,

PREDICTING THE ON-TIME GRADUATION OF UNIVERSITY STUDENTS BASED ON THEIR STUDY PERFORMANCE

Lappeenranta-Lahti University of Technology LUT
Master's Programme in Business Analytics
Master's thesis
2023
Katja Hynynen
Supervisors: Associate professor Jan Stoklasa
Professor Pasi Luukka

Abstract

Lappeenranta-Lahti University of Technology LUT

LUT Business School

Business Administration

Katja Hynynen

Predicting the on-time graduation of university students based on their study performance

Master's thesis

2023

67 pages, 20 figures, and 4 tables

Examiners: Associate professor Jan Stoklasa and professor Pasi Luukka

Keywords: Educational data mining, machine learning, classification, k nearest neighbor, support vector machine, decision tree, on-time graduation

Objectives and expectations of universities have changed from autonomic Humboldtian institutions offering education for a small elite, to institutions driven not only by science, but also government and business world, and providing education for large masses. The lifelong study right is history, and the purpose of the universities is to educate the students efficiently and pursue to get them to graduate on time. The funding of the universities is dependent on the number of graduates and emphasizing on-time graduation.

Since the timely graduation of university students has become significant, it is important to be able to follow the progress of the students and predict their graduation times. The information can also be used for finding the students in need of additional support.

This study presents applications of educational data mining, and especially prediction of performance and characteristics of students. Methods used for prediction are regression and classification. The study reviews the procedure of data science workflow, including data collection and integration, handling missing data and outliers, as well as rescaling the data. Further, it explains exploratory data analysis, which provides understanding about the data. In model development phase, a few classifier algorithms, namely, k-nearest neighbor, decision tree, and support vector machine are introduced. Additionally, performance measures for model evaluation, such as accuracy, precision, and recall, as well as k-fold cross-validation method to be used for avoiding overfitting are addressed.

In this thesis, on-time graduation of electrical engineering students at LUT University studying consecutively bachelor's and master's degrees is studied. Data used for the classification includes performance of the students in the current studies, that is, cumulative credits after each study year and average grade.

As a result, it can be concluded, that the timely graduation of the students can already be classified with reasonable accuracy after the third study year.

Tiivistelmä

Lappeenrannan-Lahden teknillinen yliopisto LUT

LUT-kauppakorkeakoulu

Kauppatieteet

Katja Hynynen

Yliopisto-opiskelijoiden ajallaan valmistumisen ennustaminen opintomenestyksen perusteella

Kauppatieteiden pro gradu -tutkielma

2023

67 sivua, 20 kuvaa ja 4 taulukkoa

Tarkastajat: Tutkijaopettaja Jan Stoklasa ja professori Pasi Luukka

Avainsanat: Koulutusdatan analytiikka, koneoppiminen, luokittelija, lähimmän naapurin luokitin, tukivektorikone, päätöspuu, tavoiteaika

Yliopistojen tehtävä on muuttunut pienelle eliitille sivistystä tarjoavista autonomisista Humboldtilaisista laitoksista massoja kouluttaviksi instituuteiksi, joiden toiminnan ajureina toimii paitsi tiede, myös valtiovalta ja liike-elämä. Ikuinen opiskeluoikeus on jäänyt historiaan ja yliopistojen tulee kouluttaa opiskelijoita tehokkaasti ja pyrkiä saamaan heidät valmistumaan tavoiteajassa. Yliopistojen rahoitus riippuu valmistuneiden opiskelijoiden määrästä, painottaen tavoiteajassa valmistumista.

Koska opiskelijoiden ajoissa valmistuminen on tullut merkittäväksi, on tärkeää pystyä myös ennustamaan opiskelijoiden valmistumisaikoja. Tätä tietoa voidaan käyttää paitsi hallinnollisiin tarkoituksiin, myös havaitsemaan tuen tarpeessa olevat opiskelijat.

Tämä tutkimus esittelee koulutukseen liittyvän datan hyödyntämiseen liittyviä sovelluksia ja erityisesti opiskelijoiden suoriutumisen ja ominaisuuksien ennustamista, missä tkäytettäviä menetelmiä ovat regressio ja luokittelu. Tutkimuksessa käydään läpi, miten kerätty data tulee esikäsitellä luotettavien mallinnustulosten varmistamiseksi. Lisäksi esitellään kokeellisen data-analyysin menetelmiä, joiden tarkoitus on antaa ymmärrystä datasta. Mallinkehitysvaiheessa esitellään muutama luokittelualgoritmi: lähimmän naapurin luokitin, päätöspuu ja tukivektorikone. Myös mallin suorituskyvyn mittareita sekä mallin ylisovittumisen välttämiseen tarkoitettuja validointimenetelmiä käydään läpi.

Tässä opinnäytetyössä tarkastellaan LUT-yliopiston sähkötekniikan opiskelijoiden ajallaan valmistumista. Mukana ovat opiskelijat, jotka opiskelevat peräkkäin sekä kandidaatin että maisterintutkinnot. Luokittelussa käytetty data sisältää opiskelijoiden vuotuiset kumulatiiviset opintopistekertymät sekä tutkinnon keskiarvon.

Tutkimuksen tuloksena voidaan todeta, että opiskelijoiden ajallaan valmistumista voidaan ennustaa riittävällä tarkkuudella jo kolmannen opiskeluvuoden jälkeen.

Acknowledgements

This thesis work was started as a part of the learning analytics research project AnalytiikkaÄly, and continued later as an individual study.

I wish to express my gratitude to my supervisors associate professor Jan Stoklasa and professor Pasi Luukka for their guidance, support, and fruitful discussions. I also want to thank associate professor Harri Eskelinen and university lecturer Katriina Mielonen for the valuable discussions during the AnalytiikkaÄly project.

Special thanks to my colleague and fellow student Ahti Jaatinen-Värri for the numerous shared study sessions throughout our study time. Without such a great fellow student, the journey would have been not only tougher, but also more boring.

Last but not least, I am grateful to my husband Tero and daughters Hilla and Tuuli for the patient attitude toward my enthusiasm for continuous studies. I also want to thank Tuuli for the language revision of the thesis.

Lappeenranta, December 4th, 2023

Katja Hynynen

Contents

Nomenclature

Roman letters

b	bias
d	difference or distance of observation couple
D	data set containing predictor and target variables
k	number of folds, number of neighbors to be considered
K	Kernel function
n	number of observations or observation couples or classes
p	risk of mistake when rejecting the zero hypothesis
r	(Pearson) correlation coefficient
t	studentized correlation coefficient
x	predictor variable
$\bar{x}$	mean value of predictor variables
$\mathbf{x}$	predictor or regression vector
y	target variable
$\bar{y}$	mean value of target variables
w	weight coefficient

Greek letters

α	significance level, weight
σ	kernel parameter

Subscripts

i	index number
j	index number
E	Euclidean
p	number of features
S	Spearmann
x	predictor variable
y	target variable
0	new or to be predicted

Abbreviations

ACC	accuracy
AUC	area under ROC curve
CART	classification and regression trees
CSV	comma separated values
DT	decision tree
ECTS	European credit transfer and accumulation system
EDA	exploratory data analysis
EDM	educational data mining
ePSP	electrical personal study plan
FN	false negative
FP	false positive
GPA	grade point average
IEDMS	International Educational Data Mining Society
KNN	k nearest neighbor
LA	learning analytics
LOO	leave-one-out cross-validation
LOOCV	leave-one-out cross-validation
RBF	radial base function
ROC	receiver operating characteristics
SoLAR	Society for Learning Analytics
SQL	structured query language
SVM	support vector machine
TN	true negative
TP	true positive

1 Introduction

This chapter presents the motivation, background, and aim of the thesis. First, development of the objectives of universities in Finland are reviewed. Then, the performance targets and financing models of the universities, as well as study time of students are explained. Further, statistics of graduation times are introduced. Finally, the research aim and questions of the thesis are provided followed by the outline of the thesis.

1.1 Motivation and background

In this chapter, changing objectives of Finnish universities are presented. How the universities free from the societal pressure and with lifelong study right, turned into institutions serving economic life, where students should be encouraged to graduate quickly to serve society? Additionally, funding model and performance targets set to universities are introduced.

1.1.1 Changing objectives of the university

University education in Finland started at the University of Turku, originally the Royal Academy of Åbo, in 1640. The conception of university education was based on German humanism, later known as Humboldtian ideas. (Välimaa 2004; 31, 33) Humboldtian model of higher education emphasized the scientific education and moral growth of the students by combining research and teaching. Both students and universities had academic freedom making the universities free of the interest of the society and business, as well as offering the students lifelong study right. (Tirronen 2006; 126, 132) In the beginning, the university education was meant only for the elite and the purpose was to educate public servants and supervise the operation of society together with the church. At that time, the economic role of universities was minor. (Ahola 1995, 36-37, 151)

With industrialization, the importance of university education for economic growth was understood, and universities started to educate professionals for the needs of growing industrial and economic enterprises. The first special colleges, Helsinki Polytechnic School and Helsinki Business College were founded in 1872 and 1881, respectfully. Polytechnic School was raised to college of technology in 1908 and further to Helsinki University of Technology in 1960's. Other fields of education followed similarly. (Välimaa 2004, 37-38).

After the Second World War, universities began to be seen as a part of social policy, economic growth and developing welfare society. One of the important factors in welfare society was to provide equal possibilities for education, also higher education, for all the citizens. (Merenluoto 2009, 13)

In the late 1950s, a mass higher education system started to be formed in Finland. In the first phase, in 1950-1960, the purpose was to provide more degrees for increasing educational demand and educate more teachers for lower educational levels. The second phase in 1960-1970 was related to the welfare-developing politics. One of the main objectives was to provide equal educational opportunities for everybody, also at higher educational level. Together with regional politics, this led to founding of several universities in the major provinces of Finland in 1970s and 1980s. Some of the universities were created by raising their status from college to university, such as University of Jyväskylä from the former teacher training college, and University of Tampere from the College of Social Sciences. The third phase in 1970-1980, was a techno-economical phase, where the purpose was to provide more labor force for the private sector promoting production and economy. (Välimaa 2004, 38-39; Lampinen 1998, 115-116)

Still in 1997, the university act expressed the mission of universities as follows:

"The mission of the university shall be to promote free research and scientific and artistic education, to provide higher education based on research, and to educate students to serve their country and humanity." (Universities Act 1997, section 4)

However, initially autonomic Humboldtian institutions had become closer to operation of society, their activities were no longer driven only by science, but also government and business world (Tirronen 2006, 125). The orientation was from academic freedom and lifelong study right of Humboldtian model to effective Anglo-Saxon model with tighter study rights and systematic study guidance. In Anglo-Saxon model, the universities are also responsible of the graduation of students. (Eriksson & Mikkonen 2003, 21) Since the expectations of university education and research had changed significantly from the Humboldtian idea, the mission of the universities in University Act was also completed in 2005 as follows:

"The mission of the university shall be to promote free research and scientific and artistic education, to provide higher education based on research, and to educate students to serve their country and humanity. In carrying out their mission, the universities shall interact with the surrounding society and promote the societal impact of research findings and artistic activities." (Universities Act 1997, section 4, amendment 715/2004)

1.1.2 Long graduation times became a problem

The long graduation times were considered a problem since 1960s. The reason was that students studying for longer times were out of labor market, and thus, brought losses for the national economy. (Pajala & Lempinen 2001, 1) This led to a degree reform that was meant to enable graduation in four years in all the fields of university studies. However, the courses were undermeasured in relation to the requirements and real workload, which led to increasing workload and longer graduation times in 1980s. (Lehtisalo1999, 156) There was an attempt made to shorten graduation times by extending semesters, adding exams outside semesters, and decreasing the workloads. (Mikkonen 2000, 16-17) Additionally, the universities had to make their operations more efficient and improve their performance in terms of graduation grades, as well as take up the performance evaluation system, and report about the results (Lehtisalo & Raivola 1999, 162).

A new reform was taken into use in 1990s when a two-tier degree structure was adapted to all fields apart from technical and medical degrees. In two-tier degree structure, there is first a three-year bachelor's degree followed by a two-year master's degree. The purpose was to shorten the study times to 5-6 years by improving admissions, degree structures, content of studies, and teaching methods. However, the graduation times still did not shorten. (Lehtisalo & Raivola 1999, 157-158; Opetusministeriö 1993, 8)

In 2005, two-tier degree structure was adapted also in the technical degree programmes at the same time with joining the Bologna process (University Act 426/205, 5§). However, in practice, the students have been able to study master level courses before graduating as bachelors. This has led to long study times in bachelor's studies and in the worst case, the students have been studying the last basic bachelor's courses when the master level studies have already been finalized. At LUT University, tighter two-tier degree regulation was put into practice in autumn 2021 such that students can start their master's studies only after finalizing their bachelor's degree, or they are missing no more than 12 ECTS. (LUT University 2023 a)

1.1.3 Performance targets and funding model

Universities' core funding model changed in the beginning of 2021. The new model can be seen in Figure 1.1. 42% of the funding is based on educational achievements, 34% based on research, and 24% on considerations of other educational science policies. 19% of the

funding is delivered based on the number of graduates in master's programmes and 11% of graduates in bachelor's programmes. (Ministry of Education and Culture 2021 a)

42% Education	30% Bachelors's degrees (11%) and Master's degrees (19%)
	5% Continuous learning
	4% Number of employed graduates and quality of employment
	3% Student feedback
34% Research	8% PhD degrees
	14% Scientific publications
	12% Competitive research funding
24% Other education and science policy considerations	15% Strategic development
	9% National duties

Figure 1.1. Universities core funding model from 2021. (Ministry of Education and Culture 2021 a)

Additionally, the number of degrees is determined based on the graduation time using the following weighting factors:

- graduated on time, weighting factor 1.5
- graduated at the maximum 12 months after the target time, weighting factor 1.3
- graduated over 12 months after the target time, weighting factor 1 (Ministry of Education and Culture 2019, 1 §)

According to the University Act (2009, section 48), universities agreed on the university specific targets and follow-up indicators with Ministry of Education and Culture based on government programme, action plan of government, vision for higher education and research, and strategy of the university. The yearly targets for graduates at LUT University during 2021-2024 are 890 master's, of which 630 in technical programmes and 260 in business administration, and 800 bachelor's. (Ministry of Education and Culture 2021 b, 5)

1.1.4 Target time and scope of a degree

In the two-tier degree structure, there are the lower bachelor's degree and higher master's degree. Target time for the bachelor's studies in the field of technology is three years and for the master's studies two years. In Finland, students admitted for bachelor's programmes have right to continue directly to master level studies within same degree programme. Target time for both bachelor's and master's studies is thus five years. However, the study right is two years longer than target time, that is, in this case seven years.

The scope of bachelor's studies is 180 and master's studies 120 ECTS (European credit transfer and accumulation system) credits. In order to graduate in target time, student is supposed to study 60 ECTS credits yearly. One ECTS credit equals approximately 27 working hours for the student. (LUT University 2023 b;5, 10, 20-21) That means that yearly workload is 1620 hours. When dividing that throughout the academic year from the beginning of September to the middle of May (35 weeks), the required weekly working time becomes 46 hours. It may be possible at some universities and in some fields of studies to take courses during summertime, and thus, divide the workload more evenly during the year. However, at least in the technical field, the students mainly work during the summers.

1.1.5 Statistics of graduation times

In this subchapter, statistics of graduation rates in different fields of university studies in Finland are reviewed. Students studying successively both bachelor's and master's degrees are considered. Target time for graduation may vary depending on the field of studies, but in engineering it is five years.

Figure 1.2 shows the graduation rate of students that started their studies at Finnish universities during 2005 – 2017 (and graduated by 2021). The students are from educational sciences, social sectors, business administration and law, natural sciences, engineering, as well as health and wellness sectors. It can be seen that in some fields of studies, it is not unusual to complete the studies faster than in five years. However, it can also be seen that in natural sciences and especially in engineering, it is very rare. In engineering, only 5% of the students have graduated in five years. After six years the number goes up to 17% and after seven years up to 38% of the students have graduated. It can also be seen that a large share of the students haven't graduated in 10 years. In practice, not all the students starting the studies graduate, but many of them change to other fields of studies, change degree programme, or even transfer university.

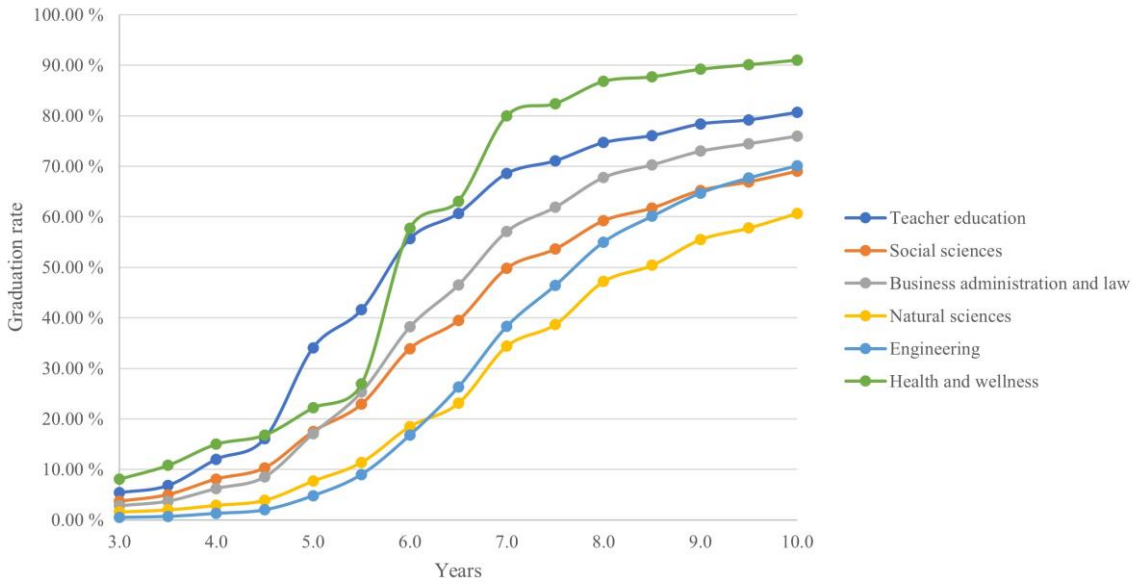


Figure 1.2. Graduation rate of university students in different fields of science during 2005 – 2021. (Vipunen 2023)

Figure 1.3 shows the graduation rate of engineering students that started their studies and graduated at Finnish universities during 2005 – 2021. Graduates are from chemical and process engineering, environmental protection technology, electrical and energy technology, electronics and automation, and mechanical engineering. It can be seen that for the students of environmental protection technology, the study times are slightly shorter compared with others, and chemical and process engineering students are the second fastest. Electronics and automation students proceed the slowest from the fields considered. In the target time of five years, 6% of environmental protection technology students and 3% or electronics and automation students have graduated. After six years, the shares are 26 and 8%, respectively.

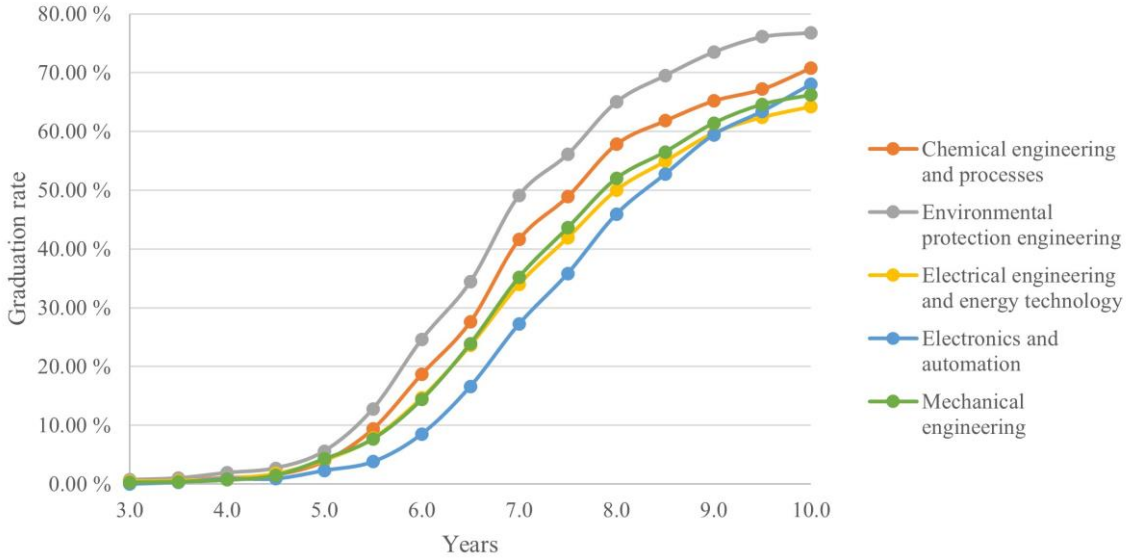


Figure 1.3. Graduation rate of university students in different fields of engineering during 2005 – 2021. (Vipunen 2023)

Figure 1.4 shows the graduation rate of engineering students that started their studies and graduated at LUT University during 2005 – 2021. Graduates of chemical and process engineering, environmental protection technology, electrical and energy technology, and mechanical engineering are included. Figure shows that students of environmental protection technology graduate faster than others at LUT University as well, and students of chemical and process, and mechanical engineering proceed the slowest. In the field electrical engineering and energy technology, further considered in this study, 6% of the students have graduated in the target time of five years. After six years, 22% and after seven years, 46% of the students have graduated.

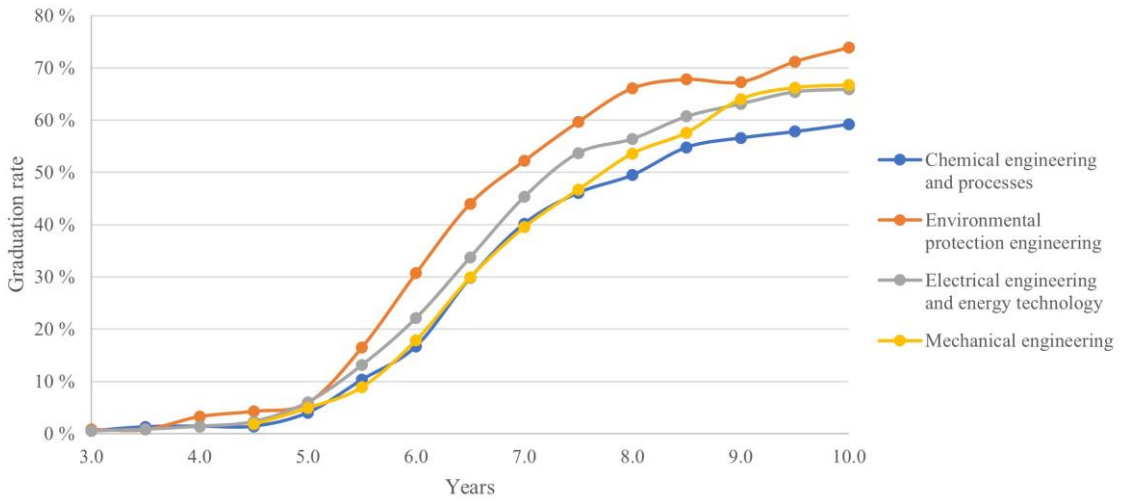


Figure 1.4. Graduation rate of the students in different engineering fields at LUT University during 2005 – 2021. (Vipunen 2023)

Figure 1.5 shows the evolution of the graduation rate of the students graduating in six years in different fields of sciences and at all Finnish universities. Students that started their studies during 2005 and 2015 are considered. It can be seen that the graduation rates have been quite stable until the academic year 2011-2012 in other fields of sciences except for the health and wellness. After that, the graduation rates have started to increase.

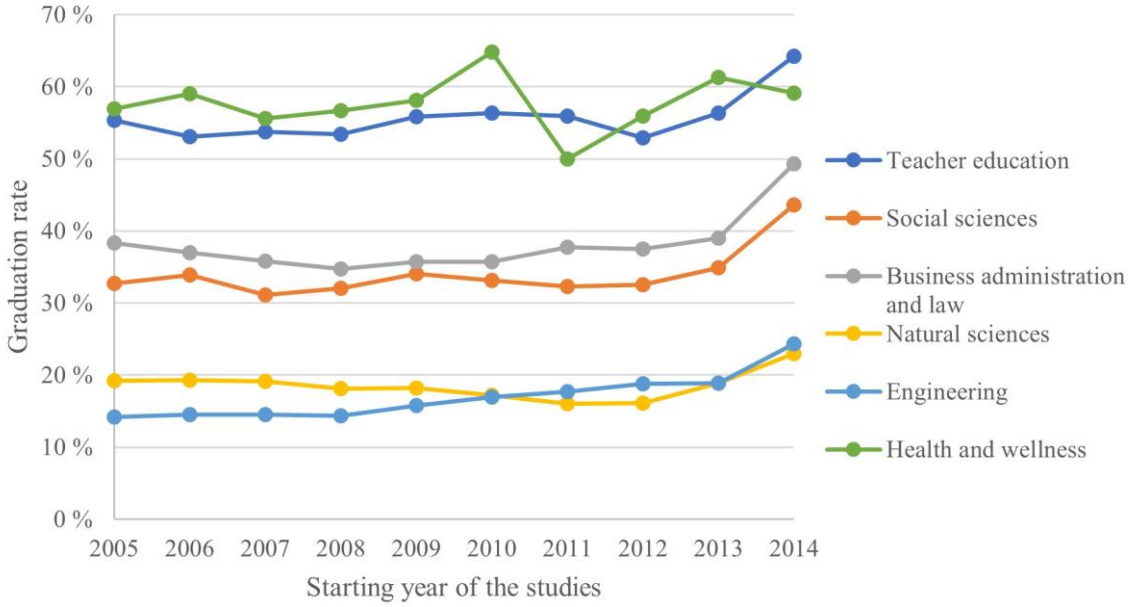


Figure 1.5. Evolution of graduation rates of the students graduated in six years at Finnish universities. (Vipunen 2023)

When considering the graduation rates after the target study time of five years, same kind of evolution hasn't occurred, but the graduation rates have stayed quite equally at the levels shown in Figure 1.2. Same trends can be seen in the evolution of the graduation rates in the different engineering fields at LUT University.

1.2 Research aim and research questions

Changed objectives of universities and the funding model connected to the performance targets have forced universities to consider more carefully how to make sure that students are able to not only gain the learning outcomes of the programmes, but also graduate in target time. There is need for more systematic follow-up on performance of students and predicting their graduation times. This information can be used not only for administrative prediction of graduates, but also for finding the students in need of additional support. The information may also help to find possible structural challenges in the study programmes.

The purpose of this study is to design a model that classifies whether the university students graduate on time or behind schedule. The model will help administrative personnel to make yearly estimates of the number of graduating students and students graduating on time, as well as find students in need of additional support in their studies.

The main research question of this study is as follows:

At which point of the studies, is it possible to reliably predict whether the students will graduate on target time or not?

Minor research questions are:

1. Which of the available features correlate with the graduation time?
2. Which classifier provides the most reliable results?

The research questions are studied using the following methods:

1. Correlation analysis to find linear correlations between features used and graduation time.
2. Classifier algorithms for predicting on-time versus over on-time graduation time, and common evaluation methods to measure their performance.

The data used is obtained from base register OODI formerly used at LUT University, as well as from register of DIA selection. The data is from 2010 to 2020.

In this study, the following delimitations, limitations, and assumptions are made:

- The analysis is delimited in the students of electrical engineering at LUT University studying successively bachelor's and master's degrees during 2010 – 2020.
- Data used, only includes graduated students.

- Data used, is chosen such that it would be available in student register and easy to process also in the future, that is, data already continuously collected during the studies and admission, and further, quantitative data.
- It is assumed, that the structure and workload of all the students is approximately the same regardless of slight changes in the study structures during the considered years.

1.3 Outline of the thesis

The thesis is divided into five chapters with the following outline:

Chapter 1 provides the motivation, background and objectives of the thesis. Development of the objectives of universities in Finland, the current performance targets and financing models of the universities are explained. Further, study time and statistics of graduation times are reviewed.

Chapter 2 presents the research fields of educational data mining and learning analytics. Different methods related to educational data mining are reviewed, and deeper look into one of the methods, namely predicting student performance and especially timely graduation, is addressed.

Chapter 3 deals with the procedure of general data science workflow, concentrating in data preparation, data analysis and model development. Data preparation contains collecting, aggregating, and preprocessing data. The purpose of data analytics is to find interesting patterns from the data using statistics and visualization. Further, in model development, machine learning-based prediction, especially classification algorithms are presented. And finally, evaluation and validation of the model are discussed.

Chapter 4 presents the results of the study. First, the data used is presented using descriptive statistics and visual graphs. Linear relationship between the predictive and target variables is determined using correlation analysis. Further, binary classifiers are trained to predict the on-time graduation of students. Finally, discussion of the results is provided.

Chapter 5 summarizes the results and gives suggestions for future work.

2 Data analytics in education

This chapter reviews the previous studies published in the field of data analytics in education. It starts from the history of establishing the communities of Educational Data Mining and Learning Analytics, and how they started to develop knowledge and tools for the challenges aroused in the field. Further, methods related to educational data mining are presented. Finally, deeper look into one of the educational data mining methods, namely, prediction of on-time graduation is addressed.

2.1 Educational data mining

Educational information systems have increased and diversified during the last several years. There are information systems for registering the basic information of students together with the studied credits and grades from the courses, systems allowing students to make their (electrical) personal study plan (ePSPs) and register for the courses. Additionally, there are information systems for planning and publishing study guides, and lecture and exam schedules. Further, along with increasing use of e-learning tools, educational software and other web-based educational systems allow collecting and storing huge amounts of information related to study, learning and development of students. All this information enables administrators and teachers to improve the study programmes, courses, and counseling of students. However, there is a huge amount of data in versatile formats. It is not possible to transform it to an understandable and useful form manually. (Romero & Ventura 2013, 12; Romero & Ventura 2020, 1)

Two research communities have grown to develop knowledge and tools for the challenges described: International Educational Data Mining Society (IEDMS) and Society for Learning Analytics (SoLAR). The roots of IEDMS are in the first workshop on Educational Data Mining that was organized in Pittsburgh in 2005. The workshop was further organized and developed as International Conference on Educational Data Mining in 2008. SoLAR was founded in 2011 after the First International Conference on Learning Analytics & Knowledge in Canada. (Siemens & Baker 2012, 252)

Educational Data Mining society determines educational data mining (EDM) as follows: “Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they

learn in." (Educational Data Mining 2023) EDM methods include the standard data mining methods, such as, visualization, prediction, clustering, outlier detection, relationship mining, causal mining, social network analysis, process mining and text mining. However, since the data available have intrinsic information, relationships with other data, and multilevel hierarchies, methods such as distillation of data for human judgment, discovery with models, knowledge tracing and non-negative matrix factorization are also used. (Baker & Yacef 2009, 4; Romero & Ventura 2020, 14)

Learning analytics (LA) includes measuring, collecting, analyzing and reporting data related to students and their learning performance in order to better understand and optimize their learning. The aims of LA applications include improving the performance of students and whole faculty, improving understandability of course material, improving the finding and monitoring the students with challenges, improving the teaching and assessment as well as using the educational resources more effectively. (Calvet Liñán & Juan Pérez 2015, 103)

Educational data mining (EDM) and Learning analytics (LA) have many common aims and interests. They both aim to improve the planning and decision-making in education based on data obtained from educational processes. However, the ideologies, methodologies, and also technologies of EDM and LA differ. EDM emphasizes analysis of individual system components and their relationships, whereas LA emphasizes the understanding of the general view. In EDM, automated fitting or prediction and discovery is important, whereas in LA, automated adaptation may be used as a tool, but the final decisions are often made by human judgement. (Siemens & Baker 2012, 252-253)

The rest of this chapter concentrates on Educational Data Mining, first reviewing its applications in general, and focusing further on one of its applications, namely, predicting performance of students and timely graduation.

2.2 Applications of educational data mining

Applications of educational data mining can be categorized in several ways. In this study, categorization presented by Bakhshinategh et al. (2018) is chosen. Other types of categorizations can be found, for example in research of Baker & Yacef (2009, 6-8) and Romero & Ventura (2010, 603). Bakhshinategh (2018) divides applications of EDM in five categories: Student modeling, decision support system, adaptive system, evaluation, and scientific

inquiry. Student modeling and decision support systems are further divided in several subcategories, as can be seen in Figure 2.1. (Bakhshinategh et al. 2018, 541)

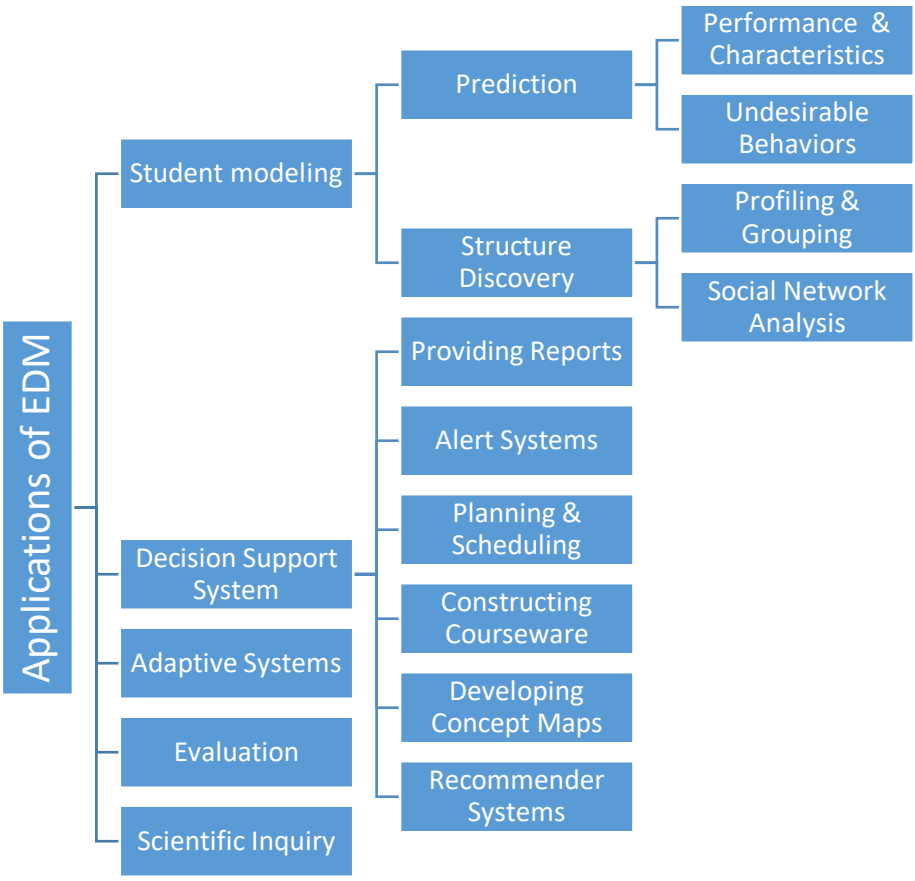


Figure 2.1. Application categories of educational data mining. (Bakhshinategh et at. 2018, 541)

Student modeling is the most studied category with two subcategories: prediction and structure discovery. In prediction, the characteristic of interest is usually known, whereas in structure discovery, it is either not known at all, or it may be known just as a structure. The difference between the two categories is not always clear. (Bakhshinategh et al. 2018, 542) The applications predicting performance of students concentrate on estimating some characteristic describing the student. It may be, for example, academic performance, motivation, satisfaction, or learning style. The methodologies used for predicting students' performance are regression and classification. (Romero & Ventura 2010, 607) Since this study considers prediction of students' performance and characteristics, the topic is further explored in

Subchapter 2.3 Predicting graduation times. Applications predicting undesirable behavior are similar to the previous category, but they concentrate on identifying students with some kind of challenges, such as low motivation, misuse, cheating, or dropping out. The most commonly used methods are clustering and classification, but feature selection and outlier detection have also been used. (Bakhshinategh et al. 2018, 542-543)

Applications of profiling and grouping tend to group the students based on their personal characteristics and knowledge such that the students inside the group would complement each other. For example, when forming groups for interdisciplinary project work, students with different skill profiles are required for each team. Suitable methods for profiling and grouping, are for example, clustering and feature selection. In the social network analysis, the aim is to model, not only individuals as in the other student modeling applications, but also relationships between them. Examples of social network analyses are modeling of group dynamics and cohesion of students. (Bakhshinategh et al. 2018, 543)

Decision support systems provide reports mainly for teachers, but also for students and administrators usually based on the results of student modeling applications. Providing reports contains data analytics and visualization of the results obtained in the applications of student modelling. Most commonly used methods are statistics and visualization. (Bakhshinategh et al. 2018, 544; Romero & Ventura 2010, 604) Results of alert systems are usually based on student characteristics or undesirable behavior, and the purpose is to find possible cases of, for example, low motivation, misuse or cheating. Commonly used methods are statistical clustering and classification. Both alert systems and providing reports serve teachers and administration. (Romero & Ventura 2010, 608) Applications of planning and scheduling help teachers and administrators, for example, in developing curriculum, courses and counseling processes. Also, course enrollment planning that serves students, belongs in this category. Various methods, such as clustering, classification, and discovery with models have been used. Courseware construction applications help teachers to create course material, such as videos and tests, automatically. Methods used are association rule mining and collaborative filtering. (Bakhshinategh et al. 2018, 544-545) Concept mapping applications are meant to help teachers automatically construct a concept map. Methods used include association rules and text mining. (Romero & Ventura 2010, 609) Recommender systems can be used, for example, to recommend courses for the students or test items for the teachers based on the personalized information of students. Commonly used methods are collaborative

filtering, content-based methods, association-rule based algorithms, and combinations of them. (Bakhshinategh et al. 2018, 545)

Adaptive systems can be used in online learning systems to provide personalized learning experience for students. Adaptation may take place in amount of instruction and tips, course material, and tests. Evaluation applications provide help for teachers in evaluation of ill-defined domains when the definitive solution is missing, or it is dependent on the problem's conception. Scientific inquiry refers to testing existing theories to different data, as well as developing new theories. (Bakhshinategh et al. 2018, 546)

2.3 Predicting graduation times

Low graduation rate and long study times has been a challenge worldwide, and prediction of drop-out rates and timely graduation serves purpose in several studies.

Many studies consider drop-out rate of the students, that is, whether or not they graduate. Typically, demographic data, performance in previous as well as current studies are considered as significant predictor variables when predicting the retention of students. (Reason 2003, López Guarín, León Guzmán & González 2015, Hannaford et al 2021)

Considering the topic of this study, it is more interesting to consider earlier research in predicting timely graduation versus slower graduation. Graduation times of students have tend to been predicted with several machine learning algorithms, typically support vector machine (SVM), decision tree, k nearest neighbor (KNN), and Naïve Bayes. The range of predictors have varied from the personal information of the students, demographic data, information from the previous studies and entrance exam, as well as performance in the current studies.

Mohammad Suhaimi et al. (2019) studied whether the engineering and science students graduate on time, by using five classifier algorithms: support vector machine both with poly kernel and radial base function (RBF) kernel, decision tree, random forest and naïve bayes. The predictors used in the study were mostly students' personal information, such as gender, age, race, and permanent address, but also admission group and cumulative grade point average (GPA). The accuracies differed between the study programmes. For students in the science programme, the best accuracy of classification was nearly 85% and for engineering programme 83%, both obtained using support vector machine poly kernel. (Mohammad Suhaimi et al. 2019; 130, 136)

Lesinski, Corns & Dagli (2016) studied graduation success of US Military Academy students based on high school data, admission test scores, extra-curricular activity score, and parent's educational status using neural network with 50 hidden neurons. The model classified with 95% accuracy, whether the students graduated on time, late, or if they didn't graduate at all. (Lesinski et al. 2016; 375, 378, 380)

Tampakas et al. (2019) studied graduation times of students at the School of Health & Social Welfare of Technological Institute of Western Greece with a purpose of finding students at risk of dropping off. Graduation time was predicted in tree classes (within 4, 5 or 6 years) using six classifiers, namely Naïve Bayes, neural networks, support vector machine, decision tree, rule-learning technics, and instance-based learner. Demographic information and performance in selected courses during the first two years of studies were selected as predictor variables. The accuracy of predictions was 60.8 – 77.39%. The best classifier was decision tree. Accuracy of support vector machine was 66.33%. (Tampakas et al. 2019; 1, 3-4)

Pang et al. (2017) predicted graduation time of students using support vector machine classifier with RBF kernel having gamma and cost parameters optimized with a few different methods. Predictor variables consisted of about 100 features including students' demographic data, high school and college performance, as well as personal thoughts and attitudes toward the studies. The best accuracy obtained for the classification was 80.59%. The precision and recall were 80.71% and 87.85%, respectfully. (Pang et al. 2017; 1,4,6)

Peling et al. (2017) studied the timeliness of veterinary students using Naïve Bayes algorithm. Predictors used in the study were entrance path, gender, and credits studied per semester. Accuracy was used for performance testing of the results, and the accuracy of classifying whether the students graduated on time was 86%. (Peling et al. 2017, 55-56)

3 Machine learning-based prediction

This chapter reviews the procedure of data science workflow and concentrates especially on data preparation, data analysis and model development. Data preparation deals with collecting and aggregating data, as well as preprocessing, such as handling outliers. Then, data analytics deals with finding interesting patterns from the data using statistics and visualization. Further, model development dealing with prediction and especially classification algorithms are presented. Finally, evaluation and validation of the model are discussed.

3.1 Procedure of data science workflow

The procedure of data science workflow includes the following steps:

1. Problem definition
2. Project organization
3. Problem knowledge acquisition
4. Data preparation
5. Data analysis
6. Model development
7. Model deployment
8. Model maintenance

Problem definition includes objectives of the project, desired products, and boundaries. Running a project requires taking care of organizational issues, such as finding team members, necessary hardware and software, as well as making the initial schedule and budget plan. Problem knowledge acquisition includes searching for previously published information about the topic. Data preparation includes collecting, aggregating and preprocessing the data used. After the data preparation, the data is ready for analysis. Data analysis gives insight into the data and includes a selection of variables used, as well as feature definition including both feature extraction and creation, and data visualization. Model development step includes building and validating a mathematical model providing a solution to the problem defined in step 1. In model deployment step, the mathematical model is resettled from the development environment to the production environment after which it is ready for the final users. Model maintenance ensures that the model developed is functioning correctly and provides support for its users. (Kordon 2020, 192-195) The process is not necessarily

straightforward and may require several iterations and returning back to the previous steps. (Fawcett & Provost 2021, 27)

This chapter concentrates on steps 4-6. Steps 1 and 3 have been explained in Chapters 1-2. This study does not include model deployment, and thus, steps 7-8 are out of its scope. However, the further use of the model is considered in Conclusions.

3.2 Data preparation

Purpose of data preparation is to gather the required data and prepare it for the analysis. There are slightly different ways to present the steps and their order depending on the reference. In this study, data preparation is divided in three main steps as shown in Fig. 3.1, combining the procedures used in Kotsiantis, Kanellopoulos & Pintelas (2006), Kordon (2020) and Romero, Romero & Ventura (2014, 40-56).

Fig. 3.1. Data preparation procedure.

Data can be collected in many ways and from several sources because it may be generated, not only in different places, but also at different times (Romero, Romero & Ventura 2014, 41). Data is divided in structured, semi-structured and unstructured data. Structured data is data for which the structure is known. It is stored, for example, in relational database and it is easy to search for using structured query language (SQL). Unstructured data, on the other hand, has no defined structure. An example of unstructured data is data from social media. Semi-structured data is unstructured data that includes some structured elements, for example, a photo with a time stamp. (Kordon 2020, 222)

Data integration and aggregation combine the data from different sources and possibly with different format structures into the same database (Romero, Romero & Ventura 2014, 43)

Data preprocessing or data cleaning includes, for example, handling missing data and outliers, and rescaling data. This phase of data preparing is very important, because the quality of its output defines the quality of the results of further data analysis. (Kordon 2020, 220)

Missing data means that for some reason, there is no data for certain observation of a variable. When talking about educational data, it may be because the student hasn't passed a course or activity. Another reason for missing data may be a faulty sensor or sampling process, cost reasons or restrictions in gathering process. The options for handling the missing data are either to remove or to replace it. If there are enough data, the missing data is usually removed. (Romero, Romero & Ventura 2014, 45) There are several methods to replace the data, for example, to use most common value or most common value for the class, to use the mean value of all data or a class, or to use some regression or classification methods to determine the values. (Kotsiantis et al. 2006, 113)

Outliers are observations that behave unlike most of the data. They may occur because of a measurement error or simply be a natural variability not being an error but leading to unexpected distribution. In case of educational data, outliers are often true observations. All the students do not behave or succeed as average, there always exist the ones who make less of an effort or fail. When searching for outliers, clustering can be used when the observations staying outside the clusters can be considered as outliers. (Romero, Romero & Ventura 2014, 46)

Some of the analysis methods require the use of rescaled data, because large differences in the variances influence the result since the variable with a high variance is dominating. Commonly used rescaling methods are normalization and standardization. In the normalization, variables are scaled in the interval 0-1 (or -1 to +1 or even $-\infty$ to $+\infty$). In the standardization, the variables are scaled such that they have zero mean and unit variance. A disadvantage of normalization is, that in case there are outliers, normal data is distributed in a very narrow interval. (Kordon 2020, 240) With educational data, min-max normalization is also often used. Other commonly used transformation methods include improving the distributions of the data closer to the normal distribution using power transforms, such as logarithm. Additionally, the data may be discretized, that is, divided in categoric classes. This decreases the number of possible values and often gives a more understandable view of the data. Generally, discretization smooths the influence of disturbances and enables more simple models, thus decreasing the possibility of overfitting. Discretization can be used, for example, for

age (young, adult, middle-aged,…) or success of the students (fail, pass, good, excellent). Finally, the data is transformed into the format in which it will be used, for example, DAT format or comma separated values (CSV). (Romero, Romero & Ventura 2014, 52-55)

3.3 Data analysis

The purpose of data analysis, or exploratory data analysis (EDA), is to provide understanding about the data and its structure. It includes visualization of the data in order to find possible relationships between predictor and target variables. EDA may also include preliminary selections of appropriate model development methods. And finally, it includes selection and possibly creation of relevant variables.

Descriptive statistics are non-graphical methods especially suitable for categorical, but also used for numerical data. Graphical methods useful for describing the distribution of the variables are histogram and box plot. They can also help with finding outliers. Methods for finding relationships between two variables include 2D scatter plot, and covariance and correlation as non-graphical methods. Correlation analysis is further described in Subchapter 3.3.1. For the visualization of higher dimensional data, 2D and 3D scatterplots with possibly several colors can be used. (Komorowski et al. 2016, 185-187)

The last task of EDA is to select the relevant variables out of all the available ones. If predictor variables are highly correlated, it is not necessary to choose them all. However, sometimes two or several variables together give valuable information that none of them could provide alone. Important variables can be found, for example, by decision tree method described in Subsection 3.4.3. Finally, data filtering means that only the relevant subset is chosen from the large data. For example, a certain student group, time interval, courses, or event. (Romero, Romero & Ventura 2014, 48-50)

3.3.1 Correlation analysis

Correlation analysis describes linear relationship and its strength between two variables. Correlation analysis is often used as a preliminary step for other analyses, such as linear regression. Correlation coefficient varies between [-1, +1], where +1 means perfect positive linear relationship, and -1 perfect negative linear relationship. Value 0 indicates that there is no linear relationship between the variables. As a rule of thumb, it can be said that values 0 – +0.3 indicate weak positive correlation, +0.3 – +0.7 moderate and +0.7 – +1 high positive correlation. And further, -0 – -0.3, -0.3 – -0.7, and -0.7 – -1 indicate weak, moderate and

high negative correlation, respectively. However, this depends on the case. If correlation coefficient is, for example, +0.85, it means that 85% of the variation in the target variable is explained in the variation of the predictor variable and further, when predictor variable increases, the target variable also increases. A correlation coefficient -0.2 means that 20% of the variation in the target variable is explained in the variation of the predictor variable, and when predictor increases, target variable decreases. (Durivage 2015; 27, 29)

The most commonly used correlation coefficient is Pearson product moment correlation coefficient. It is a parametric coefficient and can be used when both variables have at least interval scale and more or less normal distribution. Pearson correlation coefficient is determined as follows,

$$r_{\mathrm{y,x}} \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}, \tag{3.1}$$

where

x_i is i^{th} observation of a predictor variable,

y_i is i^{th} observation of a target variable,

$\bar{x}$ is mean value of predictor variables, and

$\bar{y}$ is mean of target variables.

When Pearson correlation is not suitable for the analysis, Spearmann's coefficient, or rank correlation coefficient can be used. It is suitable for the variables with at least ordinal scale, and it has no demands for their distribution. Spearmann's correlation is determined as follows,

$$r_{\mathrm{S}} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \tag{3.2}$$

where

d_i = difference of ordinal numbers of i^{th} observation couple and

n = number of observation couples. (Nummenmaa, Holopainen & Pulkkinen 2019; 215, 221, 224-225)

Both Pearson's and Spearman's correlation coefficients explain the correlation of the sample. In order to find out whether the dependency of the variables holds true for whole

population, significancy test is performed. In the test, test hypotheses are as follows: H0: Correlation coefficient of the population is 0, and thus, there is no correlation between the variables. H1: Correlation coefficient of the population differs from zero and there is correlation between the variables. The decision about keeping or rejecting the test hypothesis is made by studentizing the correlation coefficient and comparing it with critical limits of Student's *t* distribution. The correlation coefficient is studentized as follows,

$$t = \frac{r_{y,x}\sqrt{n-2}}{1-r_{y,x}^2}. \tag{3.3}$$

The critical limits are $\pm t_{\frac{\alpha}{2};n-2}$, where α is significance level and n-2 degrees of freedom used for the test. If the determined t value passes the critical limits, test hypothesis H0 is rejected. Significance level is the probability for type I error, that is, probability that H0 is rejected even if it shouldn't. Commonly used significance levels are 0.1%, 1%, and 5%.

Most of the statistical software directly determine a *p* value that is the probability of the risk of mistakenly rejecting the zero hypothesis. If *p* value is smaller than chosen significance level, H0 is rejected. (Nummenmaa, Holopainen & Pulkkinen 2019; 175-176, 224-225)

3.4 Model development

Many systems and processes can be modelled based on, for example, physical or economical theories. However, there are processes that are difficult to model in any other way than based on experimental data, or that may include crucial characteristics or disturbances caused by environment that are difficult to model theoretically. Machine learning can be a choice for modelling this kind of processes. Machine learning algorithms can learn to develop a model and make predictions based on the training data without rule-based programming. It can be seen as a process that automatically builds a model by learning it from the structure of the training data. (Akerkar & Sajja 2016, 53-54)

The most common types of machine learning categories are supervised and unsupervised machine learning. In supervised learning, both predictor and target variables are known. The purpose is to construct a model between the independent predictor variables and dependent target variable using training data set. The obtained model is used for predicting the target variables of further data sets containing only predictor variables. Supervised machine learning can be further divided in two subcategories: classification and regression. (Qamar & Summair Raza 2020, 41)

In classification, the purpose is to divide the data in previously determined classes based on the features. The target variable is categorical and often binary. Predictor variables can be either numerical or categorical. (Qamar & Summair Raza 2020, 87-88) Common classification algorithms are logistic regression, k nearest neighbor, decision tree, support vector machine, naïve bayes, and neural networks. Junk mail filter could be named as an example of a classifier. Regression differs from the classification in such a way that the target variable is continuous, whereas predictor variables can be either continuous or discrete (Qamar & Summair Raza 2020, 42).

In unsupervised machine learning, only the predictor variables of the process are known and there are no target variables. Therefore, the data is unlabeled. One of the most common unsupervised machine learning techniques is clustering. (Qamar & Summair Raza 2020, 42) Now, the data is grouped based on similar characteristics such that the similarity is large enough within the subgroups (clusters), while making sure there is enough difference among them. Clustering enables, for example, finding different customer profiles based on their interests, and targeting specified marketing campaigns for them. (Akerkar & Sajja 2016, 80)

This study concentrates on supervised machine learning, and further, classification which is suitable for answering the research question under consideration, that is, predicting whether the university students graduate on time or not. Three commonly used and easily applicable classification algorithms suitable for non-normally distributed data are chosen, namely k nearest neighbor, support vector machine (SVM), and decision tree. Further, evaluation and validation methods used for machine learning algorithms are presented.

3.4.1 K nearest neighbor

K nearest neighbor (KNN) algorithm is a simple supervised machine learning algorithm that can be used for both classification and regression, although classification is more common. K nearest neighbor provides a nonparametric, black box model, which means that it is not possible to make a meaningful equation between predictor and target variables as can be done, for example, with linear regression model. (Steele, Chandler & Reddy 2016, 279)

KNN is well suited for many machine learning applications. The algorithm assumes that similar properties are close to each other. If the assumption holds sufficiently enough, KNN is efficient and functional algorithm also for noisy data. However, calculation time may be long if the training set is large and the data includes several variables, because distances must

be calculated separately for each variable and each training sample. (Boateng, Otoo & Abaye 2020, p. 341)

The idea of the k nearest neighbor classifier is that when attempting to classify new data (x_0, y_0), it finds the k most similar or closest data in the known data set $D = \{(y_1, x_1), \dots, (y_n, x_n)\}$ and makes the decision based on that. (Steele, Chandler & Reddy 2016, 282-283) New observation is classified same as most of its k nearest neighbors. (Boateng, Otoo & Abaye 2020, 345) Figure 3.2 explains the idea of KNN algorithm. The predicted data X is compared with k nearest objects. If k = 2, the two closest objects are dark blue squares, and thus, the prediction will be that X belongs to class 3. If k = 5, three of five closest objects are light blue triangles, and thus, X is predicted to belong to class 1.

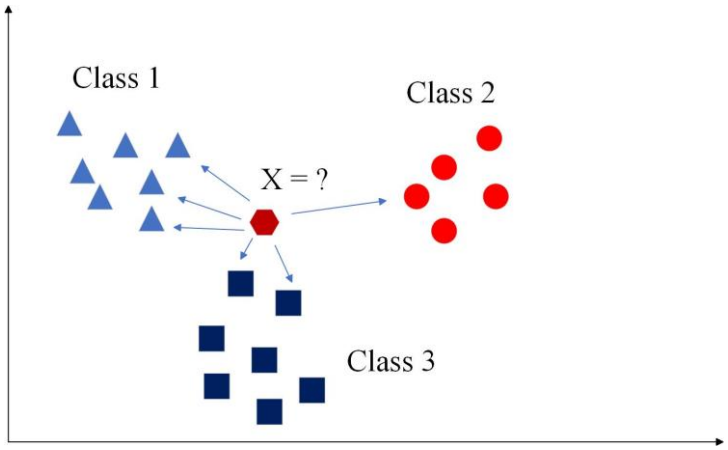


Fig. 3.2. A simple visualization of the idea of KNN algorithm. X belongs to the same class with k closest neighbors. (Boateng, Otoo & Abaye 2020, 346)

A common method to determine the distance is Euclidean distance

$$d_{\mathrm{E}}(x_i, x_0) = \left[\sum_{j=1}^{\mathrm{p}} (x_{i,j} - x_{0,j})^2\right]^{1/2}, \tag{3.4}$$

where p is number of the features, $x_i = [x_{i,1} \cdots x_{i,\mathrm{p}}]^T$ is a known predictor vector containing p features of the i^{th} observation , $x_0 = [x_{0,1} \cdots x_{0,\mathrm{p}}]^T$ is the predictor vector of the new observation used to predict y_0. (Steele, Chandler & Reddy 2016, 283) Euclidean distance function works well enough with categorial and numerical data. However, mixed type datasets require other distance functions, such as cosine, Minkowski correlation, or Chi square distance. (Hu et al. 2016, 8)

3.4.2 Support vector machine

Support vector machine (SVM) is another commonly used classification algorithm. Its operating principle is to make hyperplanes between the classes such that obtained margins between the hyperplanes and data closest to them are maximized. (Qamar & Summair Raza 2020, 105)

The benefits of SVM are the solid statistical learning theory it is based on, direct control on model complexity, an algorithm based on global optimization, and the repeatable results (Kordon 2020, 98-99). Weaknesses of SVM are its complexity, solving multi-class problems and dealing with unbalanced data sets. The complexity of SVM increases the training time of the algorithm for large data sets. SVM is originally a binary classifier solver. For multi-class problems, SVM must be divided in several binary classification problems, which further increases the calculation time. (Cervantes et al. 2020, 195-197)

Hyperplane is a decision boundary separating two or more classes located in different sides of the plane in the feature space. In the simplest linear case with two features, the hyperplane is a line. With several features, it may be a two-dimensional space. There are usually several possible hyperplanes. For example, Figure 3.3 a shows three hyperplanes separating classes of triangles and circles. The best possible hyperplane is determined such that parallel hyperplanes with same distance on both sides of it and close to the decision classes, provide maximum margin between the classes. The data points, or vectors closest to the parallel hyperplanes, are called support vectors. Figure 3.3 b shows parallel hyperplanes for the hyperplane H2 and support vectors of both classes as a filled triangle and a circle. (Qamar & Summair Raza 2020, 105-106)

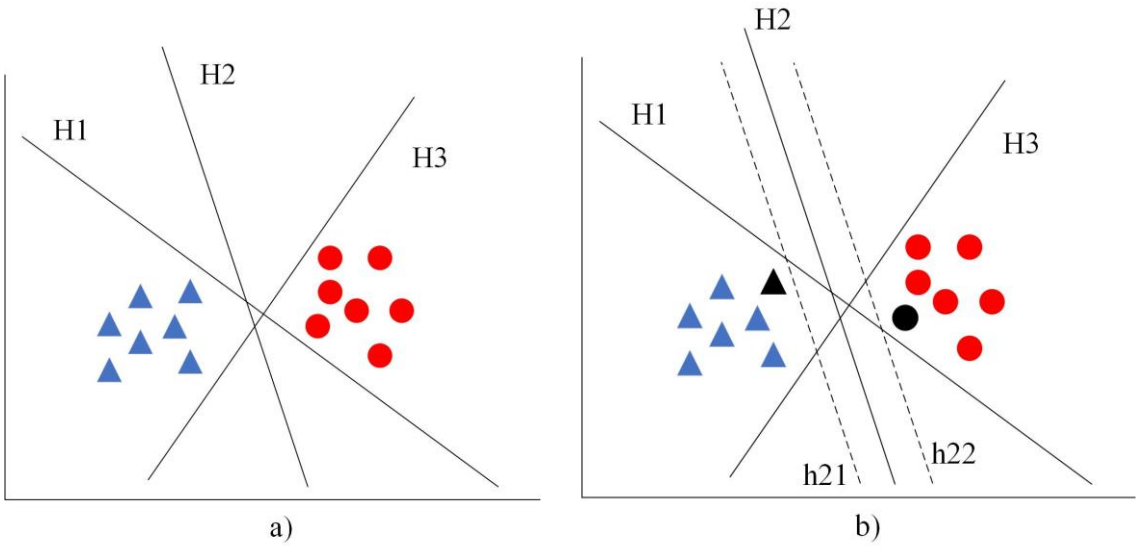


Fig. 3.3 a) Three hyperplanes separating triangles and circles, and b) parallel hyperplanes for H2 and support vectors of classes. (Qamar & Summair Raza 2020, 106)

Linear hyperplane is determined as a weighted linear combination of regression vector **x** as follows,

$$w_1x_1 + w_2x_2 + \ldots + w_px_p + b = b + \sum_{i=1}^{p} w_ix_i \quad (3.5)$$

where w_i are weight coefficients and b is a bias. Regression vector is same as predictor vector used in Equation (3.4). Weight coefficients are updated iteratively until an optimum has been reached. (Jo 2021, 169) In the model construction phase, the purpose is to find the above-mentioned support vectors that provide maximum margins between the classes of the training data.

The classes cannot always be separated using linear hyperplane, as can be seen in Figure 3.4 a, where the triangles are inside and the circles outside of the feature space. However, the data can be mapped to a higher feature space, after which it is possible to use a linear hyperplane, that is, the classes will be linearly separable. In Figure 3.4 b, the data points of Figure 3.4 a are mapped to a new feature space where $z = x^2 + y^2$. The original circular hyperplane is mapped as a horizontal line. (Qamar & Summair Raza 2020, 107-108) Mapping onto another feature space is performed using Kernel functions. Commonly used Kernel functions are linear, polynomial, gaussian, gaussian radial base function (RBF), and sigmoid kernel. (Cervantes et al. 2020, 193) Gaussian Kernel function, chosen to be used in this study, is determined as follows,

$$K(x_i, x_j) = \frac{\|x_i - x_j\|^2}{2\sigma^2}, \tag{3.6}$$

where σ is a kernel parameter determining the dispersion of the kernel in the original feature space. The parameter σ must be chosen carefully. Too large σ makes the kernel behave like a linear kernel, and too small σ makes the algorithm behave like KNN. (Varewyck & Martens 2011, 332)

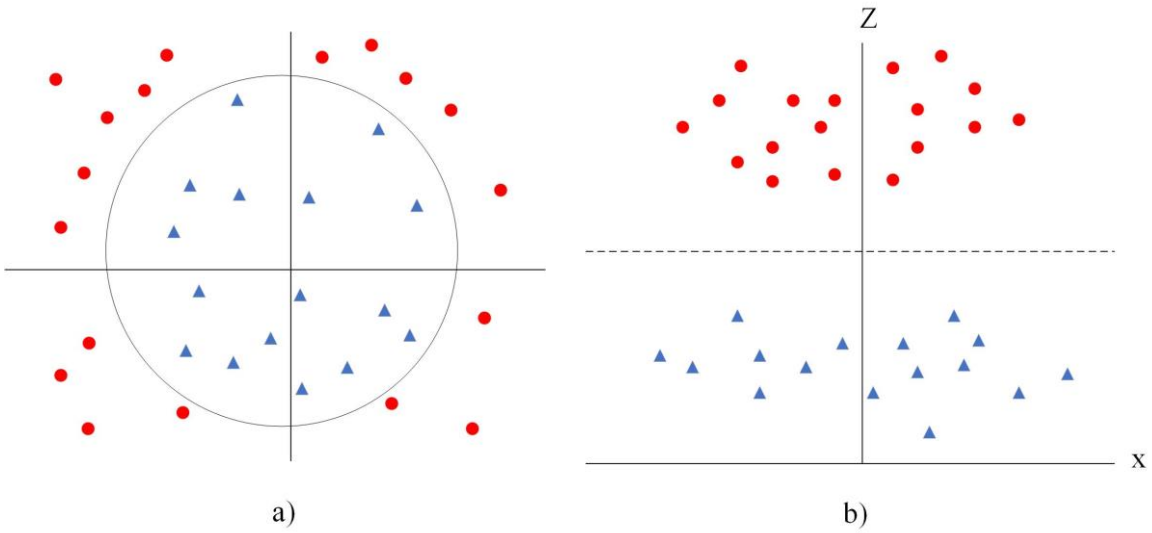


Fig. 3.4. a) Nonlinearly distributed data points in the original feature space, and b) same data points mapped in higher feature space. (Qamar & Summair Raza, 108)

3.4.3 Decision tree

Decision tree is a common and powerful method suited for both regression and classification, and for both numerical and categorical data. Decision tree model is based on rules, and thus, is easy to understand. Data is divided in two or more groups based on the most significant features. (Akerkar & Sajja, 58-59)

Decision tree is simple to make, and the model is easy to understand and follow. It is suited for multiclass problems, as well as for both numerical and categorical data. In case of simple problems, it competes in accuracy with other classifier algorithms. However, decision tree may result in underfitting the model when using too little data, or overfitting when the training data is too large. Additionally, it may not work optimally with numerical data. (Qamar & Summair Raza 2020, 51)

The name of the method comes from how it is presented visually. It resembles an upside-down tree, roots up and the branches and leaves down, see Figure 3.5. The tree starts from the root node containing the whole data set. Based on a decision, the data is split in branches and child nodes. The child nodes may contain new decisions, or they may be end-nodes containing the final prediction of the class they belong. Decision tree of Figure 3.5 describes decision making of whether to play or not based on the weather forecast. The first decision in the root node is made based on the outlook. If it is overcast, the decision is to play. If the weather is sunny, there is a child node with a new decision based on the windiness. If it is windy, the decision is not to play, and if it is not windy, the decision is to play. Similarly, if the weather is rainy, a new decision is made based on humidity. If humidity is too high, the decision is not to play, otherwise it is to play.

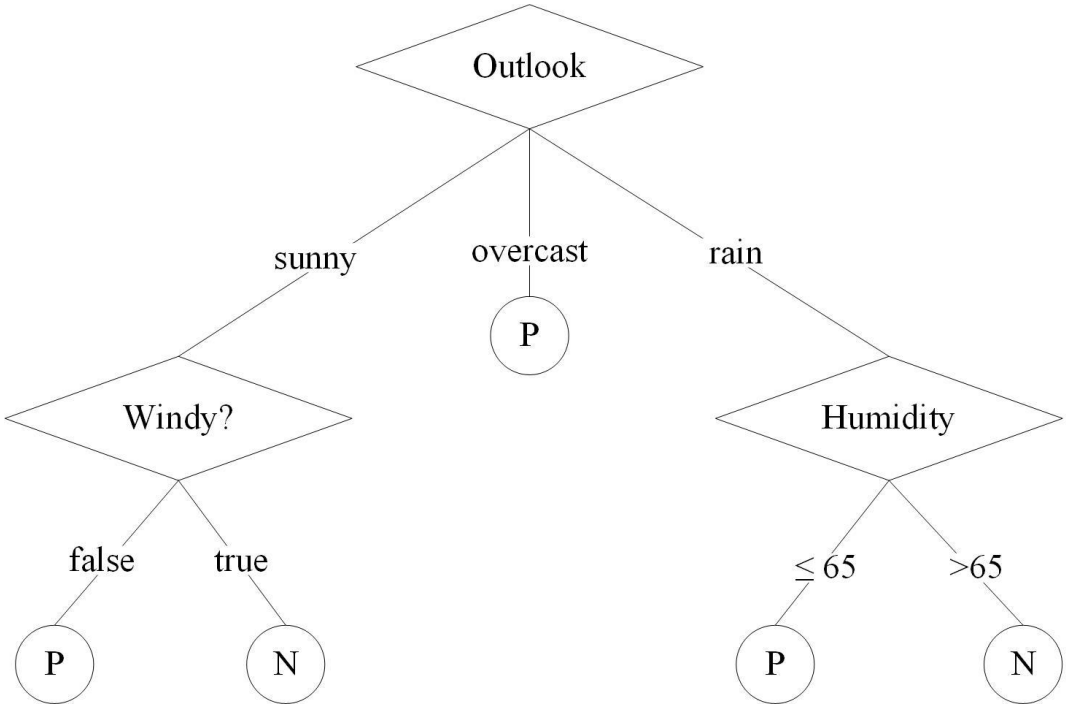


Fig. 3.5. Decision tree for choosing whether to play or not based on the weather forecast. (Quinlan 1986, 87)

One of the most commonly used decision tree algorithms is CART (Classification and Regression Trees), which is also used by Matlab. Using the training data, the algorithm generates a tree, where the root node is first split in two decision branches, and the produced child nodes are further split in new branches and nodes until the stopping criteria is met and the process stops. The tree that forms, is often overfitting. The overfitting can be reduced and model performance improved by pruning, where the lower branches and nodes of the tree

are removed. In CART algorithm, a Gini index explaining the probability of possible misclassification in each node is determined. Based on the probabilities, an optimal tree size is chosen. (Joshi 2020, 54, 57)

3.5 Model evaluation

The performance of classification models is commonly measured using accuracy, confusion matrix and metrics determined based on it, as well as receiver operating characteristics (ROC) graph together with area under ROC curve (AUC) measure.

The simplest measure is the accuracy determining the ratio of number of correctly classified and total cases as follows,

$$accuracy = \frac{\text{number of correctly classified cases}}{\text{total number of cases}} \tag{3.7}$$

However, disadvantages of accuracy are that it does not differentiate between the types of errors, and its dependency on the distributions of classes. For example, there are significantly fewer observations in one class compared with another. In such a case, it is possible that the accuracy is very high, but the classifier totally misses the other cases. (Novaković et al 2017; 3, 5)

Confusion matrix presents how the trained model predicts each class. Figure 3.5 shows a confusion matrix for a binary classifier with classes named positive and negative, but it can also be made for *n*-class classifier. Columns represent actual classes and rows predicted classes. Cells on the diagonal of the matrix, are correctly predicted. True positive (TP) is the number of correctly predicted positive cases and true negative (TN) correctly predicted negative cases. False negative (FN) is the number of cases predicted as negative, although they are actually positive. Similarly, false positive (FP) is the number of cases predicted as positive that are actually negative. (Fawcett 2006, 862)

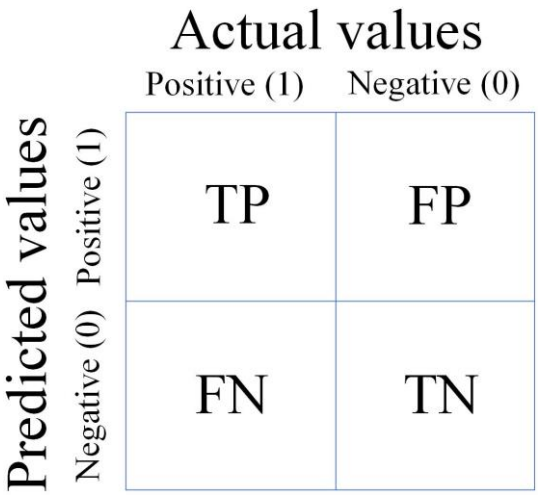


Figure 3.5. Confusion matrix. (Fawcett 2006, 862)

Confusion matrix can be further used for determining different performance measures, such as accuracy, sensitivity, specificity, precision, recall and *F* measure.

Accuracy that was already determined in Equation (3.7), can also be presented using the cases of confusion matrix as follows,

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \tag{3.8}$$

Sensitivity and specificity are measures adopted from the medical domain where the actual positive cases are important to find, and a false alarm is not so critical. Sensitivity represents the number of positive cases correctly predicted,

$$sensitivity = \frac{TP}{TP+FN} \tag{3.9}$$

and specificity number of negative cases correctly predicted,

$$specificity = \frac{TN}{FP+TN}. \tag{3.10}$$

In some fields, for example, in e-commerce recommendation systems, it is more important to catch false positive cases. There, precision is used to indicate, how many predictive positive cases are actually positive,

$$precision = \frac{TP}{TP+FP}. \tag{3.11}$$

A measure used along with precision, is recall that is same as sensitivity,

$$recall = \frac{TP}{TP+FN}. \quad (3.12)$$

Further, *F* measure is used to combine precision and recall, and it is a weighted harmonic mean of the two, as follows,

$$F_\alpha = \frac{(1+\alpha)(precision \cdot recall)}{\alpha \cdot precision + recall}. \quad (3.13)$$

The variable α is the weight chosen based on how recall and precision are weighted. It determines how much recall is weighted in comparison with precision. For example, if both are weighted equally, $\alpha = 1$, and

$$F_1 = \frac{2(precision \cdot recall)}{precision+recall}. \quad (3.14)$$

F_1 is used when false positive and false negative are equally costly or true negative is high. Further, if $\alpha = 2$, recall would be weighted twice as much as precision. α must be chosen based on the acceptability of error type I and II. (Japkowich & Shah 2014; 86, 95-96, 101, 103)

Another commonly used method is a ROC graph, which presents visually relative tradeoff between the benefits (TP) on y axis and costs (FP) on x axis. Commonly, AUC is used to compare the performance of classifiers. AUC gets values between zero and one. A diagonal line from (0, 0) to (1, 1) would have an area of 0.5, which corresponds to random guessing. Thus, AUC should always be more than 0.5. (Fawcett 2006; 862, 868)

3.6 Model validation

In data analytics, the purpose is to construct a generalized model using a finite amount of data. Without caution, the model obtained may fit very well with training data but won't be generalizable. This is called overfitting. Overfitting means that the model is too complex and includes too many details describing the data used for training, but no longer describe the general features of the population. (Provost & Fawcett 2013; 111-112, 124)

The most typical validation method to observe possible overfitting is a hold-out method where the data is divided in two parts, training data and test data. The model is constructed using the training data and its generalizability is validated with test data. A challenge with the hold-out method is that it provides only a single estimate for the performance of the model, and the performance is highly dependent on the selection of the training set. (Provost

& Fawcett 2013; 113, 124). Additionally, cutting the data in smaller pieces reduces the size of the training data, and thus, increases the variance (Liu & Özsu 2018, 681).

More sophisticated validation method would be to use cross-validation with several individual holdouts. It provides a possibility to determine statistics for the performance of the model, such as mean value and variance or standard deviation of accuracy, precision, recall and other possible performance measures. (Provost & Fawcett 2013, 124) In cross-validation, the model-building and performance testing are repeated several times crossing over the training and test sets in successive analyses. The basic cross-validation method is k-fold cross validation. In k-fold cross-validation, data is divided in *k* approximately equal parts. Then, *k* times training and testing is performed such that training and testing sets differ in each round. (Liu & Özsu 2018, 678) Figure 3.6 shows an example of k-fold cross-validation when $k = 5$. The original data is divided in 5 approximately same-sized sets, after which the modeling and testing is performed 5 times. In each round, different set is chosen as the test set, and other four are used as training sets. An evaluation (see Subchapter 3.5) is determined for each individual test. In the end, their mean value and standard deviation can be determined.

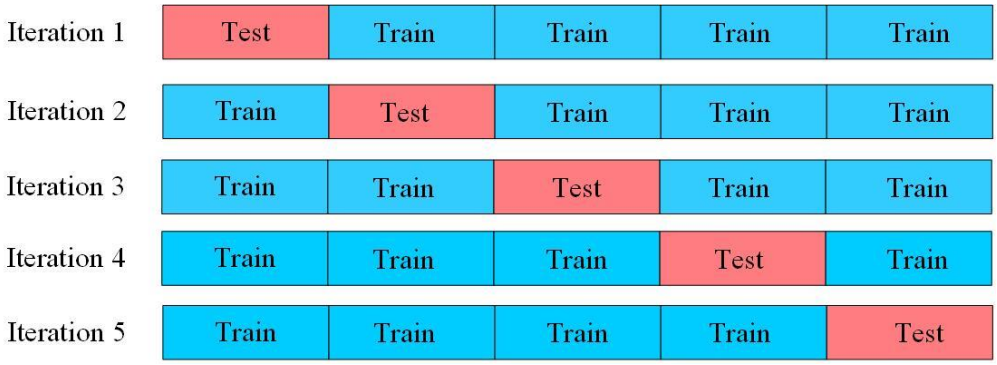


Figure 3.6. 5-fold cross-validation. The original data is divided in five approximately same-sized sets, after which the model is trained and tested five times (five iterations).

Before dividing, or folding, the data in *k* parts, it is typically stratified, that is, organized, such that each set is describing the totality as well as possible. For example, in case of binary classification, both training and test set contain percentually the same number of observations from both classes. Optimally, both training and test sets should be independent between each round. In practice, however, a satisfactory result is reached when only the test sets are independent. (Liu & Özsu 2018, 679-681)

Other cross-validation methods, such as leave-one-out cross-validation (LOOCV, LOO) and repeated k-fold cross-validation are special cases of k-fold cross-validation. In LOOCV, $k = n - 1$, that is, there are n iterations where training data consists of almost all the data, except for the one that is left for testing. However, LOOCV is computationally expensive, and it has a high variance and overfitting. (Marcot & Hanea 2021, 4) In repeated k-fold cross-validation, k-fold cross-validation is repeated several times. (Nakatsu 2020. 52).

In practice, it is nearly impossible to find a cross-validation method that would optimize both statistical performance (bias and variance) and computational cost, so the solution must be a trade-off between them (Arlot & Celisse 2010, 68-69). 5-fold (Nakatsu 2020) and 10-fold cross-validation are commonly used in machine learning (Nakatsu 2020, 51; Marcot & Hanea 2021, 5).

4 Results and discussion

This chapter presents the data used in the study, both descriptive statistics and visual graphs. Correlation analysis is performed in order to find possible linear relationships between the variables. Further, in model development phase, binary classifiers are trained to predict on-time graduation of students. Finally, discussion of the results is provided.

4.1 Data used

This subchapter presents the data used in the study, the original variables, their preprocessing and creation of new variables. Descriptive statistics of the variables are addressed, and visual data analytics performed, after which the variables are chosen for the model development. Additionally, a data analytics method, correlation analysis, is chosen for analyzing the linear correlation between the variables.

Data used in the study are collected from two different sources. The basic information of the students and data concerning their university studies are from the base register OODI formerly used at LUT University.[1] Application data are from the register of DIA selection.[2] Data was provided for the thesis by study administration of Lappeenranta-Lahti University of Technology.

The data was provided in two different files: Basic information about the graduated students, and attendance status and studied credits per calendar semesters. All the data was provided between autumn 2010 and autumn 2020. The data was prefiltered such, that it only included students of degree programmes in electrical engineering, and further, students who had studied consecutively bachelors's and master's programmes, for which the target graduating time is five years. All the data was in CSV format, and thus, structured data. Names and student numbers were excluded from the data, and thus, the data is anonymous. Both files included index numbers generated for this study and their purpose was to enable combining the different data files.

[1] OODI is a base register for information related to studies, managing information on studies, students, study rights, and study credits. OODI was replaced by SISU base register in 2020.
[2] DIA selection is a common application system used for admission of bachelor's degree in technology and architecture at Finnish universities.

The basic information of the students includes:

- student index
- birthday
- gender
- application year
- graduation date for both bachelor's and master's degree
- total credits studied in both bachelor's and master's degrees
- average grade of bachelor's degree
- average grade of master's degree

Attendance status and credit data includes:

- student index
- status of attendance: attending or non-attending autumn and spring semesters from 2010 to 2020
- number of credits studied during autumn and spring semesters from 2010 to 2020

The two data files are further integrated with the help of student indices. After the integration, the data only contains students who have both started their studies and graduated during the years 2010 -2020.

All the analysis is performed using Matlab software.

4.1.1 Data preprocessing and descriptive analysis

After combining the data, there are 95 students who have both applied to a bachelor's programme at LUT University and graduated consecutively with bachelor's and master's during the time in consideration, 2010 - 2020. A few students are excluded from the data because there are some features missing for them, their graduation time is suspiciously short, or they have obtained credits before their application year. Further, students with abnormally high total credits, students for whom the accumulation of credits differed considerably from typical are excluded, because such characteristics may indicate that students have started in another degree programme at LUT University and changed to electrical engineering in the middle of their studies. Additionally, all the students that applied to the university in 2015 are excluded, because there is something wrong in the application points, and the application group is not known. After the data cleaning, the data includes 82 students.

A few new variables are generated. The original data includes information about studied credits in certain calendar years (or semesters). New variables are generated such that they include information about credits completed by study years, that is,

- cumulative credits after first autumn semester,
- cumulative credits after first spring semester,
- cumulative credits after second spring semester,
- cumulative credits after third spring semester,
- cumulative credits after fourth spring semester, and
- cumulative credits after firth spring semester.

The autumn credits of the first year are included in the study, because it is interesting to see how the students start their studies. However, several courses last the whole academic year, and thus, credits completed during the autumn don't necessarily correspond to the studies carried out, but are lower, which would cause misleading results in this study.

Other new variables generated are

- starting year, determined assuming that the students started the studies when the first credits were obtained
- number of semesters used for studies. Possible absence semesters in between the attendance semesters are excluded because they are not counted in graduation time.
- number of absence semesters in between the study semesters.
- age of the students when starting their studies. Age is determined based on their birthday
- total number of credits is the sum of credits studied in bachelor's and master's degrees

There are 15 numerical features chosen for the initial analysis. The numerical features and their descriptive statistics are presented in Table 4.1. All the numerical features are of ratio scale. Additionally, there are two categorical variables: gender and application group. 78 of the students are men and 4 are women. Application group contains three categories: selection based on baccalaureate (high school), combined selection based on baccalaureate and entrance examination, and selection based on entrance examination. Number of students in each category is 42, 31, and 9, respectively.

Table 4.1. Descriptive statistics of numerical features used in the study. Application points are normalized in order to compare the points among the different application groups. Normalization is explained on page 44.

	Average	Median	Min	Max	Standard deviation
Age	20	20	18	22	0.70
Application points (scaled)	0.46	0.45	0.23	0.86	0.15
Number of semesters	12.3	12	8	18	2.1
Cumulative credits after 1st autumn	21	22	3	37	7
Cumulative credits after 1st spring	58	59	14	81	12
Cumulative credits after 2nd spring	115	116	50	155	17
Cumulative credits after 3rd spring	172	173	93	227	23
Cumulative credits after 4th spring	227	228	105	300	34
Cumulative credits after 5th spring	272	275	155	329	34
Grade of bachelor	3.18	3.14	2.13	4.61	0.52
Grade of master	3.43	3.40	2.39	4.40	0.52
Total number of credits	313	308	300	379	14.8
Number of absence semesters	0.3	0	0	4	0.7
Starting year	2012	2012	2010	2015	1.4
Graduating year	2018	2019	2015	2020	1.3

It can be seen from Table 4.1, that the students considered in the study, have started their studies during 2010 and 2015 and graduated during 2015 and 2020. Both average and median of semesters required for completing both bachelor's and master's studies is about 12 with the standard deviation of 2.1 semesters. Median number of absence semesters is 0. Average of total number of credits is 308, but there is quite a large variance between 300 and 379 credits. Average grades of the bachelor's and master's degrees are 3.18 and 3.43, respectively, with the standard deviation of 0.52 in both. Average of the cumulative credits after first autumn semester is 21, and in the end of each year of study from first to fifth, 58, 115, 172, 227 and 272 credits, respectively. Standard deviation of cumulative credits increases along with study years. Average age for starting the studies is 20, with the minimum age of 18 and maximum of 22. There is quite a large variance in the application points. The average points were 0.46 with the standard deviation 0.15, but the maximum points reach 0.86.

Figure 4.1 shows bar plots of a) starting year and b) graduation year of the students in consideration. Most of the students, have started their studies during 2011 - 2014, and graduated during 2017 – 2020. All the students started their studies in autumn semester. A reason for low number of students starting in 2015 is that the application data obtained for this study was faulty for that year and all the students starting the studies that year have been accepted to the university already in previous years and most likely performed military service before starting their studies. Another probable reason is that the data used is only from the time between 2010 and 2020, and very few students have graduated in five years. This is also a probable reason for the small number of graduates in 2015 – 2016.

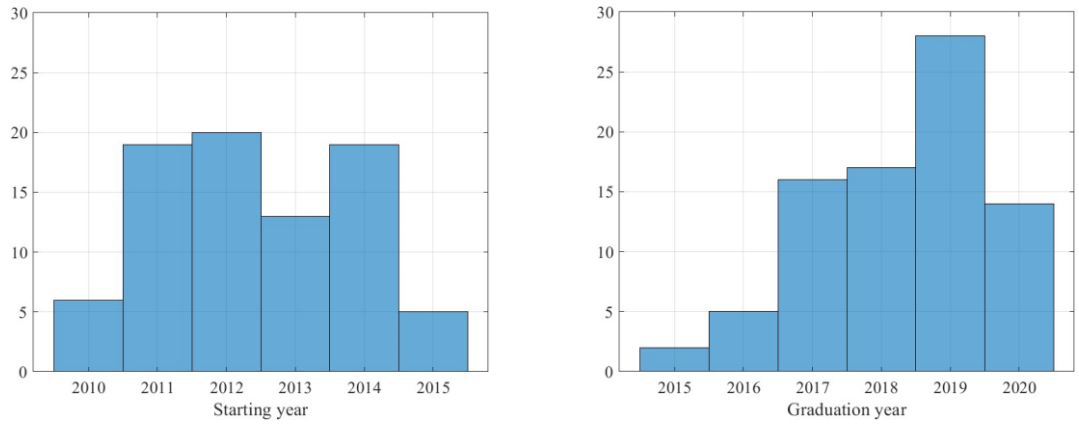
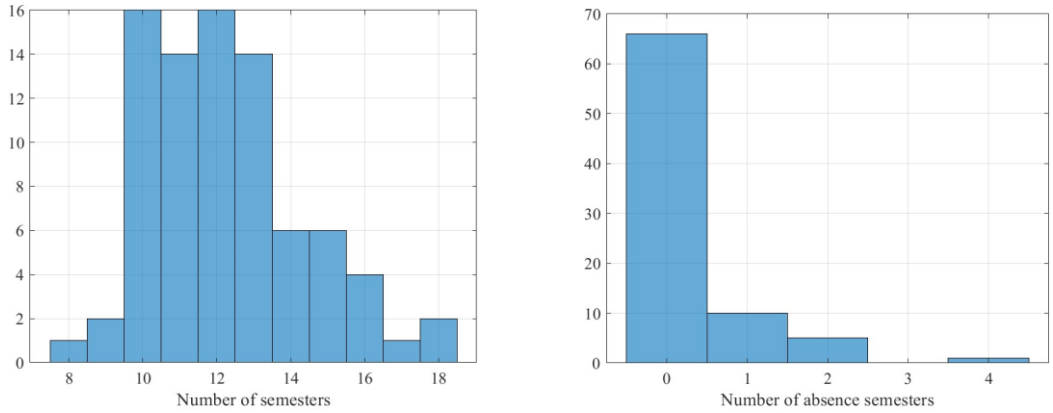


Figure 4.1. Bar plots of a) starting year and b) graduation year of the students analyzed.

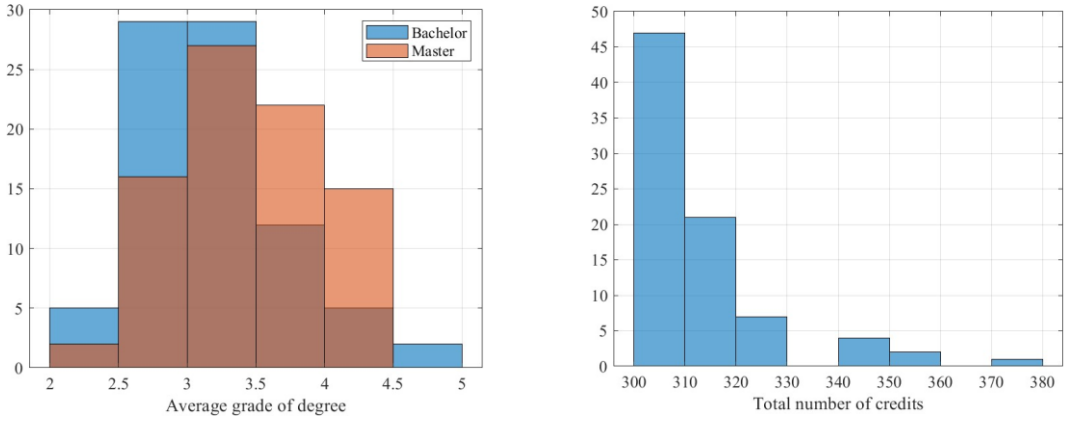
Figure 4.2 a) shows the distribution of semesters used for the studies. It can be seen that most of the students (73 %) take 10 – 13 semesters, that is, 5 – 6.5 years, to finish their studies. 23% (19 students) graduate in the target time of 10 semesters and 60% (49 students) in 12 semesters. Figure 4.2 b) shows the absence semesters in between the study semesters. Many students are absent during two semesters after being accepted for the studies and before starting them because of military service. However, this cannot be seen in the data, because only the absence semesters in between the study semesters have been considered. It can be seen that most of the students don't have breaks during the studies. Ten students have one semester break and five students have a break of two semesters. Additionally, one student has an absence of four semesters. Typical reasons for the absence semesters are, for example, military service or exchange studies.



a) b)

Figure 4.2. a) Number of semesters studied and b) number of absence semesters in between the study semesters.

In Figure 4.3 a), the distributions of the average grades of bachelor's and master's degrees can be seen. The figure shows that the grades of the master's degree are on average slightly higher compared with bachelor's degree. The same is seen in the descriptive statistics of Table 4.1. Figure 4.3 b) shows the distribution of the total number of credits studied in both bachelor's and master's degrees. The minimum number of credits for the degrees combined is 300 (bachelor 180 and master 120 credits), and most of the students have studied that or only a few credits more. However, it can be seen that some of the students have studied 40 – 60 or even 85 credits over the required amount.



a) b)

Figure 4.3. a) Distribution of the average grades of bachelor's and master's degrees. b) Distribution of total credits studied in both bachelor's and master's degrees.

Distributions of cumulative credits after each study year from first to fifth are shown in Figure 4.4. It can be seen that the peaks of the distributions are near the target numbers of credits for each year, namely 60, 120, 180, 240 and 300. When comparing the distributions to the statistics of Table 4.1, it can be seen that the peaks don't coincide with the average values. This is because the distributions are slightly skewed.

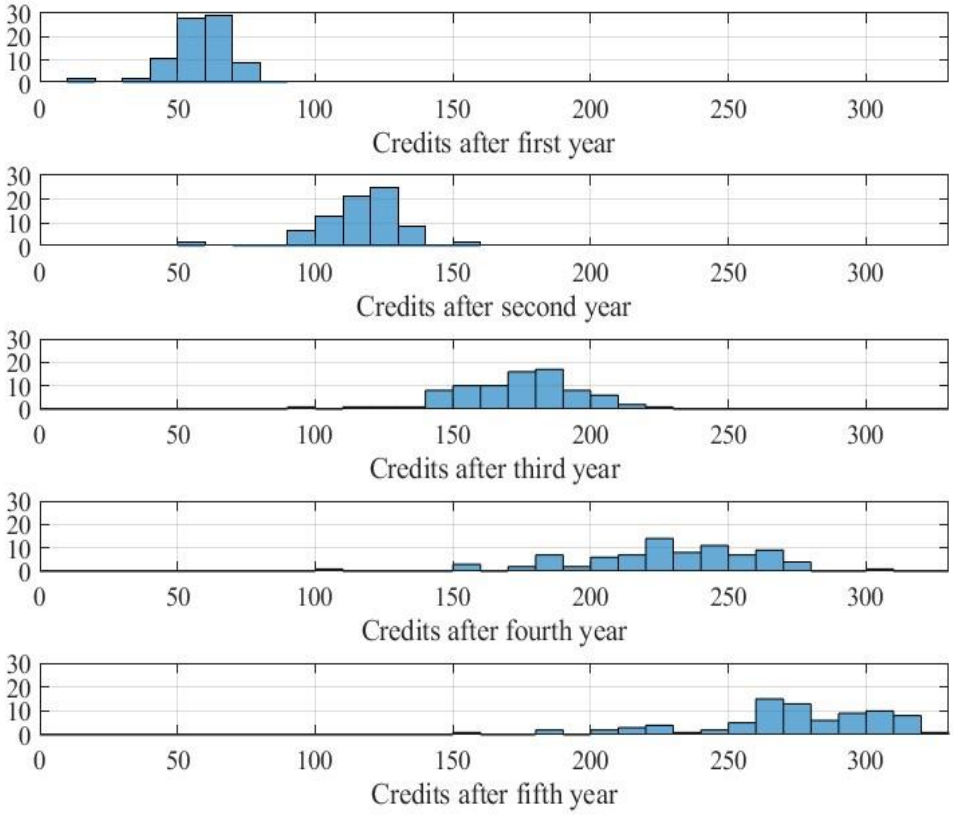


Figure 4.4. Distributions of cumulative credits after each study year from first to fifth.

In Figure 4.5 a), distribution of the students' ages at the start of their bachelor's studies is shown. It can be seen that a clear majority, 98%, of the students are 19 – 21 years old, 59% are 20 years old, and only few are 18 or 22 years old. The same can be seen in the descriptive statistics of Table 4.1.

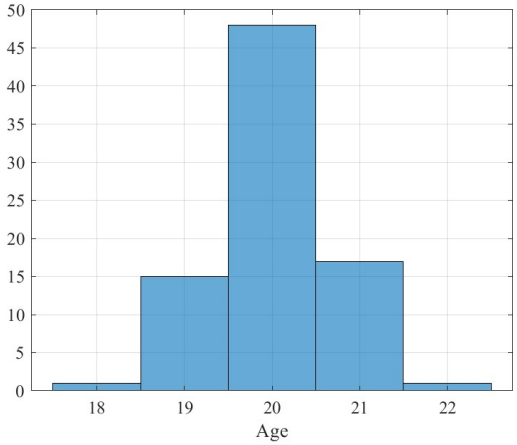
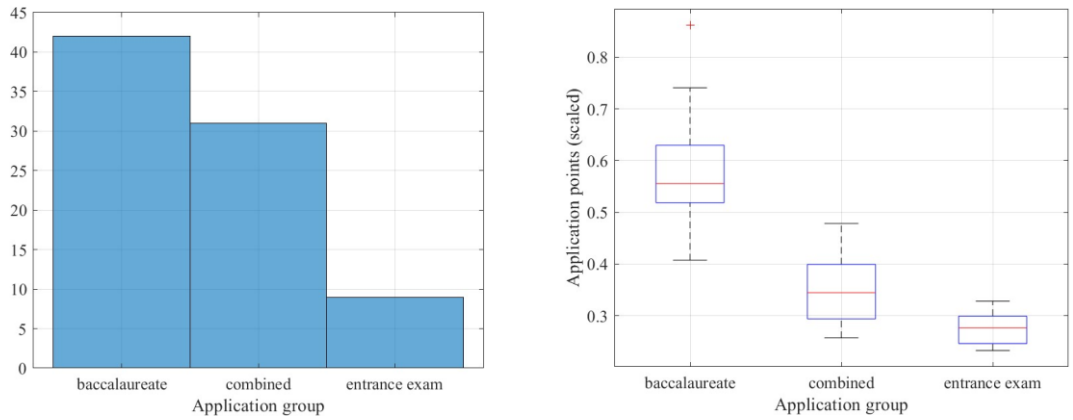


Figure 4.5. Distribution of the students' ages at the start of their bachelor's studies.

Figure 4.6 a) shows distribution of the application groups. It can be seen, that half (52%, 42 students) of the students are selected based on the baccalaureate, 38% (31 students) are selected based on combined selection of the baccalaureate and entrance examination, and 11% (9 students) based on the entrance examination alone. Figure 4.6 b) shows a boxplot describing the relations of application points and application groups. The scale of application points is different between the application groups, and varies between the years. Thus, the application points are normalized between zero and one such that they are comparable between each group. The normalization is performed separately for each year and application group such that the lowest possible points are set to zero and the highest possible points to one. It can be seen, that students selected based on baccalaureate, have considerably higher application points compared with the other selection groups. The second higher points are in the selection group of combined baccalaureate and entrance examination, and the lowest points in the group of entrance examination.



a) b)

Figure 4.6 a) Distribution of the application groups and b) boxplot of application points and application groups.

Finally, the following 14 features are chosen for further analysis:

- age of students in the beginning of studies
- application points
- application group
- number of semesters used for studies
- number of absence semesters
- cumulative credits after first autumn semester
- cumulative credits after first spring semester
- cumulative credits after second spring semester
- cumulative credits after third spring semester
- cumulative credits after fourth spring semester
- cumulative credits after firth spring semester
- average grade of bachelor's degree
- average grade of master's degree
- sum of total credits of bachelor's and master's degrees

Before further analysis and model development, data is normalized in the interval 0-1. Application points are normalized separately for each year and application group such that the lowest possible points are set to zero and the highest possible points to one. Other variables are normalized such that the minimum data value is set to zero and maximum data value to one.

4.1.2 Correlation analysis

Most of the data used isn't normal distributed, as could be seen in the distributions in Subchapter 4.1.1. Further study shows that there are only two normal distributed features, namely, application points and cumulative credits after first autumn semester. In the correlation analysis, Pearson correlation (Equation (3.1)) is only used for correlation of two mentioned features, and for others, Spearmann's correlation (Equation (3.2)) is used, because it suits for non-normally distributed data.

The null hypothesis is that there is no correlation between any variable. Alternative hypothesis is that there is correlation between two variables.

Table 4.2 shows correlation analysis and *p* values for application points, average grade of bachelor's degree, average grade of master's degree, and cumulative credits after first autumn semester. It can be seen that application points correlate moderately with average grade of bachelor's degree (0.444) and cumulative credits after first autumn semester (0.492). Further, average grades of bachelor's and master's degrees have moderate positive correlations with first autumn credits, 0.664 and 0.492, respectively. All the mentioned correlations have *p* value < 0.01 and thus, the correlations are significant and null hypotheses are rejected. None of the features correlate with the number of semesters used for the studies.

Table 4.2. Correlation analysis with *p* values for application points, average grade of bachelor's degree, average grade of master's degree, cumulative credits after first autumn semester, and number of terms used for studies.

		Application points	Grade of BSc	Grade of MSc	Credits of 1st autumn	Number of terms
Application points	corr	1	0.444	0.289	0.492	0.004
	p	*1*	*0.000*	*0.009*	*0.000*	*0.970*
Grade of BSc	corr	0.444	1	0.664	0.460	-0.122
	p	*0.000*	*1*	*0.000*	*0.000*	*0.274*
Grade of MSc	corr	0.289	0.664	1	0.386	-0.102
	p	*0.009*	*0.000*	*1*	*0.000*	*0.363*
Credits of 1st autumn	corr	0.492	0.460	0.386	1	-0.098
	p	*0.000*	*0.000*	*0.000*	*1*	*0.382*
Number of terms	corr	0.004	-0.122	-0.102	-0.098	1
	p	*0.970*	*0.274*	*0.363*	*0.382*	*1*

In Table 4.3, correlation analysis and *p* values for the correlation of application points, cumulative credits after first autumn semester and all of the spring semesters from first to fifth, as well as the number of semesters used for studies are shown. There is high positive correlation between the cumulative credits after successive spring semesters (0.777 – 0.862).

After second spring semester, there is moderate positive correlation (0.649) with the cumulative credits of fifth spring semester credits. All the correlations between the cumulative credits after semesters have *p* values < 0.01, and thus, are significant. Application points have only moderate correlation with the cumulative credits of first autumn semester. After that, the correlation between application points and cumulative credits becomes weak. Negative correlation between cumulative credits after each spring semester and semesters used for studies increases when the studies progress, but after five years, the correlation is still only moderate (-0.580).

Table 4.3. Correlation analysis with *p* values for application points, cumulative credits after first autumn semester and all of the spring semesters from first to fifth, and number of terms used for studies.

		Application points	1st autumn credits	1st spring credits	2nd spring credits	3rd spring credits	4th spring credits	5th spring credits	Number of terms
Application points	corr	1	0.492	0.252	0.276	0.192	0.207	0.203	0.004
	p	*1*	*0.000*	*0.023*	*0.012*	*0.084*	*0.063*	*0.0675*	*0.970*
1st autumn credits	corr	0.492	1	0.459	0.494	0.321	0.372	0.3492	-0.098
	p	*0.000*	*1*	*0.000*	*0.000*	*0.003*	*0.001*	*0.001*	*0.382*
1st spring credits	corr	0.252	0.459	1	0.777	0.539	0.477	0.412	-0.184
	p	*0.022*	*0.000*	*1*	*0.000*	*0.000*	*0.000*	*0.000*	*0.098*
2nd spring credits	corr	0.276	0.494	0.777	1	0.824	0.743	0.649	-0.432
	p	*0.012*	*0.000*	*0.000*	*1*	*0.000*	*0.000*	*0.000*	*0.000*
3rd spring credits	corr	0.192	0.321	0.539	0.824	1	0.862	0.733	-0.551
	p	*0.084*	*0.003*	*0.000*	*0.000*	*1*	*0.000*	*0.000*	*0.000*
4th spring credits	corr	0.207	0.372	0.477	0.743	0.862	1	0.833	-0.550
	p	*0.063*	*0.001*	*0.000*	*0.000*	*0.000*	*1*	*0.000*	*0.000*
5th spring credits	corr	0.203	0.349	0.412	0.649	0.7325	0.833	1	-0.580
	p	*0.068*	*0.001*	*0.000*	*0.000*	*0.000*	*0.000*	*1*	*0.000*
Number of terms	corr	0.004	-0.098	-0.184	-0.432	-0.551	-0.550	-0.580	1
	p	*0.970*	*0.382*	*0.098*	*0.000*	*0.000*	*0.000*	*0.000*	*1*

Additionally, it is concluded that the number of absence semesters correlates moderately with age (-0.517, $p = 0.000$), which may be because students starting the university studies right after the high school, perform military service somewhere in the middle of the studies. Age and number of absence semesters do not correlate with any other features. Further, total number of credits does not correlate with any feature.

4.2 Binary classification of on-time graduation

In this subchapter, on-time graduation of the students is predicted using a few classification algorithms. Six cases are studied: how reliably is it possible to predict the timely graduation after first autumn semester, first spring semester, second spring semester, third spring semester, fourth spring semester, and fifth spring semester. Students graduating on time, would optimally graduate for bachelor's after third spring semester and for master's after fifth spring semester.

Target variable of the classification is binary: whether the student graduates within the target time five years (positive) or later (negative). The five-year threshold is based on the funding model of the universities that was presented in Subchapter 1.1.4.

As could be seen in Figure 4.2 a, only 23% of the students have graduated on time. Since the classes are so unbalanced, obtaining reliable results would be challenging. In order to rebalance the classes, new observations for the class of on-time graduates are generated. Data generation is performed such that each 19 original observations are tripled. Random variation of integer numbers is added to selected variables of the artificial observations as follows:

- cumulative credits after first autumn semester, integer numbers between -1 and +1
- cumulative credits after first spring semester, integer numbers between -2 and +2
- cumulative credits after second spring semester, integer numbers between -3 and +3
- cumulative credits after third spring semester, integer numbers between -4 and +4
- cumulative credits after fourth spring semester, integer numbers between -5 and +5

After generation of new variables, the total number of observations are 120, of which 57 timely graduated and 63 late graduates.

Initially, predictor variable contains the features chosen in Subchapter 4.1.2 (page 44). However, further study of different feature combinations shows that the best prediction is obtained with only average grade and cumulative credits until the time considered, for example, in the end of third academic year, cumulative credits after first autumn and first to third springs, are included. Average grade of bachelor's degree is used for first four cases (study years 1-3) and of master's degree for the last two cases (study years 4-5). This information is not usually known when predicting the future graduation time. However, it is assumed that the average grade stays approximately the same during the degree programme studies,

and thus, average grade of the current moment of studies can be used to replace the final grade of the degree programme used in this thesis.

Three classifiers are compared in the study: k nearest neighbor, support vector machine and decision tree. KNN uses Euclidean distance, and SVM algorithm uses Gaussian kernel.

For the validation of the results, *k*-fold cross-validation is used. Based on the references mentioned in Subchapter 3.6, number of folds has been chosen as ten.

Results are evaluated using measures determined based on confusion matrix, namely accuracy, precision, recall, and F1. Number of neighbors used in KNN algorithm is optimized between 2 – 10 separately for each six cases based on F1 measures.

Figure 4.7 shows the evaluation parameters of accuracy, precision, recall and F1 measure determined for the test data of each case and algorithm. Since 10-fold cross-validation is used, the measures are mean values of the ten rounds. Additionally, the numerical values can be seen in Table 4.4. Accuracies for training data and optimized parameters *k* used in KNN are also presented.

It can be seen that accuracies for classifying whether of not the students graduate on time mainly increase alongside the study time. Only with decision tree, there is a bending after fourth spring. After the first autumn semester, the accuracy for training data is 74-88%, of which the best accuracy with decision tree and lowest with SVM. This means that 74-88% of the predictions are predicted correctly. Decision tree and KNN give accuracy of over 90% and SMV of 86% already after the first spring. After the third spring, the accuracies are 95-100%. For test data, the accuracies after the first autumn are 62-71%, of which the best with KNN and worst with SMV. Test data also gives accuracies of 82-89% after the second study year, and 84-94% after the third year. With decision tree, the accuracies of test data in some cases are relatively low compared with training data, which is probably because of overfitting.

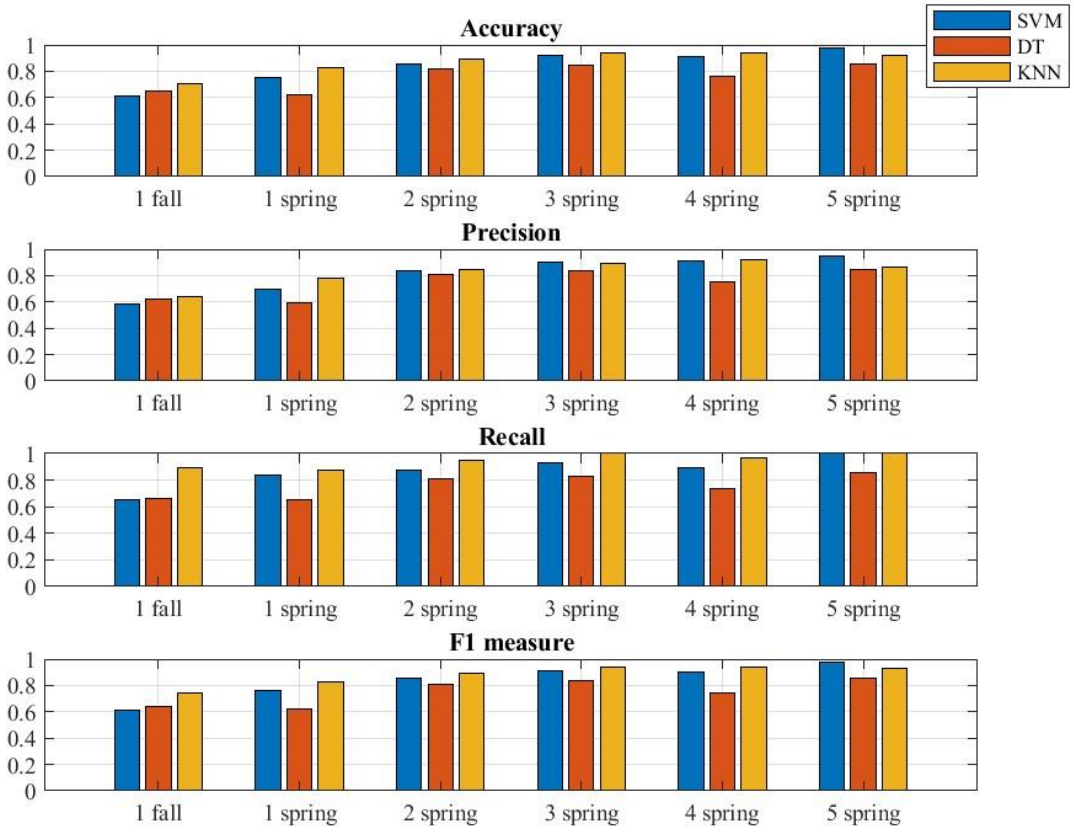


Fig. 4.7. Evaluation values accuracy, precision, recall, and F1 measure of test data using three machine learning algorithms: support vector machine, decision tree, and k nearest neighbor. Values are averages of ten cross-validation rounds.

Table 4.4. Key evaluation values of prediction using three different machine learning algorithms: Support vector machine, decision tree, and k nearest neighbor. Values are averages of ten cross-validation rounds.

	Case	Accuracy train	Accuracy test	Precision	Recall	F1	KNN *k*
SVM	1 autumn	74%	62%	59%	65%	62%	---
	1 spring	86%	75%	70%	84%	76%	---
	2 spring	95%	86%	83%	88%	85%	---
	3 spring	98%	92%	90%	93%	91%	---
	4 spring	100%	91%	91%	89%	90%	---
	5 spring	100%	98%	95%	100%	97%	---
Decision tree	1 autumn	88%	65%	62%	67%	64%	---
	1 spring	91%	63%	60%	65%	62%	---
	2 spring	92%	82%	81%	81%	81%	---
	3 spring	95%	84%	84%	82%	83%	---
	4 spring	93%	76%	75%	74%	74%	---
	5 spring	95%	86%	84%	86%	85%	---
KNN	1 autumn	84%	71%	64%	89%	74%	8
	1 spring	99%	83%	78%	88%	83%	7
	2 spring	99%	89%	84%	95%	89%	7
	3 spring	100%	94%	89%	100%	94%	6
	4 spring	100%	94%	92%	96%	94%	5
	5 spring	100%	93%	86%	100%	93%	10

Precision indicates how many of the students predicted to graduate on time have actually done so. It can be seen that after second year, the precisions of over 80% are obtained with each algorithm. This means that over 80% of the students predicted to graduate on time, actually graduate on time, while less than 20% of them graduate later. After the third year, the precision is 90% with SVM and 89% with KNN. With SVM, the precision continues increasing toward the fifth year but with KNN, the precision after fifth year decreases to 85%. This is probably because after the fifth year it has become clear that some of the students predicted to be graduated on time didn’t do so.

Recall indicates how many of students actually graduating on time, are also predicted to do so. With SVM, recall is 88% after second and 93% after third year, decreasing to 89% after fourth year. This means that of the students who actually graduate on time, 88% are correctly predicted to do so after the second study year, and 12% are erroneously predicted to be graduating late. KNN gives recall of 95% after second and 100% after third year, also decreasing to 96% after fourth year. Both SVM and KNN give recall of 100% after fifth year. Recall of decision tree stays below 90% in all the considered cases.

F1 measure combines precision and recall and weights them equally. SVM gives F1 values 85% after second, 91% after third, and 97% after fifth study year. Corresponding values for

KNN are 89, 94, and 93%, respectively. For decision tree, F1 stays below 86% in each considered case.

4.3 Discussion

Graduating in the target time of five years is clearly a challenge for the students in electrical engineering, only 23% of the postgraduates graduated on time. 60% graduated in six years, and the average graduating time was slightly above six years. On the other hand, according to Figure 1.4, share of graduates in energy technology and electrical engineering in five years is 6% and in six years 22%. Some of the differences could be explained with combined data with energy technology students or slightly different time scales of the data (2005-2021 in Vipunen data and 2010-2020 in this study). However, the main reason for the significant differences is most likely due to the fact that data of Vipunen also contains the students who never graduated, whereas in this study, the non-graduates are not considered.

The largest correlation (moderate negative correlation) for study time after two to five years is found to be with cumulative credits after each spring semester. Application points, on the other hand, have moderate correlation with cumulative credits after first autumn and grade of bachelor's thesis. However, they do not correlate with graduation time.

In classification, best models are obtained by using cumulative credits after the first autumn semester and each spring semester, as well as grade of current degree programme. Even if there is no linear correlation between the grade and study time, classification results are better with grade included. Both SVM and KNN perform very similarly for both accuracy, precision, and recall. KNN gives only slightly better performance classifying the on-time graduation with 89% accuracy already after second, and 94% accuracy after third year of studies. After second year, precision is 84%, recall 95%, and F1 89%. After third year, the corresponding values are 89, 100, and 94%, respectively which means that 11% of the students predicted to graduate on time, graduate late, and all the students predicted to graduate late actually do so. There is no significant improvement in the classification results after the fourth study year compared with the third year. And the result after fifth year is not relevant in means of prediction, because the actual outcome of timely graduates is already known.

Based on the result, it can be concluded that it is important to offer counselling not only for the students expected to be delayed with their studies but also for the students whose studies

seam to proceed well. This is a possible way of catching the 11% of students assumed to be graduated on time after the third study year, but in risk of delay in studies after that.

A challenge in this study comes from small amount of data, only 82 students. Additionally, there is clear unbalance between the classes in the original data, only 23% are graduated on time. However, data rebalancing is used in the classification by generating artificially more observations in the minority class of on-time graduates. It should be noted that the use of artificially generated data may influence in the classification result giving somewhat better result compared with fully authentic data in balanced classes. Further, a challenge comes from the data including only graduated students. The prediction model obtained, cannot be expected to predict students dropping out when the training data only consists of survivors.

The accuracies reached with each three algorithms are comparable with the classifiers studied in previous studies (see Subchapter 2.3). The best accuracy of the classifier, 95%, is obtained by of Lesinski et al. (2016). In the study, accuracy of 94% was obtained after third year of studies with KNN algorithm, and of 92% with SVM. Only in one of the previous studies precision and recall are also determined, 80.71% and 87.85%, respectively. They remain lower compared with this study, precision 89% and recall 100% with KNN and 90% and 93% with SVM after third year of studies, respectively. Most of the previous studies mentioned, use demographic data, admission data and performance of current studies as predictive variables. Lesinski et al. (2016) also uses data from extra-curricular activities and Pang et al. (2017) personal thoughts and attitudes towards studies.

However, it should be noted that human factors may influence the timely graduation, and this influence is not necessarily linear. There are students who are working during the academic year, some of them have family, are active athletes, or active in organizations, and cannot use 46 hours per week for the studies, as would be required in order to graduate in target time. Another relevant issue is the fact that along with increasing number of students, their heterogeneity has increased, especially in technical field. The prerequisites of mathematics and physics skills of the students are not equal, and the students with lower than expected mathematics and physics skills have to work harder. (Rantanen & Liski 2009, 14-15) Additionally, study profiles of the students influence the graduation time. Students with deep approach to studies, obtain credits and graduate faster, as well as with higher grades compared with their surface approached and unorganized colleagues. (Haarala-Muhonen et al. 2017, 952-957)

Information related to human factor is not possible to be acquired easily and automatically from existing educational registers but requires additional surveys for which not necessarily all the students answer at all. That is a reason why in this study, only continuously and automatically collected student register data is chosen to be used. Thus, it must be accepted, that the predicted result is not perfect.

5 Conclusions

In this master's thesis, a few classification algorithms have been applied in order to predict the on-time graduation of university students based on their performance in their studies.

The use of e-learning tools, educational software and other web-based educational systems has increased. Two research communities concentrated on EDM and LA were founded in the early 2000s to develop knowledge and tools to transform the data from several different formats into understandable and useful information about students and their learning in order to develop courses and programmes, as well as counseling for students.

One of the EDM applications, namely student modelling, and further, prediction of performance and characteristics of students concentrates on estimating certain characteristics, such as, academic performance, study time, motivation, or learning style. Methods for prediction are regression and classification.

Classification is a supervised machine learning method, in which an appropriate model is developed using both predictor and target variables. The obtained model is used for predicting the target variables of new data with only the predictor variables known. In order to obtain reliable results in modeling a classifier model, data collected is preprocessed, which includes handling the possible missing data and outliers, as well as rescaling the data. Further, exploratory data analysis consisting descriptive statistics and visualization, provides understanding about the data and its structure, following the preliminary selection of appropriate model development method and relevant variables.

In this study, three classifier algorithms suitable for non-normally distributed data were modelled, namely KNN, SVM, and decision tree. Performance of the models was evaluated using several performance measures, such as, accuracy, precision, recall, and F measure. Model validation was taken care of by using k-fold cross-validation, in which the same modelling and testing process is performed on different data sets, and the performance measures are determined as mean values of the folds.

In this study, the research question was, at which point of the studies, is it reliably possible to predict whether the students will graduate on target time or not? Students of electrical engineering studying consecutively bachelor's and master's degrees, for which the target graduation time is five years, were considered. Based on the results obtained, it was shown that the on-time graduation of the students can be predicted reliably enough after the third

year of studies with the accuracy of 94% with KNN algorithm and accuracy of 92% with SVM. Predictive features used in the classification were cumulative credits after first autumn and all of the spring semesters from first to fifth, as well as average grade of the current degree of studies. Based on the correlation analysis, it was shown that the cumulative credits after two to five years have moderate negative correlation with study time.

The results obtained correspond well with the other studies conducted on the topic. Thus, the result of the study and classification accuracy can be considered good enough to predict the on-time graduation of university students. However, it should be noted that rebalancing of classes using artificially generated data may influence in the classification result giving somewhat better result compared with fully authentic data in balanced classes. Another thing to consider when using the model, is the fact that it is developed using only data of graduated students. Model is not able to predict students in risk of dropping out, but only differentiate timely graduation from late graduation.

The resulted model can be put into use and make it easier to follow the progress of the students and predict their on-time graduation. It can be used for administrative prediction of timely graduation but also finding the students in the need of additional support. The first steps in the utilization of the model will be taken in the degree programme in electrical engineering, and as a Matlab-based tool. In future, when there are more data of graduates available, the model can be improved. Additionally, the model can be tested in other degree programmes in LUT School of Energy Systems. And further, the model can be modified to predict the on-time graduation of the students applying directly to the master's programmes, where the target graduation time is two years.

References

Ahola, S. 1995. Eliitin yliopistosta massojen korkeakoulutukseen – Koulutuksen muuttuva asema yhteiskunnallisen valikoinnin järjestelmänä. Väitöskirja. Turku: Painosalama.

Akerkar, R. & Sajja, P. S. 2016. Intelligent Techniques for Data Science. Springer International Publishing Switzerland. ISBN 978-3-319-29206-9 (eBook). DOI: 10.1007/978-3-319-29206-9.

Arlot, S. & Celisse, A. 2010 A survey of cross-validation procedures for model selection. Statistics Surveys. Vol. 4. pp. 40-79. DOI: 10.1214/09-SS054.

Bakhshinategh, B., Zaiane, O. R., ElAtia, S. & Ipperciel, D. 2018. Educational data mining applications and tasks: A survey of the last 10 years. Education and Information Technologies. Vol. 23, pp. 537-553. https://doi.org/10.1007/s10639-017-9616-z.

Baker, R. S. J. D. & Yacef, K. 2009. The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining. Vol. 1, no. 1, pp. 3-17. https://doi.org/10.5281/zenodo.3554657.

Boateng, E., Otoo, J. & Abaye, D. 2020. Basic Tenets of Classification Algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. Journal of Data Analysis and Information Processing. Vol. 8, pp. 341-357. DOI: 10.4236/jdaip.2020.84020.

Liñán, C., Juan Pérez, L. & Alejandro, Á. 2015. Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. RUSC. Universities and Knowledge Society Journal. Vol. 12, no. 3, pp. 98-112. doi: http://dx.doi.org/10.7238/rusc.v12i3.2515.

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L. & Lopez Asdrubal. 2020. A comprehensive survey on support vector machine classification: Application, challenges and trends. Neurocomputing. Vol. 408, pp. 189-215. doi: https://doi.org/10.1016/j.neucom.2019.10.118.

Durivage, M. A. 2015.Practical Engineering, Process, and Reliability Statistics. American Society for Quality (ASQ) Press.

Educational Data Mining. 2023. Web pages of Educational Data Mining Society. Available: https://educationaldatamining.org/ [referred 31.10.2023].

Eriksson, I. & Mikkonen, J. 2003. Opiskelijat ja opiskelu yliopistossa. Teoksessa Eriksson, I. & Mikkonen, J. (toim. 2006.). Opiskelun ohjaus yliopistossa. Helsinki: Edita.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recognition Letters. Vol. 27, issue 8, pp. 861-874. doi: https://doi.org/10.1016/j.patrec.2005.10.010.

Haarala-Muhonen, A., Ruohoniemi, M., Parpala, A., Komulainen, E. & Lindblom-Ylänne, S. 2017. How do the different study profiles of first-year students predict their study success, study progress and the completion of degrees? Higher Education. Vol. 74, pp. 949–962. doi: https://doi.org/10.1007/s10734-016-0087-8.

Hannaford, L., Cheng, X. & Kunes-Connell, M. 2021. Predicting nursing baccalaureate program graduates using machine learning models: A quantitative research study. Nurse Education Today. Vol. 99, no. 3. doi: 10.1016/j.nedt.2021.104784.

Hu, L.-Y., Huang, M.-W., Ke, S.-W. & Tsai, C.-F. 2016. The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus. Vol. 5, no. 1. doi: 10.1186/s40064-016-2941-7.

Japkowich, N. & Shah, M. 2014. Evaluating Learning Algorithms: A Classification Perspective. Cambridge University Press, New York.

Jo, T. 2021. Machine Learning Foundations – Supervised, Unsupervised, and Advanced Learning. Springer Nature Switzerland AG. ISBN 978-3-030-65900-4 (eBook)

Joshi, A.V. 2020. Machine Learning and Artificial Intelligence. Springer Nature Switzerland AG. ISBN 978-3-030-26622-6 (eBook).

Komorowski, M., Marshall, D. C., Salciccioli, J. D. & Crutain, Y. 2016. Exploratory Data Analysis. In: MIT Critical Data. Secondary Analysis of Electronic Health Records. Springer Cham. ISBN 978-3-319-43742-2 (eBook). pp. 185-203.

Kordon, A.K. (2020) Applying Data Science, How to Create Value with Artificial Intelligence. Springer Nature Switzerland. ISBN 978-3-030-36375-8 (eBook).

Kotsiantis, S. B., Kanellopoulos, D. & Pintelas, P. E. 2006. Data Preprocessing for Supervised Learning. International Journal of Computer Science. Vol. 1, no. 1, pp. 111-117.

Lampinen, O. 1998. Suomen koulutusjärjestelmän kehitys. Tampere: Tammerpaino Oy.

Lehtisalo, L. & Raivola, R. 1999. Koulutus ja koulutuspolitiikka 2000-luvulle. Juva: WSOY.

Lesinski, G. Corns, S. & Dagli, C. 2016. Applications of and Artificial Neural Network to Predict Graduation Success at the United States Military Academy. Procedia Computer Science. Vol 95, pp. 375-382.

Liu, L. & Özsu, M. T. (eds.). 2018. Encyclopedia of Database Systems. 2nd ed. Springer, New York. doi: https://doi.org/10.1007/978-1-4614-8265-9.

López Guarín, C. E., León Guzmán, E. & González, F. A. 2015. A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje. Vol. 10, no. 3, pp. 119-125. doi: 10.1109/RITA.2015.2452632.

LUT University. 2023 a. Degree regulations of the Lappeenranta-Lahti university of technology LUT. Available: https://elut.lut.fi/en/completing-studies/rules-and-regulations/degree-regulations [19.10.2022].

LUT University. 2023 b. Study rights : Duration of the right to study for a bachelor's and master's degree and Total duration of the study right. eLUT info page for LUT students. Available: https://elut.lut.fi/en/completing-studies/planning-your-studies/study-rights [19.10.2023].

Marcot, B. G. & Hanea, A. M. 2021. What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis? Computational Statistics. Vol. 36, no. 3, pp. 2009-2031.

Merenluoto, S. 2009. Menestyksekkäät yliopiston pelaajat? Tutkimus nopeasti ja nuorena valmistumisesta. Turun yliopiston julkaisuja, sarja C, osa 286.

Mikkonen, J. 2000. Opintoviikon ongelmat. Helsingin yliopisto. Opintoasiainosaston julkaisuja 20/2000.

Ministry of Education and Culture. 2019. Opetus- ja kulttuuriministeriön asetus yliopistojen perusrahoituksen laskentakriteereistä. Available: https://www.finlex.fi/fi/laki/alkup/2019/20190119 [referred: 1.11.2023]

Ministry of Education and Culture. 2021 a. Universities Core Funding From 2021. Available: https://okm.fi/documents/1410845/4392480/UNI_core_funding_2021.pdf/a9a65de5-bd76-e4ff-ea94-9b318af2f1bc/UNI_core_funding_2021.pdf?t=1608637262540 [referred 1.11.2023]

Ministry of Education and Culture. 2021 b. Agreement between Lappeenranta-Lahti Technical University and Ministry of Education and Culture for years 2021 – 2024. (In Finnish). Available: https://okm.fi/documents/1410845/3992814/Lappeenrannan+teknillinen+yliopisto+sopimus+2021-2024.pdf/fcd553da-a005-2986-1067-6ef11a94bb44/Lappeenrannan+teknillinen+yliopisto+sopimus+2021-2024.pdf?version=1.1&t=1611841175096 [referred: 1.11.2023]

Mohammad Suhaimi, N., Abdul-Rahman, S., Mutalib, S., Abdul Hamid, N.H., Md Ab Malik, A. 2019. Predictive Model of Graduate-On-Time Using Machine Learning Algorithms. In: Berry, M., Yap, B., Mohamed, A., Köppen, M. (eds). Soft Computing in Data Science. SCDS 2019. Communications in Computer and Information Science. Vol. 1100. Springer, Singapore. https://doi.org/10.1007/978-981-15-0399-3_11.

Nakatsu, R. T. 2020. An evaluation of four resampling methods used in machine learning classification. IEEE Intelligent Systems. Vol. 36, no. 3, pp. 51-57.

Novaković, J.D., Veljović, A., Ilić, S.S., Papić, Ž. & Tomović, M. 2017. Evaluation of Classification Models in Machine Learning. Theory and Applications of Mathematics & Computer Science. Vol. 7, no. 1, pp. 39-46.

Nummenmaa, L., Holopainen, M. & Pulkkinen, P. 2019. Tilastollisten menetelmien perusteet. Sanoma Pro Oy. ISBN 978-952-63-6337-0.

Pajala, S. & Lempinen, P. 2001. Pitkä tie maisteriksi : Selvitys 1985, 1988 ja 1991 yliopistoissa aloittaneiden opintojen kulusta. Opiskelijajärjestöjen tutkimussäätiö Otus rs 22/2001. Yliopistopaino, Helsinki.

Pang, Y. Judd, N., O'Brien, J. & Ben-Avie, M. 2017. Predicting Students' Graduation Outcomes through Support Vector Machines. IEEE Frontiers in Education Conference (FIE), Indianapolis, IN, USA. pp. 1-8. doi: 10.1109/FIE.2017.8190666.

Peling, I. B. A., Arnawan, I. N., Arthawan, I. P. A. & Janardana I. G. N. 2017. Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm.

International Journal of Engineering and Emerging Technology. Vol. 2, no. 1, pp. 53-57. Available: https://ojs.unud.ac.id/index.php/ijeet/article/view/34457. [referred 31.10.2023]

Provost, F. & Fawcett, T. 2013. Data Science for Business. O'Reilly Media, United States of America. ISBN 978-1-449-36132-7.

Qamar, U. & Summair Raza, M. 2020. Data Science Concepts and Techniques with Applications. Springer Nature Singapore. ISBN 978-981-15-6133-7 (eBook).

Quinlan, J. R. 1986. Induction of decision trees. Machine Learning. Vol. 1, pp. 81-106. https://doi.org/10.1007/BF00116251

Rantanen, E. & Liski, E. 2009. Valmiiksi tavoiteajassa? Teknillistieteellisen alan opiskelijoiden opintojen eteneminen ja opiskelukokemukset tekniikan kandidaatin tutkinnossa. Teknillisen korkeakoulun Opetuksen ja opiskelun tuen julkaisuja 3/2009. Available: http://lib.tkk.fi/Raportit/2009/isbn9789512297740.pdf [referred 31.10.2023].

Reason, R. D. 2003. Student Variables that Predict Retention: Recent Research and New Developments. Journal of Student Affairs Research and Practice. Vol. 40, no. 4, pp. 704-723. doi: https://doi.org/10.2202/1949-6605.1286.

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. Vol. 40, no. 6, pp. 601-618.

Romero, C. & Ventura, S. 2013. Data mining in education. WIREs Data Mining and Knowledge Discovery. Vol. 3, issue 3, pp. 12-27. doi: https://doi.org/10.1002/widm.1075

Romero, C., Romero, J. R. & Ventura, S. 2014. A Survey on Pre-Processing Educational Data. In: Peña-Ayala, A. (eds) Educational Data Mining. Studies in Computational Intelligence. Vol 524. Springer, Cham. doi: https://doi.org/10.1007/978-3-319-02738-8_2

Romero, C &Ventura, S. 2020. Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery. Vol. 10, issue 3. doi: https://doi.org/10.1002/widm.1355

Siemens, G. & Baker, R. S J.d. 2012. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, April 2012, pp. 252-254. doi: https://doi.org/10.1145/2330601.2330661

Steele, B., Chandler, J. and Reddy, S. 2016. Algorithms for Data Science. Springer. Switzerland. ISBN 978-3-319-45797-0 (eBook). doi: 10.1007/978-3-319-45797-0.

Tampakas, V., Livieris, I. E., Pintelas, E., Karacapilidis, N. & Pintelas, P. 2019. Prediction of Students' Graduation Time Using a Two-Level Classification Algorithm. In: Tsitouridou, M., A. Diniz, J., Mikropoulos, T. (eds) Technology and Innovation in Learning, Teaching and Education. TECH-EDU 2018. Communications in Computer and Information Science. Vol 993. Springer, Cham. doi: https://doi.org/10.1007/978-3-030-20954-4_42

Tirronen, J. 2006. Kolme näkökulmaa yliopistoon: tutkimusta, opetusta vai palvelua? Jyväskylä: Koulutuksen tutkimuslaitos.

Universities Act. 1997. Universities Act 645/97. Amendments up to 1453/2006. Unofficial translation. Available: https://www.finlex.fi/en/laki/kaannokset/1997/en19970645_20061453.pdf [referred 1.11.2023]

Universities Act. 2009. Universities Act 558/2009. Amendments up to 644/2016 included. Translation from Finnish. Available: https://www.finlex.fi/en/laki/kaannokset/2009/en20090558_20160644.pdf [referred 1.11.2023]

Opetusministeriö. 1993. Koulutuksen ja korkeakouluissa harjoitettavan tutkimuksen kehittämissuunnitelma vuosille 1991–1996. Helsinki. Opetusministeriö.

Varewyck, M. & Martens, J.-P. 2011. A Practical Approach to Model Selection for Support Vector Machines With a Gaussian Kernel. IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics. Vol. 41, no. 2, pp. 330-240.

Vipunen. 2023. Vipunen – opetushallinnon tilastopalvelu (Education Statistics Finland). Yliopistot läpäisy, prosentit. Opetushallinnon ja Tilastokeskuksen tietopalvelusopimuksen aineisto. Available: https://vipunen.fi/fi-fi/_layouts/15/xlviewer.aspx?id=/fi-fi/Raportit/Yliopistot%20l%C3%A4p%C3%A4isy%20-%20yliopisto%20prosentit.xlsb. [referred 19.10.2023].

Välimaa, J. 2004. Nationalisation, Localisation and Globalisation in Finnish Higher Education. Higher Education Vol. 48, no. 1, pp. 27-54.