy process. Moreover, a thorough documentation of the research methodology and the use of random seeds (see Appendix 11) throughout the model building phase guaranteed that the results are repeatable and reproducible. Interested researchers are able to replicate and verify the outcome of the thesis, therefore providing reliability to the research findings.

The main limitation of the study is that the resulting models would only be able to correctly predict the study path selection of students with predisposed affinity to selecting Software Development or Digital Service Design paths. Future iteration of the thesis topic could be made better by comprehensively analysing and adding more factors related to characteristics of Business IT students. Additionally, the study path prediction should be expanded to cover the two other paths in the programme i.e. ICT Infrastructure and Business ICT as well. The next research should also consider including the students from the Business IT programme in Finnish to get more data points. Scope expansion into university-wide prediction of degree programme selection is also another topic worth considering in the future.

The resulting models for Logistic Regression and Decision Trees were exported as PMML format which makes the deployment of the models in other tools possible. A continuation of this study, possibly a thesis topic for other students, could be done by developing a study path predicting app that makes use of the logistic regression model derived from this research. The application could be implemented as part of the orientation week services to incoming students of BITe. The app could recommend the possible appropriate study paths for the first-year students, thereby offering guidance and lessening their confusion in which courses to take in the future.

The classification rule obtained from the classification model could be used as baseline classifiers and could be used as a guide when developing learning materials and assignments, bearing in mind the motivational goals and mastery orientation of the students. This could help improve the teaching pedagogy in the university.

Actions to improve the data collection about Business Information Technology students and making these data readily accessible are also recommended to be initiated. Availability of data and accessibility to these types of information in the university are important to encourage more data mining initiatives in the future, which in turn would only mean improvements in educational institutions.

Finally, the researcher recommends further exploring other machine learning algorithms and tools to implement more data mining researches that tackle gaps in educational settings. Current digital learning services in the university could be examined and developed based on individual student behaviour, characteristics, needs and learning styles, thus improving student performance and possibly lowering dropout rates and/or increasing graduation rates. The insights about these factors could be acquired from EDM using machine learning models.

References

Adeyemo, A., Wimmer, H. & Powell, L. 2018. Effects of Normalization Techniques on Logistic Regression in Data Science. Proceedings of the Conference on Information Systems Applied Research, 11, 4813. URL: http://proc.conisar.org/2018/pdf/4813.pdf. Accessed: 24 April 2019

Almahadeen, L., Akkaya, M. & Sari, A. 2017. Mining Student Data Using CRISP-DM Model. International Journal of Computer Science and Information Security Feb 2017, 15, 2, pp.305-316. URL: https://sites.google.com/site/ijcsis/vol-15-no-2-feb-2017. Accessed: 3 April 2019

Amarillo, M. 2018. Is KNIME (A Machine Learning Platform With ZERO Coding Involved) Suitable For You/Your Business?. Blog. URL: https://medium.com/@matthew.mh.wong/is-knime-a-machine-learning-platform-with-zero-coding-involved-suitable-for-you-your-business-10f4b2864e1d. Accessed: 3 March 2019

Azevedo, A. & Santos, M.F. 2008. KDD, SEMMA and CRISP-DM: a parallel overview. URL: http://hdl.handle.net/10400.22/136. Accessed: 30 March 2019

Baker, R. & Inventado, P. 2014. Educational Data Mining and Learning Analytics. URL: https://www.researchgate.net/publication/278660799_Educational_Data_Mining_and_Learning_Analytics. Accessed: 18 March 2019

Bansal, R., Mishra, A. & Singh S.N. 2017. Mining of Educational Data for Analysing Student's Overall Performance. 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence, pp. 495-497. URL: https://ieeexplore-ieee-org.ezproxy.haaga-helia.fi/document/7943202. Accessed: 3 April 2019

Bhilegaonkar, A. 2016. Machine Learning and Cognitive Computing: A Proposed Framework to Navigate the Opportunities. Master Thesis. Massachusetts Institute of Technology. Massachusetts. URL: http://hdl.handle.net/1721.1/107589. Accessed: 10 March 2019

Breiman, L. 2001. Random Forests. Machine Learning, 45, pp. 5-32. URL: http://dx.doi.org/10.1023/A:1010933404324. Accessed: 30 March 2019

Chatti, M.A., Dyckhoff, A.L., Schroeder, U. & Thus, H. 2012. A Reference Model for Learning Analytics. International Journal of Technology Enhanced Learning 4, 5/6, pp.

318-331. URL: https://www.thues.com/upload/pdf/2012/CDST12_IJTEL.pdf. Accessed: 19 March 2019

Clark, A. 2018. The Machine Learning Audit—CRISP-DM Framework. Information Systems Audit and Control Association Journal 2018, 1. URL: https://www.isaca.org/Journal/archives/2018/Volume-1/Pages/the-machine-learning-audit-crisp-dm-framework.aspx. Accessed: 30 March 2019

Ding, F. 2018. Click Prediction with Machine Learning Tools. Master Thesis. University of California. Los Angeles. URL: https://escholarship.org/uc/item/0cf772s3. Accessed: 4 March 2019

Dirin, A. 15 May 2018. Students' Educational Performance and Drop Out Factors in the Degree Programme in Business Information Technology (BITE). eSignals, Haaga-Helia Online Working Paper. URL: https://esignals.haaga-helia.fi/en/2018/05/15/students-educational-performance-and-dropout-factors-in-the-degree-programme-in-business-information-technology-bite/. Accessed: 4 March 2019

ElAtia, S., Ipperciel, D. & Zaiane, O. 2016. Data Mining and Learning Analytics: Applications in Educational Research. Wiley. New Jersey.

Finch, W.H. 2018. Bootstrapping. The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation, pp. 218-220. URL: https://dx.doi.org/10.4135/9781506326139. Accessed: 19 April 2019

Goodman, S.N., Fanelli, D. & Ioannidis, J.P.A. 2016. What does research reproducibility mean?. Science Translational Medicine, 8, 341, pp. 341ps12 URL: https://dx.doi.org/10.1126/scitranslmed.aaf5027. Accessed: 24 April 2019

Guazzelli, A. 2010. What is PML? Explore the power of predictive analytics and open standards. IBM Corporation. URL: https://www.ibm.com/developerworks/library/ba-ind-PMML1/ba-ind-PMML1-pdf.pdf. Accessed: 29 April 2019

Haaga-Helia University of Applied Sciences 2019a. Degree Programme in Business Information Technology. Webpage. URL: http://www.haaga-helia.fi/en/education/bachelor-degree-programmes/degree-programme-business-information-technology?userLang=en. Accessed: 31 March 2019

Haaga-Helia University of Applied Sciences 2019b. Study Profiles. URL: http://www.haaga-helia.fi/en/students-guide/degree-programmes/degree-programme-business-information-technology-pasila-campus-182015/profiles?userLang=en. Accessed: 31 March 2019

Hamedi, A. & Dirin, A. 2018. A Bayesian approach in students' performance analysis. 10th annual International Conference on Education and New Learning Technologies. Palma de Mallorca (Spain). URL: https://doi.org/10.21125/edulearn.2018.2498. Accessed: 19 April 2019

Herold, J. 2013. Data Mining Students' Ordinary Handwritten Coursework. Doctoral Dissertation. University of California. Riverside. URL: https://escholarship.org/uc/item/98p905xs. Accessed: 30 March 2019

Hsieh, J. 2012. .Receiver Operating Characteristic (ROC) Curve. Encyclopedia of Epidemiology, pp. 896-898. SAGE Publications. URL: https://dx.doi.org/10.4135/9781412953948. Accessed: 19 April 2019

International Educational Data Mining Society (IEDM), 2019. Home. Website. URL: http://educationaldatamining.org/. Accessed: 30 March 2019

Journal of Educational Data Mining (JEDM), 2019. Home. Website. URL: https://jedm.educationaldatamining.org/index.php/JEDM. Accessed: 30 March 2019

Kai, S., Almeda, M. V., Baker, R., Heffernan, C. & Heffernan, N. 2018. Decision Tree Modeling of Wheel- Spinning and Productive Persistence in Skill Builders. JEDM | Journal of Educational Data Mining, 10, 1, pp. 36-71. URL: https://jedm.educationaldatamining.org/index.php/JEDM/article/view/210. Accessed: 30 March 2019

Kampakis, S. 2016. Performance Measures: Cohen's Kappa Statistic. URL: https://thedatascientist.com/performance-measures-cohens-kappa-statistic/. Accessed: 24 April 2019

KDnuggets. 2014. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. URL: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html. Accessed: 29 March 2019

KNIME. 2019. KNIME Open Source Story. URL: https://www.knime.com/knime-open-source-story. Accessed: 10 April 2019

Landis, J.R.; Koch, G.G. 1977. The measurement of observer agreement for categorical data. Biometrics, 33, 1, pp. 159–174.

Lemm, K. 2012. Stratified Sampling. Encyclopedia of Research Design, pp. 1452-1454. URL: https://dx.doi.org/10.4135/9781412961288. Accessed: 24 April 2019

Leventhal, B. 2010. An introduction to data mining and other techniques for advanced analytics. Journal of Direct, Data and Digital Marketing Practice, 12, 2, pp 137–153. URL: https://doi.org/10.1057/dddmp.2010.35. Accessed: 3 March 2019

Long, M., Ferrier, F. & Heagney, M. 2006. Stay, play or give it away? Students continuing, changing or leaving university study in first year. Centre for the Economics of Education and Training. Monash University – Australian Council for Educational Research. Victoria. URL: http://www.monash.edu/education/non-cms/centres/ceet/docs/2006stayplayorgiveit-away.pdf. Accessed: 18 April 2019

Make, G. 2018. Implementing KNIME Analytical Platform for visualizing data in educational context. Bachelor's Thesis. Haaga-Helia University of Applied Sciences. Helsinki. URL: https://www.theseus.fi/handle/10024/157624. Accessed: 21 January 2019

Menard, S. 2013. Model Specification, Variable Selection, and Model Building. Logistic Regression: From Introductory to Advanced Concepts and Applications, pp. 105-124. SAGE Publications. URL: https://dx.doi.org/10.4135/9781483348964. Accessed: 24 April 2019

Ministry of Education and Culture 2019. Vision 2030. URL: https://minedu.fi/en/vision-2030. Accessed: 5 March 2019

Moreira, J., Carvalho, A. & Horváth, T. 2018. Appendix A: A Comprehensive Description of the CRISP‐DM Methodology. URL: http://dx.doi.org/10.1002/9781119296294.app. Accessed: 3 April 2019

Mohri, M, Rostamizadeh, A & Talwalkar, A. 2012. Foundations of Machine Learning, MIT Press, Cambridge.

Muenchen, R. 2019. The Popularity of Data Science Software. URL: http://r4stats.com/articles/popularity/. Accessed: 3 April 2019

Nadali, A., Kakhky, E.N. & Nosratabadi, H.E. 2011. Evaluating the Success Level of Data Mining Projects Based on CRISP-DM Methodology by a Fuzzy Expert System. International Conference on Electronics Computer Technology, 6, pp. 161-165. URL: http://dx.doi.org/10.1109/ICECTECH.2011.5942073. Accessed: 3 April 2019

Nisbet, R., Miner, G. & Elder, J. 2009. Handbook of Statistical Analysis and Data Mining Applications. Academic Press. London.

Niemivirta, M. 2002. Motivation and Performance in Context: The Influence of Goal Orientations and Instructional Setting on Situational Appraisals and Task Performance. Psychologia - An International Journal of Psychology in the Orient. URL: https://doi.org/10.2117/psysoc.2002.250. Accessed: 21 January 2019

Official Statistics of Finland (OSF) 2019. Discontinuation of Education 2017. URL: http://www.stat.fi/til/kkesk/2017/kkesk_2017_2019-03-14_tie_001_en.html. Accessed: 5 March 2019

Oreski, D., Pihir, I. & Konecki, M. 2017. Crisp-dm Process Model In Educational Setting. Varazdin: Varazdin Development and Entrepreneurship Agency (VADEA). URL: https://search-proquest-com.ezproxy.haaga-helia.fi/docview/2070395138?accountid=27436. Accessed: 3 April 2019

Otaris. 2018. Data Analysis, Modeling and Reporting. Services. URL: https://www.otaris.de/gb/datenanalysen-modellbildung-reporting/. Accessed: 3 April 2019

Palacios, H. J. G. 2017. A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change. Advances in Science, Technology and Engineering Systems, 2, 3, pp. 598-604. URL: http://dx.doi.org/10.25046/aj020376. Accessed: 10 April 2019

Papamitsiou, Z. & Economides, A. 2014. Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. Journal of Educational Technology & Society, 17, 4, pp. 49-64. URL: https://www.researchgate.net/publication/267510046_Learning_Analytics_and_Educational_Data_Mining_in_Practice_A_Systematic_Literature_Review_of_Empirical_Evidence. Accessed. 18 March 2019

Park, J.H. 2015. Synthetic Tutor: Profiling Students and Mass-Customizing Learning Processes Dynamically in Design Scripting Education. Doctoral Dissertation. Massachusetts Institute of Technology. Massachusetts. URL: http://hdl.handle.net/1721.1/101544. Accessed: 5 March 2019

Rangra, K. & Bansal, K. 2014. Comparative Study of Data Mining Tools. International Journal of Advanced Research in Computer Science and Software Engineering, 4, 6, pp. 216-223. URL: http://ijarcsse.com/Before_August_2017/docs/papers/Volume_4/6_June2014/V4I6-0145.pdf. Accessed: 29 March 2019

Rebolledo-Mendez, G., Boulay, B., Luckin, R. & Benitez-Guerrero, E. 2013. Mining Data from Interactions with a Motivational-aware Tutoring System Using Data Visualization. Journal of Educational Data Mining, 5, 1, 72-103. URL: https://jedm.educationaldatamining.org/index.php/JEDM/article/view/31. Accessed: Accessed: 30 March 2019

Reddi, S. 2017. New Optimization Methods for Modern Machine Learning. Doctoral Dissertation. Carnegie Mellon University. Pittsburgh, Philadelphia. URL: https://figshare.com/articles/New_Optimization_Methods_for_Modern_Machine_Learning/6720833. Accessed: 9 March 2019

Roiger, R. 2017. Data Mining – a Tutorial-based Primer. Second edition. CRC Press. Data Mining and Knowledge Discovery Series. Florida

Romero, C. & Ventura, S. 2010. Educational Data Mining: A Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 40, 6, pp. 601-618. URL: http://ieeexplore.ieee.org.ezproxy.haaga-helia.fi:2048/stamp/stamp.jsp?tp=&arnumber=5524021&isnumber=5593938. Accessed: 18 March 2019

Romero, C., Ventura, S., Pechnizkiy, M. & Baker, R. 2010. Handbook of Educational Data Mining. Chapman & Hall/ CRC Press. Data Mining and Knowledge Discovery Series. Florida.

Ryan, R. & Deci, E. 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. Contemporary Educational Psychology, 25, pp. 54-67. URL: http://dx.doi.org/10.1006/ceps.1999.1020. Accessed: 4 March 2019

Sarra, A., Fontanella, L. & Zio, S. D. 2018. Identifying students at risk of academic failure within the educational data mining framework. Social Indicators Research, pp. 1-20. URL: http://dx.doi.org.ezproxy.haaga-helia.fi:2048/10.1007/s11205-018-1901-8. Accessed: 4 March 2019

Shearer, C. 2000. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, 5, 4, pp. 13-22. URL: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf. Accessed: 30 March 2019

Smith, R. 2011. Examining the Motivating Factors that Influence Students with an Associate's Degree to Complete a Bachelor's Degree at a Private University. Doctoral Dissertation. National Louis University. Chicago. URL: https://digitalcommons.nl.edu/diss/41/. Accessed: 4 March 2019

Song, Y. 2018. Stock Trend Prediction: Based on Machine Learning Methods. Master Thesis. University of California. Los Angeles. URL: https://escholarship.org/uc/item/0cp1x8th. Accessed: 4 March 2019

Spoon, K., Beemer, J., Whitmer, J., Fan, J., Frazee, J., Stronach, J., Bohonak, A. & Levine, R. 2016. Random Forests for Evaluating Pedagogy and Informing Personalized Learning. JEDM | Journal of Educational Data Mining, 8, 2, pp. 20-50. URL: https://jedm.educationaldatamining.org/index.php/JEDM/article/view/JEDM2016-8-2-2. Accessed: 30 March 2019

Stegmann, R. 2016. Student Performance in an Online Learning Platform - Predicting Grades given Student Features and Behaviour. Master Thesis. Aalto University. Espoo.

Tuominen-Soini, H. 2012. Student Motivation and Well-being – Achievement Goal Orientation Profiles, Temporal Stability, and Academic and Socio-Emotional Outcomes. Doctoral Dissertation. University of Helsinki. URL: http://urn.fi/URN:ISBN:978-952-10-8201-6. Accessed: 18 April 2019

University of Oulu. 2018. Admissions. URL: https://www.oulu.fi/university/apply. Accessed: 21 January 2019

Vipunen. 2018. Students and degrees. URL: https://vipunen.fi/en-gb/polytechnic/Pages/Opiskelijat-ja-tutkinnot.aspx. Accessed: 29 March 2019

Winne, P. & Baker, R. 2013. The Potentials of Educational Data Mining for Researching Metacognition, Motivation and Self-Regulated Learning. Journal of Educational Data Mining, 5, 1, pp. 1-8. URL: https://jedm.educationaldatamining.org/index.php/JEDM/article/view/28. Accessed: 30 March 2019

Wirth, R. & Hipp, J. 2000. CRISP-DM: Towards a standard process model for data mining. URL: https://www.semanticscholar.org/paper/Crisp-dm%3A-towards-a-standard-process-modell-for-Wirth/48b9293cfd4297f855867ca278f7069abc6a9c24. Accessed: 30 March 2019

Witten, I. & Frank, E. 2005. Data Mining, Practical Machine Learning Tools and Techniques 2nd Edition. Morgan Kaufmann Publishers. The Morgan Kaufmann Series in Data Management Systems. California.

Zhu, X. 2017. Agile Mining - a Novel Data Mining Process for Industry Practice Based on Agile Methods and Visualization. Master Thesis. University of Technology Sydney. URL: https://opus.lib.uts.edu.au/bitstream/10453/123178/1/01front.pdf. Accessed: 30 March 2019

Appendices

Appendix 1. Topics of Interest Used in EDM (Stegmann 2016)

Topic	Description
Generic frameworks and methods	To develop tool, frameworks, methods, algorithms, approaches, and so forth, specifically oriented to educational data mining research.
Mining educational data	Mining assessment data, mining browsing or interaction data, mining the results of educational research.
Educational process mining	To extract process-related knowledge from event logs recorded by educational systems.
Data-driven adaptation and personalization	To apply data mining methods for improving adaptation and personalization in educational environments and systems.
Improving educational software	Many large educational data sets are generated by computer software. Can we use our discoveries to improve the software's effectiveness?
Evaluating teaching interventions	Student learning data provides a powerful mechanism for determining which teaching actions are successful. How can we best use such data?
Emotions, affect, and choice	The student's level of interest is critical. Can we detect when students are bored and uninterested? What other affective states or student choices should we tract?
Integrating data mining and pedagogical theory	Data mining typically involves searching a large volume of models. Can we use existing educational and psychological knowledge to better focus our research?
Improving teacher support	What types of assessment information would help teachers? What types of instructional suggestions are both feasible to generate and would be welcomed by teachers?
Replications studies	To apply previously used techniques to a new domain, or to reanalyse an existing data set with a new technique.
Best practices	Best practices for adaptation of data mining, information retrieval, recommender system, opinion mining and question answering techniques to educational context.

Appendix 2. Research Questionnaire

Part 1. Study Path

1. I am currently in my …
   - ☐ 1st semester
   - ☐ 2nd semester
   - ☐ 3rd semester
   - ☐ 4th semester
   - ☐ 5th semester
   - ☐ 6th semester
   - ☐ 7th semester
   - ☐ 8th semester and over

2. My **main** specialization path is… (choose only one)
   - ☐ Software Development
   - ☐ Digital Service Design
   - ☐ Business and ICT
   - ☐ ICT Infrastructure

*Part 2. Please rate your agreement for each statement from a scale of **1** (strongly disagree) to **7** (strongly agree).*

1. I like to learn new languages.
   ☐1- Strongly disagree ☐ 2-Disagree ☐ 3-Slightly Disagree ☐ 4-Neutral ☐ 5-Slightly Agree ☐ 6-Agree ☐ 7-Strongly Agree

2. I am interested in technology.
   ☐1- Strongly disagree ☐ 2-Disagree ☐ 3-Slightly Disagree ☐ 4-Neutral ☐ 5-Slightly Agree ☐ 6-Agree ☐ 7-Strongly Agree

3. I want to understand how people use technology.
   ☐1- Strongly disagree ☐ 2-Disagree ☐ 3-Slightly Disagree ☐ 4-Neutral ☐ 5-Slightly Agree ☐ 6-Agree ☐ 7-Strongly Agree

4. It is important that I can be creative in my work.
   ☐1- Strongly disagree ☐ 2-Disagree ☐ 3-Slightly Disagree ☐ 4-Neutral ☐ 5-Slightly Agree ☐ 6-Agree ☐ 7-Strongly Agree

5. I solve problems specifically with the end goal in mind.
   ☐1- Strongly disagree ☐ 2-Disagree ☐ 3-Slightly Disagree ☐ 4-Neutral ☐ 5-Slightly Agree ☐ 6-Agree ☐ 7-Strongly Agree

6. An important goal for me is to do well in my studies.
   ☐1- Strongly disagree ☐ 2-Disagree ☐ 3-Slightly Disagree ☐ 4-Neutral ☐ 5-Slightly Agree ☐ 6-Agree ☐ 7-Strongly Agree

7. I enjoy working in an environment where there is always something new going on.
   ☐1- Strongly disagree ☐ 2-Disagree ☐ 3-Slightly Disagree ☐ 4-Neutral ☐ 5-Slightly Agree ☐ 6-Agree ☐ 7-Strongly Agree

8. I am interested in designing archetypes/prototypes.
   ☐1- Strongly disagree ☐ 2-Disagree ☐ 3-Slightly Disagree ☐ 4-Neutral ☐ 5-Slightly Agree ☐ 6-Agree ☐ 7-Strongly Agree

9. To acquire new knowledge is an important goal for me in school.
   ☐1- Strongly disagree ☐ 2-Disagree ☐ 3-Slightly Disagree ☐ 4-Neutral ☐ 5-Slightly Agree ☐ 6-Agree ☐ 7-Strongly Agree

10. I want to work with my hands.
□1- Strongly disagree □ 2-Disagree □ 3-Slightly Disagree □ 4-Neutral □ 5-Slightly Agree □ 6-Agree □ 7-Strongly Agree

11. My goal is to succeed in school.
□1- Strongly disagree □ 2-Disagree □ 3-Slightly Disagree □ 4-Neutral □ 5-Slightly Agree □ 6-Agree □ 7-Strongly Agree

12. I always keep myself with up-to-date information on new technological innovations.
□1- Strongly disagree □ 2-Disagree □ 3-Slightly Disagree □ 4-Neutral □ 5-Slightly Agree □ 6-Agree □ 7-Strongly Agree

13. I would rather work with people than to work with machines.
□1- Strongly disagree □ 2-Disagree □ 3-Slightly Disagree □ 4-Neutral □ 5-Slightly Agree □ 6-Agree □ 7-Strongly Agree

14. An important goal for me is to learn as much as possible.
□1- Strongly disagree □ 2-Disagree □ 3-Slightly Disagree □ 4-Neutral □ 5-Slightly Agree □ 6-Agree □ 7-Strongly Agree

15. Career development and promotions are important for me.
□1- Strongly disagree □ 2-Disagree □ 3-Slightly Disagree □ 4-Neutral □ 5-Slightly Agree □ 6-Agree □ 7-Strongly Agree

16. I would like to invent and develop new devices and applications.
□1- Strongly disagree □ 2-Disagree □ 3-Slightly Disagree □ 4-Neutral □ 5-Slightly Agree □ 6-Agree □ 7-Strongly Agree

17. I enjoy coming up with new solutions to problems.
□1- Strongly disagree □ 2-Disagree □ 3-Slightly Disagree □ 4-Neutral □ 5-Slightly Agree □ 6-Agree □ 7-Strongly Agree

18. Salary means a lot to me.
□1- Strongly disagree □ 2-Disagree □ 3-Slightly Disagree □ 4-Neutral □ 5-Slightly Agree □ 6-Agree □ 7-Strongly Agree

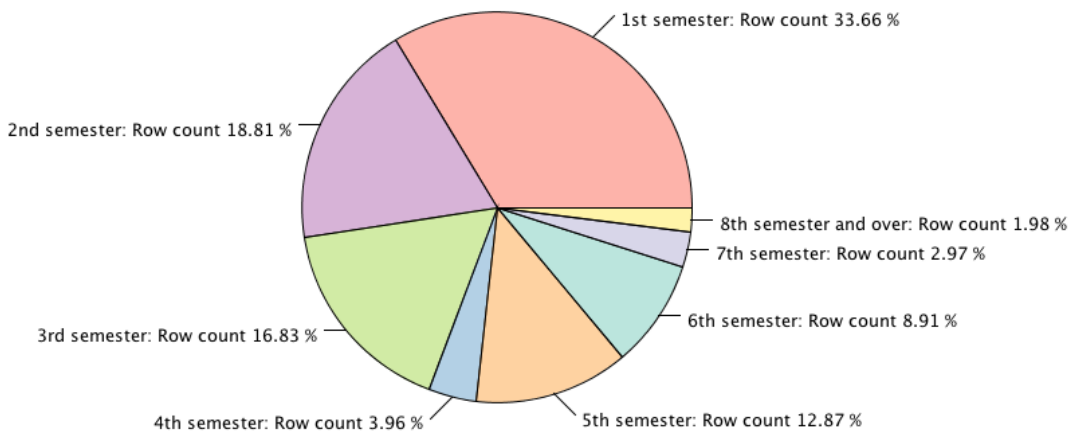
Part 3. Demographic details

1. Age
   - □ 17 to 22 years old
   - □ 23 to 28 years old
   - □ 29 to 34 years old
   - □ 35 years old and over

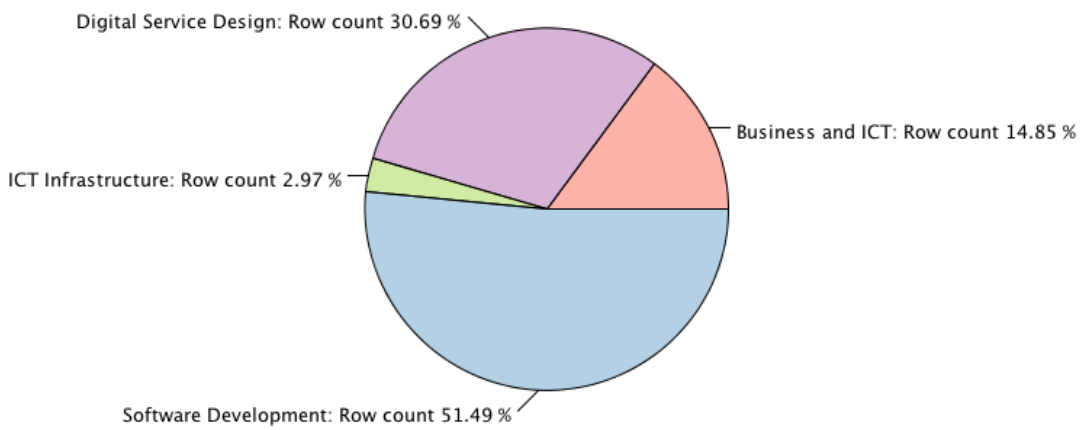
2. Gender
   - □ Male
   - □ Female

3. I am from …
   - □ Finland
   - □ Europe (other than Finland)
   - □ Asia and Oceania
   - □ Africa
   - □ North America
   - □ South America

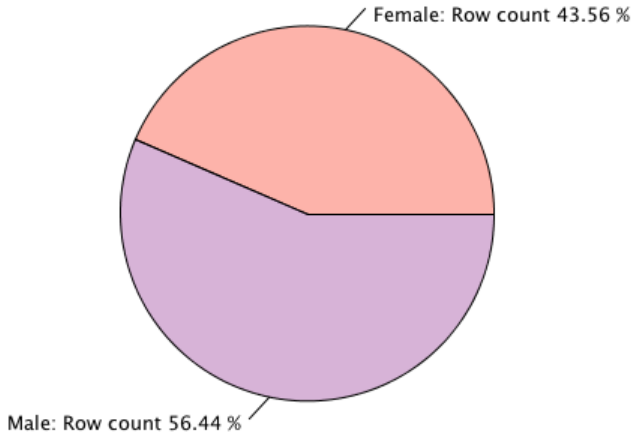
Appendix 3a. Respondents by Semester Level



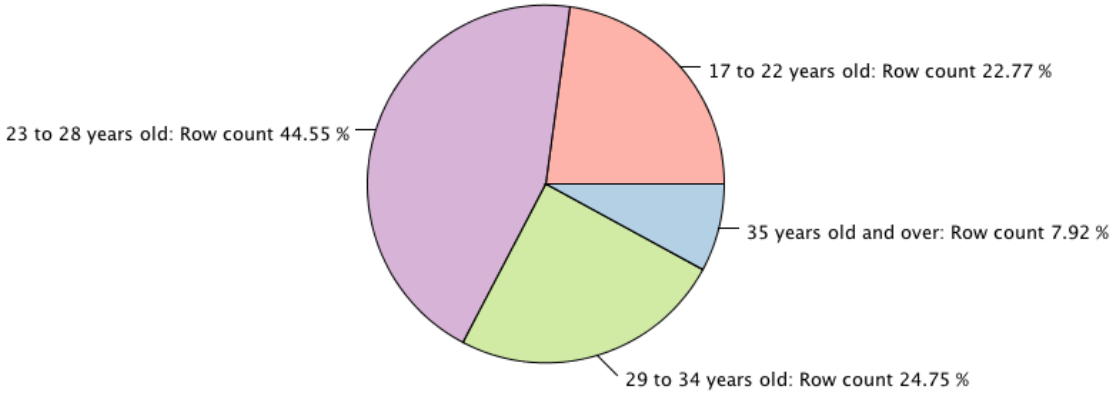
Appendix 3b. Respondents by Study Paths



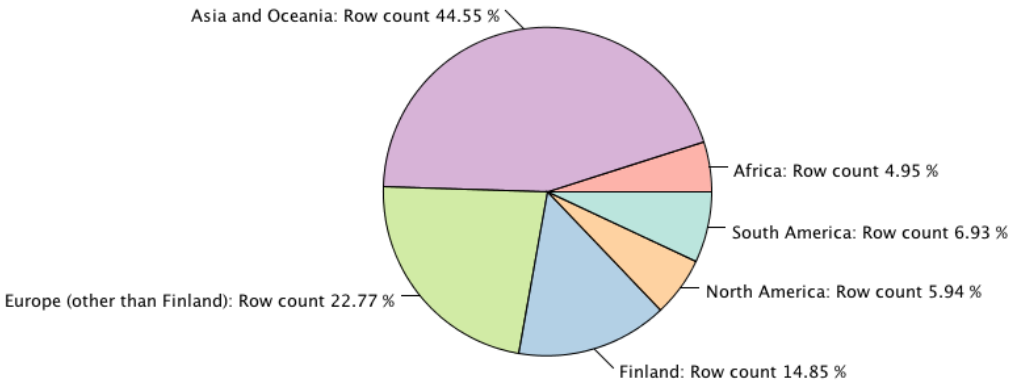
Appendix 3c. Respondents by Gender



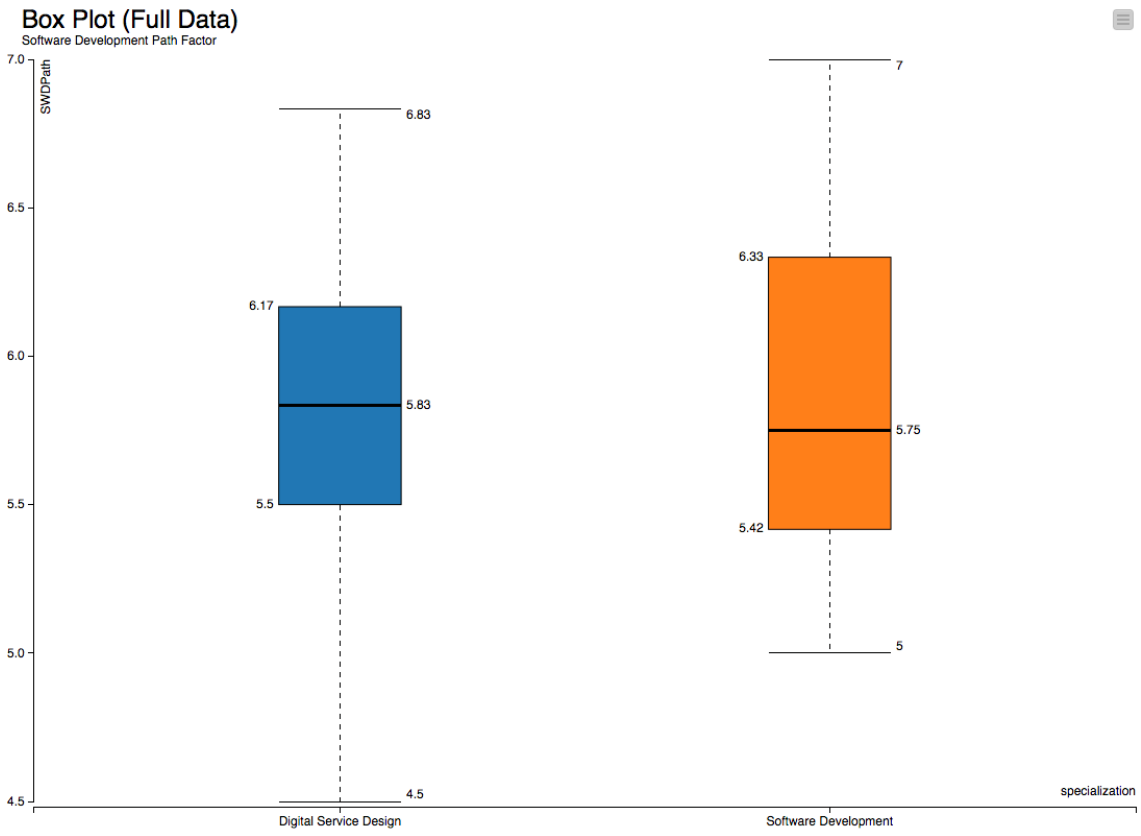
Appendix 3d. Respondents by Age Group



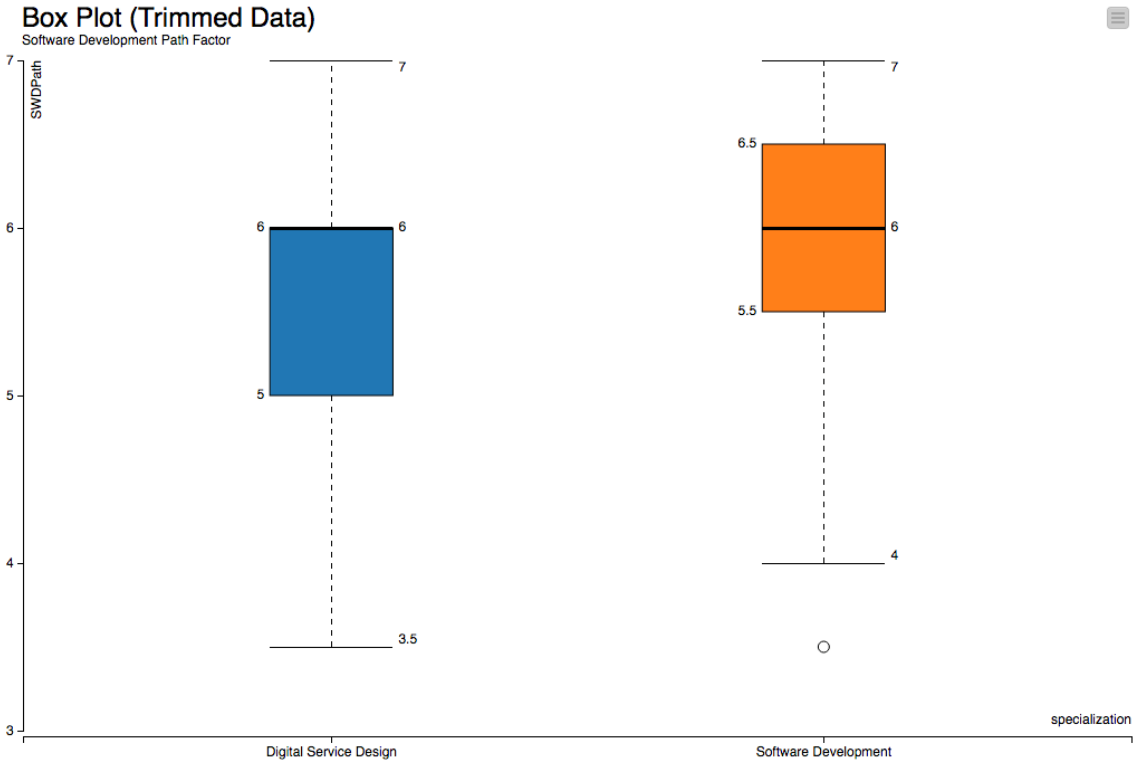
Appendix 3e. Respondents by Geographical Area of Origin



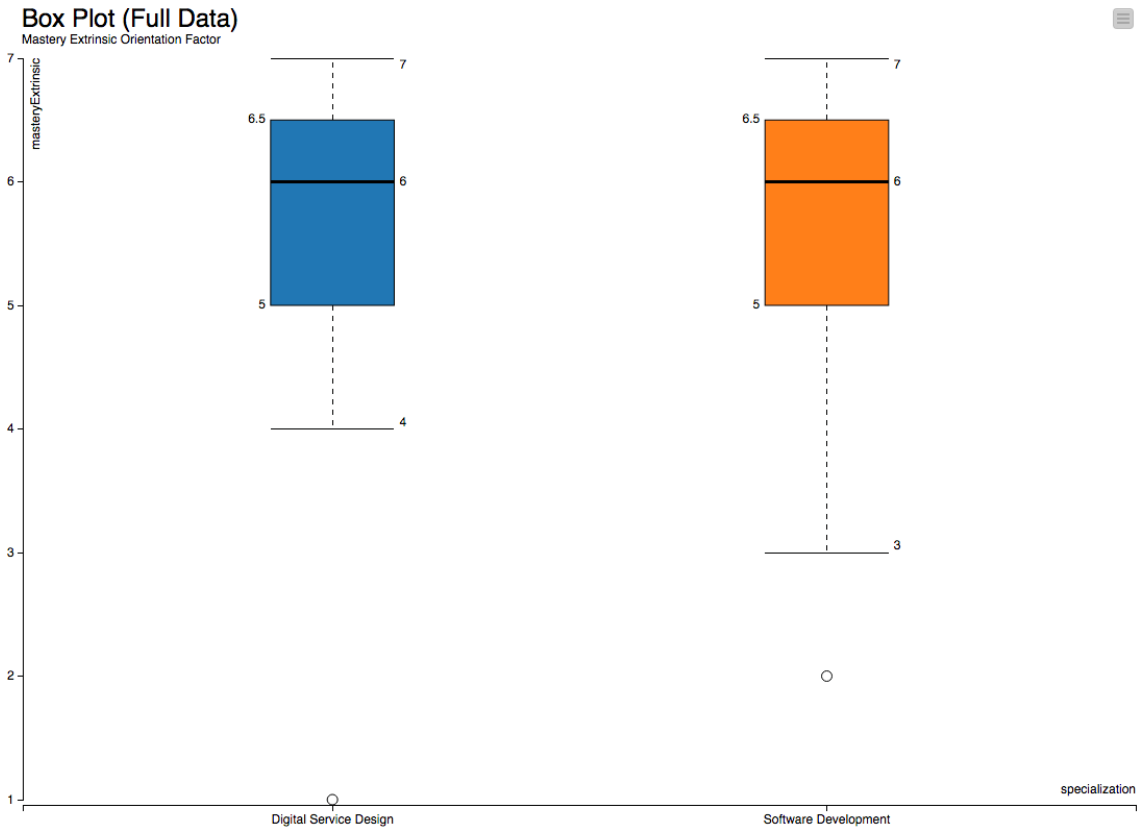
Appendix 4a. Boxplot of SWD Factor by Study Path (Untrimmed set)



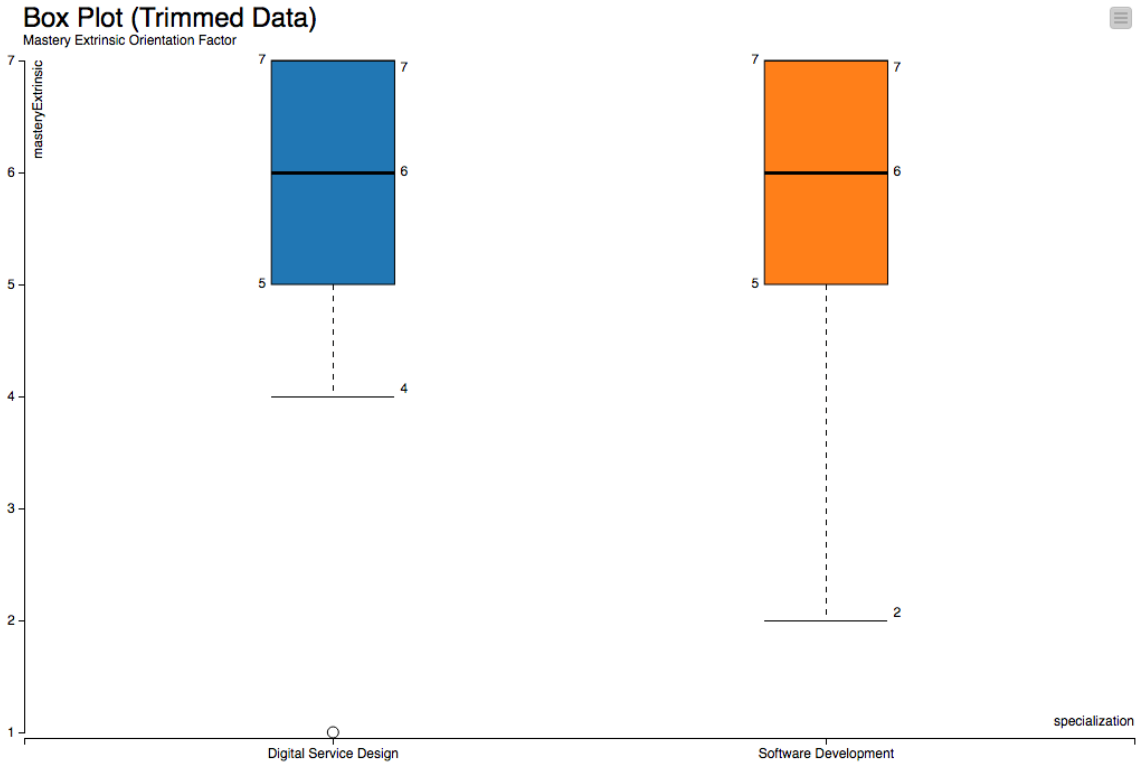
Appendix 4b. Boxplot of SWD Factor by Study Path (Trimmed set)



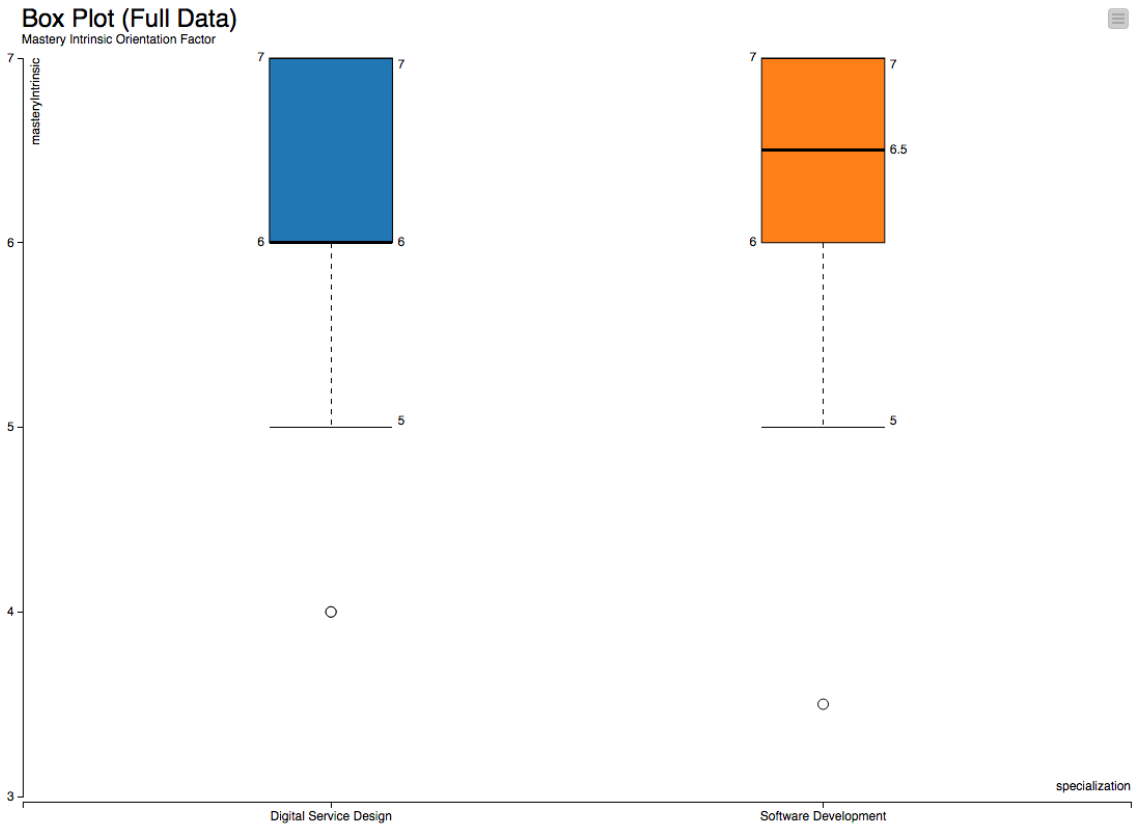
Appendix 5a. Boxplot of MEO by Study Path (Untrimmed set)



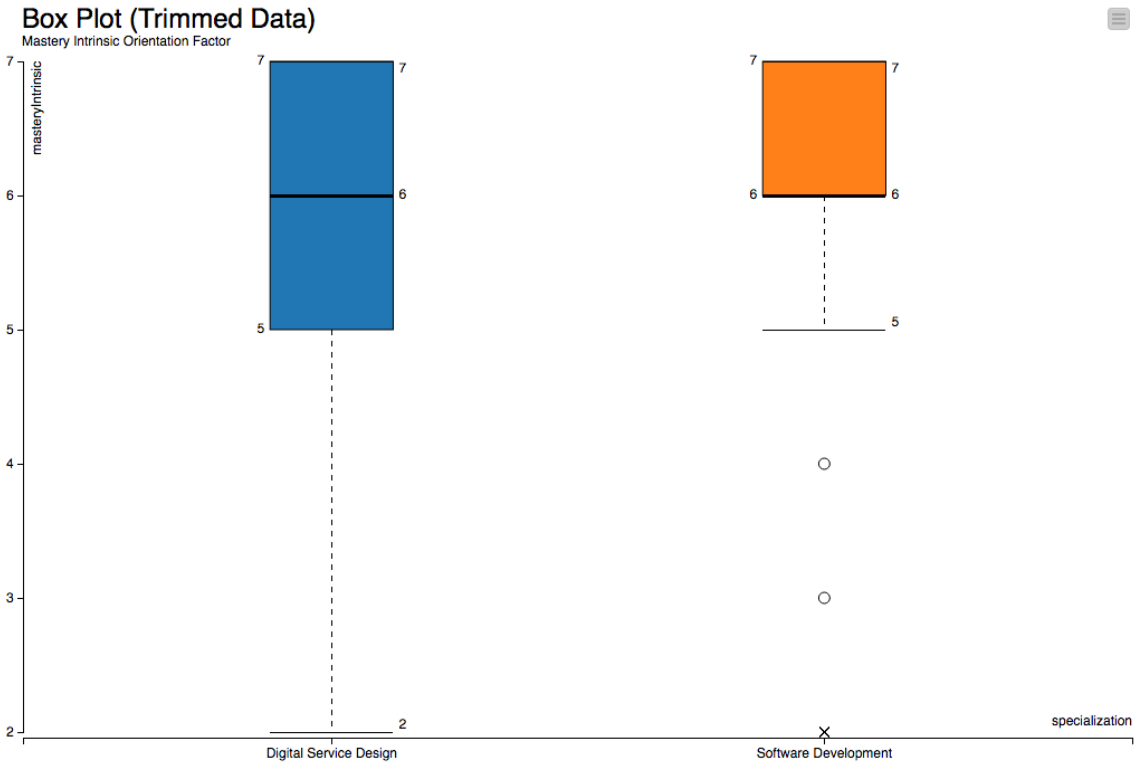
Appendix 5b. Boxplot of MEO by Study Path (Trimmed set)



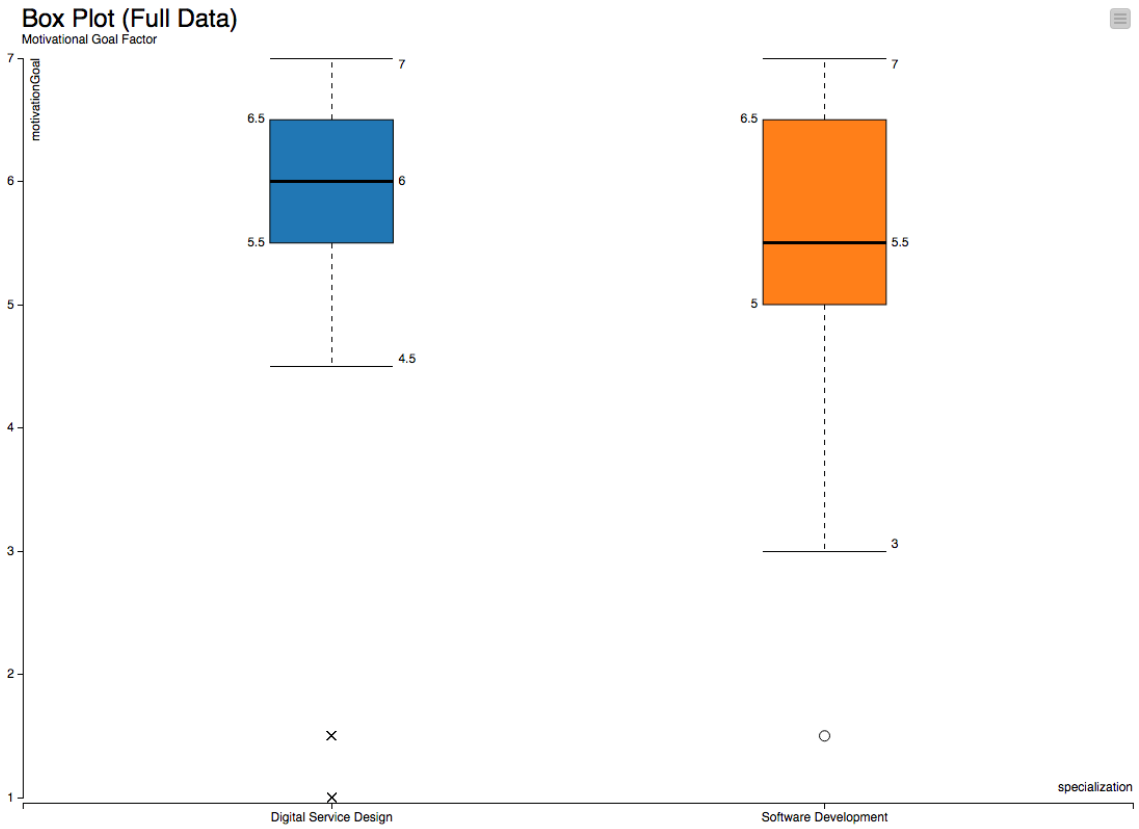
Appendix 6a. Boxplot of MIO by Study Path (Untrimmed set)



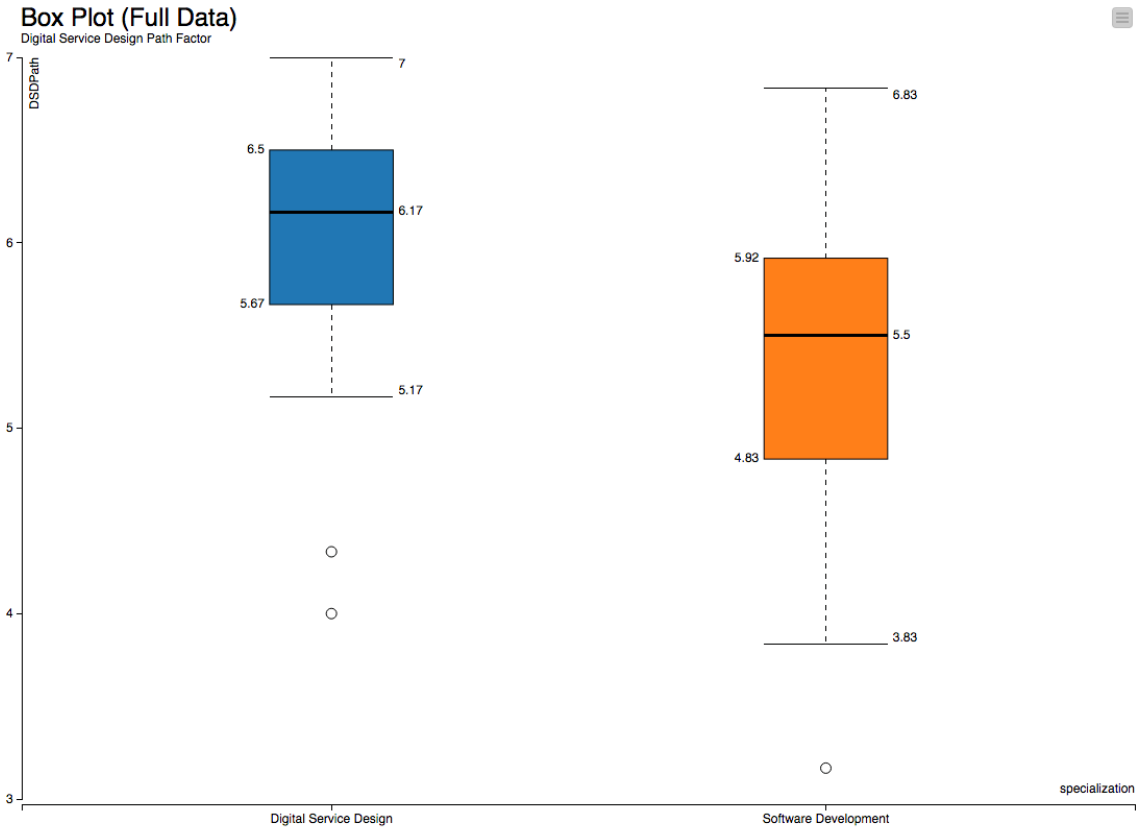
Appendix 6b. Boxplot of MIO by Study Path (Trimmed set)



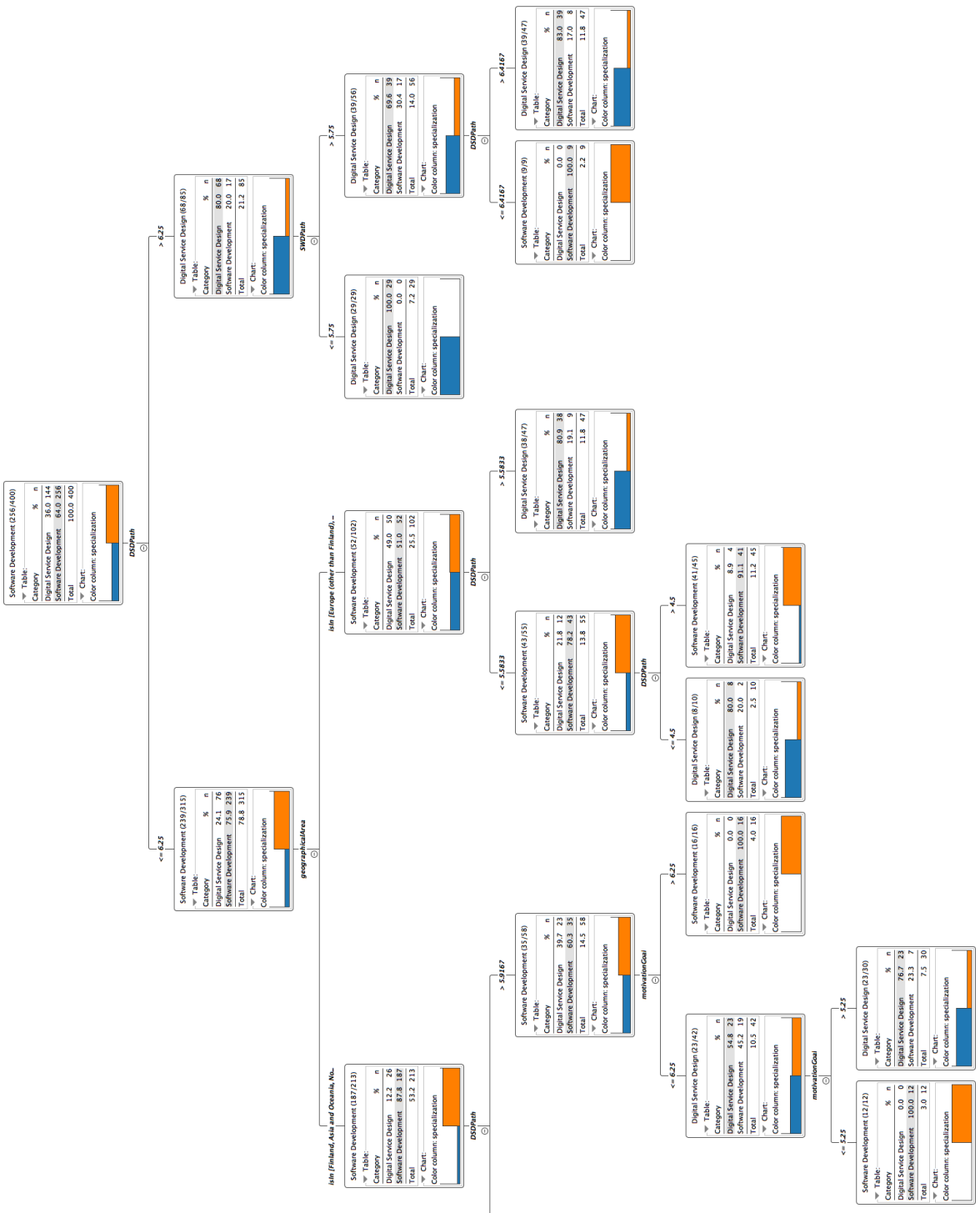
Appendix 7. Boxplot of MG by Study Path



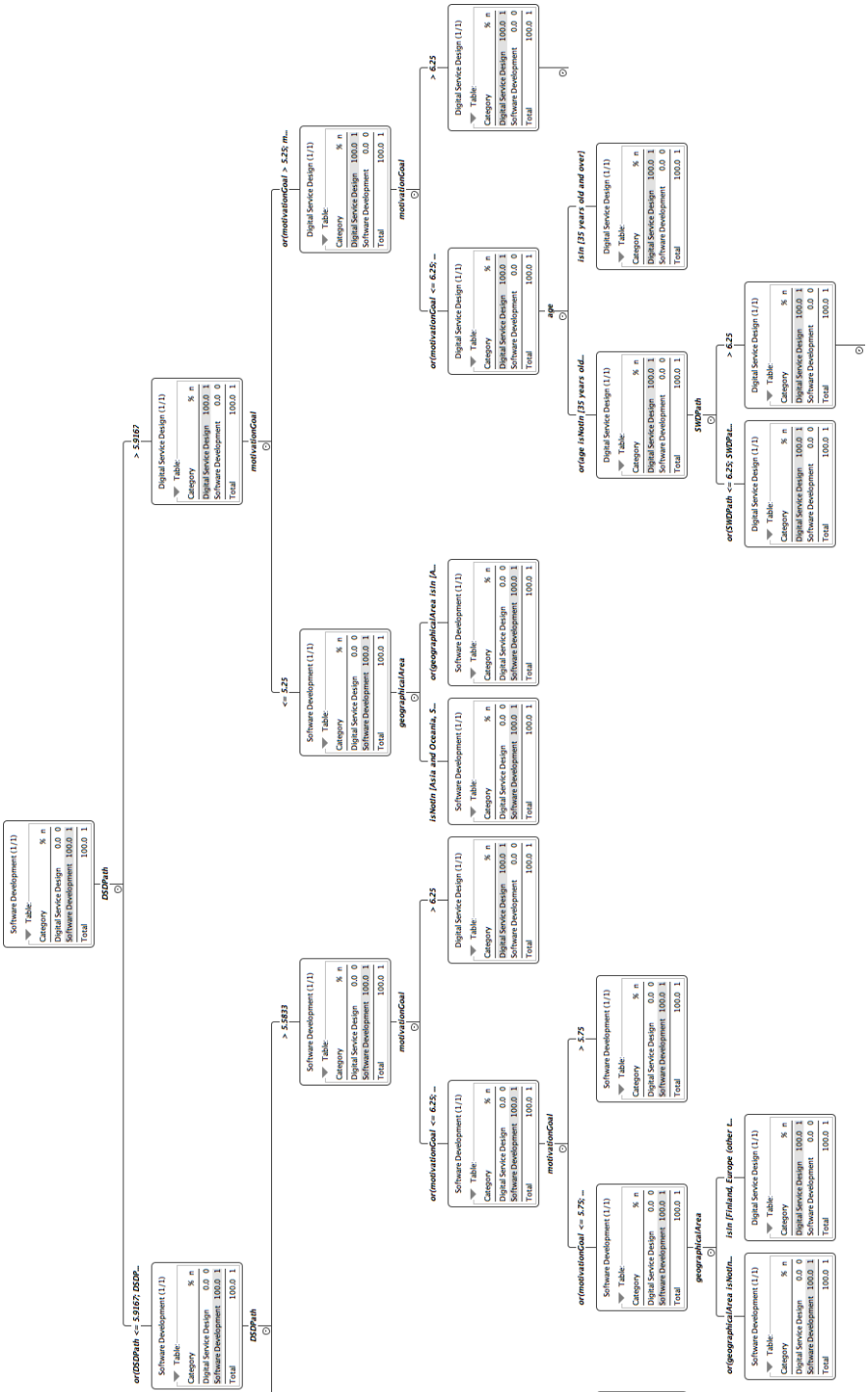
Appendix 8. Boxplot of DSD Factor by Study Path



Appendix 9. Full Decision Tree View of DTModel 26



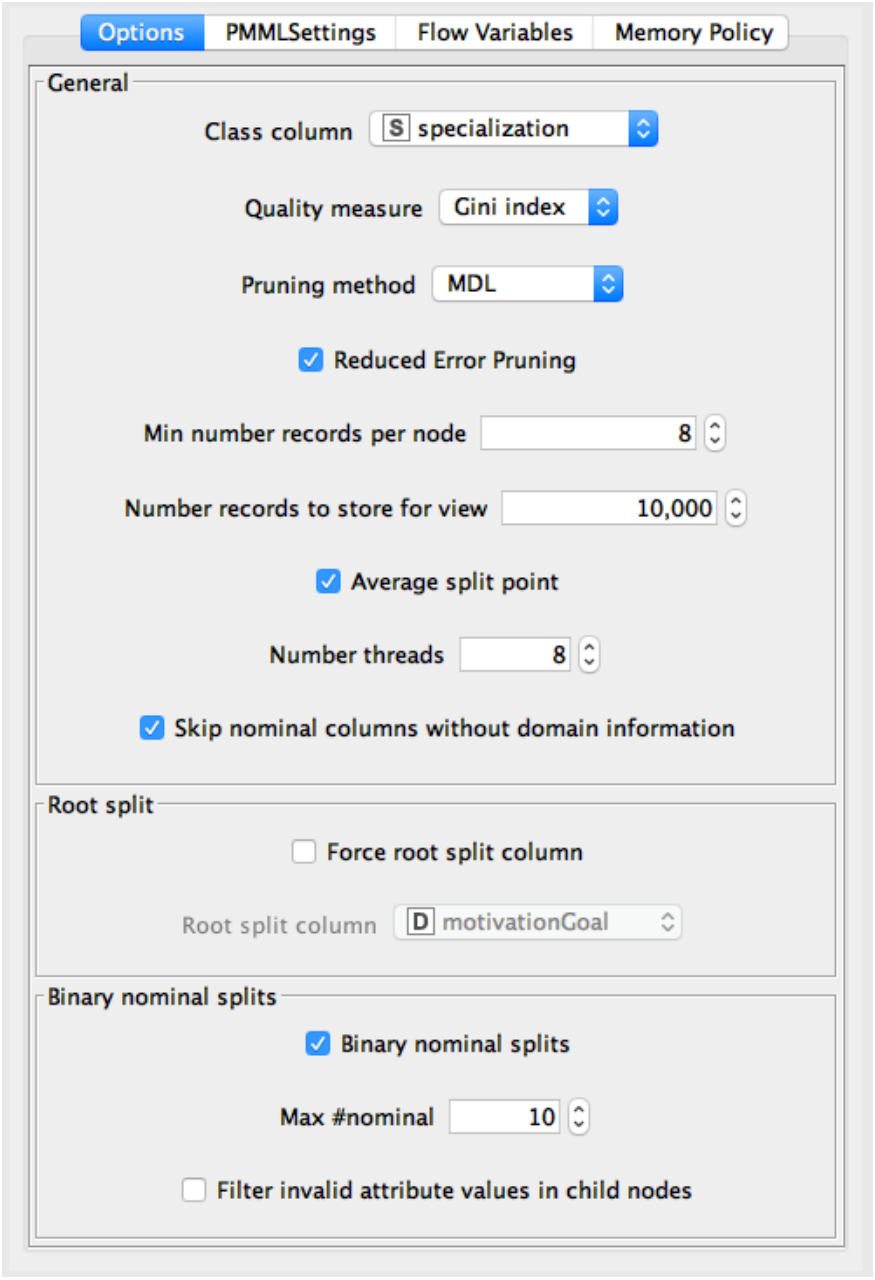
Appendix 10. A Sample Tree View of RFModel 22



Appendix 11. Random Seed Numbers Used in KNIME Nodes

KNIME Node	Random Seed
Bootstrap Sampling	1555809226459
Partitioning	1555809504806
Logistic Regression Learner	1557466433582
Random Forest Learner	1555812156666

Appendix 12. Decision Tree Learner Configuration



Appendix 13. Random Forest Learner Configuration