TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Kristin Ehala
166759 IVSM

CONTEXT BASED STUDY PROGRESS MONITORING MODEL

Master's thesis

Supervisors: Juhan-Peep Ernits
Ph.D
Sven Nõmm
Ph.D

Tallinn 2018

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Kristin Ehala
166759 IVSM

KONTEKSTIPÕHINE ÕPPEEDUKUSE MONITOORIMISE MUDEL

magistritöö

Juhendajad:	Juhan-Peep Ernits
	Doktorikraad
	Sven Nõmm
	Doktorikraad

Tallinn 2018

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Kristin Ehala

16.05.2018

Abstract

The amount of data that surrounds our everyday and professional life is growing in a rapid rate. People are making decisions not only based on their intuition and experience but on facts and data [1]. The same is happening in the academic area, but there is no uniformed method to put data into use, and make it beneficial for both the academic institutions and to students. This is because all courses and curriculums are different from each other. In this thesis two prediction methods are proposed to predict student progress at Tallinn University of Technology. First method proposed is called context based method, where data is modified according to the context it is needed. The context in the case of this thesis is the curriculum a student is studying on and a semester under analysis. The second method proposed is a general method, which takes into account descriptive features of students. General method does not fit the classification model for each of the students separately and does not use courses as features. The advantages and disadvantages of both models are brought out and analysed. The hypotheses are tested and possible use cases are described. It is stated in the thesis that context based method predicts graduation with the accuracy of 82% and general model predicts graduation with 91% of accuracy. Predicting graduation by semester yields to accuracy of nearly 96%.

This thesis is written in English and is 55 pages long, including 6 chapters, 17 figures and 11 tables.

Annotatsioon

Kontekstipõhine õppeedukuse monitoorimise mudel

Inimesi ümbritsev andmehulk kasvab iga päevaga. Sellest tulenevalt tehakse aina enam otsuseid põhinedes andmetel ja analüüsil, mitte ainult kõhutundel ja kogemusel. Sarnane tendents on toimumas ka haridusmaastikul, kus saavutamaks paremat haridust ja jätkusuutlikkust on ülikoolid hakanud kasutama endale kättesaadavaid andmeid hariduse kvaliteedi parendamiseks. Selles magistritöös esitatakse kaks meetodit, mille abil on võimalik TTÜ ÕIS-i andmete põhjal jälgida ning ennustada tudengite õppeedukust ülikoolis. Esimene meetod on kontekstipõhine meetod, mis kasutab iga tudengi tehtud sooritusi ja vastavalt neile ennustab tema jätkamist ülikoolis ja kooli lõpetamist. Kontekst selle töö mõistes on õppekava ja semester, millel tudeng õpib. Teine esitatav meetod on üldine meetod, mis kasutab agregeeritud näitajaid tudengi kohta ja muid kättesaadavaid kirjeldavaid atribuute. Üldine mudel ei ole individuaalne igale tudengile. Magistritöös on välja toodud arendamisel esinenud takistused ning mõlema meetodi analüüs ja võrdlus. Lisaks on kirjeldatud, kuidas mõlemat meetodit oleks eesmärgipärane kasutada ning kuidas saaks tulemusi potentsiaalselt parendada. Tulemustes tuleb välja, et kontekstipõhine mudel saavutab lõpetamise ennustamisel täpsuse 82% ning üldine mudel 91%. Ennustades semestri haaval tudengi lõpetamist võib ennustustäpsus küünida kuni 96% juba viie semestri andmete põhjal. Erinevused tulevad välja erinevate õppekavade analüüsimisel. Lisaks on kirjeldatud, et kontekstipõhine meetod on edukas ennustamaks tudengi õppimise jätkamist järgneval semestril. Töös on kirjeldatud ka õppetunnid ja tulevikus uuritavad probleemid ja küsimused, mis kerkisid tehes seda magistritööd.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 55 leheküljel, 6 peatükki, 17 joonist, 11 tabelit.

List of abbreviations and terms

API	Application programming interface
CART	Classification and Regression Tree
CHAID	Chi-square automatic interaction detection
Context based method	Method that predicts student graduation based on context (grades and curriculum)
CSV	Comma-separated values
Dataframe	DataFrame is a 2-dimensional labelled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects [2]
EAP	*Euroopa ainepunktisüsteemi ainepunkt.* Same as ECTS.
ECTS	The European Credit Transfer and Accumulation System. Same as EAP.
EDM	Educational Data Mining
EWS	Early Warning System
General method	Method that predicts student graduation based on general features like curriculum, stipend, sex etc.
IDE	Integrated development environment
KKH	*Kaalutud keskhinne*, Average grade
ML	Machine Learning
Original dataset	Students from the original time period under analysis. 2012 - 2018
Pre-reform dataset	Students from experimental analysis dataset. 2008 - 2012
SAIS	Sisseastumise Infosüsteem (Admission information system)
SSH	Secure Shell
TTU	Tallinn University of Technology
VÕTA	*Varasemate õpingute ja töökogemuse arvestamine*; The meaning is equal to RPL = Recognition of Prior Learning; APEL = Accreditation of Prior and Experiential Learning
ÕIS	Study Information System, Õppeinfosüsteem

Table of contents

List of figures

List of tables

1 Introduction

The amount of data that surrounds our everyday and professional life is growing in a rapid rate. This means that more and more decisions are made using data-driven and evidence-based methods using automated tools. One of the areas where the evidence based approach would be useful is understanding the reasons why students drop out from school by predicting it early so that something could be done to mitigate it.

Universities and colleges have been trying to pinpoint why students drop out of their institutions for a while. It has been under the discussion, what are the main reasons why students leave their programs and also how can institutions increase the student retention rates [3]. The discussion of developing so-called early warning systems is also important regarding secondary and high schools [4]. In an article by “American Graduate DC”, it is said that in every 26 seconds one student drops out of high school [5]. This shows an immerse need for systems to be developed to reduce the dropout rate and increase the percentage of graduation.

According to an analysis conducted in 2015 by Estonian Ministry of Education and Research, the number of students who will get a higher education is going to decrease about 4% a year during the years 2010-2020 [6]. It is also described that the number of students who will continue their higher education in Estonia straight after high school also decreases. It is believed that the reduction is due to more young people going to study abroad.

Tallinn University of Technology (TTU) is a university in Estonia that was established in 1918. In 2016 the university had 11070 students and 106 different curricula [7]. In TTU the issue of high student dropout rate has been under focused attention for several years. According to a report published in 2016 the student dropout rate was 17-23% from the total number of current students in the university. After the first year of studies, about 15% of the students who started will be exmatriculated. About a third of them were capable students with mathematics state exam result of over 75 points out of a hundred [8]. This shows the need for analysing and improving the system. In 2016 TTU started a development program called ASTRA that aims to develop a system for higher education institutions to use online platforms in order to reduce the student dropout rate [8] [9]. ASTRA is an Estonian program financed by the European Cohesion Fund [10].

1.1 Goals of the Thesis

In this thesis we will focus on the analysis of the data provided by the Study Information system at TTU (ÕIS). We will develop a context based machine learning classifier model to predict whether a student is in line of graduating the university or more likely to drop out or quit the university studies. The context in the case of this thesis is the curriculum a student is studying in. Context based method means that the fitted machine learning model will be modified to work individually on all undergraduate and 2nd level (MSc) curriculums of the Tallinn University of Technology, providing the information that is related to a specific curriculum and students taking the curriculum. The results of a context based model are compared to the results of a more general model, trained on features common to most students. We restrict ourselves to using interpretable models such as decision trees [11] [12] [13] because the interest is not only to predict dropout, but understand the reasons behind it and help find ways to increase student retention.

We seek answers to the following research questions:

1. What is the most feasible method for predicting dropouts: context based approach based on curriculum or a general common student feature based approach?

2. How well can we predict dropout based solely on Study Information system data?

3. How does the prediction accuracy change over period of different consecutive semesters?

To answer the questions, we seek to achieve the following objectives:

1. To build relevant datasets from the data gathered from ÕIS database, so it is possible to use them for classification.

2. Find out if and when it is possible to use the context based model on student's data, to predict their dropout.

3. To provide an automated approach to finding out what are the key courses in each curriculum that have the highest correlation to student retention.

4. Compare a general prediction method, that uses general data about a student to a curriculum context based method, and bring out advantages and disadvantages of both.

5. To investigate according to the advantages and disadvantages which model is most suitable to be used in practice.

1.2 Research Method and Tools

The research method applied in this thesis is experimental empirical research. The hypotheses are tested and the results are compared to bring out advantages and disadvantages of the methods. We follow the ideas in [12] to validate the results by being able to interpret them.

For developing Python 3.5 programming language is used. Python libraries used for data cleaning, transformation and machine learning are Pandas (version 0.22.0), NumPy (version 1.14.2) and Scikit-learn (0.19.1). Pandas is a Python library used for data analysis, providing simple data structures to manipulate the data [14]. NumPy is a Python package for scientific computing [15]. Scikit-learn is a machine learning library, built on NumPy and SciPy. Scikit-learn has built in tools for data analysis, data cleaning and machine learning [16].

For visualising the data libraries Matplotlib, Seaborn (version 0.8.1) and Excel are used. Matplotlib is a plotting library for Python [17]. Seaborn is Python vizualisation library, built on Matplotlib. It provides a simpler interface when working with visualizations [18]. The IDE for development is Spyder 2.3.8.

In addition to following the data protection laws in force at the time of writing the thesis, we followed practices guided by the General Data Protection Regulation (GDPR) [19]. This involved working with anonymized data stored in a secure server. Access to the server was provided from the campus network. During data processing the data never left the secure server.

In this thesis:

1. Two different methods (context based method, general method) are being used to predict student graduation. The methods are compared.

2. The context based method uses the grades of the student to assess his or her current progress at the university. The model will predict “0” if the resemblance of the progress is more to students who have quit their studies. This means that the student under analysis might also follow the pattern and drop out.

3. The models for context based datasets select meaningful features by themselves, according to the student that the prediction is being done for. The features vary for every prediction and are personalised for a student.

4. From data, some derived variables are composed using feature engineering. These are expected to add extra knowledge for the model. The features can easily be added to the datasets for any student.

5. Context based models takes into account the concept drift, that happens overtime when curriculums change and students improve.

6. The predictions for context based model are expected to work by semester, considering how far the student is with his or her studies.

Additionally, three working hypotheses are generated. With this thesis these hypotheses should be proved or refuted. It is important to keep in mind, that assumptions made beforehand may be wrong and it is important to get the answers from the data.

Working hypothesises are:

1: If context based method is being used for predicting student dropout, the results are more accurate than using a more general model.

2: Context based method is better at predicting continuation at studies by semester and worse at predicting actual graduation or dropping out.

3: General model predicts better when the data is grouped by faculties, compared to when data is not grouped by faculties.

1.3 Overview of the Thesis

In the related works section, I introduce what similar works and researches have been done with the topic previously.

In the theoretical background section, I describe the knowledge that is necessary to understand the thesis and its methods used.

In the data preparation section, I give an overview on how the data is mined and held. I give a brief introduction of the data and explain the background of it. I also describe how the datasets are constructed and how the data is cleaned. Additionally, the data transformation to fit the machine learning models is described. Problems encountered during the process are also described.

In the results and analysis section I experiment with the methods developed in the data preparation section and analyse the results. I bring out disadvantages and advantages of the methods and also point out some interesting findings that emerged from the data. Ideas for future work, based on the results are brought out.

In the conclusion I bring out all the outcomes of the research. The results for research questions and hypothesis are presented.

2 Related Work

In the past, the process of investigating issues related to student dropout rate was manual. It included many questionnaires, interviews and in many ways, a privacy invasion into a student life. The process was time consuming and invasive. An example of it is a study made in 1966, which was conducted among the engineering students at Three Midwestern Universities [20]. This study was done using only statistical methods and models, which are good to try to understand the features involving the problem.

A guide [4] on how to build early warnings systems brings out four factors that must be taken into account when selecting the indicators, that influence students' progress at school. It claims that the right indicators must be valid for intended purpose, actionable by schools, meaningful and easily understood and aligned with district and school priorities. This guide also indicated that it would be more reasonable to select indicators that are connected with educational institutions and leave characteristics that are important outside of school out of the scope, even though they have high correlation rate with the outcome. The paper also stated that using a small set of factors lets the system to be more efficient at first [4].

Selecting meaningful indicators is also crucial for machine learning algorithms and methods that predict the progress of students'. In 2013 a research team used Apiori algorithm to predict the student outcome for certain subjects. With this algorithm they calculated confidence and support for a student getting a specific grade [21]. Their solution was not automatic and the data about students had to be inserted manually. They also noticed that the results for male and female students varied. The outcome of their research was a REPS – Result Prediction System, which allowed academy staff to query information about student through a GUI and see how likely they were to get a specific grade for a specific course.

In a research published in 2015 by Sung-Po Lin [1] decision tree method CHAID was used. The research developed a EWS system for Taiwan University and aimed to understand what are the main factors that influence student dropout rate. The research included all divisions: bachelor, masters and PhD. The results gave an interesting overview of how different variables, such as personal, family, academic, health, work and economic factors can influence the final outcome and whether the student graduates university or not. The final decision tree was done using vertices sex, place of residence, college, division, admission channel, category of identity, tuition waiver, amount of student loan and learning status [1]. In this paper more effort was put of understanding the data before using it for analysis and it was interesting to see that some not obvious features were used to analyse dropout.

In a research published in the beginning of 2017, a research group claims to be the first one to use various different ML techniques to solve the problem of student dropout rate [22]. Their work focused on three majors from the University of Barcelona, and compared five different classifier methods in order to answer to a question "Is it possible to predict if a student will enrol to university in the second or third year given information of the first academic year?" [22]. The researchers decided not to use any descriptive features in the model, besides the grades, to make the results independent from external factors. The methods used in that research were logistic regression, support vector machine, random forest, Gaussian Naïve Bayes and adaptive boosting. ML methods reached accuracy of nearly 90%, but because the models were applied to different majors separately, they reached slightly different results.

A paper [23] focused on the relationship between student characteristics from before college and the progress at the university and investigated the causes of student retention. The work states that the precollege characteristics do play a significant role in predicting student's retention, but do not explain every reason for it. In the research they wanted to illustrate the use and advantages of survival analysis that was performed on a retention data and bring out factor that influenced the student retention at Oregon State University. The research also brought out some implications that could be used to help with the retention.

In the paper published in 2015 [24] the researchers were looking for answers to questions, which variables are the most discriminative in terms of predicting student retention and what features are the most predictive in terms of student success and retention over the course of several semesters. It was also under analysis, if the measures and variables of the data change with time. The research looked into motivational factors.

In 2017 a bachelor’s thesis [25] was defended at TTU analysing student retention at TTU with various machine learning methods, using student data from Informatics BSc curriculum. It was stated that the accuracy with various models and datasets yielded to be between 60-90%. From the results it was visible that the best method used was the decision tree method. It is also stated that the biggest correlation between graduation and features is EAP sum, stating that if a student has 151 or more EAP in total, he or she is more motivated to continue. It was also stated that speciality courses were more influential in terms of dropout.

3 Theoretical Background

In this paragraph the theoretical background related to the thesis is described. Methods used in this thesis are explained. Brief overview is given about educational data mining, feature selection, machine learning methods used and methods for interpreting and assessing machine learning models.

3.1 Educational Data Mining

Data mining is analytical process designed to explore data. A term EDM is used to describe the process of data gathering from educational environment. It is educational data mining. A method uses data coming from educational environments in order to better understand students and their academic surroundings [26]. It helps to make sense of the data that is created by learning-related environments. The popularity of this field has increased in the past ten years and first official conferences about the topic were held in 2007 and 2008 by Educational Data Mining Society [26]. According to Wikipedia it aims to predict students future learning behaviour, improve and discover domain models, study the effect of educational support and advance scientific knowledge about learning and learners [27].

3.2 Machine Learning Method Used

In the context of the current thesis, we are solving the classification problem. There are many different approaches to solving the classification problem, but we limit our choice to the methods that support interpretability. While linear classifier can be considered interpretable as well and has been used in e.g. [23] to predict student retention, we restrict the approach in the current thesis to decision trees. Those have support for interpretability and in previous research in predicting student retention at TTU, decision trees performed very well [25].

3.2.1 Decision Tree

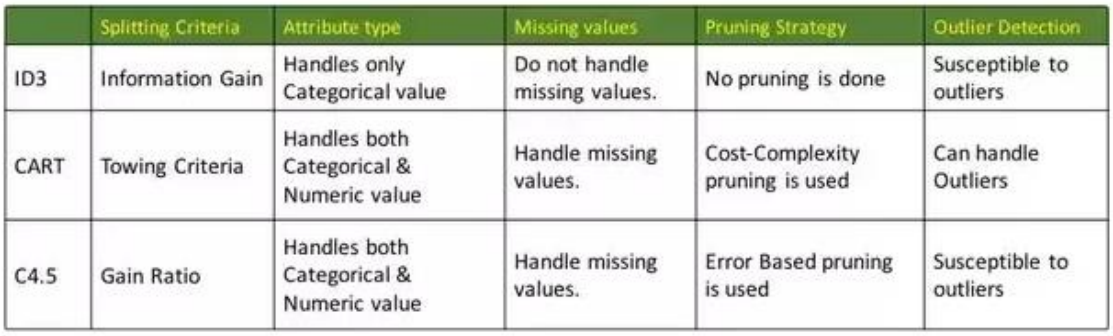
Decision tree is one of the most popular methods used in ML, data mining and statistics. It a predictive method that is used for classification the result label and is part of the supervised machine learning methods. It is a helpful method that can demonstrate visually how the decisions are made and what are the key factors on how the target result has been reached.

Regression tree and classification tree can be used for solving regression and classification tasks. There are also additional so-called combinational methods such as boosted trees and bootstrap aggregated including random forest.

During the construction of decision trees it is possible to use various algorithms to choose the next feature to make the next choice. The most common ones are ID3 and C4.5. CART, CHAID, MARS and Conditional Interface Trees represent alternative tweaks to the ways decision trees are constructed. Both CART and CHAID were used in [1] to implement EWS system in university and predict the student progress. To facilitate the understanding of the approach taken in related work, ID3, CHAID and CART are described in more detail.

ID3 stands for Iterative Dichotomiser 3 and is an algorithm used in machine learning and especially in the natural language processing domain. In the process of ID3 an entropy of each leaf is calculated during each step to find out the leaf with the smallest entropy. The next step is to split the set of attributes into subset according to the resulting entropy. This is repeated on every subset until there are not attributes left. Disadvantages are that numeric and missing values are not supported. In comparison to C4.5, in ID3 both numeric and categorical values are supported.

CHAID method is a chi-squared automatic interaction detection. It uses Chi-Square test to determine the best next split at each step when dealing with classification problem. F test is used when dealing with regression problem. If the test result for predictive categories is not significant, the categories will be merged and this step will be repeated. After that the predictor variable that will yield the most significant split will be chosen until there are no splits left. This method requires quite a large scale sample size, because the method by default naturally defines the classes and the distribution is expected to be fairly equal.



	Splitting Criteria	Attribute type	Missing values	Pruning Strategy	Outlier Detection
ID3	Information Gain	Handles only Categorical value	Do not handle missing values.	No pruning is done	Susceptible to outliers
CART	Towing Criteria	Handles both Categorical & Numeric value	Handle missing values.	Cost-Complexity pruning is used	Can handle Outliers
C4.5	Gain Ratio	Handles both Categorical & Numeric value	Handle missing values.	Error Based pruning is used	Susceptible to outliers

Figure 1. Comparison of different decision tree methods [28].

CART are types of decision trees that produces either classification or regression trees depending on the input variable. CART constructs a binary tree, meaning that each node has 2 outgoing branches. Compared to ID3, this method can handle missing values. In Scikit Learn, an optimised version [29] of CART is used. For this thesis, CART type decision trees are being used.

3.3 Interpretation of Machine Learning Models

In this thesis interpreting machine-learning models correctly, plays a big role. Since the data is gathered from educational background and will be used to benefit the same field, it is important to clearly state why students drop out, what are the main bottleneck courses and how this problem could be mitigated. In terms of this thesis, since machine learning model is made especially for each curriculum and each student, the interpretability is different for each. In an article "Interpreting machine learning models" it is said: "Interpretability of data and machine learning models is one of those aspects that is critical in the practical 'usefulness' of a data science pipeline and it ensures that the model is aligned with the problem you want to solve. [30]". The topic is also widely discussed in [11, 13, 12].

One of the most interpretable machine learning models is decision tree, since it is possible to clearly demonstrate why a certain decision was made. When using random forest, the model will lose its interpretability significantly, but it is still possible to analyse it. This is the reason why for this thesis it is decided to use only decision tree model, even though the prediction results for other methods could be better.

In addition to interpreting the model, it is important to evaluate the results on how well a model predicts an outcome. For this feature engineering is used to modify the features to carry the most meaningful data and methods and metrics like accuracy, precision and recall will be used.

3.3.1 Feature Engineering

Feature engineering is part of the data preparations that is done before any model is applied to the data. The idea is to manually modify the features, so they would carry more meaning in the result of the model [31]. For this thesis I have used two techniques of feature engineering: feature selection and feature construction. The difference between feature selection and feature construction is that while the first aims to select the most meaningful features from data, to therefore speed up the model and make it more interpretable, feature construction tries to take what features are in data, and modify them to carry a more meaningful value. In the article [31] it is stated that “Much of the success of machine learning is actually success in engineering features that a learner can understand. The flexibility of good features will allow you to use less complex models that are faster to run, easier to understand and easier to maintain.”

To be able to go through a classification process, it is important to use feature selection methods to be able to select the most meaningful features to be present in the machine learning model. It is important to note, that when the complexity of the model is increased, meaning more features are used, the interpretability of the machine learning model decreases significantly. Feature selection helps to evaluate what features bring more meaning to the data and carry more information regarding to the classification label that we are trying to predict. This is particularly important, since for this work data training sets with over fifty features can be used.

Feature selection has three main categories: filter models, wrapper models and embedded models. With filter models a class-sensitive discriminative criterion is used. The criterion is used to filter out irrelevant features and keep only the most discriminative ones [32]. Wrapper models detect the possible interactions between variables [33]. Since this method uses subset of features, the computation time is expected to be greater than with filter methods. Embedded methods are considered to be a combination of both previous ones. The knowledge about the features is embedded into the classification problem [32].

In this thesis, filter models are being used. Three main methods considered as filter methods in feature selection are entropy, GINI index and Fisher score.

3.3.2 Entropy

"Entropy is defined as the average amount of information produced by a stochastic source of data" [34]. It is used to understand how much information a feature carries when compared to the feature that will be predicted. It is a method that has its roots in the sound information-theoretic principles and has a similar goal as does GINI index [32]. A higher result implies that the feature carries so called mixed signals, and a result of 0 shows the best discriminative power of a feature.

$$E(v_i) = -\sum_{j=1}^{k} p_j log_2(p_j)$$

3.3.3 GINI Index

Gini index is also used to measure the discriminative power of a categorical feature. The smaller the index is, the better the feature is in a relation to the predictive feature. In this thesis, Gini index is used to evaluate feature importance in a decision tree classifier, provided by Scikit Learn.

$$G = \sum_{i=1}^{r} n_i G(v_i)/n$$

3.3.4 Fisher Score

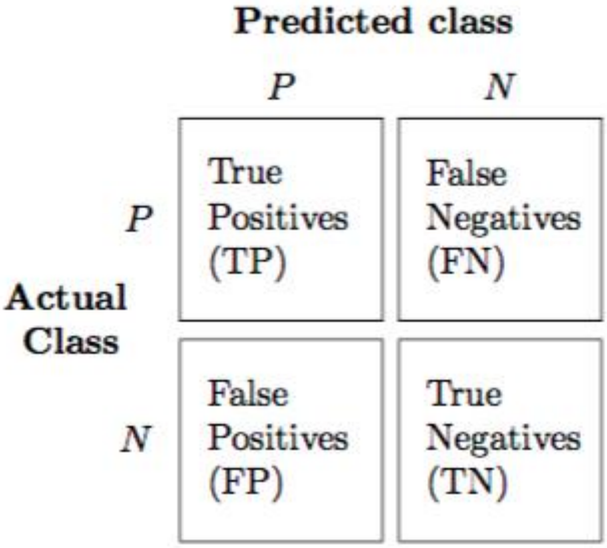
Fisher score is used to demonstrate the discriminative power of a numerical feature. The greater the score, the better discriminative power a feature has. Normalized fisher score has a value between {0; 1}. In this thesis Fisher score is used to evaluate the discriminative power of features in context based model.

$$F = \frac{\sum_{j=1}^{k} p_j(\mu_j - \mu)^2}{\sum_{j=1}^{k} p_j \sigma_j^2}$$

3.3.5 Confusion Matrix

A popular way to evaluate how correctly a classification machine learning model predicts, is to build a confusion matrix and calculate rates that are used for evaluation.

Confusion matrix is 2x2 matrix (in case of two labels), that gives an overview of true positives, true negatives, false positives and false negatives. True positives (TP) represent cases where both the prediction and original value are True, meaning the feature that is being predicted is true in both cases. True negative (TN) shows that both prediction and original value are False. False positive (FP) represent cases when prediction has made a decision that the value is True, whereas in reality the value is False. This is also known as "Type I error". False negatives (FN) represent cases when prediction is that the value is False, but in reality, the value is True. This is known as "Type II error" [35].



		Predicted class	
		P	*N*
Actual Class	*P*	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Figure 2. Example of a confusion matrix [36].

When using confusion matrix for an interpretation method, it is important to understand the domain you are predicting for, since sometimes it is better when there are more true positives than true negatives and vice versa. For this thesis the FP and FN demonstrate students about whom the predictions are performed wrongly and they play a significant role for predicting continuation of studies at the university.

There are different rates that are calculated using confusion matrix, to give more information about the model.

Accuracy is calculated to give an overview on how often the classifier classifies correctly.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision describes how often the prediction is correct, when it has predicted True.

$$Precision = \frac{TP}{TP + FP}$$

Recall is also known as a true positive rate (TPR) and it demonstrates the rate the model predicts True, when the value actually is True.

$$Recall = \frac{TP}{TP + FN}$$

4 Preparation of ÕIS Data

In this paragraph I describe what is needed to be done in order to gain the expected value and outcome of this thesis. I describe the process of gaining access to the data and steps taken to clean and prepare the data for analysis. I give an overview how the data is protected to mitigate a possible data breach. Next, the preliminary data analysis, data cleaning and data transformation are described, with the experiments that are vital to understand the data and use it in the models the best possible way. It is also described how a specific model is built for each curriculum and for predicting with a more general dataset that is not curriculum based. Most of the time spent for this thesis focused on constructing relevant datasets and dataframes and cleaning the data.

4.1 Data Acquisition

In this thesis I am investigating and using the data from the Study Information System (ÕIS) of the Tallinn University of Technology. The database system is developed by a software company AS Fujitsu. All student, studies and curriculum information is being held in that database.

The data used is from three database schemas ISIK_1, OIS2 and OIS3. ISIK_1 includes tables about student information. From there information about the student sex, activity (*aktiivsus*) and graduation (*lopetanud*) is collected. OIS2 schema includes tables regarding information about studies. This includes the grades and course declarations the students have done. In the OIS3 schema, the information about structures of curriculums and units of the university is kept.

4.1.1 Data Protection

Any type of personal information must be held securely, especially when dealing with data that is not your own. This has become especially important with the data protection regulations that are applied to all European companies starting from May 2018 (GDPR). For that, various security layers must be applied to the dataset, to be sure that the data is safe and not recognisable or reachable for outside user. Personal data must and will be anonymised. Among the anonymized data is the student id, student surname and family name and also the title of his or her thesis, since it publicly available and therefore can be used to de-anonymize the data. The data must also be handled in a secure environment to minimize the risk of any possible data breach.

ÕIS system data for this thesis is being held on a Linux virtual server. The data anonymization is done on the Integration layer, that is on top of the Linux analysis server. Data anonymization and building a virtual environment is trusted to the TTU IT department to mitigate any possibility of data leak and to divide the risk and responsibility.

For communication with the server we use SSH public and private key based authentication. The client used is PuTTY [37]. It is an open source SSH client for the Windows platform. SSH is a cryptographic network protocol for operating network services securely over an unsecured network [38]. The best-known example application is for remote login to computer systems by users. All testing and analysing is done on the secure Linux server. The files and data are also kept there. In addition, Cygwin and Xterm are used to visualise and work with the data on the server. Cygwin supports X server and enables X to render graphical user interface in Windows [39].

4.1.2 Data Construction

The access to a selected set of database tables and API is enabled by TTU IT department. Since they are responsible for the anonymizing of personal data, the process of gaining data is time consuming. Each table that contains any personal information, must be anonymized and the analysis of the database schema slowed down the process of accessing data. We were granted access to an anonymized schema that mirrors the actual schema of ÕIS. In [25] data had been pre-processed and made much simpler to access. The data tables are accessible through an API built between the Linux Server and ÕIS database. Tables queried through the API are then saved into CSV files and made accessible for specific users. The tables are copies of the current state of the database, but anonymized. The database tables queried are described in Table 1.

Table 1. ÕIS database tables queried.

DB table queried	Schema	Description
c_tudeng	ISIK_1	Academic data about a student
c_dokument	ISIK_1	Student related directives and orders.
s_str_yksus	OIS3	Structural units
f_sooritus	OIS2	Students grades
f_opinguk_aine	OIS2	Subject-lecturer pairs added to the curriculum
f_aine_opetamine	OIS2	Subject-lecturer pairs
f_sooritus_kavas	OIS2	Results for a course in a curriculum.
a_aine	OIS3	Course/subjects data
f_avaldus	OIS2	Student statements
k_oppekava	OIS3	Curriculum data
k_kava_versioon	OIS3	Curriculum versions

To be able to create correct dataset for analysing, the help of the IT department was required again, since they have the domain knowledge of what information is kept in what tables. To be able to generate relevant datasets, I had to understand what features were necessary and how to join them into datasets all together. The TTU IT team provided the needed information for querying the data. For this I used the Python Pandas dataframe functions and decided not to use a local database. If the approach is going to be used in the future in an online system to predict student retention, it makes sense to build most of the database queries into the integration layer.

From the tables queried in Table 1, initial datasets are joined. This is important in order to gain the datasets that carry the most relevant information for this thesis. Initially most of the tables joined together carry also categorical features, to better describe the numerical features. This can be demonstrated with for example course version id that is unique and used for the ML model, and also the course name, that is categorical and not unique and helps us get a better understanding and meaning behind the numbers. The overview of the tables joined can be seen in Appendix 1 – ÕIS Database Tables Used.

The initial tables are saved as dataframes that will directly be used in the process of developing the model for predicting student dropout. Dataframes are not saved into a csv files. This means that all ÕIS database tables are read in every time and when new data is queried from the database, the method need to be modified to read in the new generated CSV files.

In Table 2 it is visible, what dataframes were constructed, what features each of them has, and how many rows they have. Table 2 also describes what initial database tables were necessary to create the dataframe.

Table 2. Dataframes developed for analysing the data.

Dataframe	Description	Features	Rows
All_students	Dataframe includes all students provided in TTU database. Tables joined: ISIK_1.c_tudeng, ÕIS3.s_str_yksus.	54	125230
All_results	Dataframe includes information about all results that have been done by the students at TUT. Tables joined: OIS2.f_sooritused, "OIS2.f_opinguk_aine ", "OIS2.f_sooritus_kavas", "OIS2.f_aine_opetamine", "OIS3.a_aine".	16	2322427
Curriculums	Dataframe includes information about all curriculums and curriculum versions. Tables joined: OIS3.k_oppekava, OIS3.k_kava_versioon.	6	703
Five_years_students	Segment of students from the all_students dataframe, who have started their studies after August 2012.	54	26905
Five_year_results	Segment of results that have been done by students in the five_years_students dataset.	16	461331

4.1.3 Preliminary Data-Analysis

After the process of constructing the datasets, it was important to gain an understanding of other features that might become useful when building the machine learning model and to understand the background of the data.

From the tables mentioned in Table 1 we selected a 5-year time period to use for analysis. These are dataframes Five_years_students and Five_year_results from Table 2. The data was gathered from 1st of August 2012 to 29th of March 2018. The scope of the data includes all TTU faculties and all courses within the curriculums. During the time of the thesis, in TTU there are 5 faculties, 152 different curriculums, with total number of 300 versions of curriculums and 4264 courses in total. There are 26905 students described, who have started their studies after the 1st of August 2012 and there are 461331 different declarations of courses done during this time period.

In Figure 3 it is visible that from the 26905 of students who have started their studies at TTU after 1.08.2012, only 5283 have graduated, 11832 have dropped out or decided to quit their studies, and 9790 are still active students. Overall there were 13432 active students at the TTU at the time of accessing the data. School of Engineering has 8196, School of Business and Governance has 7959, School of Information Technologies has 7396, Estonian Maritime Academy, 1757 and School of Science has 1597 students. Additionally, there are students studying at TTU open university who are not used for predicting any results in this thesis.

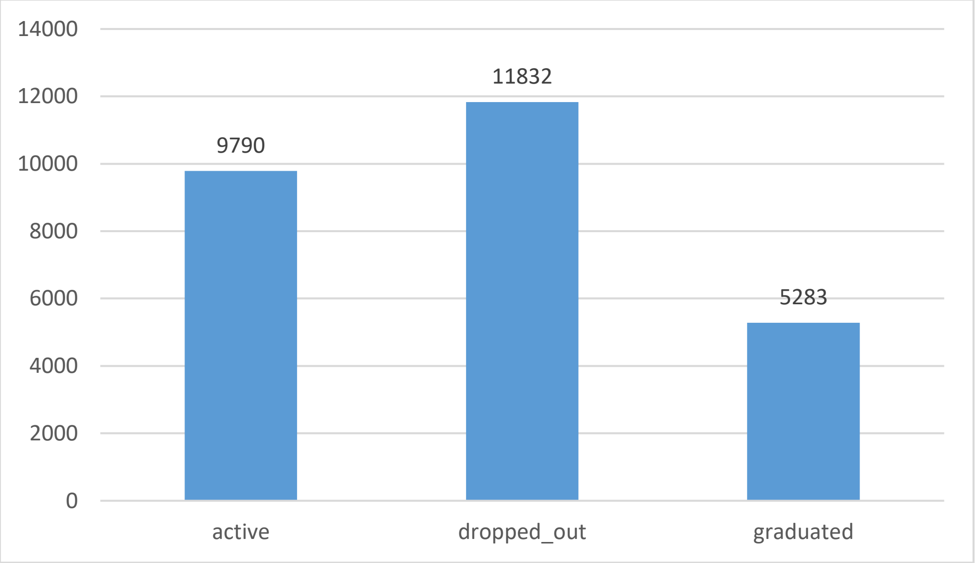


Figure 3. Distribution of students by type in TTU from 2012 – 2018.

The diagram in Figure 4 shows the actual numbers of students who graduate and who for one reason or another, decide to not to do so. The relation of men and women graduating, dropping out or studying is also depicted.

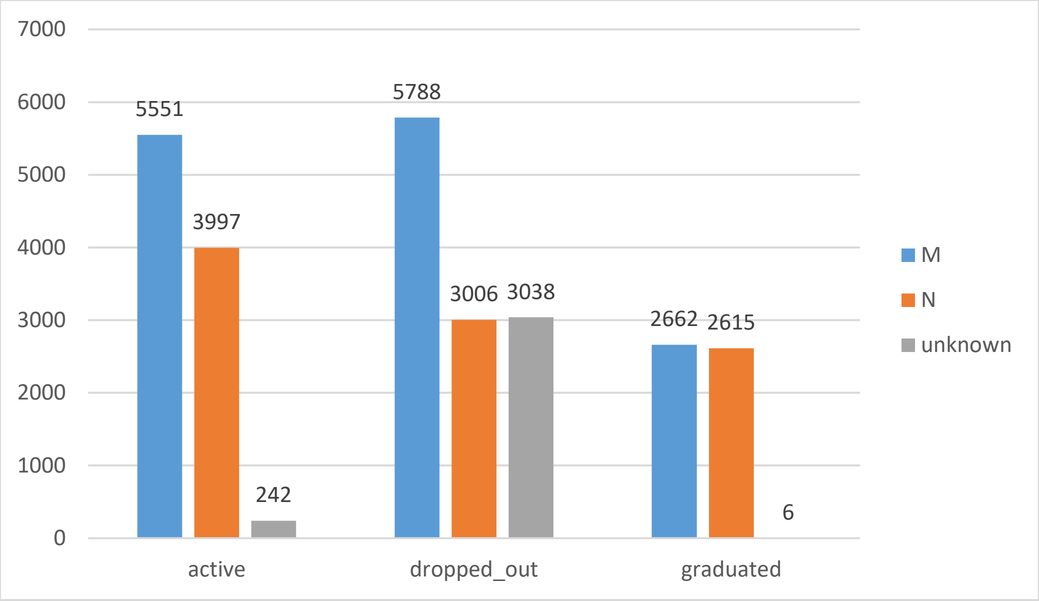


Figure 4. Distribution of students in the data under analysis in TTU 2012 – 2018.

What is interesting, is that even though there are many more men studying actively at the university, the number of graduates for men and women is fairly similar, i.e. the retention among female students is much higher than in male.

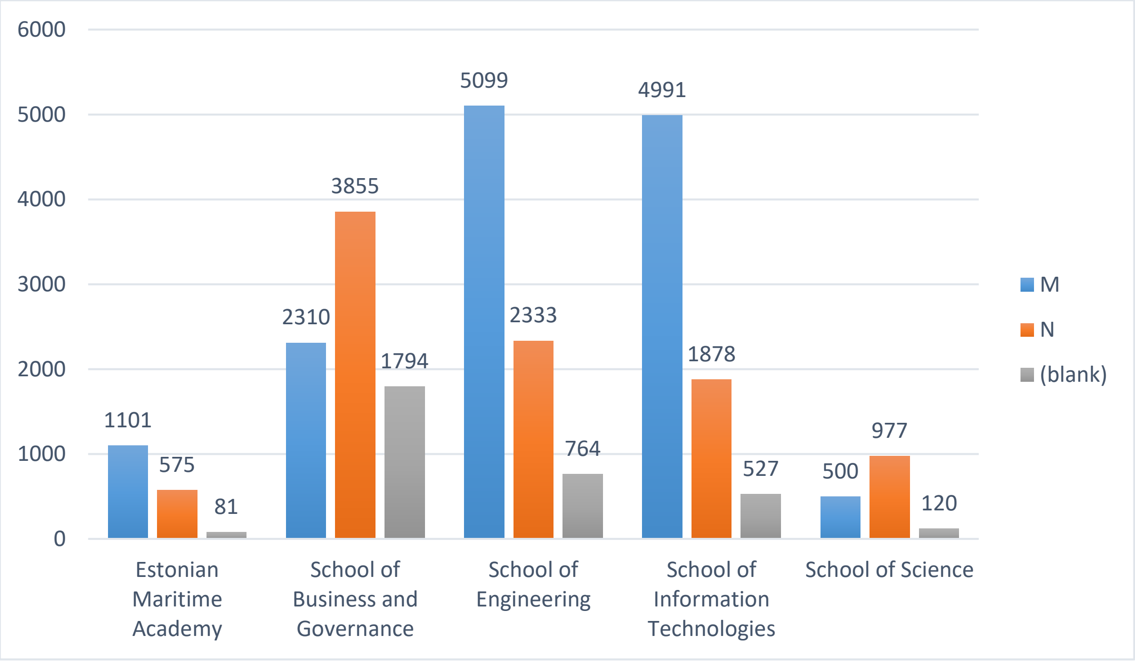


Figure 5. Distribution of students in different faculties in TTU, 2012 – 2018.

It is also interesting to assess how faculties differ from one another regarding the distribution of students. This is demonstrated in Figure 5 and on Figure 6. It is visible that the School of Science has different features than other faculties on a Figure 6. The distribution of students between graduated, active and dropped out is close to uniform.

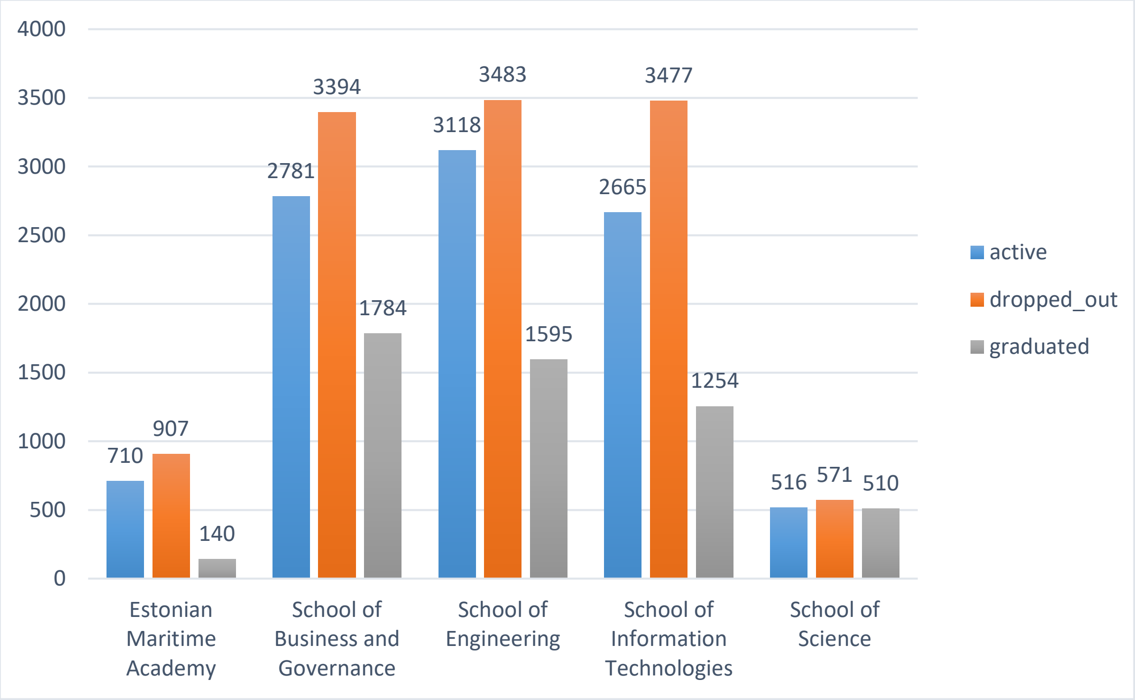


Figure 6. Distribution of student types among the faculties of TTU, 2012 – 2018.

4.2 Relevant Dataset Generation

In this thesis, two types of datasets are generated for two different methods. The first ones are the context based datasets. These datasets are created for each of the curriculum separately and include all grades for courses in the curriculum. In total there can be 300 context based datasets generated, one for each curriculum. Each of the created context based datasets is later modified to suit the needs of an individual student under analysis. The context based datasets are not saved into a dataframe or CSV, but are created during the process of predicting.

Another type of dataset includes more general features of a student's progress at Tallinn University of Technology. The feature selection for this dataset is motivated by related works, and includes more descriptive features of a students, as sex and faculty, as well as average grade etc. Features used will describe the students overall progress, when compared to all students at TTU. The general dataset does not include grades for each course. Instead it has aggregated values for numerical data and there is one dataset for all students.

Both generic and context based model is divided for testing into different subsets of training and test data, depending of the tests performed. Both datasets are always regenerated and modified based on the time period that is used for analysing. For this thesis, the time period under analysis starts from 1. august 2012 and ends with 29. March 2018. This means that students that have started their studies since then are used for training and testing the models. Students who have started and graduated within this time period, are used for training data. Some testing can be performed on active students, but due to not having their final outcome of studying, predicting graduation for those students is not reasonable. The testing accuracy on active students could be tested when currents students have graduated.

It is important to keep in mind that students who have been prolonging their studies at university and do not follow the traditional curriculum plan provided by the university are harder to predict for. The outcome that would apply to a nominal student will not apply to them and the model learns different patterns. Keeping that in mind, those students are still used in the datasets, but are additional factors on adding variety and fluctuation to the accuracy of predictions.

4.2.1 Context Based Dataset

Context based datasets are subsets of the "*sooritused*" (results) database table (Table 1). The idea is to generate a testing dataset for one context e.g one curriculum. From the "five_year_studens" dataframe all the students that are in one curriculum are selected, and from "*sooritused*" all the results and declarations of courses, that are done by previously selected students, are then selected.

In the beginning of constructing the relevant context dataset, each row in the merged dataframe represents one result in one course per student. This means that one student has the same number of rows in the initial dataset, of how many courses he or she as ever declared. There are duplicates for courses the student has failed and declared again. These attempts are held as two different occasions of declaration. The table also includes courses that the student has declared but has not finished.

To be able to use context based dataset for ML models, it is necessary to structure it so that each row represents one student, and each column represents a feature that might be important for predicting. These features include all the courses students from the curriculum have taken. To achieve this desired structure, dataframe is transposed into a pivotal table. After that the dataset has one row per student and the number of features is determined by the courses declared by all the students in the curriculum. For courses the student has not declared, the missing values are imputed with “-2”. To make the dataset more structured, only the courses that are marked mandatory in “*sooritused*” dataset are selected. Other features that describe a student can then be merged from other datasets onto the transposed dataset. For ML model training, only the passive students of the dataset are used. These are students who have “*aktiivsus*” (active) as “ei” (no).

When creating the datasets for different curriculums, I encountered various problems that emerged from the data. These are for example cases, when a person has been marked as student in the last 5 years, but his or her courses are finished years before that. Usually these results for courses are transferred with a VÕTA system. The decision whether these results should be included to the final dataset had to be made. In this case I decided not to include those courses, since courses that are not taught anymore could be transferred to the dataset and they do not add any meaningful information and therefore could disrupt the results of prediction.

An additional obstacle that also merged was the amount of duplicated declarations of a course in the results table. Previously it was possible to declare a course, and for it to be valid for three semesters, but currently the systems have been changed so that the declarations are valid for only one semester. This change has increased the amount of duplicate declarations for courses by the same student. It came out as a significant issue when creating context based datasets for analysis, because pivotal table does not accept more than one student-course pair in the table. To solve this, the most recent results had to be selected. The double declarations can later be used an additional feature for predicting.

In the process, there were also data mistakes, that hindered the analysis and had to be removed. These issues were related with data insertion and getting the data via the API built between server and database. A significant reason why the data has a lot of missing values, can be justified with the fact that the university systems and table structures seem to be restructured in the past, and older data points are inserted according to the old schema. Newer data points are entered with the new structure, but database schemas carry both to accommodate all data. One additional reason for this could also be, that TTU has recently been through many structural changes, with the merger of Estonian Maritime Academy and IT College of TTU. This also explains why in Figure 4 and Figure 5, there is missing data for students' sex.

An interesting issue I noticed, came out with grades for courses, when I wanted to merge "*hinne*" (grade) column with "*arvestatud*" (accounted) column to gain one feature from the results. Typically, when a course has a grade as a result, it does not have "A" or "M" in the "*arvestatud*" (accounted) column and vice versa. But to my surprise, for some results, it had values in both columns and the values were essentially different and contradictory.

Data cleaning was done on multiple occasions and in many iterations, because different data was tested and various data mistakes raised when testing the prediction methods. Filling in missing values was only done for prediction purposes. Since most of the data is categorical, imputing missing values would add incorrectness to the data.

Other data mistakes that were fixed included:

- Date inserted as "*kp*" (date) instead of actual date
- Meaningless words inserted instead of actual values
- Date being from too far from the future
- Data parsing mistakes ([, ", ",])
- Different data types in similar columns, that are required for table joining
- Filling in missing values
- Removing duplicates
- Escaping unicode to get the values in Estonian alphabet.

Since students who have finished their studies at TTU in the time period under analysis are used for training the prediction models, it was necessary to have a certain number of students at the training dataset to be able to train ML model. For some curriculums, the amount of already graduated or quitted students is fairly large due to the demand in those areas in the working field (e.g Informatics, Business Information Technology, Business, Law etc) and the amount of students that are accepted to the curriculum. For others getting the required sample of training data, was more difficult, since these were the curriculums that either are studied by very few students, the curriculum is a new one or the graduation rate is extremely low. This can be demonstrated by Figure 6.

To solve this issue some decisions had to be made. To increase the amount of student in a context based dataset, especially among the passive students that are used for training, one possibility was to merge the students from one curriculum, but different versions of it. For example, the cases when a curriculum was created in 2013, and enhanced with changes in 2014. This would mean there are two versions of the same curriculum. Since the courses and structures are somewhat similar the students from both curriculum versions could be merged. This solution raised the number of students in the context based dataset but did not completely eliminate the problem.

After solving the issues mentioned before, I ended up with 300 context (curriculum) based datasets. demonstrated in the Table 10 in Appendix 5 – Maximum Rows in Context Based Training Dataset. This table shows the maximum rows that could be used for predicting the dropout. In the process of developing a context based model to fit one student and his or her results, the number of rows used from the original dataset decreases significantly. This table carries information only about the curriculums that are demonstrated in the data under analysis. After constructing all context based datasets it was still evident that for some curriculums the amount of training data was very low and varies a lot.

After some initial testing with the data, I understood that students who have changed their major in-between their studies, carry with themselves results from two or more different curriculums. This came out when I analysed why some context based models have many almost empty features in the dataset. Since the structuring of a pivotal table is built as such, that it selects students from curriculum, and then picks the results and courses that have been made by those students, it is evident, that some students bring courses from few different curriculums, in case they have attended other curriculums before. These results and courses have been marked mandatory due to being mandatory in a previous curriculum. This will add unnecessary features to the context training dataset and might slow down the calculation process. This issue can be solved by checking if the course is mandatory for a curriculum the context based table is being created for. In this thesis I decided not to solve the issue, since later Fisher score will eliminate meaningless features for model fitting and those extra courses in pivotal dataset will have lower Fisher score due to not playing a significant role for the curriculum under analysis.

In order to split the data into test and training set, a division has to be made according to the columns "*lopetanud*" (graduated) and "*aktiivne*" (active). Since each row in the dataset represents a student, then the students having column "aktiivne" as "jah" (yes), are removed from the initial dataset used for training a model. Those students do not have the final results of their studies and can be used as test dataset. After this removal, it leaves the training dataset with students who have graduated, having "*lopetanud*" feature as "*jah*" and feature "*aktiivne*" as "*ei*" (no), and students who have quit their studies, having feature "*aktiivne*" as "ei" and feature "*lopetanud*" as "*ei*".

To minimize the number of unnecessary columns, feature "*aktiivne*" is then removed from the training dataset, because it carries no extra information regarding the graduating of the student, because all students in this dataset are not active.

4.2.2 General Dataset

The general dataset is mostly generated from the ISIK_C1 table and "Sooritused" database table. Some of the features like ['fk_oppija_id', 'kava_vers_id', 'eelm_kooli_id', 'eelm_lop_a', 'aktiivsus', 'sisse_a', 'akad', 'lopetanud', 'fk_oppekeel_kood', 'fk_tunnus_kood', 'fk_fin_kood', 'fk_oppevorm_kood', 'fk_koormus_kood', 'kehtib', 'sugu', 'str_yksus_id', 'oppekava_id', 'fk_oppetase_kood', 'kava_versioon_id'] are directly selected from the original ISIK_C table, and some aggregated values are calculated to the dataset one by one from "Sooritused" table for each individual student.

Since for the general dataset it was more important to get an understanding whether overall aggregated features can be used for predicting dropout, then developing functions for additional features were done under the development of general dataset. These features were: sum of EAP, sum of active semesters, average grade and average EAP per semester.

The most complicated part of developing a general dataset, was finding ways to predict graduation for students who are in-between their studies. This means that those students do not have the final amount of EAP and grades. To come up with features that would be suitable for student in every stage of their studies, the features had to be influenced by the time the student has studied in the university. It also became evident that if the concept drift was important to be considered, I had to integrate students aggregated progress on the semester basis to the general dataset. This means that features as EAP in first, second, third etc semester had to be added. Additionally, a possibility to calculate average grades per semester was developed.

This calculation followed the formula that is used in TTU to calculate average grade (KKH) for a student [40]. It is important to note that the formula does not take into account courses that the student has failed or not finished. EAPs that are collected from "*arvestatud*" (accounted) courses are also not used.

$$Average\ grade = \frac{\sum_{n=1}^{k}(EAP_n * \ Grade_n)}{\sum_{n=1}^{k} EAP_n}$$

To achieve the additional features, the best way was to calculate EAP and average grade for each of the semesters the student has declared any courses on. The aggregated sum of EAP was grouped by semester and by student and added as an additional feature that could be added to any dataset for more information. Since for each student, their first semester varies from another students' first semester, the method has to generate additional conversion tables in-between the calculations to transform the data into workable form. Each calculation and data transformation is done for each student separately.

One of the main issue that raised when constructing a general dataset, was the amount of missing values. Initially the plan was to use all features from the "ISIK_1C" table with most of the features being categorical. After analysing the data further, I understood that many of the features have more than half of the values missing. Since they are categorical features, imputing the values was not possible, because it would add significant incorrectness. I decided to remove the columns, where more than half of the values were missing.

4.3 Machine Learning Model Preparation and Feature Engineering

In this paragraph I describe how and what features are selected for fitting a machine learning model. I give an overview on how different features are built and what complications raised when selecting them. I also describe the changes that had to be done to the methods when initial testing had been started with predicting.

4.3.1 Context Based Dataset

For understanding how feature selection could be done for context based dataset, I selected different curriculums to test the methods on. These curriculums varied within number of students graduated and also in the number of courses selected by students.

For data preparation all columns are turned from categorical data type to numerical data type. Grades are in-between values {-2; 5}. -2 demonstrates a missing value, -1 is the grade when a student has failed a course, 0 is not showing up to an exam and 1-5 are actual grades. Accounted courses were changed as “A” to 5 and “M” to -1. Graduation attribute is changed accordingly: “1” for “*jah*” (yes) and “0” for “*ei*” (no).

To gain an understanding, what features might play a more significant role in predicting, whether a student will drop out or no, Fisher score is used. For each column (course) in the context based dataset a fisher score is calculated, showing the importance of the features related to the target feature, which can be “lopetanud” feature (when predicting final dropout or graduation) or “continuing on next semester” feature (when predicting on a semester base). For that a Git repository “scikit-feature/skfeature” is used [41]. It is important to note that Fisher score is calculated individually for each course in the context dataset and each context dataset is personalised for a student under analysis.

To assess the correctness of the Fisher score results, the results must be evaluated from human perspective, by someone who has enough domain knowledge. According to the academic staff in IT faculty, the results seemed to be correct. This means that course that are expected to be the most difficult ones for students to pass, play the most significant role in the way of student graduating a university and therefore carry the most information related to predictive label. This is also demonstrated by the results gained from the scores, an example visible in Table 3. The initial Fisher scores were calculated for all the courses in the curriculums for a testing purpose. Since the curriculum includes many curriculum versions, the results may vary if calculated for each curriculum version. In the case when

Fisher score calculation is applied to all of the data from one curriculum, the results may describe the correlation between courses done during the final semesters and graduation label.

Table 3. Example of Fisher score results for curriculums 652, 50031 and 729.

Importance in terms of information gain	Business Information Technology (652)	Law (50031)	Product Development and Production Engineering (729)
1	Databases II	EU Competition Law and Policy	Machine Tools and Manufacturing Systems
2	Web Services	EU Internal Market Law	Thermal Engineering
3	Business Process Modeling and Automation	Non-Contractual Obligations	Integrated Product Development
4	Software Engineering	Intellectual Property Law	Machine Elements II
5	System Theory	Environmental Law	Special Studies Project
6	Databases I	Human Rights Law	Machine Automation
7	Ergonomics	European Union Law	Machine Elements I
8	System Analysis	Public International Law	Mechanics of Machines
9	Physics	Private International Law	Core Studies Project
10	Computer Networks	Legislative drafting	Strength of Materials I

To analyse a method developed, I started predicting results for graduation. With some initial testing it was visible that using only three courses gave a meaningful result in terms of information gain related to the predictive label. When I used top three courses from the Fisher score, the accuracy for predicting dropout with cross validation was almost always 100%. In this instance, I had made a test and train dataset from the context based dataset and did not predict for actual active students. This result was unexpected.

After analysis it was visible that the result can be explained. The training of a model was done on all the courses and all the students that had graduated or left from the curriculum under analysis. This means that all the students who had already graduated, had done the three courses, selected by a Fisher score, and the ones that had decided to quit their studies, had not finished all those 3 courses. This in a vague explanation means that the model checked if all three selected courses were done and then decided that the student will graduate. If some of the top three courses were not done, the student would have been marked as a dropout candidate. It was evident that the system is not sustainable like that and does not apply to students who are still for example on their first year of studies, because they would always be predicted as “dropout”.

To better understand how to predict if a student is going to graduate or not, I had to understand that most students drop out during the first 3 semesters [8]. This means that it is important to modify a training dataset just for those students to be able to predict for active students.

To solve this issue, I developed a function that generates a context based training dataset for a student under evaluation specifically, using only the courses that the students has declared for fitting a ML model. The function takes as an argument the student id (data was anonymised, so the id was a hash) and selects all results done by this student, turns it into a pivotal transposed table with one row. Then it generates a context dataset for training, including all the students from this curriculum who have finished their studies and selects only the courses that have been done by the student under analysis. This approach ensures that the student is judged on the basis of his or her performance and results and not on the basis of the absence of any important course. It also broadens the possibility of evaluating students who do not follow the nominal curriculum.

After the previous change, another issue with the data was discovered. To recall, the context based dataset is constructed of passive students of all the curriculum versions of the one curriculum under analysis. In some cases, when a curriculum has been changed, the training data for this context based dataset, does not have any information about the new courses that have been added to the curriculum in a newer version. This happens because in the beginning of the development phase it was decided to merge the students from one curriculum but different curriculum versions. This means that the context based dataset that is generated from passive students for training a model, will have some features/courses, with None values, since none of the passive students have done the new courses, provided in the newer curriculum versions. This will lead to poor predictions, because the model is training on data, that it previously has, the predictions for a student under analysis will not consider the results that the student actually has in new courses. This led to predictions getting worse over a period of semesters, because the number of students on the actual same curriculum version in training dataset decreased and the number of students of other curriculum versions increased.

Table 4. Example of training data division among curriculum versions.

Curriculum ID	1008 (Informatics)			Training dataset size
Curriculum versions	50263	50067	1010	
Passive students used in train dataset	321	343	3	667

From Table 4 it is visible that when the method generates a training dataset for example a student studying Informatics on a curriculum version 50263, then majority of the training dataset is built from students at a curriculum version 50067 and therefore the possibility of predicting wrongly is higher, when those two curriculum versions are very different. This in most cases does not play a significant role in prediction, but has to be kept in mind when further developing the solution.

To solve this, I decided to add another step in creating of the training dataset. This step removed all the rows, that had only None values in it. This way eliminating those students who mostly study on another curriculum versions from the training dataset. This mitigated the issue slightly, but also decreased the number of rows in the training dataset. Overall the issue will only be solved when more data is gathered and more students have finished the new courses, but it is also evident that as universities are constantly changing organisations, then this problem might not be solved for small curriculums that are improved frequently. The training set for predicting for those curriculums will be too small.

The overall process of getting necessary datasets and predicting student dropout for context based model, has 6 steps. The following functions are fully described Appendix 2 – Description of Methods Used in the Thesis. The process is also described in Appendix 3 – Process of Predicting with Context Based Method.

1. Preliminary data needs to be read into the system. This will speed up further calculations.

```
all_students = read_students()
curriculums = get_all_curriculums()
all_results = creating_results_dataset()
five_years_students = get_five_year_students(all_students)
five_year_results = get_five_year_results(all_results,
five_years_students)
```

2. The dataset with context based rows is generated. The dataset is created for relevant curriculum and for a specific student and includes one row, with the results of one student under analysis.

```
pivot_table_predict_one, curriculum =
create_active_test_dataset_one_student(fk_oppija_id, curriculums,
five_year_results, five_years_students)
```

3. According to the `pivot_table_predict_one` and curriculum, a training dataset, based on already graduated students is generated, carrying only the courses that the `pivot_table_predict_one` dataset has.

```
pivot_table_train = create_context_dataset_train(curriculums,
five_year_results, five_years_students, curriculum,
pivot_table_predict_one)
```

4. The training dataset is put through Fisher score evaluator method and then depending on the input 2-10 most meaningful course are selected. The selected features are applied to training dataset which is modified according to the courses selected. Data and target for prediction are assigned.

```
pivot_table_train = pivot_table_train.fillna(-2)
fisher_scores = get_fisher_scores(pivot_table_train)
courses = fisher_scores.head(fisher_number)
courses = courses.ainekood.values
pivot_table_train = pivot_table_train[(pivot_table_train[courses] != -
2).all(axis=1)]
data = pd.DataFrame(pivot_table_train, columns=courses)
target = pivot_table_train['lopetanud']
```

5. The decision tree model is fitted on the data and model evaluation is returned.

```
clf, target_test, data_test, data_train, target_train =
decision_tree_classifier(data, target)

cv_accuracy, test_accuracy, report, test_data =
prediction_evaluation(clf, data, target, target_test, data_test,
data_train, target_train)
```

6. The selected features are then used for predicting.

```
pivot_table_predict_one = pd.DataFrame(pivot_table_predict_one,
columns = courses)

pivot_table_predict_one['predicted'] =
clf.predict(pivot_table_predict_one)
```

4.3.2 General Dataset

For the general dataset, the data is very different from the context based dataset, carrying more aggregated and descriptive values. For testing a model, random subsets of data is selected and previously developed methods are tested.

When it comes to decision tree model preparations I encountered a big problem. Sklearn decision tree classifier does not accept categorical values in features. This problem in that scale was not encountered with the development of context based model, since most the features there are already numerical. Solution for the issue was to find another decision tree classifier model or encode the labels. To keep the coherence between context based and general model, decision to encode the labels was made. It is important to do this step with all the data provided, before splitting it into testing and training data, since only then all the levels of values are encoded to numerical values correctly.

Before encoding the labels, I was interested to see, how much variety each feature carries. Datetime features were the most scattered and since for testing it was not important to have the exact time and date as a feature, the decision to modify the features was done. Datetime features were modified to carry information only about the year or the year-month combination. Date and time aspects of the feature were discarded. This lowered the amount of variety a lot in these features.

For label encoding I decided to use Sklearn method called `LableEncoder`. `LableEncoder` is a function from pre-processing library of Sklearn and labels with value between 0 and n_classes-1 [42] . Before I was able to use the method, I had to fill in NaN values for the features that were categorical. After label encoding NaN values for already numerical features had to be replaced as well.

To enable the understanding of the data, I added the encoded features as new features to the dataset with a suffix “_numerical” and did not replace the original features. For predicting only the new numerical features are used. These are ‘kava_vers_id', 'eelm_kooli_id', 'eelm_lop_a', 'str_yksus_id', 'oppekava_id', 'kava_versioon_id', 'eap_sum', 'active_nr_semesters', 'eap_semester', '1 sem', '2 sem', '3 sem', '4 sem', '5 sem', '6 sem', '7 sem', '8 sem', '9 sem', '10 sem', 'aktiivsus_numeric', 'sisse_a_numeric', 'akad_numeric', 'lopetanud_numeric', 'fk_oppekeel_kood_numeric', 'fk_tunnus_kood_numeric', 'fk_fin_kood_numeric', 'fk_oppevorm_kood_numeric', 'fk_koormus_kood_numeric', 'sugu_numeric', 'faculty_numeric', 'curriculum_name_numeric', 'fk_oppetase_kood_numeric'.

The final process of developing a general dataset for predicting has 6 steps:

1. Preliminary data needs to be read into the system to speed up further calculations.

```
all_students = read_students()
curriculums = get_all_curriculums()
all_results = creating_results_dataset()
five_years_students = get_five_year_students(all_students,
curriculums)
five_year_results = get_five_year_results(all_results,
five_years_students)
all_active, active_five_years = get_active_students(read_students())
```

2. Relevant testing dataset has to be generated. The dataset depends on what students are fed into it. The “fk_oppija_id” is added as index to a dataframe, to enable the adding of aggregated values in next step.

```
general_dataset = create_general_dataset(five_years_students)
general_dataset = general_dataset.set_index(['fk_oppija_id'])
```

3. Additional features are calculated to the general dataset. These are sum of EAP, average grade per semester, number of active semesters and EAP per semester. Average grade and sum of EAP also need an attribute to describe how many semesters are under evaluation.

```
general_dataset = add_active_semesters(general_dataset,
five_year_results)
general_dataset = add_sum_of_eap(dataset, five_year_results, 'all')
general_dataset = add_average_grade(general_dataset,
five_year_results, 'all')
general_dataset = add_eap_per_semester(general_dataset)
```

4. Data is divided into passive and active students, to enable testing.

```
passive_general = general_dataset [general_dataset.aktiivsus_numeric
== 0]
active_general = general_dataset [general_dataset.aktiivsus_numeric ==
1]
```

5. Target and data are selected, and the test and training set are composed. The decision tree classifier is fitted.

```
target = passive_general.lopetanud_numeric
X = passive_general.drop(['aktiivsus_numeric', 'lopetanud_numeric'],
axis=1).fillna(-1)
X = X.replace([np.inf, -np.inf], np.nan)
X = X.fillna(-1)

clf, target_test, data_test, data_train, target_train =
decision_tree_classifier(X, target)
```

6. The prediction is done on testing dataset.

```
target_test = pd.DataFrame(target_test)
target_test['prediction'] = clf.predict(data_test)
```

4.3.3 Analysis Dataset

For the comparative analysis purpose, I decided to create a dataset, that uses data from August 2008 to August 2012. This dataset in this thesis is called pre-reform dataset (in code it is marked as drift dataset) [43]. Into pre-education reform dataset students from 2008 – 2012 are selected, who are marked as inactive (passive). This means that they have either graduated university or have quit their studies at TTU, and are not actively studying anymore.

For the previously developed methods to work on this dataset as well, a percentage of students is manually changed from passive to active, thus imitating the current active students. The percentage of students that were randomly made into “active” students was the same as it is in the original dataset (students from August 2012 to March 2018) to keep the comparability.

All functions mentioned in previous sub-paragraph work for this analysis dataset as well or newer modified versions of those functions are developed. The methods are described in Appendix 2 – Description of Methods Used in the Thesis.

5 Results and Analysis

In this paragraph I describe the experiments with different datasets, curriculums and features that are done with general and context based methods. I compare the results between pre-reform dataset and original dataset to understand if the patterns of predicting are similar with the change of time. I first have a look at the general method and then follow up with the context based method. The advantages and disadvantages of both methods and situations in which either model works better are brought out.

5.1 General Method

5.1.1 Predicting Graduation

To start analysing the general model I first wanted to see what was the overall accuracy of the model, when it is applied to all of the data.

The model was first trained using all passive students from the dataset and there were no limitations made to the features. Initially I was trying to predict graduation. I used 80% of the passive students, and predicted for the 20% of "passives" to be able to measure the accuracy. For pre-reform dataset, the accuracy of prediction was 81%. For the original dataset the accuracy was 91.4%. These results were quite promising, considering that no pruning of the decision tree was done.

For analysis purposes, Master's and Bachelor's degree students were selected separately. It came out that when predicting for Bachelor's students, the accuracy increased, and yielded to nearly 93-96%. On the other hand, Master's result dropped significantly in the accuracy, yielding only on average 84%. For both testing and training dataset size decreased to about 4000 students. For Bachelor's six semesters worth of aggregated data was used, and for Master's four semesters.

I assumed that the accuracy would get higher, compared to overall accuracy, when the data was segmented into faculties, because the data would be more similarly grouped. I also expected that less variety would enhance the predictions and the model would have more patterns to fit the data on. Same analysis was performed on pre-reform dataset.

From Figure 7 it is visible that the accuracy did not have high fluctuation compared to the accuracy in all faculties. For some faculties the results got better, and for some the accuracy in predictions decreased. Similar pattern is visible for pre-reform dataset.

The results mean that the faculties do carry some significant information that groups the features together. It demonstrates that for future work it is better to segment the data into chucks, but still it is necessary to keep the amount of data large, to limit overfitting on small number of examples. This will allow the decision tree models to generalize over a larger group of data. With these results the third working hypothesis can be rebutted, because the results are not always better when grouped by faculties.

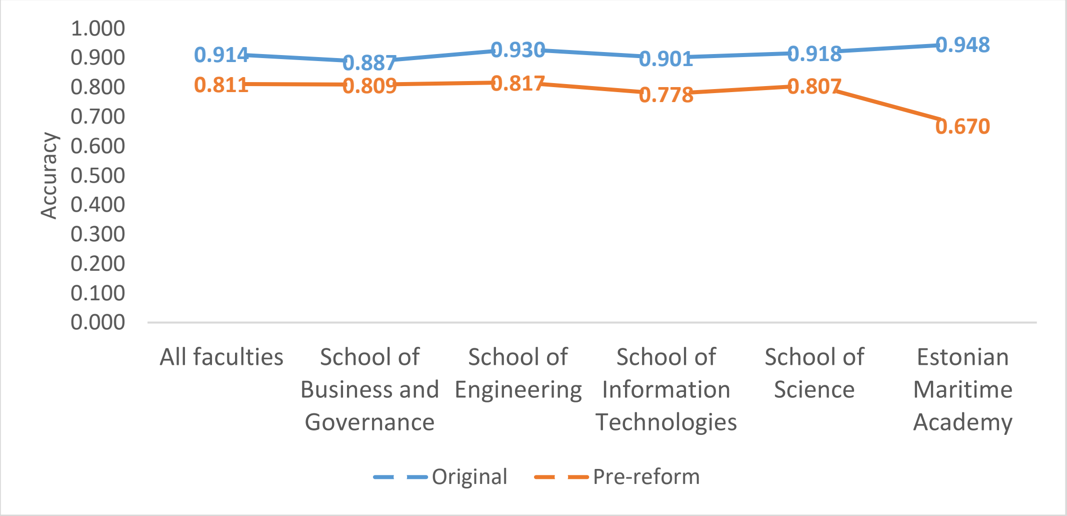
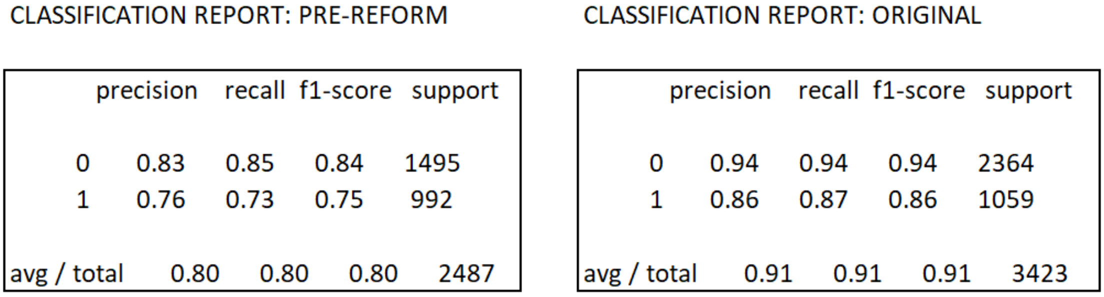


Figure 7. Accuracy of predicting graduation, general model.

It is also interesting to see, that the prediction was extremely low for the Estonian Maritime Academy in the pre-reform dataset. In the original dataset the result for Estonian Maritime Academy increased significantly. This can be explained with not having enough data about Estonian Maritime Academy in the pre-reform dataset. After Estonian Maritime Academy merged with TTU (2014), the data collecting has followed the same structure as other faculties in TTU.

In the Figure 7 it is also demonstrated how the accuracy changed and followed similar pattern in pre-reform and original dataset. The accuracy for original dataset is 0.13 points higher than in the pre-reform dataset. This can be explained with the fact that in 2012 the education reform came into force and students were pressured to finish their studies in a nominal time and follow a better structure. Only students who have started in August 2012 are used for the original dataset. This explains why it was easier for a model to generalize better on data to predict the drop out and graduation more accurately in the original dataset. Similar pattern also came out with the context based model described in the next chapter.

What was interesting as well, was that the model was always predicting dropout better than graduation. This is demonstrated with the classification report provided by Sklearn, showing precision of predicting 0 and 1 (Figure 8). The data used for the analysis is biased towards more dropped out students, as demonstrated on the Figure 3, and has had more data to learn the patterns of students who have prematurely ended their studies.



CLASSIFICATION REPORT: PRE-REFORM

```
             precision    recall  f1-score   support

          0       0.83      0.85      0.84      1495
          1       0.76      0.73      0.75       992

avg / total       0.80      0.80      0.80      2487
```

CLASSIFICATION REPORT: ORIGINAL

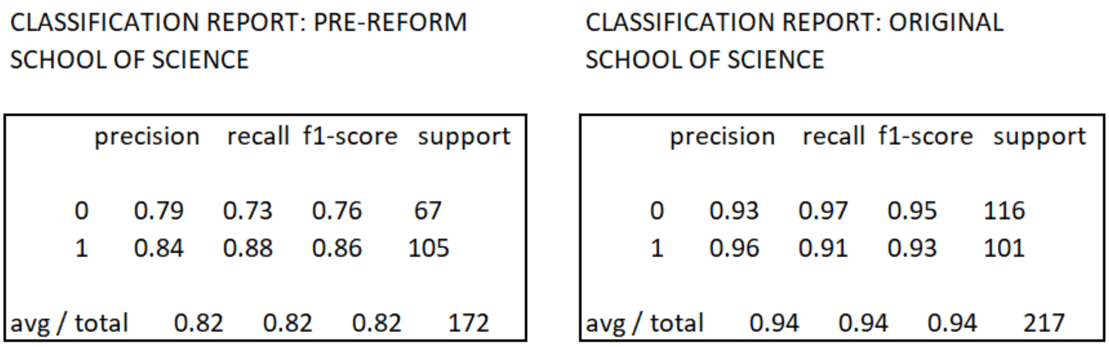
```
             precision    recall  f1-score   support

          0       0.94      0.94      0.94      2364
          1       0.86      0.87      0.86      1059

avg / total       0.91      0.91      0.91      3423
```

Figure 8. Classification report example for all faculties.

This was an exception in the case of School of Science. For that faculty, both the pre-reform dataset and original dataset the model predicted graduation slightly better than dropout. This is explainable with the fact as demonstrated on a Figure 6: the graduates, active students and dropped out students are on a similar level in the School of Science. The assumption that graduation students follow a better structure for model to learn on can be made.



CLASSIFICATION REPORT: PRE-REFORM
SCHOOL OF SCIENCE

```
             precision    recall  f1-score   support

          0       0.79      0.73      0.76        67
          1       0.84      0.88      0.86       105

avg / total       0.82      0.82      0.82       172
```

CLASSIFICATION REPORT: ORIGINAL
SCHOOL OF SCIENCE

```
             precision    recall  f1-score   support

          0       0.93      0.97      0.95       116
          1       0.96      0.91      0.93       101

avg / total       0.94      0.94      0.94       217
```

Figure 9. Classification report example, School of Science.

The analysis also showed, that there is no possible way to predict anything for students in the Open University of TTU, because they are not studying on any curriculum and are always marked as “*lopetanud*” == “*ei*” (graduated == no).

5.1.2 Predicting Graduation by Semester

To understand how more generic features could be used for predicting student graduation, tests by semester were also done. It was expected that the accuracy of predicting dropout would get more accurate when more semesters were added to the data. This means that with every additional semester, the average grades for each semester and EAP amount for each student is recalculated, considering the student as if he or she was still studying.

The tests are done with one bachelors and one masters curriculum, to demonstrate if the accuracy would change as expected. Curriculum 1008 is Informatics and curriculum 50052 is Human Resource Management/Personnel and Development. The two curriculums also differ from the number of students they have in the training dataset. Informatics has 605 students among passive students and Human Resource Management/ Personnel and Development 194. The test is not performed with pre-reform dataset, since the random accuracy results were much lower and from that an assumption can be made, that the students follow different patterns.

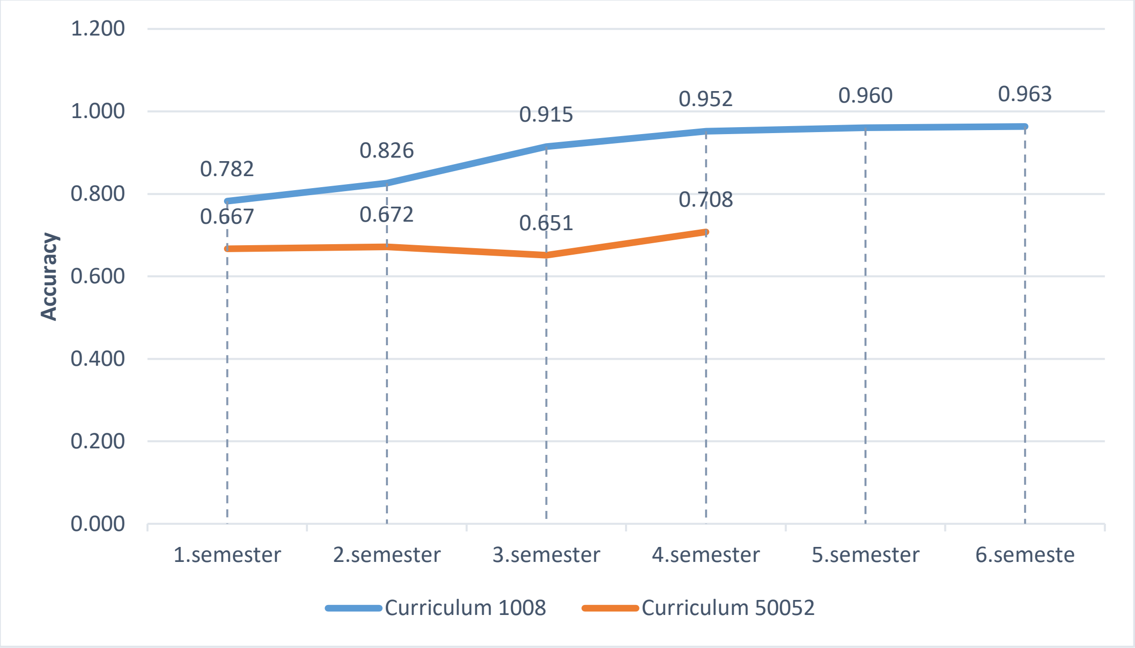


Figure 10. Graduation prediction by semester, general model, curriculums 1008 and 50052.

From Figure 10 it is evident that the accuracy rises when more features are added to the dataset, that describe the students' progress over time. It is also visible that the prediction is better for the Bachelor's curriculum, and with the 3rd semesters data it is already more 90% predictable whether the student is going to graduate. When analysing Master's curriculum, it is visible that the accuracy is quite low using generic method for prediction in curriculum 50052.

From these tests it clear that general method can be used for predicting graduation and should be investigated in more depth. It is also visible, that predicting for 2nd level students the accuracy is not good. This might mean that this particular curriculum does not follow a great pattern or that the students behave more unexpectedly during their master studies. More analysis and tests about Master students should be performed, to find better explanations.

5.1.3 Feature Analysis

In [25] it was stated that the most meaningful feature in terms of predicting graduation was sum of EAP. For random accuracy predictions done in paragraph 5.1.1, it became evident that the most important feature is EAP sum during the 3rd semester. It was followed by "*koormus kood*" (full- or part time studies) and "*oppetase kood*" (level of studies). The fourth feature was EAP sum. The importance of "*oppetase kood*" can be explained with the fact, that prediction accuracy for Masters and Bachelor students was very different, and depending of the level of studies the results are different. The fifth most important feature was "*eelm_lop_a*", stating the time of the graduation of last school.

When the prediction was done for Informatics curriculum, the most important features changed, when more data about additional semesters was added to the tests. Most important feature, when using all the data (all six semesters), was EAP sum during 5th semester. It was followed by "*koormus kood*", average grade of 4th semester, EAP sum on 1st semester and EAP sum on 3rd semester. From 32 features that were provided to the model, only 10 yielded with any importance. Features like "*sugu*" (sex), "faculty" and EAP sum were not used. It was also interesting that for prediction by 1st or 2nd semesters, feature called "*sisse_a*" had quite high importance related to prediction. Later this feature did not add any information to the prediction. The same is true for "*sugu*" feature. From the results it is understandable, that the model substituted missing grade information with other relevant features during the first few semesters.

When predicting for Human Resource Management curriculum, the feature with highest importance was EAP sum on 3rd semester, when all data about 4 semesters was used. This was followed by average grade on 2nd semester and EAP sum. The fourth feature was the id of previous school. Only 13 features from 28 were used. What is interesting, is the fact that while predicting for Bachelor's degree, the number of features used decreased, but for Master's degree the number of features increased. This could be explained with the fact that the predictions for curriculum 50052 were very unstable and the training set was small, therefore it needed as much information that was possible to reach from the data, even if the importance of a feature was very low.

5.2 Context Based Method

5.2.1 Predicting Graduation

To start analysing context based method I first wanted to gain an understanding what would be the overall accuracy of the model, if I applied it on random data, and whether the accuracy was higher with lower number of courses or higher number of courses. The course selection was done using Fisher score.

I expected the accuracy to be quite low, since the context based model was more meant to suit the needs of specific student and be personalised just for him or her and would have high fluctuation for overall accuracy. The context based model would choose the courses it predicts on according to a student, and this would vary for each person and each curriculum, also depending of the amount of training data that is accessible for each curriculum.

Prediction and model evaluation was first applied to the pre-reform data (2008 - 2012). The system randomly selected 100 students of that dataset. This sample size was selected to be able to manually validate the results. The test was made with different number of courses suggested by Fisher score, and repeated three times on random datasets.

Table 5. Example of context based methods accuracy predicting final graduation on a random pre-reform dataset. F10 – ten most informative courses are selected; F2 – two most informative courses are selected.

	F10	F8	F6	F4	F2	Average
Sample 1	0.761	0.773	0.731	0.803	0.818	0.777
Sample 2	0.697	0.662	0.731	0.731	0.727	0.71
Sample 3	0.683	0.698	0.762	0.641	0.683	0.693

From Table 5 it is visible that the accuracy is very unstable, with the average being about 0.726. The number of different courses selected by the Fisher score also brought instability to the accuracy. It is visible that the accuracy is better with lower number of courses (green and white cells). The results can be explained with the fact that the 100 students selected to a sample dataset are very different. This also might mean that for some students the accuracy of prediction is really high, and for others it might not be. The results differ from the sample data selected.

The same test was performed with the original dataset (students from 2012 - 2018) and results are demonstrated in the Table 6. It is evident that the accuracy was much higher than in pre-reform dataset yielding to about 0.826 on average, but the results are still very unstable and depend of the sample gathered. Results being 0.1 points higher with the original dataset can again be explained with the education reform that was enforced in 2012. It means that even from the grades of the students, it is visible that they follow a better pattern, as it came out with general method.

Table 6. Example of context based methods accuracy predicting final graduation on a random original dataset. F10 – ten most informative courses are selected; F2 – two most informative courses are selected.

	F10	F8	F6	F4	F2	Average
Sample 1	0.814	0.794	0.804	0.814	0.825	0.81
Sample 2	0.847	0.773	0.796	0.814	0.857	0.818
Sample 3	0.844	0.856	0.848	0.839	0.862	0.85

For both datasets, the accuracies varied a lot and this was the result I was expecting for a random dataset, when context based method is applied to it.

Since the context based model is meant to be more personalised and work on a more granular level, I decided to analyse graduation on a curriculum level. I assumed the accuracy of graduation prediction to get better. These tests were performed on original dataset (2012 - 2018), since the previous analysis proved that the pre-reform dataset yielded to lower results. Three curriculums were selected for result demonstration. Two of the curriculums chosen were the biggest, in terms of training data size. The third curriculum chosen for example was a smaller one, having only 67-121 rows to train on, depending of a student in the sample data (Food Technology and Development). The prediction was performed on all the data and not done by semester.

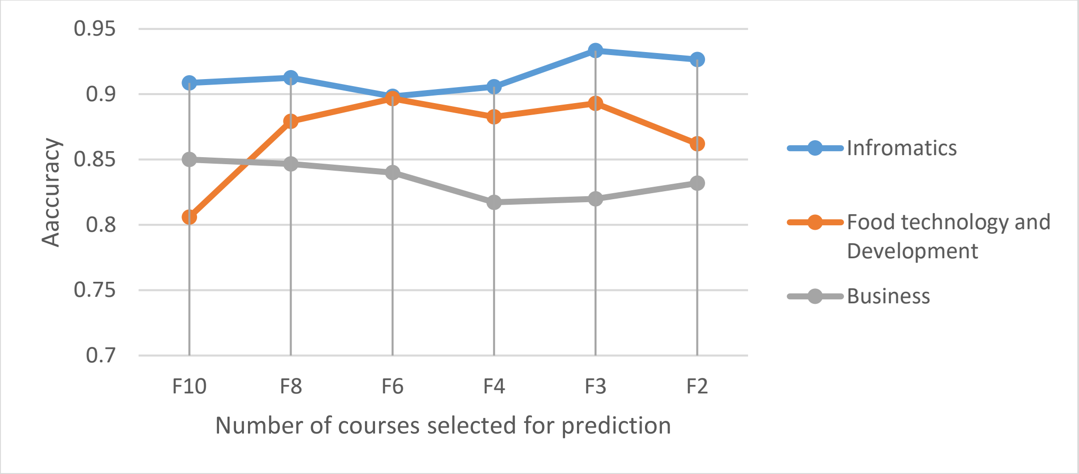


Figure 11. Example of curriculum based graduation prediction, context based method.

From Figure 11 it is visible that the results have high fluctuation in accuracy. It is also evident that the accuracy is higher compared to random accuracy, that was on average 0.826. It is interesting that curriculums follow a different pattern when different number of courses are selected. From these results it is evident that when predicting on a more grouped data, the results follow a better pattern and are more reliable, but each curriculum acts very different. Context based model does have extremely high variety when predicting students' final outcome, and is therefore unreliable model to use for that purpose in this format. The reason for that might be, that for each student the method chooses the most important courses he or she has done, and this does not always reflect the overall most important courses of a curriculum.

Predicting with context based model also brought out corner cases, when it is not possible to use this method. These were the cases when a student is studying on a curriculum, but has not taken any mandatory courses. For those students it is not possible to generate training dataset. It is also reasonable to assume that these students will not graduate this specific curriculum they are on. Context based method cannot be used for students studying in Open University TTU, because they are not following any curriculums. It is important to note that for curriculums with about 150 students in training data, the standard deviation for the accuracy will have a high fluctuation. This means that the prediction can still be correct, but it is necessary to be sceptic about the results and look into them more in depth and analyse manually.

5.2.2 Predicting Graduation by Semester

To understand how the context based methods accuracy would change with semesters, I analysed 2 curriculums in more depth. Similarly to generic method, the students in the curriculum were treated as they were still studying and their results were recalculated and selected based on the semester the prediction was done. The tests were performed with the same curriculums used for general method: 1008 (Informatics) and 50052 (Human Resource Management).

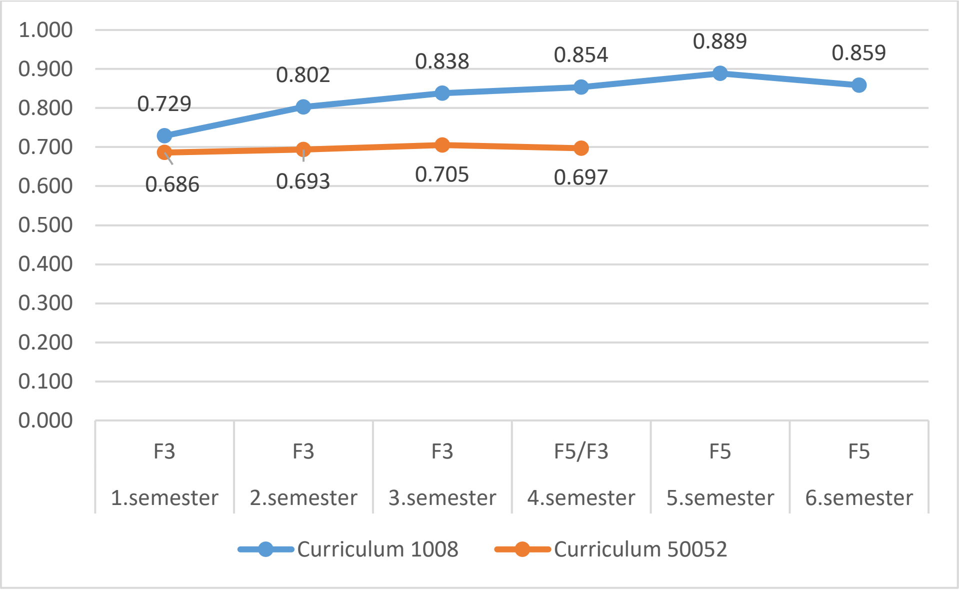


Figure 12. Graduation prediction by semester, context based model, curriculums 1008 and 50052. Semesters 1-3 are predicted using 3 courses. Semesters 4-6 are predicted using 5 courses.

When the predictions are done by semester, it is visible from Figure 12 that the context based model predicts with less accuracy when compared to general model, that yield to 96% during the 5th and 6th semester. It is evident, that context based model also predicts the best during 5th semester, but the accuracy is lower than with general model. It is important to state, that starting from 4th semester, the number of courses selected for prediction was increased from 3 to 5 (3F to 5F). This was done in order to increase the accuracy. With lower number of courses used the tendency in prediction stayed the same or decreased. Explanation for this tendency is that when more features are added to the prediction dataset, it becomes more difficult to make a prediction only based on very few of the courses. This happens after the third semester, when the prediction dataset starts with around 18 courses the student has declared.

Compared to the general model, the average accuracy for predicting graduation on original dataset was 0.1 points lower with context based method. If the Fisher selection filter were to be removed and all the courses done by a student would be used for fitting a model, the model would need more data to learn on, since the number of features would wary from 1- 45, depending on a student. Since the context model predicts for each student separately, this method is more time consuming when compared to general model.

To conclude graduation prediction, it is evident that context based model is less accurate choice when predicting final drop out or graduation. It is evident that the second working hypothesis can also be rebutted, since general model predicts more accurately by semester and on random data. In order to better understand the behaviour of context based model and the affect on course number change, more tests should be performed and each curriculum should be analysed separately. This is necessary because every curriculum is very different, including various courses and different number of students.

5.2.3 Predicting Continuation

Since it was visible that the context based model does not suit predicting graduation, I wanted to analyse how the method would work for predicting student continuation by semester. For that new methods had to be developed and the predictive feature had to be changed from "*lõpetanud*" (graduated) to "will continue next semester".

I analysed the original dataset (2012 - 2018) and predicted student continuation during next semester, on passive students, selected randomly from the data. Because the training model is built on passive students as well, I had to remove the student under analysis from the training dataset. The tests were performed with 1-5th semester, having selected course number as three (F3). The results are visible in the Figure 13.

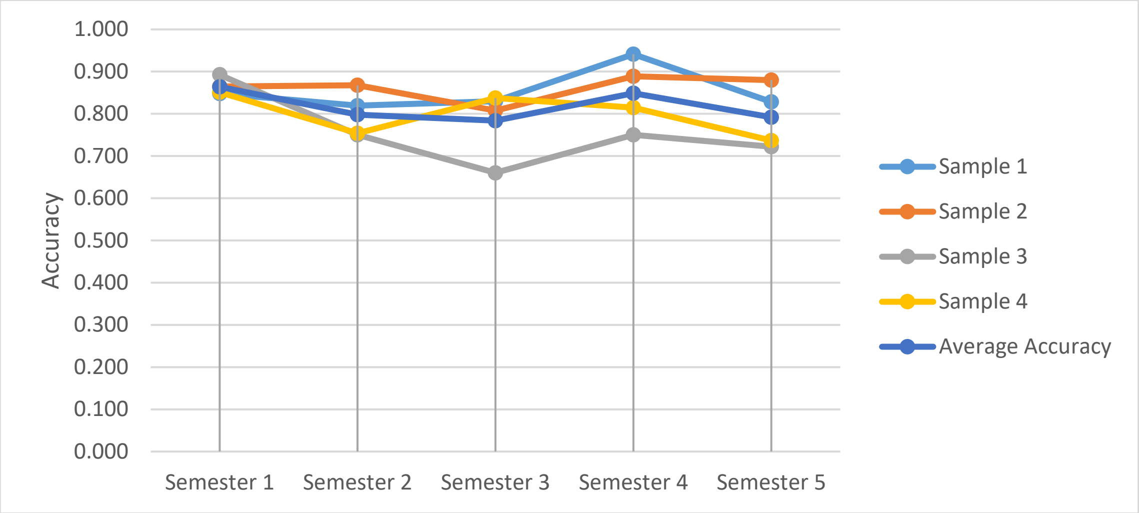
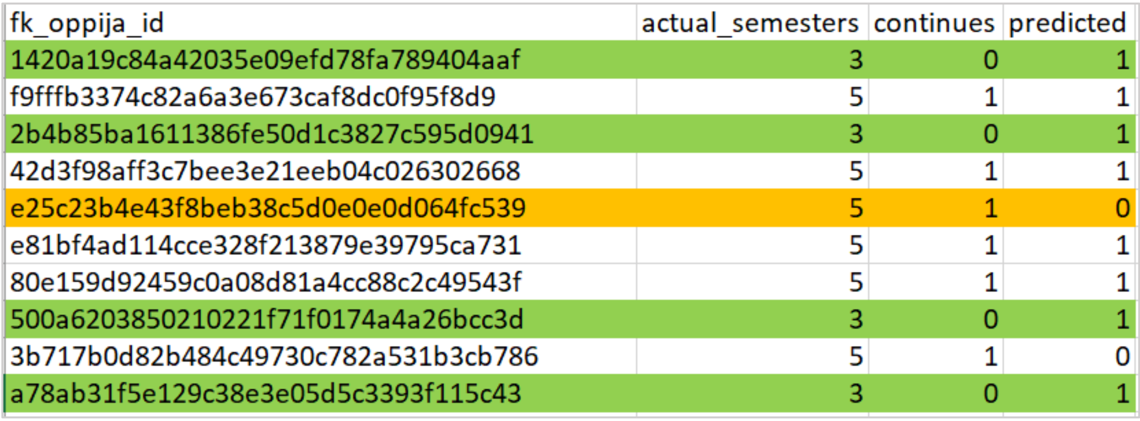


Figure 13. Random accuracy for predicting continuation at studies, context based model, 2012 - 2018.

In the Figure 13 it is visible that the accuracy is still quite unstable and usually drops on the 3rd and 5th semester. This can be explained with the fact this sample dataset includes students from all levels of education, including masters and bachelors who by the nominal curriculum, have different number of semesters. From the previous paragraph it was visible that context based model had very low accuracy when predicting for a Masters curriculum (50052). Taking this into account the decreasing in accuracy after 3rd semester can be explained.

To demonstrate the decrease in accuracy after 3rd semester a test was performed with curriculum 50263. In the Figure 14 an example of students who are predicted to continue their studies based on the grades of their third semester (green), and students, who are predicted to drop out, marked with orange, have been brought out. The model predicts "0" when the student is expected to not continue the next semester, and "1", when according to the grades the student should have no difficulties continuing the studies. From these cases it is visible that according to the performance, the model would assume different result for a student. The actual result, whether a student continues or not, is influenced by the students' decisions and factors that are not evident from the data. When looking at the grades of these students, it was visible that in the most informative courses, selected by Fisher score, the results were quite low, when the prediction was "0".



fk_oppija_id	actual_semesters	continues	predicted
1420a19c84a42035e09efd78fa789404aaf	3	0	1
f9fffb3374c82a6a3e673caf8dc0f95f8d9	5	1	1
2b4b85ba1611386fe50d1c3827c595d0941	3	0	1
42d3f98aff3c7bee3e21eeb04c026302668	5	1	1
e25c23b4e43f8beb38c5d0e0e0d064fc539	5	1	0
e81bf4ad114cce328f213879e39795ca731	5	1	1
80e159d92459c0a08d81a4cc88c2c49543f	5	1	1
500a6203850210221f71f0174a4a26bcc3d	3	0	1
3b717b0d82b484c49730c782a531b3cb786	5	1	0
a78ab31f5e129c38e3e05d5c3393f115c43	3	0	1

Figure 14. Segment of predictions for 3rd semester, curriculum 50263.

To better understand the results, measuring accuracy is not the best way to evaluate the context based model for predicting continuation. This is evident due to a fact that students do tend to quit their studies due to outside factors there is no data about and this would lead to a drop in accuracy. Examples of those cases are demonstrated in green on the Figure 14. A more informative way to get truthful values would be to analyse the confusion matrix. In terms of this thesis the confusion matrix depicts the continuation of studies as such: 0 – not continuing studies next semester; 1 – continuing studies next semester. This is demonstrated in Table 7.

Table 7. Confusion matrix explanation in the context of this thesis. 0 = drop out; 1 = continues.

<table>
<tr><td colspan="2" rowspan="2"></td><th colspan="2">Predicted</th></tr>
<tr><th>0</th><th>1</th></tr>
<tr><th rowspan="2">Actual</th><th>0</th><td>TP: Student who according to data will not continue and are also not predicted to continue.</td><td>FN: Students, who have no more semesters of data, but are predicted to continue according to the grades. In this segment of results are also students who for one reason or another decide not to study further, because of outside factors that are not evident from data accessible currently. (work etc)</td></tr>
<tr><th>1</th><td>FP: Students who according to the data will continue, but the grades are resembling student who will drop out.</td><td>TN: Student, who continue and are predicted to continue.</td></tr>
</table>

For four samples demonstrated on Figure 13, the sum of FP and FN are calculated. Figure 15 demonstrates how the percentage of FP and FN from all samples among all the values changes. It is evident that as the semesters increase, the FN percentage gets lower. This means that the number of students who decide to quit their studies, for other reasons than poor grades, gets lower. It is also visible that FP percentage gets higher. This means that more students will be predicted as potential dropouts, but students decide to not quit their studies and are more motivated to graduate.

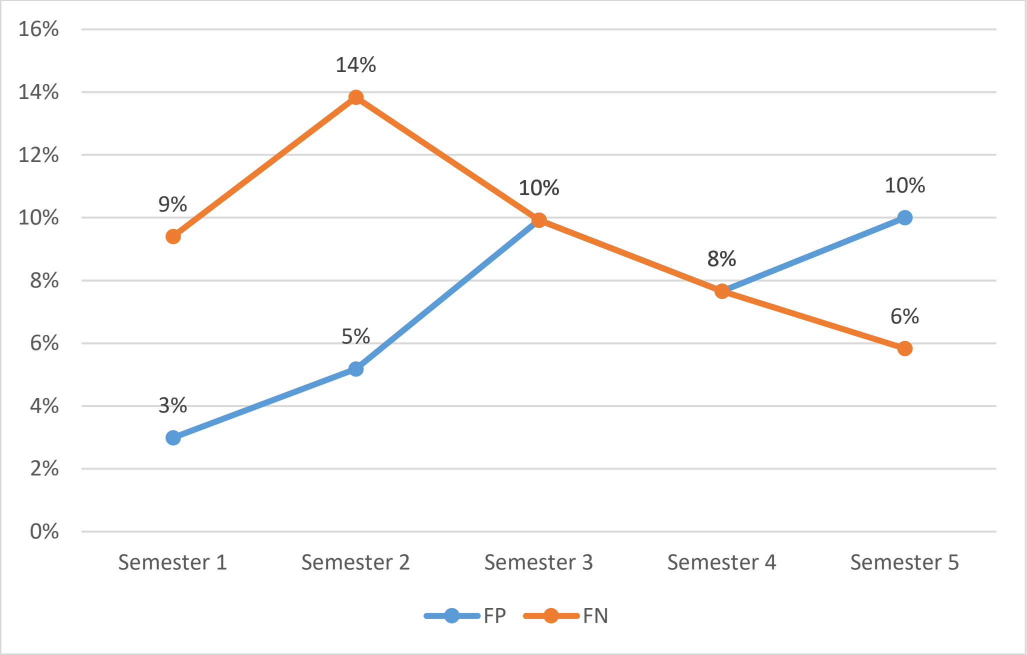


Figure 15. Percentage change for FP and FN rate in sample dataset, context based model.

An implication can be made, that FP can be used to understand how many students are performing worse than expected and how this is related to how far the students are with their studies. Looking at FN gives an understanding of the students who are performing well and are expected to continue the next semester. Analysis of the FP and FN could give an insight into what curriculums are more difficult in the university overall.

From this data it is evident that the context based model is efficient in bringing out students who are on the path to drop out and students who are doing well with their studies and are expected to continue their studies the next semester. However, the prediction of continuing the studies does not demonstrate the actual numbers of students who will continue. These decisions are done by the students themselves. The model gives a beforehand notice about students who are struggling and about students who are doing well. With these results the second hypothesis of the thesis must be rebutted as well, because according to the accuracy the prediction of continuation is not better than the prediction of graduation with context based model. It is evident, that context based method can be used for bringing out students who would need some assistance.

5.3 Future Work

For future work, I believe that the results could be enhanced with adding additional features for both general and context based dataset. In related work many researches used performance measures from before university. Similar results can be taken from SAIS systems and added to the datasets used in this thesis. Other features could be used as well, like duplicated declarations that for this research were eliminated.

The methods and functions developed for this thesis could be enhanced and modified to bring out faulty data automatically to have a better understanding why the methods do not work on some students. There are also many various comments and errors that could be built in, to inform when a model can be used on data and when not. Currently the single cases of error had to be investigated manually. To maximise the usability of the methods, the dates and csv files should be changed to global variables instead of being hard coded to the methods.

It would also be interesting to predict continuation of studies with the nonmandatory courses and see how this would affect the predictions and what patterns would emerge from the data. Additionally, a context based model with less complexity and personalization could be tested.

To gain some additional insight it would be interesting to see what sort of students generally perform better at the university and maybe bring out the correlations between those features. Additionally, it would be interesting and informative to investigate the complexity of each curriculum. This would demonstrate what curriculums are more difficult and what are easier to graduate. An interesting viewpoint would also be to get an understanding how the difficult courses have been distributed on a curriculum.

The data and the ideas investigated in thesis will be additionally analysed and the ideas presented in previous paragraphs will be looked into.

6 Conclusion

The experiments in this thesis brought out some interesting patterns and ideas that have to be considered when working with students' data to be able to predict his or her progress at the university and whether he or she is in line of graduating or dropping out.

In the thesis I wanted to get answers to three questions. I was first interested to see which method is more feasible for predicting dropout and how well can we predict dropout based solely on Study Information system data? From the results it was visible that the general method was more accurate while predicting final graduation or dropping out and yielded to accuracy over 90% after 4th semesters worth of data. The accuracy of predicting graduation with context based methods was on average 0.826 with the original dataset. When predictions were made by semester, for two different curriculums (1008 and 50052), it became evident that accuracy gets higher, when more data about semesters is added to the model, but the results were still lower with context based method than with general model. The average accuracy with general model was 91%, and with the full set of features with general method during 5th and 6th semester, the accuracy yielded to 96%. My third question was regarding the change on accuracy over a period of time. On Figure 10 and Figure 12 it is demonstrated how the accuracy changes over a period of different consecutive semesters. It is visible that for bachelor's degree, on a curriculum 1008, the accuracy of predictions gets higher, for both general and context based methods. This is a clear demonstration that with more data, the predictions get better.

In this thesis I had 5 objectives I wanted to reach. Firstly, I was able to build relevant dataset for predicting from the data gathered from ÕIS database. In hindsight building datasets and cleaning data were the most time-consuming steps of the work, because the data included many corner cases, that had to be looked through. The dataframes and methods had to be modified accordingly. The methods were changed and tested in many iterations.

Second objective was to find out if and when it was possible to use the context based model on student's data, to predict their dropout or graduation. The accuracy of predicting graduation with context based methods was on average 0.826 with the original dataset. Predictions done by semester for master's degree 50052 yielded to very low results in accuracy and it was evident that for that curriculum the method did not work. The same was evident with general method. The result can be extended to other curriculums with a little training data.

During the process of performing tests it came out that context based mode is efficient in terms of predicting whether a student is going to continue his or her studies next semester or he or she is having some complications and would need assistance with the studies. This yielded in accuracy for first semester being more than 0.9 (varied for different curriculums). When it comes to predicting actual graduation or dropping out, context based model has quite high accuracy, but is unreliable due to having high variance in terms of curriculums and the students that are in the training dataset.

The third objective was to bring out courses that play significant role in student's graduation. These courses were analysed with Fisher score evaluation and are brought out in the Appendix 6 – Most Informative Courses Selected by Fisher Score. For each curriculum the most informative course is demonstrated. It is important to state that the results were only brought out for the curriculums that had students in the data under analysis (2012-2018). Fisher score selects courses that carry the most information regarding the graduation label.

Fourth objective was to compare the general and context based model. The advantages, disadvantages and differences of them are described in the Results and Analysis paragraph. The main differences were, that context based model had high accuracy in predicting drop out and graduation, but also high variety, which made it unreliable. On the other hand, context based model did well predicting students who will continue the studies next semester and brought out students who were struggling with their studies. The general model had a high accuracy of predicting student's graduation and drop out based on all of the data. The prediction accuracy for curriculum 1008 yielded to 96%. General method is therefore more reliable when predicting final graduation.

Fifth objective was to get an understanding which model would be more meaningful to look into and analyse more. From the results I would suggest continuing using and developing both methods, because context based model is better for predicting continuation, and general model at this moment seems to be better and more reliable when predicting final graduation or dropping out.

The three working hypotheses for this thesis were all rebutted. This goes to show that the initial assumptions were not correct, and the actual result based on data is different.

A big part of this thesis was the constructing of the datasets and data cleaning. This is also the part of the thesis that took the most time. A lot of time went on testing models on the pre-reform dataset and the discovery that the data is a lot more scattered than the recent five years. Because of students are very different a lot of time was spent on performing individual analysis on students and their results and prediction. The methods for generating tests were modified multiple times and structured to collect meta information regarding each prediction to make the later analysis more efficient.

6.1 Lessons Learned

During the process of developing the methods and datasets, I learned and discovered many ideas that I would be doing differently next time.

I would use local database, to improve the joining of big tables and getting access to them faster. I also in practise understood that data cleaning, preparation and structuring takes the most time. It was something I knew before, but going through the process gives a better understanding of the scale of the work. Fixing unexpected mistakes and data faults is time consuming and has to be worked on during many iterations. I gathered ideas how to look through data before any predictions are done with it and experienced what are the main issues that might rise when working with such amount of data. I also in many cases underestimated the grand scheme of the task ahead. Creating diagrams requires structuring data in a very specific way to be able to provide visually appealing diagrams and this also requires a level of understanding of the data, methods and time. I learned that to be able to predict or visualize data, a lot of effort goes into structuring the data so it would be possible to use it.

References

[1] S.-P. Lin, “Using EDM for Developing EWS to Predict University Students Drop Out,” *International Journal of Intelligent Technologies and Applied Statistics,* vol. 8, no. 4, pp. 365-388, 2015.

[2] “Pandas: Intro to Data Structures,” [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/dsintro.html. [Accessed 2017].

[3] S. M. Gratiano and W. J. Dr. Palm IV, “Can a Five Minute, Three Question Survey Foretell First-Year Engineering Student Performance and Retention?,” in *Jazzed about Engineering Education,* New Orleans, LA, 2016.

[4] S. Frazelle and A. Nagel, “A practitioner’s guide to implementing early warning systems,” The National Center for Education Evaluation and Regional Assistance, Washington, DC, 2015.

[5] “American Graduate DC,” [Online]. Available: http://www.americangraduatedc.org/dropout_crisis.

[6] Haridus- ja Teadusministeerium , “Haridus- ja Teadusministeeriumi aasta-analüüs,” Haridus- ja Teadusministeerium, Tartu, 2015.

[7] “Tallinn University of Technology,” [Online]. Available: https://www.ttu.ee/university/ttu-in-brief/about-university/.

[8] Tallinna Tehnikaülikool, “Kokkuvõte Tallinna Tehnikaülikooli õppetegevusest 2015/2016. õppeaastal,” Tallinna Tehnikaülikool, Tallinn, 2016.

[9] Haridus- ja teadusminister, “Institutsionaalne arendusprogramm teadus- ja arendusasutustele ja kõrgkoolidele,” 2015. [Online]. Available: https://www.riigiteataja.ee/akt/110042015004.

[10] “Cohesion Fund,” [Online]. Available: http://ec.europa.eu/regional_policy/en/funding/cohesion-fund/.

[11] A. Saabas, “Diving into data: Interpreting random forests,” 19 October 2014. [Online]. Available: http://blog.datadive.net/interpreting-random-forests/. [Accessed 2018].

[12] A. Vellido, J. D. Martin-Guerrero and P. J. Lisboa, “Making machine learning models interpretable,” in *European Symposium on Artificial Neural Networks, Computational Intelligence,* Bruges, 2012.

[13] F. Doshi-Velez and B. Kim, “Google AI: Towards A Rigorous Science of Interpretable Machine Learning,” 2017. [Online]. Available: https://ai.google/research/pubs/pub46160.

[14] “Python Data Analysis Library,” [Online]. Available: https://pandas.pydata.org/.

[15] “NumPy,” [Online]. Available: http://www.numpy.org/.

[16] “Scikit-learn: Machine Learning in Python,” [Online]. Available: http://scikit-learn.org/stable/.

[17] “Matplotlib,” [Online]. Available: https://matplotlib.org/.

[18] "Seaborn: statistical data visualization," [Online]. Available: https://seaborn.pydata.org/.

[19] "General Data Protection Regulation," [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679. [Accessed 2018].

[20] R. D. Augustine, "Persistence And Attrition Of Engineering Students, A Study Of Freshman And Sophomore Engineering Students At Three Midwestern Universities," Michigan State University, Michigan, 1966.

[21] A. A. Aziz, J. A. Jusoh, W. M. R. W. Idris and H. Hassan, "Implementing Aproiri Algorithm for Predicting Result Analysis," *GSTF Journal on Computing,* vol. 2, no. 4, p. 87, 2013.

[22] S. Rovira, E. Puertas and L. Igual, "Data-driven system to predict academic grades and dropout," University of Barcelona, Barcelona, 2017.

[23] P. A. Murtaugh, L. D. Burns and J. Schuster, "Predicting the Retention of University Students," *Research in Higher Education,* vol. 40 , no. 3, 1999.

[24] W. D. Slanger, E. A. Berg, P. S. Fisk and M. G. Hanson, "A Longitudinal Cohort Study of Student Motivational Factors Related to Academic Success and Retention Using the College Student Inventory," *Journal of College Student Retention: Research, Theory & Practice,* vol. 17, no. 3, pp. 278-302, 2015.

[25] B. Uga, "TTÜ tudengite väljalangemise ennustamine: Tõenäosuse arvutamine masinõppe meetodite abil ning tulemuste kuvamine veebirakenduses," Tallinn, 2017.

[26] "Wikipedia, Educational Data Mining," [Online]. Available: https://en.wikipedia.org/wiki/Educational_data_mining.

[27] "Educational Data Mining," [Online]. Available: http://educationaldatamining.org/about/. [Accessed 2017].

[28] C. Pradhan, "Quora," [Online]. Available: https://www.quora.com/What-are-the-differences-between-ID3-C4-5-and-CART. [Accessed 22 January 2018].

[29] "Scikit learn: Decision Trees," [Online]. Available: http://scikit-learn.org/stable/modules/tree.html. [Accessed 2018].

[30] L. Hulstaert, "Towards Data Science," [Online]. Available: https://towardsdatascience.com/interpretability-in-machine-learning-70c30694a05f. [Accessed January 2018].

[31] J. Brownlee, "Machine Learning Mastery," [Online]. Available: https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/. [Accessed February 2018].

[32] C. C. Aggarwal, Data Mining The Textbook, Switzerland: Springer, 2015.

[33] "Wikipedia: Feature Selection," [Online]. Available: https://en.wikipedia.org/wiki/Feature_selection#cite_note-M._Phuong,_Z_pages_301-309-30. [Accessed January 2018].

[34] "Wikipedia: Entropy (information theory)," [Online]. Available: https://en.wikipedia.org/wiki/Entropy_(information_theory). [Accessed January 2018].

[35] K. Markham, "Data School: Simple guide to confusion matrix terminology," [Online]. Available: http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/. [Accessed March 2018].

[36] “Confusion Matrix,” [Online]. Available: https://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix_files/confusion_matrix_1.png. [Accessed 2018].

[37] “PuTTY,” [Online]. Available: https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html. [Accessed 2017].

[38] “The Secure Shell (SSH) Protocol Architecture,” [Online]. Available: https://tools.ietf.org/html/rfc4251. [Accessed 2017].

[39] “Cygwin/X,” [Online]. Available: https://x.cygwin.com/. [Accessed 2017].

[40] “Õppekorralduse eeskiri: Õpitulemuste hindamine,” [Online]. Available: https://www.ttu.ee/tudengile/oppeinfo/oppekorraldus/oppetegevuse-juhendid-ja-oigusaktid/oppee/#14_kulalisuliopilane. [Accessed 2018].

[41] J. Li, “Github: jundongl/scikit-feature,” [Online]. Available: https://github.com/jundongl/scikit-feature/blob/master/skfeature/function/similarity_based/fisher_score.py. [Accessed 2018].

[42] “Sklearn,” [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html. [Accessed April 2018].

[43] Riigikogu, *Ülikooliseaduse, rakenduskõrgkooli seaduse ja teiste seaduste muutmise seadus,* Vabariigi President, 2012.

Appendix 1 – ÕIS Database Tables Used

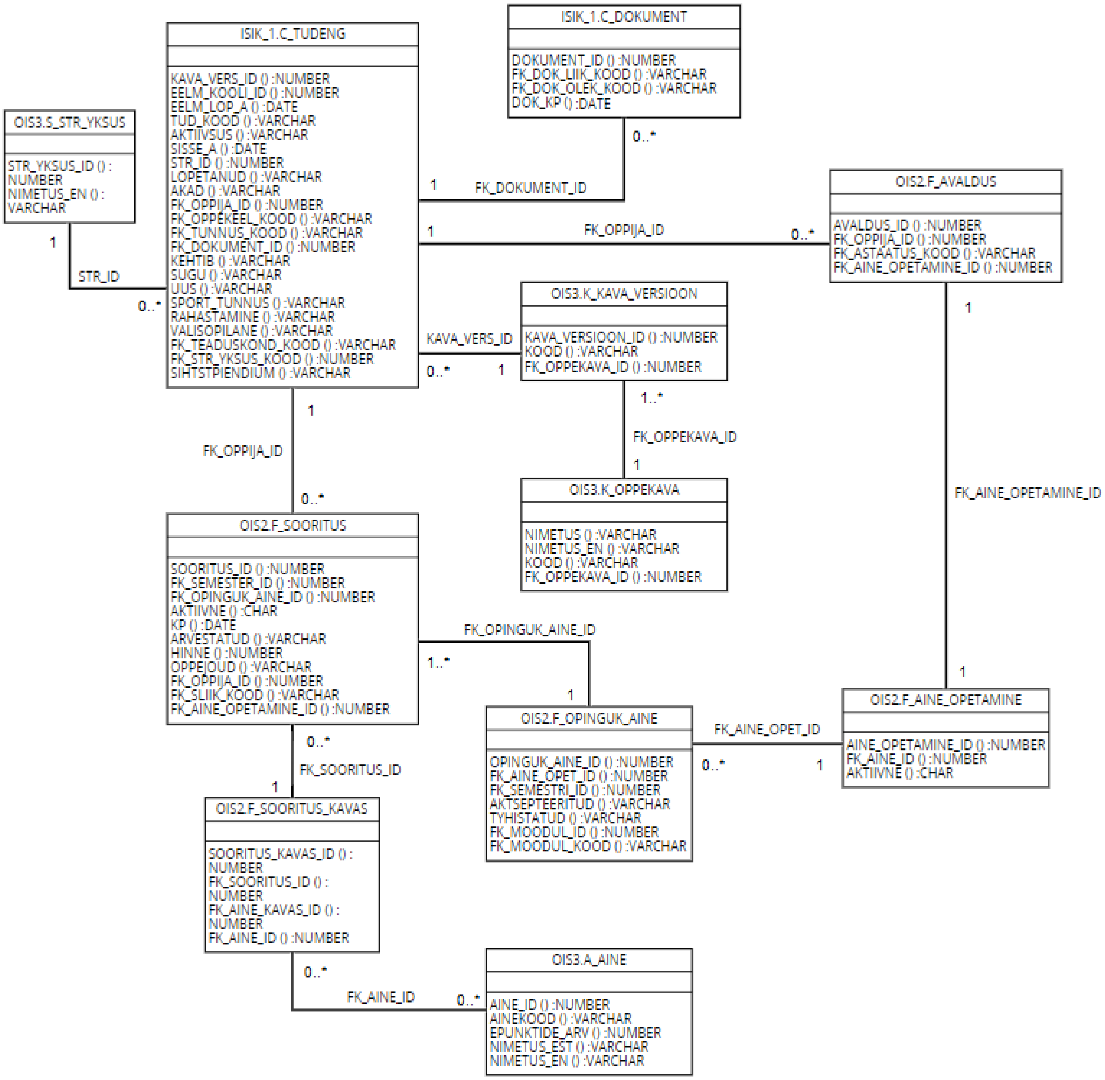


Figure 16 Overview of the main database tables used for this thesis.

Appendix 2 – Description of Methods Used in the Thesis

The code can be found at https://gitlab.cs.ttu.ee/Kristin.Ehala/MasterThesis

Table 8 Methods developed for the usage of this thesis.

Method	Description
get_all_curriculums()	This method reads in curriculum data from CSV file and returns the data. The dataframe is merged with "curriculum version" dataset to add some additional information.
read_students()	This method reads in student data provided in the CSV file. The method takes only "valid (kehtiv)" rows from the dataset. Student data is merged with "struktuuri üksused" dataset, to provide information about faculties. The dataset is the merged with curriculums dataset to add information about a specific curriculum the student is studying at.
get_graduated_students_last_five_years()	This method returns student who have graduated their studies at TTU in a positive way. The method merges data from student data CSV file and "document" data CSV file.
get_five_year_students(all_students, curriculums)	The method returns dataframe of students from the time period under analysis in this thesis (1. August 2012 - 29. March 2018). format is mm-dd-yyyy. Missing curriculum versions are imputed with value 0.
get_active_students(all_students)	This method returns all active students from the data and all active students that have started their studies after the period under

	analysis (1.August 2012) (aktiivsus == 'jah'). The method returns two dataframes.
get_dropout_students(all_students)	This method returns all dropped students from the data and all dropped out students that have started their studies after the period under analysis (1.August 2012) (aktiivsus =='ei')(lopetanud == 'ei') The method returns two dataframes.
creating_results_dataset()	This method returns all results that have been done by the students at TTU from the beginning of time. The CSV soorutused is read in and merged with "F_OPINGUK_AINE", "f_sooritus_kavas", "f_aine_opetamine", "a_aine". Method "change_counted_to_grade" is applied to the final dataset to change "arvestatud" courses and "graded" courses to one scale. This method adds the feature "grade".
create_internship_dataset()	This method selects all internship related declarations from the"sooritused" dataset. "Sooritused" CSV file is merged with "f_avaldus", "f_aine_opetamine", and "a_aine" CSV file.
create_thesis_dataset()	This method returns thesis done by the students. The data is anonymised.
create_vota_dataset()	This method returns course declaratiosn that are done through VÕTA system. The data is covered with "sooritused" dataset. The "sooritused" CSV is merged with "f_sooritus_kavas" and "a_aine".

get_courses()	This method returns dataframe of all unique courses taught at TTU.
get_five_year_results(all_results, five_years_students)	This method returns returns dataframe of all results "sooritused" that have been done after the time period that is under analysis (later than 1.August.2012)
create_context_dataset(curriculums, five_year_results, five_years_students, curriculum)	This method creates context dataset, returning a dataframe of all the results that have been done by students in one curriculum. The method needs 4 attributes: all curriculums, five year results, five year students and the specific curriculum under analysis. All curriculum versions of one curriculum are merged into one dataset. The method returns dataset where each row represents one student with all of their results. Only students who are NOT active are used in the making of the pivotal dataset. The courses selected are MANDATORY.
create_context_dataset_train(curriculums, five_year_results, five_years_students, curriculum, pivot_table_predict)	This method changes the dataframe created in previous method "create_context_dataset " according to the dataframe that will be used for predicting. The method returns training dataset for context based model.

create_active_test_dataset(curriculums, five_year_results, five_years_students, curriculum)	This method creats context dataset, returning a dataframe of all the results that have been done by ACTIVE students in one curriculum. The method needs 4 attributes: all curriculums, five year results, five year students and the specific curriculum under analysis. All curriculum versions of one curriculum are merged into one dataset. The method returns dataset where each row represents one student with all of their results. Only students who are NOT active are used in the making of the pivotal dataset. The courses selected are MANDATORY.
create_active_test_dataset_one_student(fk_oppija_id, curriculums, five_year_results, five_years_students)	This method creates pivot dataset according to one student. This method is used for testing context based model and the returned dataframe is used to change the context dataset according to a student. The method returns two outcomes: a dataframe generated for a student and the specific curriculum that the student is studying for.
create_general_dataset(five_years_students)	This method creates dataframe for general model. The method takes as an input previously made "five_years_students" dataframe and drops some uninformative columns from the data. Columns where over half of the data is missing are removed.These cannot be imputed.

create_semester_with_student()	This method creates dataframe where all semesters when student has been active are brough out. CSV files "f_oppimine_sem" is merged with "f_semester" file. The data is then cleaned.
remove_duplicate_declarations(dataset)	This method removes duplicated declarations from the results dataset.
apply_corresponding_grade(value_in_series)	This method changes 'arvestatud' courses to 'grades'. The function takes in a series in dataframe.
change_counted_to_grade(dataset)	This method changes "arvestatud" courses into grades. Method takes in a dataset and returns a dataset with a new feature.
replace_comma_with_dot(value_in_series)	This method changes some data mistakes and returns a numerical series.
remove_uninformative_columns(dataset)	This method removes features from a dataset that are not informative.
add_sum_of_eap(dataset, results, semester)	This method calculates EAP sum for each student in the dataset. Student fk_oppija_id has to be an index in dataset. Results is the dataset, that includes all the results the student has done. Semester show for how many semesters the sum of EAP has to be calculate for.
add_average_grade(dataset, results, semester	This method calculates average grade for each student in the dataset. Student fk_oppija_id has to be a index in dataset. Results is the dataset, that includes all the results the student has done. Semester show for how many semesters the grade has to be calcluated for.

add_active_semesters(dataset, results)	This method calculates number of unique active semesters for each student. Student fk_oppija_id has to be a index in dataset. Results is the dataset, that includes all the results the student has done.
add_eap_per_semester(dataset)	This method calculates average EAP per semester for each student. Student fk_oppija_id has to be a index in dataset. Results is the dataset, that includes all the results the student has done.
add_semester_to_sooritus(dataset, all_fk_oppija_id, five_year_results_drift)	This method calculates sum of EAPs per semester for each student. Student fk_oppija_id has to be a index in dataset. Results is the dataset, that includes all the results the student has done. all_fk_oppija_id is the series that includes all the fk_oppija_id-s in the dataset.
add_remaining_semesters(test_data, semester)	This method calculates sum of EAPs per semester for each student. Student fk_oppija_id has to be a index in dataset. Results is the dataset, that includes all the results the student has done. all_fk_oppija_id is the series that includes all the fk_oppija_id-s in the dataset.
get_fisher_scores(pivot_table)	This method takes a pivot_table and uses the data in table to evaluate what courses/features in the table carr the most knowledge related to the final label. The predictive label has to be 'lopetanud'. Method uses fisher_score

	(fisher_data, target) from Git repository “scikit-feature/skfeature”
get_general_fisher(data_for_prediction)	This method finds fisher score for general dataset. The target feature has to be 'lopetanud_numeric'. Method uses fisher_score(fisher_data, target) from Git repository “scikit-feature/skfeature”
decision_tree_classifier(data, target)	This method takes in data and target and splits it into test and training data and returns fitted Decision Tree classifier.
prediction_evaluation(clf, data, target, target_test, data_test, data_train, target_train)	This method uses test and train data to evaluate the accuracy of the classifier.
prediction_evaluation_real(target, prediction)	This method returns only classification report.
get_statistics_datasets()	This method generaes statistics about how much training data is available for each curriculum. It shows the maximum amount that is possible. The method requires: all_students, curriculums, five_years_students and five_year_results as preliminary dataframes.
get_fisher_statistics()	This method generaes statistics about top X coures selected with fisher score. Method loops through all curriculums that are visible in the data and have enough information in them.
context_origial_100(active_five_years, fisher_number, sample_data)	This method uses original dataset to provide prediction /accuracy info on sample of the data. The

	method returns two dataframes, both with different meta data provided by the prediction. Fisher_number decides how many courses are selected for prediction.
drift_statistcs_100(five_year_students_drift, fisher_number, sample_data)	This method uses pre-reform/drift dataset to provide prediction/ accuracy info on sample of the data.
predict_by_semester(sem_students, fisher_number, sample_data, semester)	This method predicts continuation by semester for next semester.
graduation_predict_by_semester(sem_students, fisher_number, sample_data, semester)	This method predicts graduation by semester, taking account fisher_number worth of courses done by the student.

Appendix 3 – Process of Predicting with Context Based Method

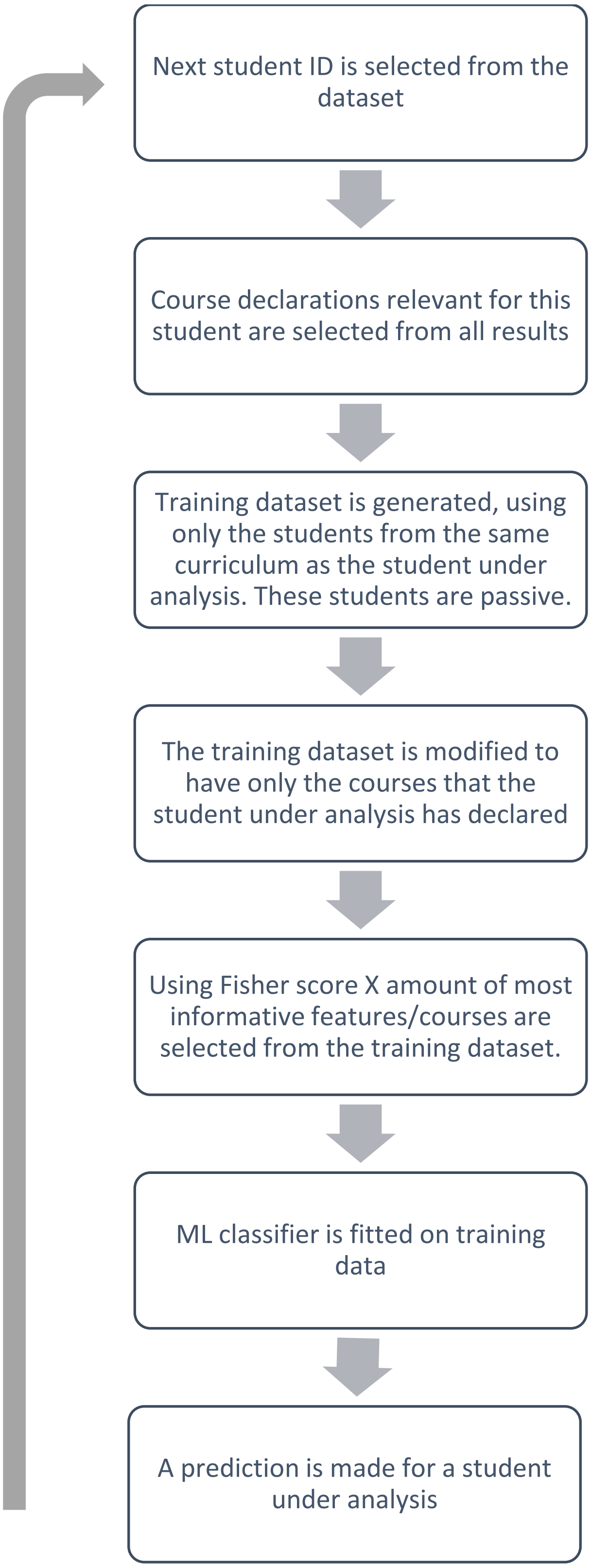

Next student ID is selected from the dataset
Course declarations relevant for this student are selected from all results
Training dataset is generated, using only the students from the same curriculum as the student under analysis. These students are passive.
The training dataset is modified to have only the courses that the student under analysis has declared
Using Fisher score X amount of most informative features/courses are selected from the training dataset.
ML classifier is fitted on training data
A prediction is made for a student under analysis

Figure 17 Process of predicting with context based model.

Appendix 4 – Prediction on 3rd Semester for Curriculum 50263

Table 9 Prediction for continuation on 3rd semester for curriculum 50263.

fk_oppija_id	**Actual semesters**	**Continues**	**Predicted**	**Training accuracy**	**Test accuracy**
1420a19c84a42035e09efd78fa789404aaf	3	0	1	Accuracy: 0.98 (+/- 0.07)	0.9762
f9fffb3374c82a6a3e673caf8dc0f95f8d9	5	1	1	Accuracy: 0.95 (+/- 0.11)	0.9773
2b4b85ba1611386fe50d1c3827c595d0941	3	0	1	Accuracy: 0.98 (+/- 0.08)	0.9762
42d3f98aff3c7bee3e21eeb04c026302668	5	1	1	Accuracy: 0.99 (+/- 0.05)	0.9750
e25c23b4e43f8beb38c5d0e0e0d064fc539	5	1	0	Accuracy: 0.95 (+/- 0.09)	0.9767
e81bf4ad114cce328f213879e39795ca731	5	1	1	Accuracy: 0.96 (+/- 0.09)	0.9773
80e159d92459c0a08d81a4cc88c2c49543f	5	1	1	Accuracy: 0.98 (+/- 0.06)	0.9545
500a6203850210221f71f0174a4a26bcc3d	3	0	1	Accuracy: 0.98 (+/- 0.08)	0.9524
3b717b0d82b484c49730c782a531b3cb786	5	1	0	Accuracy: 0.95 (+/- 0.06)	0.9545
a78ab31f5e129c38e3e05d5c3393f115c43	3	0	1	Accuracy: 1.00 (+/- 0.00)	1.0000
f19e007c66d2e101382adc7bca73e698935	6	1	1	Accuracy: 0.95 (+/- 0.08)	0.9545
c24278de5bf7ac1848aa61289747bef9db3	6	1	1	Accuracy: 0.94 (+/- 0.10)	0.9773
4f41df3ba2182a6f13c6e61bab14bd54634	3	0	0	Accuracy: 0.97 (+/- 0.06)	0.9545
0c5cd6ae6d6f67ed5e8a8f5382c9cb52a7c	3	0	1	Accuracy: 0.97 (+/- 0.06)	1.0000
e533c348b0604e484b72bd59d204a107652	5	1	1	Accuracy: 0.97 (+/- 0.09)	0.9318
e626f53da8663a7ba795b788711c12c40c0	3	0	1	Accuracy: 0.98 (+/- 0.05)	0.9524
983fe10d4a7e18231f271605764d9d7596c	5	1	1	Accuracy: 0.97 (+/- 0.09)	0.9545
360778e1998fbaba917ac2dd4c29dd4a1e6	6	1	1	Accuracy: 0.99 (+/- 0.05)	0.9744
d909e710dff7d948721b2d9251ec21de0fe	5	1	1	Accuracy: 0.98 (+/- 0.05)	0.9318
7af253bc4ea124b6e26f1e547cd497d4595	3	0	1	Accuracy: 0.96 (+/- 0.08)	1.0000
6d3ca09bce66a4fa77089a1bc91e1b6b31d	3	0	1	Accuracy: 0.97 (+/- 0.06)	1.0000
771e257d2c4d75bbf033952a29eb1b7a068	5	1	1	Accuracy: 0.97 (+/- 0.06)	1.0000

0fbe28f28312eb9e5e46bb7d1f0e3df9cea	3	0	1	Accuracy: 0.98 (+/- 0.08)	1.0000
9ee4558247765633991e38f6a81ca022064	7	1	1	Accuracy: 0.93 (+/- 0.09)	1.0000
433cdc03e664d067df6399ebd1dacd5ae96	3	0	0	Accuracy: 0.82 (+/- 0.18)	1.0000
7aae086a9db632f18a525e106c8097cc896	6	1	1	Accuracy: 0.97 (+/- 0.06)	0.9545
eed3778f65cc9d934d07217a7fa430293e2	3	0	1	Accuracy: 0.97 (+/- 0.06)	0.9762
b71cbaaf3e9a469d3ae9c9733ca4fd9128e	3	0	1	Accuracy: 0.97 (+/- 0.06)	1.0000
244bbd5115f2404a5e30d8b295970cb91fd	5	1	1	Accuracy: 0.98 (+/- 0.05)	0.9318
b576c25ff8599f89d023728b1bd97d9b44b	6	1	1	Accuracy: 0.95 (+/- 0.12)	0.8837
bbfad170311f7964dc2cd38c42602c37646	5	1	1	Accuracy: 0.97 (+/- 0.07)	0.9318
3a8411367d36b75e86812fb786874703ed8	5	1	1	Accuracy: 0.98 (+/- 0.06)	0.9545
cf26bb7f78c6c72490167ee5521fbac8d06	5	1	1	Accuracy: 0.98 (+/- 0.05)	0.9318
76d13a9862bfdf5d5aad5ee511a7cceb7b3	5	1	1	Accuracy: 0.94 (+/- 0.09)	0.9318
f030b1a62fd4fad5af59b416525fec6eb2c	5	1	1	Accuracy: 0.94 (+/- 0.08)	0.9773
dbbbde4433858209c54ef721c69cf189057	5	1	0	Accuracy: 0.82 (+/- 0.33)	0.9286
a00b35d8a99d78c383b4a9abd7b645a1d14	3	0	1	Accuracy: 0.98 (+/- 0.06)	0.9524
9cdfdac7424b3be44394f88c021d6a74d03	3	0	1	Accuracy: 0.97 (+/- 0.06)	0.9762
31bcce5c1f806c77048fbddd0f5ac4a40f9	5	1	1	Accuracy: 0.98 (+/- 0.06)	0.9545
bc22163ae4a5d29a1e2b09601297349aeda	4	1	1	Accuracy: 0.93 (+/- 0.12)	0.9767
d3760ecbb4af490c06c005e74167423e2e9	5	1	1	Accuracy: 1.00 (+/- 0.00)	1.0000
ce13bd0c56647556d9d9fd344f7520a7a6e	5	1	1	Accuracy: 0.95 (+/- 0.08)	1.0000
c37e8ad566bcd247fb710f90428d3a2a6b0	5	1	1	Accuracy: 0.94 (+/- 0.09)	0.9773
f920afae066080ec6ed8ffc0768f9e45bc1	3	0	1	Accuracy: 0.96 (+/- 0.07)	0.9318
41f3a50be3e305e0ee1f61ca3f0c056d4fd	3	0	1	Accuracy: 0.98 (+/- 0.06)	0.9762
0d79b5d0e109a5fae5349778dcc78f781fe	3	0	1	Accuracy: 0.98 (+/- 0.05)	0.9524
1da54f9a52f5caf7185a85516aedf5cf83d	4	1	1	Accuracy: 0.94 (+/- 0.07)	0.9545

ea300bb261f06e24500cce42cb43473e6c1	3	0	1	Accuracy: 0.98 (+/- 0.06)	0.9762
3edd966f143aca485f2e3ffebb1778c8b64	5	1	1	Accuracy: 0.98 (+/- 0.05)	0.9318
2154ab345a563fa1673b0fa7c60126aca3b	5	1	1	Accuracy: 0.96 (+/- 0.07)	0.9545
6c59134914d2a216e2050c5c0e3d79087ab	5	1	1	Accuracy: 0.96 (+/- 0.07)	0.9773
bfc42d14dd8cd045061b6486748cbc2e930	5	1	1	Accuracy: 0.97 (+/- 0.07)	0.9545
c0c53300cfde03923c1f0f6854ccd18ffad	5	1	1	Accuracy: 0.97 (+/- 0.10)	0.9545
64050d81b47681085cf2f74c6454e46e274	3	0	1	Accuracy: 0.96 (+/- 0.08)	1.0000
64d37d910f79ce33f3770415396c79bfabe	6	1	1	Accuracy: 0.99 (+/- 0.05)	0.9524
098858682d45b7c629863de437d09cadf4f	6	1	1	Accuracy: 0.98 (+/- 0.06)	0.9091
46da5cdc618518dd1c59c62b79fe7cd7f0f	3	0	1	Accuracy: 0.96 (+/- 0.10)	1.0000
93f098ffb51142a645f0a08ce244c6440bc	3	0	1	Accuracy: 0.94 (+/- 0.09)	0.9535
2b49866341d7387866be962c2761e057a28	3	0	1	Accuracy: 0.98 (+/- 0.07)	1.0000
1f5a3f7ebe173f83d927c495a8467f4d145	4	1	1	Accuracy: 0.98 (+/- 0.06)	0.9091
a7c6faaf9bc1732fa6c2b4d5d0bc93ce12c	5	1	1	Accuracy: 0.98 (+/- 0.06)	0.9091
d0387b453c65fa81476d23022b785514657	5	1	1	Accuracy: 0.97 (+/- 0.07)	0.9773
3b10e2c0870498e7b3733a3d1aac23434f5	3	0	1	Accuracy: 0.98 (+/- 0.05)	0.9762
85d30bb8e2475ec5c2cd06317549987be3e	5	1	1	Accuracy: 0.95 (+/- 0.08)	1.0000
8c35e0c073a2d46178e50a180647ee50dda	3	0	1	Accuracy: 0.97 (+/- 0.06)	1.0000
207cdde638c9c2efecca83c2f4f42613609	6	1	1	Accuracy: 0.99 (+/- 0.04)	0.9524
14c3cd6707333646a59cd03c484f71707e7	4	1	1	Accuracy: 0.98 (+/- 0.06)	0.9750
215e665f8fc08b89c434821a74b73ac3167	5	1	1	Accuracy: 0.96 (+/- 0.07)	0.9773
52d13e118ff3358ce08b59a8948915b29b5	5	1	1	Accuracy: 0.96 (+/- 0.08)	0.9091
4498618d246d7e048876735006c1919f6b3	3	0	1	Accuracy: 0.97 (+/- 0.08)	0.9762
f20e2c92519d27e278d36d719eab31434e8	3	0	1	Accuracy: 0.98 (+/- 0.06)	1.0000
1af1780b93acbdb44aea28d63ae2012751e	3	0	1	Accuracy: 1.00 (+/- 0.00)	1.0000

ccac86073ebe1b65774ee22c81ddc38f383	5	1	1	Accuracy: 0.90 (+/- 0.31)	0.7692
45ac81b024f73ad16b3e67b8aec2cb90f35	3	0	1	Accuracy: 0.96 (+/- 0.08)	1.0000
50f5cf184bbfe42993c5287fe15e4925e25	5	1	1	Accuracy: 0.98 (+/- 0.06)	1.0000
b8e979b64d0522f44f2a93497201b06f445	3	0	1	Accuracy: 0.98 (+/- 0.06)	0.9762
b9135bf51ce71eb0701a9345ec6b876e934	5	1	1	Accuracy: 0.98 (+/- 0.06)	0.9750
d7d3173b3f598500ecce546de20a31ec05a	5	1	1	Accuracy: 0.98 (+/- 0.06)	1.0000
0cb8f623de83469afbb831e3226e4fb1f31	7	1	0	Accuracy: 0.82 (+/- 0.26)	0.9167
92e8f4246d0d2f03708eb0abe5d9fd43158	3	0	1	Accuracy: 0.94 (+/- 0.11)	0.9767
ddfa812e29ed73d9aeaf4c87431900bec13	3	0	1	Accuracy: 0.98 (+/- 0.08)	0.9762
da02d7cd8f52d2b46c7f1be9646821dfe87	5	1	1	Accuracy: 0.95 (+/- 0.06)	0.9773
cc1b9188f43299cb09ea61c613a56f8c539	3	0	1	Accuracy: 1.00 (+/- 0.00)	1.0000
bb80c39702eba8ed5a628a346905a00c386	5	1	1	Accuracy: 0.92 (+/- 0.17)	0.7857
d0236b97a3a4d1c09ba2d29a4aff095bdf9	3	0	1	Accuracy: 0.97 (+/- 0.08)	1.0000
7522321785933a144408f4c5b2dbe31bcbb	5	1	1	Accuracy: 0.95 (+/- 0.06)	1.0000
e0a56ebf6cc412226040574a549c12618fd	5	1	1	Accuracy: 0.95 (+/- 0.08)	0.9773
e2ef9a3f32f7b36c60546d1720e00698351	3	0	0	Accuracy: 0.84 (+/- 0.22)	0.9231
bbf569def290b26e4fda65b4cb6f28e15ad	3	0	1	Accuracy: 0.98 (+/- 0.05)	0.9524
c1990f71d9a91c7c1240e035be54eef97e5	5	1	1	Accuracy: 0.96 (+/- 0.11)	1.0000
91a25099d58c0368cb50a49bad48dffc435	5	1	1	Accuracy: 0.97 (+/- 0.08)	0.9545
3d82b49df925106f772789f8c239129bd48	7	1	1	Accuracy: 0.82 (+/- 0.19)	0.8235
47a2155e6e0905ed505ccd327fe96c5c894	3	0	1	Accuracy: 1.00 (+/- 0.00)	0.9286
a7995536f5ef51411d96365c258b0d2a38a	6	1	1	Accuracy: 0.97 (+/- 0.06)	0.9545
8f02a13e4e09456a35287c05df23875aa49	3	0	1	Accuracy: 1.00 (+/- 0.00)	1.0000
2f98346c3f1616ffe58378e4c4d0f899119	7	1	1	Accuracy: 0.95 (+/- 0.08)	0.9535
072c7e81a5a560fdf539254aa58e60e10ab	5	1	1	Accuracy: 0.95 (+/- 0.08)	0.9773

75929c66b65be2cd65d785d2d307c05830f	4	1	1	Accuracy: 0.88 (+/- 0.13)	0.9583

Appendix 5 – Maximum Rows in Context Based Training Dataset

Table 10 Maximum rows in context based training dataset.

Curriculum id	Max students in trainset	Max nr of courses	Curriculum name
1008	592	60	Informatics
652	382	82	Business Information Technology
1045	330	65	Computer and Systems Engineering
775	313	72	Business
50083	301	67	International Business Administration
406	259	143	Structural Engineering and Construction Management
50031	245	100	Law
729	233	58	Product Development and Production Engineering
732	223	46	Product Development and Production Engineering
860	210	35	Business Information Technology
789	208	32	Finance and Accounting
923	203	52	Management and Marketing
1013	198	58	Computer Science
50026	192	27	Cybersecurity
50035	192	84	International Relations
50052	190	26	Personnel and Development
831	182	56	Mechatronics
754	174	64	Electrical Power Engineering
50027	147	14	Software Engineering
929	145	46	Public Administration and Governance
529	140	54	Logistics
556	136	54	Logistics
709	132	67	Applied Economics
508	129	42	Gene Technology
810	128	41	Food Engineering and Product Development
50032	122	42	Law
1599	120	40	Applied Information Technology
1055	107	15	Computer and Systems Engineering
50115	107	109	Navigation
1449	106	58	Accounting and Business Management
786	104	25	Electrical Power Engineering
875	102	66	Electrical Engineering
869	101	22	Mechatronics
948	101	49	Applied Chemistry and Biotechnology
573	100	49	Chemical and Environmental Technology
974	100	18	Business Administration

50116	99	95	Port and Shipping Management
1429	99	51	International Economics and Business Administration
1284	98	22	International Business Administration
50036	93	54	International Relations and European-Asian Studies
824	93	37	Food technology and Development
620	88	146	Indoor Climate in Buildings and Water Engineering
50096	87	23	E-Governance Technologies and Services
658	77	44	Thermal Power Engineering
50108	76	43	Integrated Engineering
5373	75	56	Industrial Automation
50103	74	46	Electronics and Telecommunications
3590	74	54	Fuel Technology
50062	73	28	Communicative Electronics
3982	72	40	Power Engineering
1729	72	17	Environmental Engineering and Management
50038	70	34	Health Care Technology
4051	69	22	Industrial Engineering and Management
488	67	19	Gene Technology
1589	67	48	Machine-building Engineering
959	67	24	Applied Chemistry and Biotechnology
496	67	85	Road Engineering and Geodesy
565	66	39	Technology of Wood and Textile
884	62	11	Energy Conversion and Control Systems
50085	60	21	Technology of Wood, Plastic and Textiles
520	59	24	Chemical and Environmental Technology
50059	57	19	Energy Efficiency of Buildings
50020	57	18	Materials and Processes for Sustainable Energetics
50109	56	99	Operation and Management of Marine Diesel Powerplants
3973	55	42	Civil Engineering
50056	55	55	Marine Engineering
718	54	31	Applied Economics
50102	53	38	Real Estate Maintenance
1193	53	31	Applied Physics
50065	52	16	Fuel Chemistry and Technology
3958	52	45	Management of Environment
50147	51	9	IT Systems Development
1222	50	22	Public Administration and Innovation
726	50	18	Energy Technology and Thermal Engineering
2144	50	43	Business and Experience Management
1200	47	59	Office Administration
3963	45	37	Industrial Ecology
50058	44	30	Design and Engineering

3942	44	58	Landscape Architecture
1117	44	27	Telecommunication
1469	40	44	Tourism and Catering Management
802	38	35	Geotechnology
50064	37	38	Earth Sciences
50004	37	22	Technology Governance and Digital Transformation
50081	36	21	Urban Planning and Building Design
50084	35	16	Biomedical Engineering and Medical Physics
50127	35	53	Building and Infrastructure Engineering
1113	35	46	Telecommunication
50071	33	12	Distributed Energy
1086	32	44	Electronics and Bionics
50113	31	22	Maritime Studies
4044	31	28	Road Engineering and Geodesy
4036	30	23	Civil and Building Engineering
50097	28	71	Architecture
1233	28	25	Geotechnology
991	28	22	Technology of Materials
50153	25	7	Informatics
50088	25	29	Cyber-Physical Systems Engineering
50129	25	37	Fisheries Technologies Management and Administration
1263	25	56	Engineering Physics
50112	25	78	Waterway Safety Management
5030	24	25	Earth Sciences
50070	24	19	Work and Organizational Psychology
1789	23	22	Public Management
50111	22	85	Refrigerating Technology
50075	22	22	Finance and Economic Analysis
3945	21	39	Landscape Architecture
50076	21	27	Law and Technology
50122	20	39	Earth Sciences and Geotechnology
1090	19	14	Electronics and Bionics
50128	19	24	Maritime Studies
50093	18	29	Building Engineering
10000003	17	12	IT Systems Development
50149	16	6	Business Information Technology
50145	15	7	Product Development and Robotics
50143	14	7	Electrical Power Engineering and Mechatronics
1369	13	9	Chemical and Materials Technology
50139	13	9	Mechanical Engineering and Energy Technology Processes Control
1326	12	10	Economics and Business Administration
1039	11	16	Biomedical Engineering and Medical Physics

50125	11	19	Earth Sciences and Geotechnology
50141	9	8	Ship Engineering
50151	9	6	Telematics and Smart Systems
50148	8	10	Computer and Systems Engineering
4058	8	18	Public Administration
50138	8	6	Environmental, Energy and Chemical Technology
1475	8	41	Real Estate Administration
50073	8	8	Vocational Teacher
50055	8	7	Maritime Studies
50150	7	7	IT Systems Administration
1310	7	5	Mechanical Engineering
1331	6	8	Civil and Environmental Engineering
50117	5	5	European Architecture
1350	5	7	Chemistry and Gene Technology
50146	5	7	Chemical Technology
50132	4	7	Applied Chemistry, Food and Gene Technology
4099	4	4	Engineering Physics
1312	3	5	Power Engineering and Geotechnology
50137	3	6	Landscape Architecture and Environmental Management
1161	2	10	Electronic Systems
50152	2	7	Cyber Security Engineering
50114	2	18	Fishing and Fish Processing Technology
50133	2	8	Materials Technology
10000002	1	5	IT Systems Administration
50136	1	8	Georesources

Appendix 6 – Most Informative Courses Selected by Fisher Score

In the Table 11 courses selected by the Fisher score as the most informative ones are selected. It must be taken into account that when the course seems to be not correct or in other ways does not make sense, then this is determined by the fact that there were very few students studying at the curriculum.

Table 11. Most informative courses selected by Fisher Score

Fisher score	Course id	Curriculum	Course name	Curriculum name
3.23	BCU4020	1449	Auditing	Accounting and Business Management
0	ICM0007	50154	Analysis and Design of Information Systems	Analysis and Design of Informations Systems
0.349	YKI0050	959	Inorganic Chemistry II	Applied Chemistry and Biotechnology
14.544	YKB1130	948	Biotechnology I	Applied Chemistry and Biotechnology
0	KTO0330	50132	Principles of Food Technology	Applied Chemistry, Food and Gene Technology
7.278	TTR0110	709	Foundations of Social Cost-Benefit Analysis	Applied Economics
3.252	TET1060	718	Intermediate Macroeconomics	Applied Economics
2.433	RAR0380	1599	Database Systems	Applied Information Technology
4.384	EMR0030	1193	Continuum Mechanics	Applied Physics
6.35	EEA7170	50097	Vertical Planning	Architecture
1.708	DBR0130	50084	Measurements in Physiology	Biomedical Engineering and Medical Physics
2.703	DBR0130	1039	Measurements in Physiology	Biomedical Engineering and Medical Physics
0	NTS0122	50093	Informatics II	Building Engineering
0.882	ETT0042	50127	Road Construction Materials II	Building and Infrastructure Engineering
1.092	TXX1110	974	Research Paper	Business Administration
0.803	TMO1130	860	Human Resource Management	Business Information Technology
0	HHP0020	50149	Self-management	Business Information Technology

7.198	IDU0230	652	Databases II	Business Information Technology
50.773	SKK0890	2144	Research Paper	Business and Experience Management
0	RAE0810	50146	Technical physics	Chemical Technology
9.374	KAT0123	573	Chemical Engineering Design Project	Chemical and Environmental Technology
1.425	KAK0046	520	Advanced Processes of Environmental Technology	Chemical and Environmental Technology
2.799	KXX9110	1369	Doctoral Course in Chemical and Materials Technology 2	Chemical and Materials Technology
1.333	YKO9030	1350	Advance Course in Spectroscopy	Chemistry and Gene Technology
73.929	RAR2172	3973	Reinforced Concrete Structures - Project	Civil Engineering
1.074	EMD0020	4036	Theory of Elasticity	Civil and Building Engineering
49.824	HHF9030	1331	Philosophy of Science	Civil and Environmental Engineering
0.359	IED3040	50062	Chip Design	Communicative Electronics
0.501	IDX5711	1013	Intelligent Systems	Computer Science
0.832	TMJ1030	1055	Entrepreneurship and Business Planning	Computer and Systems Engineering
0	HHP0020	50148	Self-management	Computer and Systems Engineering
2.858	ISS0021	1045	Automatic Control Systems	Computer and Systems Engineering
0	ICA0013	50152	Fundamentals of networking	Cyber Security Engineering
0	NTS1730	50088	Matemathical Analyses	Cyber-Physical Systems Engineering
0.303	ITX8043	50026	Foundations and Management of Cyber Security	Cybersecurity
0.334	MER0170	50058	Simulations of Products and Processes	Design and Engineering
1.37	AES0360	50071	Feasibility Studies of Distributed Power Generation - project	Distributed Energy
0.754	IDU0330	50096	Business Process Modeling and Automation	E-Governance Technologies and Services
1.244	NGG8026	5030	Earth System	Earth Sciences

0	NSO0110	50125	Baltic Sea	Earth Sciences and Geotechnology
3.525	TMJ3331	50122	Entrepreneurship and Business Planning	Earth Sciences and Geotechnology
inf	TMK9130	1326	Innovation Theory and Management	Economics and Business Administration
0.527	TMJ3330	786	Business Administration	Electrical Power Engineering
0	AES0200	50143	Strategic Development of a Power System	Electrical Power Engineering and Mechatronics
0	HHP0020	1161	Self-management	Electronic Systems
0.905	IEM0250	1090	Sensor Signal Processing	Electronics and Bionics
8.64	IED0170	1086	Electronic Circuits II	Electronics and Bionics
19.75	IED3050	50103	SPICE Simulations	Electronics and Telecommunications
1.086	TMJ3330	884	Business Administration	Energy Conversion and Control Systems
1.06	EKK0270	50059	Design Tools for Buildings Energy Balance Analyses	Energy Efficiency of Buildings
2.077	AEK8141	726	Transmission of Electrical Energy and Equipment	Energy Technology and Thermal Engineering
1.152	YFR0310	1263	Simulation of Physical Processes	Engineering Physics
inf	YMR9140	4099	Mathematics for Doctoral Students II	Engineering Physics
1.54	EKE8230	1729	Eco-Design	Environmental Engineering and Management
0	HLE0060	50138	Oral and Written Communication in Estonian	Environmental, Energy and Chemical Technology
0	EAA0010	50117	Urban Strategies [Workshop Tallinn	European Architecture
0.393	TMO1041	789	Strategic Management	Finance and Accounting
1.741	TET2409	50075	Global Economic Trends	Finance and Economic Analysis
0	KTT0100	50129	Principles of Food Processing	Fisheries Technologies Management and Administration
0	HSK3032	50114	Physical Education	Fishing and Fish Processing Technology
9.524	KAT3182	810	Design Projects in Fundamentals of Chemical Engineering	Food Engineering and Product Development

0.619	KTT0180	824	Food Innovation Management	Food technology and Development
1.224	RAS0140	50065	Instrumental Analysis - special course	Fuel Chemistry and Technology
69.38	RAR0670	3590	Engine Fuel and Lubricating Oils	Fuel Technology
0.925	YKA3411	488	Instrumental Analysis	Gene Technology
7.795	YTM0033	508	Developmental Biology	Gene Technology
0	ETG5101	50136	Geodetic Surveying I	Georesources
7.56	AKM0264	802	Mineral Economics and Course Paper	Geotechnology
1.9	ETG3733	1233	Geodesy of Mines - Measuring Practice	Geotechnology
0.606	DMK1052	50038	Research design and master's seminar II	Health Care Technology
0	IAX0010	50150	Discrete Mathematics	IT Systems Administration
0	I232	10000002	Network Administration, Routers	IT Systems Administration
0	II249	10000003	Basics of Telecommunication	IT Systems Development
0	I233	50147	Operating System Administration	IT Systems Development
1.766	NTS0310	620	Heating and Ventilation	Indoor Climate in Buildings and Water Engineering
12.816	RAR0971	5373	Database I	Industrial Automation
1.808	NTK0430	3963	Industrial ecosystems	Industrial Ecology
0.678	TMM8790	4051	Industrial Marketing	Industrial Engineering and Management
3.389	ITI0021	1008	Logic Programming	Informatics
0	ITI0101	50153	Introduction to Information Technology	Informatics
0.205	IED9050	1336	Electronics Design	Information and Communication Technology
27.307	AAV0060	50108	Electrical Drives and Power Electronics	Integrated Engineering
5.252	BCU3170	1429	Midterm paper	International Economics and Business Administration
6.489	TSX0003	50035	Research Paper in the Core Study	International Relations
0.909	TSP1060	50036	Theories of International Relations	International Relations and European-Asian Studies
0.354	BCU4470	3945	Landscape Project III	Landscape Architecture
2.23	BCU4310	3942	Small Design Elements	Landscape Architecture

0	BCU5200	50137	Fundamentals of landscape architecture	Landscape Architecture and Environmental Management
0.127	HOL7004	50032	Case Studies in Internal Market Law of EU	Law
2.984	HOL6011	50031	EU Competition Law and Policy	Law
0.557	HOL7004	50076	Case Studies in Internal Market Law of EU	Law and Technology
0.119	TMO1071	556	Organization Development and Change Management	Logistics
11.517	MEA5530	529	Information System of Transportation	Logistics
11.354	RAR3470	1589	Basics of Production Engineering	Machine-building Engineering
0.693	TMK1160	923	Methods of Analysis in Business Research	Management and Marketing
9.728	NTS0051	3958	Waste Management I	Management of Environment
105.81	SKK0520	50056	Quality and Productivity Management	Marine Engineering
13.511	NSO8042	50055	Marine Environment Protection	Maritime Studies
0.389	NSO8042	50113	Marine Environment Protection	Maritime Studies
115.632	VME0070	50128	Merchant shipping law	Maritime Studies
0	EAI0050	50133	Engineering Graphics	Materials Technology
1.605	TMM1820	50020	Science Communication	Materials and Processes for Sustainable Energetics
inf	MTX9110	1310	Special Chapters of Materials Science	Mechanical Engineering
0	RAE0960	50139	Descriptive Geometry	Mechanical Engineering and Energy Technology Processes Control
7.89	MHK0011	831	Microcontrollers and Practical Robotics	Mechatronics
0.902	MHT0014	869	Measurements in Mechanical Engineering and Mechatronics	Mechatronics
4.838	VLL0660	50115	Marine Radio Communication Practice (GOC)	Navigation
5.889	HHR0100	1200	Regional Policy and Local Self-Government	Office Administration

149.732	VLL0550	50109	Maritime English (STCW)	Operation and Management of Marine Diesel Powerplants
0.688	HPP8208	50052	Personnel Management Ethics	Personnel and Development
5.505	VMS0250	50116	Hoisting and Conveying Equipments	Port and Shipping Management
18.292	RAR2440	3982	Heat Generators	Power Engineering
inf	AAR9110	1312	Individual Learning in Industry Automation	Power Engineering and Geotechnology
inf	HLI2040	732	English	Product Development and Production Engineering
6.433	MET0030	729	Machine Tools and Manufacturing Systems	Product Development and Production Engineering
0	EAI0050	50145	Engineering Graphics	Product Development and Robotics
0.778	HHI9011	4058	Special Course in Technology Governance	Public Administration
8.792	HHX0050	929	Bachelor's Exam in Public Administration	Public Administration and Governance
2.046	HHA1510	1222	Small States	Public Administration and Innovation
2.079	HHM1170	1789	Organization Theory	Public Management
9.0072E+15	BCU4590	1475	Urban Economics	Real Estate Administration
5.911	BCU3680	50102	General Course of Building Structures	Real Estate Maintenance
inf	VAY0370	50111	Basics of Accounting	Refrigerating Technology
inf	ETT0102	496	Design of Intersections and Interchanges - Project	Road Engineering and Geodesy
7.892	ETT0011	4044	Organization of Road Construction	Road Engineering and Geodesy
0	VAY0800	50141	IT Foundations	Ship Engineering
0.126	IDY0204	50027	Software Quality and Standards	Software Engineering
inf	HLS0010	406	German	Structural Engineering and Construction Management
0.751	HHM1030	50004	Policy Skills: Strategic Management, Policy Analysis and Lesson-Drawing	Technology Governance and Digital Transformation
0.872	MER0100	991	Research & Development and Innovation	Technology of Materials

9.83	KMP0220	565	Principles of Polymer Science	Technology of Wood and Textile
3.776	TMJ3380	50085	Entrepreneurship and Small Business Management	Technology of Wood, Plastic and Textiles
0.756	IRM0100	1117	Modelling of Telecommunication Equipment	Telecommunication
0	RAH1371	50151	Estonian I	Telematics and Smart Systems
7.125	MSE0130	658	Design of Power Plant Equipment - Project	Thermal Power Engineering
12.321	SKK0350	1469	Basic Marketing	Tourism and Catering Management
0	MER0150	50081	Computer Aided Design	Urban Planning and Building Design
0	HPI0230	50073	Human Communication and Academic Writing	Vocational Teacher
inf	VMV0080	50112	Geographic Information Systems	Waterway Safety Management
2.219	HPP8320	50070	Ethics and Values in Work and Organizational Psychology	Work and Organizational Psychology