Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Computer Science

Master's Thesis

Prediction of unsuccessful completion of compulsory courses and evaluation of financial demands of teaching

Bc. Adam Johanides

Supervisor: Ing. Lukáš Zoubek

Study Program: Open Informatics

Study Branch: Software Engineering

Jan 05, 2021

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Johanides** Jméno: **Adam** Osobní číslo: **434775**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra počítačů**

Studijní program: **Otevřená informatika**

Specializace: **Softwarové inženýrství**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Predikce neúspěšného dokončení povinných předmětů a vyhodnocení finanční náročnosti výuky

Název diplomové práce anglicky:

Prediction of unsuccessful completion of compulsory courses and evaluation of financial demands of teaching

Pokyny pro vypracování:

Vytvořte nástroj, umožňující stažení dat, jejich zpracování a zobrazení.
Na základě historických dat pro vybrané studijní programy FEL ČVUT:
1) proveďte vyhodnocení minimální a reálné finanční náročnosti výuky dle metodiky Kometa pro rozdělování financí za výuku na FEL ČVUT
2) predikujte úspěšné dokončení studia a povinných předmětů, zohledňující zejména střední školu studenta, způsob přijetí, studijní program a jeho dosavadní studijní výsledky. Prezentujte zjištěné ukazatele neúspěšného dokončení studia.
3) predikujte počet studentů v jednotlivých povinných předmětech.
Implementaci otestujte pomocí integračních a jednotkových testů.

Seznam doporučené literatury:

1. Hair, J. F., et al.: Multivariate Data Analysis: A Global Perspective. 7th ed., Prentice Hall, 2009.
2. James, G. et al.: An Introduction to Statistical Learning with Applications in R., Springer, 2013.

Jméno a pracoviště vedoucí(ho) diplomové práce:

Ing. Lukáš Zoubek, Centrum znalostního managementu FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **14.02.2020** Termín odevzdání diplomové práce: **14.08.2020**

Platnost zadání diplomové práce: **30.09.2021**

Ing. Lukáš Zoubek podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	prof. Mgr. Petr Páta, Ph.D. podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

. Datum převzetí zadání

Podpis studenta

Aknowledgements

I would like to thank my supervisor Ing. Lukáš Zoubek for the valuable comments and remarks he has given me during the creation of this work.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

In Prague on Jan 05, 2021 .....

Abstract

Every year, the Faculty of Electrical Engineering (FEL) at CTU in Prague, distributes funds among its departments according to their workload. These funds have been distributed for several years using a methodology called KOMETA2 to determine the quantitative share of the departments and the equipment requirements. This work examines the influence of the methodology on the financial demand of individual full-time bachelor's programmes. Using linear programming, we determine the minimal possible financial demand of the programmes and compare the results with the real costs of students' studies. Within the created tool, we use logistic regression to predict the unsuccessful completion of study plans and their compulsory subjects, based on students' previous results.

Key Words: CTU in Prague, Faculty of Electrical Engineering, prediction, logistic regression, linear programming, financial demand of studies, minimal financial demand of studies

Abstrakt

Fakulta elektrotechnická, patřící pod ČVUT v Praze, každoročně rozděluje finance mezi své katedry podle jejich vytíženosti. Již několik let využívá pro stanovení kvantitativního podílu kateder a materiálové náročnosti metodiku jménem KOMETA2. Tato práce zkoumá vliv této metodiky na cenu jednotlivých prezenčních bakalářských programů. V práci pomocí lineárního programování stanovujeme minimální teoretickou cenu vystudování programů a porovnáváme je s reálnými náklady studentů za studium. V rámci vytvářeného nástroje dále také pomocí logistické regrese predikujeme neúspěšné dokončení studijních plánů a jednotlivých povinných předmětů studenta na základě jejich dosavadních výsledků.

Klíčová slova: ČVUT v Praze, fakulta elektrotechnická, predikce, logistická regrese, lineární programování, cenová náročnost studia, minimální cenová náročnost studia

Překlad názvu: Predikce neúspěšného dokončení povinných předmětů a vyhodnocení finanční náročnosti výuky

Contents

1 Introduction **1**
1.1 Motivation . . . 1
1.2 Thesis Structure . . . 2

2 Related Work **3**
2.1 Minimal Financial Demand . . . 3
2.2 Study and Course Completion . . . 4

3 Problem Formulation and Goal Specification **7**
3.1 Financial Demand of a Study . . . 7
3.1.1 Course and Programme Demands . . . 7
3.1.1.1 Dividing of Wage Funds . . . 7
3.1.1.2 Dividing of Non-investment Funds . . . 8
3.1.1.3 Dividing of Investment Funds . . . 8
3.1.1.4 Determination of Accredited Hours of a Course . . . 8
3.1.1.5 Accredited Hours for Lectures . . . 8
3.1.1.6 Accredited Hours for Laboratories . . . 9
3.1.1.7 Accredited Hours for Exams . . . 9
3.1.1.8 Accredited Hours for Projects and Theses . . . 10
3.1.1.9 Equipment Demand of a Course . . . 10
3.1.1.10 Complete Course Financial Demand . . . 11
3.1.1.11 Course Financial Demands per Student . . . 11
3.1.2 Minimal Demands of Study Plans . . . 11
3.1.2.1 Equivalent courses . . . 12
3.1.3 Course Relations . . . 13
3.1.4 Real Study Demands . . . 14
3.2 Prediction of Course and Study Completion . . . 15
3.3 Course Occupancy Estimation . . . 15
3.4 Goal Specification . . . 16

4 Data Understanding **17**
4.1 Data for Computation of Financial Demands . . . 17
4.2 Data for Predictions . . . 18
4.3 Data Visualisation . . . 19
4.4 Summary . . . 23

5 The Proposed Approach **25**
5.1 Teaching Demands . . . 25
5.1.1 Course relations . . . 25
5.1.2 An ILP Formulation for the MFD . . . 26
5.2 Course Completion Prediction . . . 26
5.3 Study Completion Prediction . . . 30
5.4 Course Occupancy Estimation . . . 32
5.5 Summary . . . 32

6 Results **35**
6.1 Teaching Demands . . . 35
6.1.1 Course Financial Demands . . . 35
6.1.1.1 Course Demand Comparison . . . 35
6.1.2 Minimal Financial Demands of Study Plans . . . 37
6.1.3 Real Financial Demands of Study Plans . . . 38
6.1.3.1 Accredited Hours of SIT . . . 42
6.1.3.2 Accredited Hours of OI . . . 43
6.1.3.3 Accredited Hours of KYR . . . 44
6.1.3.4 Accredited Hours of EECS . . . 45
6.1.3.5 Accredited Hours of EEM . . . 46
6.1.3.6 Accredited Hours of EK . . . 47
6.1.3.7 Accredited Hours of OES . . . 48
6.1.3.8 Accredited Hours of Bioinformatics . . . 49
6.2 Study Completion Prediction . . . 50
6.3 Course Completion Prediction . . . 53

7 Application **55**
7.1 Unit and Integration Testing . . . 57

8 Conclusion **59**
8.1 Future Work . . . 60

Bibliography **61**

A List of Abbreviations **63**

B User Guide **65**

C CD Content **67**

List of Figures

4.1 The count plot comparing the number of completed courses to the uncompleted from the studies of the selected study plans. . . . 19
4.2 The stack bar chart comparing the ratio of successfully finished courses between elective and compulsory subjects. . . . 20
4.3 The stack bar chart comparing the ratio of successfully finished courses between type of course endings. . . . 20
4.4 The stack bar chart comparing the ratio of successfully finished courses between students of different study plans. . . . 21
4.6 A comparison of compulsory course completions depending on their recommended semester. . . . 22
4.5 The stack bar chart comparing the ratio of successfully finished courses between students of different study plans. . . . 22
4.7 A comparison of successfully finished courses between students of different citizenship. . . . 23

5.1 Statistics from the prediction model for B0B01MA2. . . . 28
5.2 Precision, recall and accuracy of the B0B01MA2 classifier. . . . 29
5.3 An ROC curve of the B0B01MA2 classifier. . . . 29
5.4 K-fold Cross-Validation results of the B0B01MA2 classifier. $k = 5$ 30
5.5 Logit Regression Results for the classifier of Open Informatics' study plans. . . . 31
5.6 K-Fold test results for Open Informatics' classifier ($k = 5$). The test data set consists of students' histories only from their first semester. . . . 31
5.7 K-Fold test results for Open Informatics' classifier ($k = 5$). The test data set consists of students' histories from all semesters. . . . 32

6.1 The histogram compares accredited hours between failed and successfully finished studies since 2016. . . . 41
6.2 The histogram comparing financial demands of finished studies which started in 2016 and later. . . . 42
6.3 The histogram comparing accredited hours of students of Software Engineering and Technology. . . . 43
6.4 The histogram comparing accredited hours of students of Open Informatics. . . . 44
6.5 The histogram comparing accredited hours of students of Cybernetics and Robotics. . . . 45

6.6 The histogram comparing accredited hours of students of Electrical Engineering and Computer Science. 46
6.7 The histogram comparing accredited hours of students of Electrical Engineering, Power Engineering and Management. 47
6.8 The histogram comparing accredited hours of students of Electronics and Communications. 48
6.9 The histogram comparing accredited hours of students of Open Electronic Systems. 49
6.10 The histogram comparing accredited hours of students of Medical Electronics and Bioinformatics. 50
6.11 The average results of the course completion's classifiers. 53
6.12 The average results of the course completion's classifiers with scaled variables. 53

7.1 Layer diagram of the proposed application. 56
7.2 Sequence diagram of the classifications. 57

List of Tables

3.1 K_{zk} value options . . . 10
3.2 ZH_{ost} per one student . . . 10
3.3 Course relations [4] . . . 14

4.1 Courses with unknown financial demands and their manually added equivalent courses. . . . 18

6.1 Courses with the minimal average accredited hours per student per one credit. 36
6.2 Courses with the maximal average accredited hours per student per one credit. 37
6.3 Theoretical minimal demands per student of the selected study plans. . . . 38
6.4 Accredited hours of finished studies. . . . 39
6.5 Accredited hours of failed studies. . . . 40
6.6 Kos ids of Open Informatics study plans covered in histogram 6.4. . . . 43
6.7 Kos ids of Electrical Engineering, Power Engineering and Management study plans covered in histogram 6.7. . . . 46
6.8 Kos ids of Electronics and Communications study plans covered in histogram 6.4. . . . 47
6.9 Kos ids of Open Electronic Systems study plans covered in histogram 6.4. . . . 48
6.10 Results of classifiers on unsuccessful studies. . . . 51
6.11 Results of classifiers on successful studies. . . . 51
6.12 Results of classifiers on unsuccessful studies based on results from the first semester. . . . 52
6.13 Results of classifiers on successful studies based on results from the first semester. 52
6.14 Courses which grade is a good predictor feature of study completion. . . . 53

Chapter 1

Introduction

The Faculty of Electrical Engineering (FEE) is a faculty one of the largest universities in the Czech Republic, Czech Technical University (CTU) in Prague. Every year, CTU distributes funds to individual faculties. These funds are then further distributed among the departments according to their own methodology. The FEE strives for a fair distribution of money according to individual departments' workload, whether it is scientific work or teaching students. Since 2012, the FEE has been using a methodology called KOMETA2 [3] to distribute funds for teaching students. This methodology takes as input parameters the attributes of individual subjects, their occupancy, the number of study groups and much more. Based on these data, the number of accredited hours is given to individual departments responsible for the subject.

1.1 Motivation

Kometa2 methodology contains many fixed coefficients, which try to define the complexity of teaching the subject. However, these global variables' effects are not very well mapped with respect to the final distribution of all funds. Intuitively, we feel that study programs with fewer students are more expensive to study than those with a larger number of students. However, the real financial demands were not yet known. Thus, this work tries to get a better view of the influence of methodology on the financial demand of studying a particular program.

Part of the work is the creation of a tool that would allow a summary of the financial demands to be easily determined and displayed. As an extension, the tool tries to predict the success of individual students both in their individual compulsory subjects and in the entire study, based on their previous results. With students' predictions and known costs of unsuccessful and successful studies, it is possible to estimate the financial demands of departments in advance. Simultaneously, these predictions can work as an early-alert for students and possibly their teachers to pay more attention to a subject.

Sometimes it is also challenging to estimate interest in a subject. In the case of a poor estimate of students before the beginning of the semester, the department may lack tutors. Since the application will already have all the necessary data from the solutions of the

previous two points, we try to help determine the number of students in the subjects in the following semesters.

1.2 Thesis Structure

The thesis consists of several quite independent parts. However, the structure of the work is done in such a way that the individual parts are described together and is divided as follows:

- **Related Work** is proposed for the evaluation of the minimum financial demands and study and course completion predictions.
- **Problem Formulation and Goal Specification** provides a more detailed description of the individual problems and goals of the thesis.
- **Data Understanding** chapter presents the used data and provides simple observations.
- **The Proposed Approach** tackles the selected approach for the specified goals.
- **Results** presents the primary outcomes.
- **Conclusion** discusses the fulfilment of the individual goals and the future work.
- **Application** chapter briefly describes the proposed application and its testing, including unit and integration tests.

At the end of the thesis is located list of abbreviations and guideline for the application. Note the thesis works with students' personal data such as grades from individual courses. For this reason, the data required for the correct function of the application is not included with the work. Nor can we attach the results of individual students' predictions.

Chapter 2

Related Work

This chapter briefly introduces related work to the discussed problems. Closely related problems to the searching of minimal financial demands are introduced in the following section 2.1. The later section 2.2 focuses on related work of a successful study and course completion.

2.1 Minimal Financial Demand

Given a set of courses, their credits and their demands, we search for a subset with a minimal sum of demands, which covers required credits for a study plan graduation. This problem refers to one of the most famous problems of combinatorial optimisation - *Knapsack Problem*.

F. Furini et al. [12] describe Minimum-Cost Maximal Knapsack Packing (MCMKP) and Maximum-Profit Minimal Knapsack Cover (MPMKC) problems. They have presented dynamic programming algorithm with pseudo-polynomial time complexity. They have demonstrated equivalency between those two problems and present results of benchmarks which have shown that their DP algorithm outperforms other already known mixed-integer-programming algorithms from the literature. In the discussed MPMKC problem the profit is maximised with the minimal knapsack cover while in MCMKP we minimise the cost and the knapsack packing is maximised.

The MCMKP problem have several well known scheduling variations in the literature depending on the coefficients of the objective function. E.M. Arkin et al.[6] as the first introduce the **Lazy Bureaucrat scheduling problem**. In this problem, we suppose a worker can go home when he can not finish another task before the end of his working hours. The worker is trying to schedule his tasks in such a way that he can go home early. F. Furini et al. [11] defines the problem with a common deadline as follows:

Definition 2.1.1. Let $I = \{1, ..., n\}$ be set of jobs, such that each job $i \in I$ is assigned a duration $w_i \in \mathbb{N}$ and a profit $p_i \in \mathbb{N}$. All jobs arrive at the same time, and all have a common deadline $C \in \mathbb{N}$. The goal is to find a least profitable subset of jobs S to be executed so that the schedule cannot be improved by inserting an additional job into it. More precisely, the optimal solution $S^* \subseteq I$ solves the following problem:

$$S^* = \arg\min_{S \subseteq I} \Big\{ \sum_{i \in S} p_i \mid \sum_{i \in S} w_i \leq C \wedge \sum_{i \in S} w_i + w_j > C, \forall j \notin S \Big\} \tag{2.1}$$

This definition uses more general **weighted-sum** objective with arbitrary profits $p \geq 0$. As mentioned above, F. Furini et al. [12] have proposed an $\mathcal{O}(n^2C)$ dynamic programming algorithm for this general problem. Lazy Bureaucrat scheduling problem corresponds to **min-time-spend** objective where $p_i = w_i$ for all $i \in I$. L. Gourvès et al. [14] proposed two greedy approximation algorithms for this. The last commonly used objective function is called **min-number-of-jobs** where the goal is to minimise the number of scheduled tasks and where $p_i = 1$ for all $i \in I$. It has been shown [13] that these problems are weakly NP-hard.

Just like for MCMKP there is a name "lazy employee", MPMKC with $p_i = w_i$ for all $i \in I$ is called **Greedy Boss** [14]. Greedy boss uses a law that prohibits refusing a job if the employee have nothing to do at a current time even if the task is going to be finished after his working hours. In such case, boss is trying to maximise the exceed of the schedule as much as possible. We use more general formulation of MPMKC as follows [6]:

Definition 2.1.2. Let $C > 0$ be the capacity of a knapsack and $I = \{1, \ldots, n\}$ be a set of items with profits $p_i \geq 0$ and weights $w_i \geq 0$. The optimal solution $S^* \subseteq I$ that maximises the profit with the minimal knapsack coverage solves the problem:

$$S^* = \arg\max_{S \subseteq I} \Big\{ \sum_{i \in S} p_i \mid \sum_{i \in S} w_i \geq C \wedge \sum_{i \in S \setminus \{j\}} w_i < C, \forall j \in S \Big\} \tag{2.2}$$

L. Gourves et al. [14] propose greedy $\frac{1}{2}$-approximation algorithm for the greedy boss problem with time complexity $\mathcal{O}(n)$ if the jobs are already sorted in non-increasing order of durations.

2.2 Study and Course Completion

The topic of prediction of academic performance is widely researched. Studies usually try to identify significant factors of student drop out. It has been shown [22] that grade point average (GPA) from a high school is one of the best predictors of academic performance before students' first semester. Štuka et al. [23] have shown that GPA predicts overall success almost with the same accuracy as admission tests at the largest school of medicine in the Czech Republic. Other studies include a wide range of potential predictors such as academic achievements, personality factors, social and general intelligence, aptitude tests, demographic data, or delayed entry into higher education ([9], [18]). However, there is no uniform agreement on the specific factors among the studies.

Many Data-mining techniques and algorithms are being used for these predictions such as Decision Trees [9], Support Vector Machines algorithms (SVM) [7], C4.5 and naive Bayes classifier algorithms [24] or for example Classification and Regression Tree (CART) [17].

Some work focus on a specific part of academic studies. Nouri et al. [20] discuss predictors of bachelor theses completions. Nouri et al. conclude that the supervisors' abilities

and experience are significant for the successful thesis projects. They point out factors like the ratio of previously unfinished thesis projects and the time supervisors needed to complete thesis projects. Online education is getting more popular now, mostly because of the current Covid-19 pandemic situation. Moreno-Marcos et al. [19] propose state of the art on predictions in Massive Open Online Courses MOOCs and discuss future research directions. Online education brings along new data for possible predictions. Alamri et al. [5] have already shown on four MOOCs that course completers are more likely to learn linearly, whereas the students who tend to dropout are more likely to jump forward to later activity.

There exist many approaches to problems very similar to MFD. The most similar are knapsack and scheduling tasks. However, none of these formulations fits perfectly for our case. This leads us to more general ILP formulation used in further chapters. The prediction of course and study completions have been discussed many times. However, the most significant prediction factors can differ among the institutions. Prediction models depend on data availability. Thus, we provide specific prediction models for our faculty in the later chapters.

Chapter 3

Problem Formulation and Goal Specification

3.1 Financial Demand of a Study

Systematic distribution of the monetary funds between departments is essential to ensure the operation of a faculty. The Faculty of Electrical Engineering uses a methodology called KOMETA2 [3] since 2012. The minimal and real financial demands of graduation of a study programme depend on this methodology. Therefore, the first part of this chapter describes how the financial demands are distributed according to the methodology [3] and how we could compute demands per course in a study programme.

The second part of this chapter formulates the problem of finding a set of courses with the minimal financial demands needed to satisfy prerequisites for the successful completion of a student's study programme. These lists of courses will differ according to a student's branch and type of study, e.g. present or combined type of study.

The next subsection then describes the problem of looking for the *real* financial demands needed for a students' branch completions.

3.1.1 Course and Programme Demands

Kometa2 [3] divides financial funds into three different categories. At first, wage funds M_f, are distributed for pedagogical teaching. Secondly, investment funds INV_{pf}. Moreover, as the last category, there are non-investment funds NEI_{pf}, which ensure the maintenance of resources needed for education, e.g. academic software and HW in laboratories.

3.1.1.1 Dividing of Wage Funds

The wage funds are distributed according to so-called *accredited hours* ZH (Subsection 3.1.1.4).

Let K be a set of departments. Then the cost of one accredited hour C_{zh} is computed as follows:

$$C_{zh} = \frac{M_f}{\sum_{k \in K} ZH_k} \tag{3.1}$$

Then a wage fund M_k for a department k is:

$$M_k = ZH_k \cdot C_{zh} \tag{3.2}$$

3.1.1.2 Dividing of Non-investment Funds

Let D be the set of departments and $k \in D$. Let K_{mnk} be a coefficient of equipment requirements to of ensuring the pedagogical activity of the department k. Then the non-investment funds NEI_{pk} for a department k are defined by following equation:

$$NEI_{pk} = NEI_{pf} \cdot \frac{K_{mnk}}{\sum_{k \in D} K_{mnk}} \tag{3.3}$$

3.1.1.3 Dividing of Investment Funds

Investment funds are distributed among the departments the same way as non-investment funds. Let D be the set of departments and $k \in D$. Let K_{mnk} be a coefficient of equipment requirements of the department k. Then the investment funds INV_{pk} for a department k are defined by following equation:

$$INV_{pk} = INV_{pf} \cdot \frac{K_{mnk}}{\sum_{k \in D} K_{mnk}} \tag{3.4}$$

3.1.1.4 Determination of Accredited Hours of a Course

Accredited hours of a course are calculated as a sum of a pedagogical performance for lectures ZH_{pr}, laboratories ZH_{cv}, exams ZH_{zk} and for projects and theses ZH_{ost}. Accredited hours ZH_p for a course are then:

$$ZH_p = ZH_{pr} + ZH_{cv} + ZH_{zk} + ZH_{ost} \tag{3.5}$$

A course is usually in charge of more than one department. The accredited hours are then divided proportionally between the departments. For example, a teacher from a department a takes all lectures. Two other teachers, one from the department a and one from a department b, are in charge of laboratories equally. Then all ZH_{pr} goes to the department a, and ZH_{cv} are split in half.

3.1.1.5 Accredited Hours for Lectures

Let P_{par} be a number of parallels and H_{ps} be a number of lectures in a current semester of a course. Let K_p and K_j be real constants, where $K_p, K_j \in \mathbb{R}_{\geq 0}$. K_p represents coefficient for lectures and K_j coefficient for a language of the course. Accredited hours for lectures ZH_{pr} are then:

$$ZH_{pr} = P_{par} \cdot H_{ps} \cdot K_p \cdot K_j \tag{3.6}$$

3.1.1.6 Accredited Hours for Laboratories

Accredited hours for laboratories ZH_{cv} of a course are defined as:

$$ZH_{cv} = ZH_{cvu} + ZH_{cvp} \tag{3.7}$$

ZH_{cvu} represents the presence of a teacher or teachers on lectures.

Let P_{ss} be the number of study groups, H_{cs} be the number of laboratories in a semester. Let K_{cv} be a coefficient for laboratories, K_{pnp} an average coefficient for a pedagogical complexity of the course and K_j the coefficient for the language of the course.

$$ZH_{cvu} = P_{ss} \cdot H_{cs} \cdot K_{cv} \cdot K_{pnp} \cdot K_j \tag{3.8}$$

Method for calculation of K_{pnp} differs according to the type of study. K_{pn} symbolise how many teachers are necessary to be present in a laboratory in some week. This number is changing according to safety regulations.

Let W be a set of weeks in a semester and P_{ts} the number of weeks in a semester. K_{pnp} for a course with a **present type of study** is:

$$K_{pnp} = \min\left(1.6, \frac{\sum_{w \in W} K_{pn_w}}{P_{ts}}\right) \tag{3.9}$$

For **block** and **combined studies**, K_{pnp} is calculated as follows:

$$K_{pnp} = \min\left(1.6, K_{pn}\right) \tag{3.10}$$

ZH_{cvp} represents technical support of laboratories.

Let P_{tk} be the number of classes in laboratories or in IT classrooms and let P_{ss} be the number of study groups. The value of a technical support ZH_{cvp} for a course i is represented as:

$$ZH_{cvp} = 1.6 \cdot P_{tk} \cdot (1 + 0.15 \cdot (P_{ss} - 1)) \tag{3.11}$$

Accredited hours for technical support goes to a department which maintains the room. In other words, accredited hours for technical support does not have to go to a department which is responsible for the course.

3.1.1.7 Accredited Hours for Exams

Let P_{stud} be the number of students enrolled in a course and K_{zk} a coefficient of an examination of the course. Let K_j represents the coefficient for a language of the course. Accredited hours for the course's examination is then:

$$ZH_{zk} = P_{stud} \cdot K_{zk} \cdot K_j \tag{3.12}$$

The coefficient K_{zk} depends on a type of examination of a course. The table 3.1 shows possible values of K_{zk}.

Type of Examination	**Value**
Exam in all forms of study	0.8
Classified credit in all forms of study	0.2

Table 3.1: K_{zk} value options

3.1.1.8 Accredited Hours for Projects and Theses

Accredited hours for project and theses is given by the table 3.2:

Number of ZH_{ost} per 1 student	**BSP**	**MSP**
Individual project	12	14
Team project	6	7
	BT	**MT**
Supervisor	20	25
Opponent	3	4
State Final Examination	6	8

Table 3.2: ZH_{ost} per one student

BSP and **MSP** in the table 3.2 represents bachelor and master semestral project. **BT** and **MT** bachelor and master thesis.

Actual ZH_{ost} for a course is then multiplied by a language coefficient K_j of the course:

$$ZH_{ost} = P_{stud} \cdot K_{zk} \cdot K_j \tag{3.13}$$

3.1.1.9 Equipment Demand of a Course

The methodology Kometa2 [3] presents a table of values with a coefficient of equipment requirements K_{mnc} of laboratories. This coefficient depends on the type of room where a laboratory is. Further gives a coefficient K_{nr}, which represents the difference of requirements between bachelor and master studies.

Let H_{cs} be a number of laboratories in a semester and P_{stud} be a number of students enrolled in the course. Then the demand for the equipment requirements K_mnp of the course is:

$$K_{mnp} = K_{nr} \cdot H_{cs} \cdot K_{mnc} \cdot log(1 + P_{stud}) \tag{3.14}$$

3.1.1.10 Complete Course Financial Demand

Let C_{zh} be the cost per 1 accredited hour in a semester. Further, let C_{nk} be the amount from non-investment funds per K_{mnp} and C_{ik} be the amount from investment funds per K_{mnp}. Then the total course financial demand C of a course can be computed as follows:

$$C = C_{zh} \cdot ZH_p + (C_{nk} + C_{ik}) \cdot K_{mnp} \tag{3.15}$$

This thesis does not provide information about the actual amount of financial funds of the faculty. Thus, the costs of one accredited hour ZH and equipment requirements K_{mn} are not known. Therefore, all financial demands are described only just in terms of accredited hours ZH and equipment requirements K_{mn}.

3.1.1.11 Course Financial Demands per Student

Financial demands are not linearly dependent on the number of students in a course, because they rather depend on the number of student groups. However, we cannot decide which student is responsible for an increase of study groups in a course. Therefore, we omit this fact and demands are distributed between students equally.

Definition 3.1.1. Let ZH_p be accredited hours of a course, K_{mnp} demands for equipment requirements and P_{stud} be the number of students enrolled in the course in a semester. Then course demand per student in the semester is defined as:

$$ZH_s = \frac{ZH_p}{P_{stud}} \tag{3.16}$$

$$K_{mns} = \frac{K_{mnp}}{P_{stud}} \tag{3.17}$$

3.1.2 Minimal Demands of Study Plans

Nowadays, the faculty provides several study programmes. A student can study presently, combined, or with a block teaching. This fact creates a high number of possibilities, how a student can graduate a bachelor programme. Each of these possibilities has a different financial demand for the faculty. This section formulates the problem where we search a list of courses for each study plan, so its financial demand is minimised.

A student is required to obtain at least 180 credits in bachelor programme study plans. Also, there is a set of course groups for each study plan, where every course group is defined by its requirements and a set of courses. These requirements are a minimal number of finished courses and a minimal and a maximal number of obtained credits from this group. As an example, a course group can consist of a bachelor thesis only. This course group has the minimal and the maximal number of credits equal to 20 and the minimal number of finished courses equal to 1. Therefore, a student has to finish exactly one bachelor thesis during his study.

3.1.2.1 Equivalent courses

For many reasons, new entries are created in the catalogue of courses in the scholar system [1]. The courses represent an older course with a new course code and attributes. If the attributes were the same, there would be no reason to create a new entry. However, these courses are considered as identical to their older version for the purposes of the study plan conditions check. This relation is symmetrical, reflexive and transitive. Therefore, the courses are divided into disjunctive equivalence classes.

A student can not enrol two subjects from the same equivalence class in one semester. Also, if one course from an equivalence class is completed, all courses from the class are considered as finished as well. There exists other older course relations in the course system described later in section 3.1.3. The equivalency works as asymmetric version of **substitute** and **forbidden enrolment** relations together.

Definition 3.1.2. Let $P = \{p_1, ..., p_n\}$ be a course group of a study plan and $I = \{1, \dots, n\}$ the set of respective indices, such that each course $p_i \in P$ is assigned a demand $d_i > 0$ and credits $c_i \geq 0$. Let C_{min}, $C_{max} \in \mathbb{N}$ be the minimal and the maximal number of credits respectively from the course group. Let $K \in \mathbb{N}$ be the minimal number of courses taken from the course group for a study plan. Suppose all courses can be enrolled together during the study of the study plan for now. The goal is to find a subset of I minimising the demand, which is feasible with the given conditions. More precisely, the optimal solution $G^* \subseteq I$ solves the following problem:

$$G^* = \arg\min_{G \subseteq I} \Big\{ \sum_{i \in G} d_i \mid C_{min} \leq \sum_{i \in G} c_i \leq C_{max} \ \wedge \ |G| \geq K \Big\} \tag{3.18}$$

The problem is slightly modified if it is possible that some demands $d_i = 0$:

$$G^* = \arg\min_{G \subseteq I} \Big\{ \sum_{i \in G} d_i \mid C_{min} \leq \sum_{i \in G} c_i \leq C_{max} \wedge |G| \geq K \wedge \Big(|G| = K \vee \sum_{i \in G \setminus \{j\}} c_i < C_{min}, \forall j \in G \Big) \Big\} \tag{3.19}$$

The added condition minimises the coverage of the knapsack - a student is forced to take as little courses or credits as possible. However, in our problem we suppose all demands $d_i > 0$.

A study plan conditions consist of several course groups and minimal required credits. We have the definition of a problem, minimising a single course group demand. We can find a set of courses with the minimal demand for the study plan as a union of optimal subsets of each course group. This formulation would work only in such case, where all course groups are disjunctive, and the sum of their required credits is higher than the required credits for the study plan graduation. This would mean the study plan is defined only by mandatory and mandatory elective courses. However, our course groups are not disjunctive, and for example, a course group with elective courses has minimal required credits equal to zero. Moreover, there exist equivalent courses. Therefore, we use a different formulation 3.21 of the problem.

Definition 3.1.3. Let $P = \{p_1, ..., p_n\}$ be a set of all available courses and $I = \{1, \ldots, n\}$ a set of respective indices, such that each course $p_i \in P$ is assigned a demand $d_i > 0$ and credits $c_i \geq 0$. Let $G = \{G_1, \ldots, G_m\}$ be a set of all course groups, such that each group $G_j \subseteq P$. Let $E = \{E_1, \ldots, E_r\}$ be a set of all course equivalency classes. Let C^j_{min}, $C^j_{max} \in \mathbb{N}$ be the minimal and the maximal number of credits respectively from a course group G_j and $Z \in \mathbb{N}$ be the required credits for the study plan graduation. Let $K_j \in \mathbb{N}$ be the minimal number of courses taken from the course group G_j. Suppose all courses can be enrolled together during the study of the study plan. The set of all **feasible course sets** S_k with their sum of credits C_k for the study plan graduation is defined as:

$$S = \Big\{ S_k \mid S_k \subseteq P \wedge C^j_{min} \leq \sum_{i \in S_k \cap G_j} c_i \leq C^j_{max} \wedge |S_k \cap G_j| \geq K_j \wedge |S_k \cap E_t| \leq 1 \wedge C_k \geq Z \Big\} \tag{3.20}$$

Definition 3.1.4. Assuming all demands $d_i > 0$, the set of courses for a study plan graduation with the minimal financial demands is defined as:

$$S^* = \arg\min_{S_k \in S} \Big\{ \sum_{i \in S_k} d_i \Big\} \tag{3.21}$$

3.1.3 Course Relations

Another problem which needs to be tackled while minimising a set of courses is the fact that courses have relations between themselves. Nowadays, the scholar system KOS distinguish seven course relations described in the table 3.3 below.

Abbrev.	Name	Description
P	P prerequisite	Soft prerequisite. A course **B** is a type "P" prerequisite to a course **A** if the course **B** has to be enrolled before (in previous semesters) the course **A**. The classification or the credit from the course **B** is not required.
Z	Z prerequisite	A Course **B** is a type "Z" prerequisite to a course **A** if the credit from the course **B** is required before the enrolment of the course **A**.
A	A prerequisite	Hard prerequisite. A Course **B** is a type "A" prerequisite to a course **A** if the course **B** completion is required before the course **A** enrolment.
K	Co-requisite	A course **B** is co-requisite with a course **A** in case the course **B** must be undertaken before or simultaneously with the course **A**.
N	Forbidden enrolment	A course **B** is in "N"-relation to **A** in case the course **A** can not be enrolled if the course **B** has been already enrolled in a current or previous semesters. This relation is not symmetric.
Q	Previous classification	A course **B** is "Q"-related to a course **A** if the classification of the course **A** is possible only after the classification of the course **B**.
R	Substitute	Let a course **B** be a substitute of a course **A**. Then if a student successfully completes the course **B**, the course **A** is considered as successfully undertaken as well. This relation is not symmetric.

Table 3.3: Course relations [4]

Despite of the course relations, the formulation of feasible sets 3.20 still works for most of them. Since the order of course enrolments does not matter, we can omit all prerequisite relations. Co-requisite relations holds because of the course group conditions. Therefore, the only problematic relations are substitutions and enrolment prohibitions. The proposed approach is described in Section 5.1.1 more precisely.

3.1.4 Real Study Demands

Minimal demands could be significantly different from real demands spend for students' studies.

Definition 3.1.5. Let C be a list of courses which a student attended during his studies. Let ZH_s^i be accredited hours per student for a course i and K_{mns}^i be demands for equipment requirements per student for the course. The accredited hours and equipment demands of the student's study are:

$$ZH_s = \sum_{i \in C} ZH_s^i = \sum_{i \in C} \frac{ZH_p^i}{P_{stud}^i} \tag{3.22}$$

$$K_{mns} = \sum_{i \in C} K_{mns}^i = \sum_{i \in C} \frac{K_{mnp}^i}{P_{stud}^i} \tag{3.23}$$

Note that the same course could occur in the list C more then once if enrolled repeatedly. Every time with different financial demand.

Definition 3.1.6. Let $z \geq 0$, $e \geq 0$ defines the price of the accredited hour and the equipment demand unit. Assuming the price is not changing each semester, the real financial demand $D_s \in \mathbb{R}$ of a student's study s is:

$$D_s = z \cdot ZH_s + e \cdot K_{mns} \tag{3.24}$$

3.2 Prediction of Course and Study Completion

With a good prediction of a course completion, a teacher could be early notified that he has to devote more time to students to help them succeed in a given course. Furthermore, the prediction of the number of unsuccessful students in a given course helps to estimate the number of students who will enrol into the subject in advance. We are trying to predict the success rate for 223 different compulsory subjects for a total of 14 study plans. All data for the further predictions come from the dataset described before in Section 4.2. Our goal is to create a dataset for each compulsory course with a probability of successful completion for every student who will have to complete the course in the current or following semesters in order to meet the requirements of the study plan.

Our next goal is to estimate whether or not a student will complete his study. In this case, we do not consider a possibility the student will enrol into study again after a failure. In the case of study repetition, we estimate the new study's success rate and the student's history from the previous studies are included for the new prediction. Currently, we are trying to find the probability of successful study completion of 1784 bachelor studies. We also look for features which can predict a student's study success.

3.3 Course Occupancy Estimation

This section describes the occupancy estimation problem. The goal is to predict the number of students in a course in a future semester to help organisers know the number of necessary lecturers. Most subjects are not taught in both semesters. Thus, when determining the number of students, we usually already know the number of previously unsuccessful students. If we wanted to determine the number of students for the next semester or the year ahead, we would basically have two options. First, we could estimate the number according to the development of the subject's occupancy from history. However, this procedure would not reflect the current number of active students in the school. If we wanted a more accurate

estimate, we would have to include a lot of variables. In this second approach we would have to estimate the number of students who will not complete the course as compulsory in the current semester and determine the number of people who have a subject in their study plan in the next semester. From the group of people, we would also have to subtract those students who, for some reason, are going to cancel their study. Finally, we would have to estimate the interest of students who optionally enrol in the course.

The first procedure risks being utterly inaccurate due to a sudden significant change in the number of students in the individual study plans. The second approach would lead to a cumulative distribution function with a vast uncertainty band due to the number of variables. It would then be complicated to determine a suitable threshold for determining the number of students. We have therefore decided to simplify the problem while providing more relevant information without any estimation. We propose the approach in Section 5.4.

3.4 Goal Specification

To sum up the problems discussed in this chapter, we specify these goals:

- Present the minimal possible financial demands of full-time bachelor study plans for a successful graduation.
- Calculate the students' real financial demands of these study plans and compare the results with the minimal possible demands.
- Propose a classifier for unsuccessful study completion and present its accuracy.
- Propose a classifier for unsuccessful compulsory course completion and present its accuracy.
- Determine the most significant courses for successful study completion of the full-time bachelor study plans, according to the classifier.
- Propose a tool for a course occupancy estimation.

Chapter 4

Data Understanding

We focus on bachelor's full-time study programs. On the other hand, the data structure does not differ in any way for master's, combined and distance programs. All procedures from this work are also applicable to the other study plans. The only difference are the doctoral studies. These studies have individual study plans, and we can not apply the proposed approaches on them. We focus on 22 bachelor's full-time study plans created in 2016 and later and on the study plans which still continue. One of the study plans (EECS) is currently taught entirely in English.

4.1 Data for Computation of Financial Demands

All data used in this thesis come from views into the KOS database, except the Kometa application results. The Kometa results are usually adjusted manually at the end of the distribution of the application. Therefore, we use the adjusted real results loaded from the historical CSV files, although the proposed application can load the Kometa system's results using Rest API. Nowadays, we have distributions of financial demands since 2016 (winter semester B161).

Sometimes, there are changes across subjects' accreditations (e.g. number of laboratories or obtained credits have been changed). With these changes, the courses have to get a new course code. The equivalency relations [1] between the old and the new versions of the subjects have to be created to, for example, recognise previously completed student subjects. With the new course versions, the whole study plans have been changed too. For example, Open Informatics study plans have been adjusted in 2016 and later in 2018, but we do not have the financial demands of subjects thought before 2016. Therefore, we use the new course versions' demands to estimate the needs of the old ones.

The equivalency relations are sometimes missing for an unknown reason. We manually added the equivalencies into the Kometa results to estimate as many students' financial demands as possible from the discussed study plans. The following table 4.1 shows the subjects without the equivalences and the courses we marked as their replacements. We could not find their replacement for some of them, sometimes because these courses simply no longer exist. We skip the students with these subjects as well as students who studied on Erasmus or students with an individual study plan.

Missing	Replacement	Missing	Replacement	Missing	Replacement
A0B01LGR	B0B01LGR	A0B16MPS	B0B16MPS	A4B01NUM	B4B01NUM
A0B01MA1	B0B01MA1	A0B36APO	B0B35APO	A4B33SI	B4B36SIN
A0B01MVM	B0B01MVM	A1B14SEM		A4B77ASS	
A0B02DCE		A1B16RIP		A7B16ISP	B6B16ISP
A0B02PSF1		A1B38EMA	B1B38EMA	A7B16UFI	A1B16UFI
A0B02PT	B02PT	A2B31IN1		A7B36WMM	
A0B02SF	BV002SF	A2B32DAT	B2B32DATA	A7B39MGA	
A0B02TF1	B02TF1	A2B34MIK	B2B34MIK	AE0B16HT1	
A0B02VNP	BV002VNP	A2B34SEI		AE0B99PP4	
A0B02ZIP	B02ZIP	A3B31EOP		AE0B99PP6	AE0M99PP6
A0B04CA		A3B33OSD		AE4B33OSS	BE2M32OSS
A0B04R3		A3B99RO		B0B13ETM	
A0B13KEO	B0B13KEO	A4B01JAG	B4B01JAG	B1B14ZEL	B1B14ZEL1
A0B16HSD		A4B01MA2	B0B01MA2	B3B35APO	B0B35APO

Table 4.1: Courses with unknown financial demands and their manually added equivalent courses.

4.2 Data for Predictions

The proposed application exports the dataset for the predictions with adjusted data from the Kos views. The dataset provides information about students' course records and their results. In the case of study prediction, the classification goal is to predict whether a student will finish his study plan. On the other hand, in case of course prediction, the classification goal is whether the student will complete his course enrolment in the current or following semesters. Sincerely, the Kos views do not provide information about applications for studies from the previous years and about a student's high school. Therefore, we do not have reliable data for a prediction of the student's first semester. The only student's information we can use for the first semester are student's gender and citizenship. As expected and proposed later, these two factors are not very reliable for the prediction models.

Dataset variables

- Person (categorical: Student's KOS id)
- Study Plan Kos Id (categorical)
- Course (categorical: Course code)
- Grade (categorical: "A", "B", "C", "D", "E", "F", "/", "Z")
- Finished (binary: True means the student finished the course successfully)
- Credit (binary: True means the student obtained the credit from the course)
- Attempts (numeric: Number of undertaken attempts on examinations of the course)

- Semester (categorical)
- Mandatory (binary: True if the course is compulsory in the student's study plan)
- State (categorical: "Finished", "Failed", "Studying". Represents the state of the student's study)
- Citizenship (categorical: Citizenship of the student)
- Gender (categorical: "Muž" - male, "Žena" - female)
- Current Semester (binary: True if the course record is from the current semester)
- End (categorical: "ZK" - exam, "KZ" - classified credit, "Z" - credit)

4.3 Data Visualisation

There are 52276 successful and 14572 uncompleted course enrolments in our dataset 4.1. Our dataset is imbalanced since the ratio of successful to unsuccessful course completions is 78:22. Thus we will need to balance it in further steps.

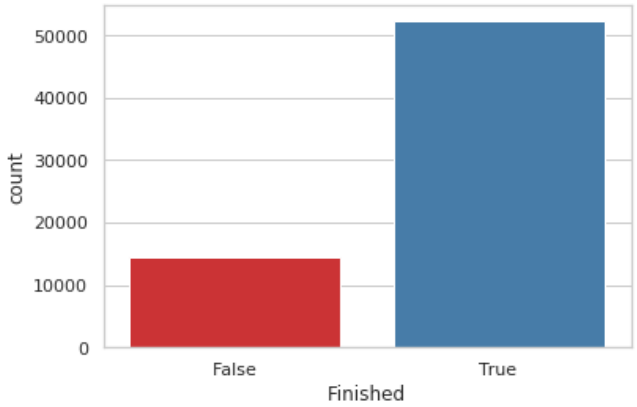


Figure 4.1: The count plot comparing the number of completed courses to the uncompleted from the studies of the selected study plans.

There are 51951 records of compulsory and 14897 of elective subject records in the dataset 4.2. It would be reasonable that the ratio of successfully finished courses would be different for these two groups. Surprisingly, the completed courses' percentage is the same for both mandatory and elective courses - 78%. Therefore, the mandatory condition cannot be a good predictor.

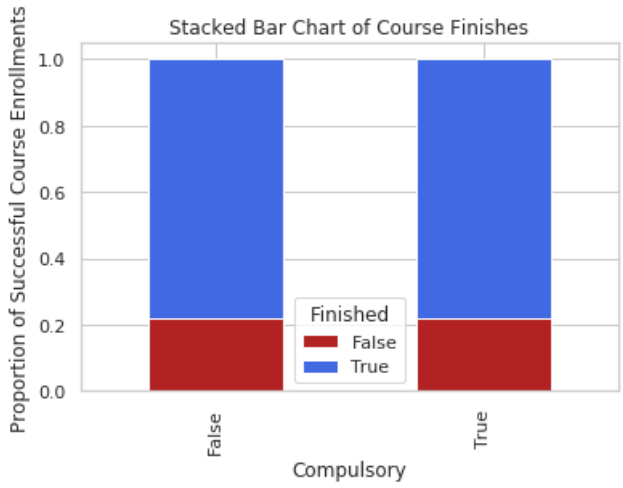


Figure 4.2: The stack bar chart comparing the ratio of successfully finished courses between elective and compulsory subjects.

The chart 4.3 below shows that the type of course ending significantly affects students' chances. The chart compares the success proportions of compulsory courses. Subjects ended by an examination have only 74% success rate. Success rates of subjects with credit and classified credit are 94% and 86%, respectively.

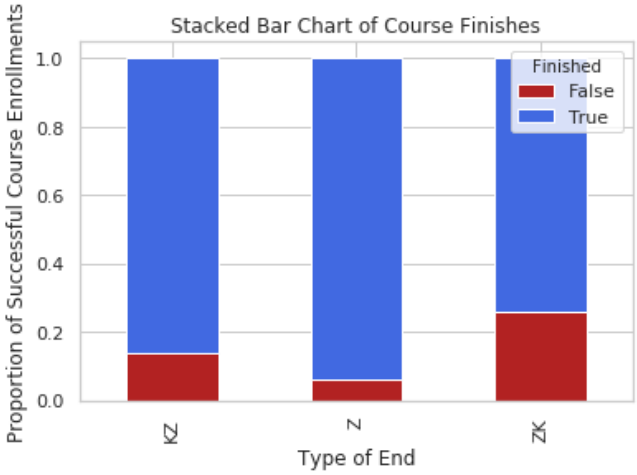


Figure 4.3: The stack bar chart comparing the ratio of successfully finished courses between type of course endings.

The following graph 4.4 compares students' success ratio from different study plans. The most successful students in course enrolments (success ratio 88%) are from OI - Artificial Intelligence and Computer Science 2018 and EEM - Applied Electrical Engineering 2016.

On the other hand, the worst course success ratios (73%) come from Electronics and Communications 2018, OES and the English study plan EECS.

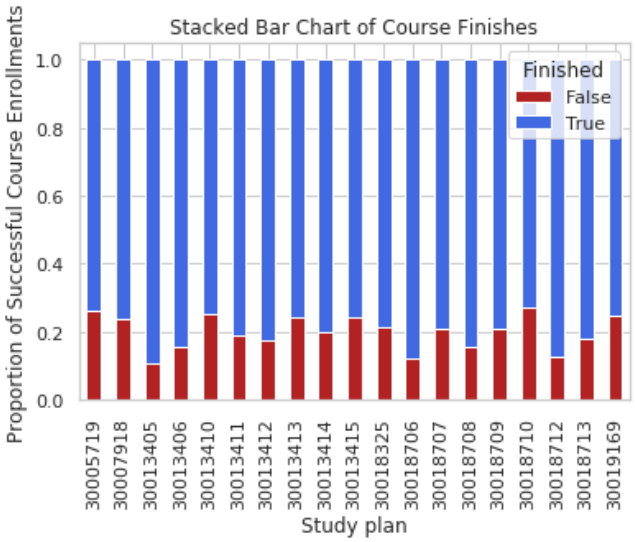
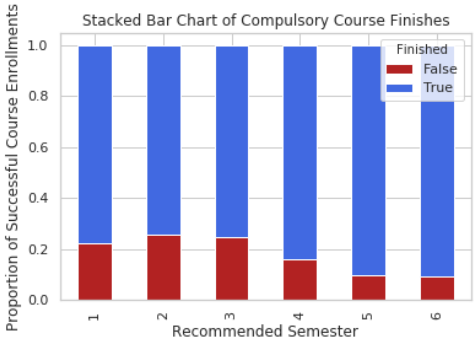
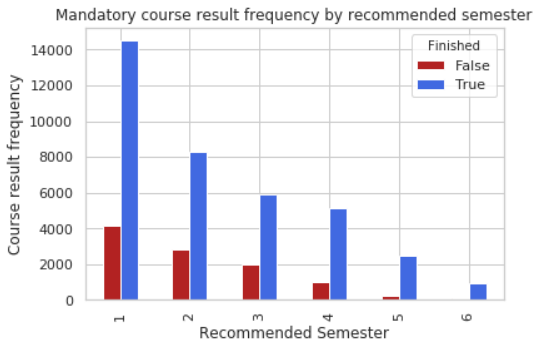


Figure 4.4: The stack bar chart comparing the ratio of successfully finished courses between students of different study plans.

The next stacked bar chart 4.5 analyse the success rate of students with at least one previous unsuccessful study. In the case of students with the first study, the percentage of completed subject enrolments is 79%. Students who enrolled in a study again have success rate only 72%. This fact might not seem as a significant difference. However, note that a student's first failed study is included between other first studies. The comparison tells us that students who failed their study before tend to fail courses again. Thus, this can be a good predictor for a course completion.



(a) A stack bar chart of compulsory course finish ratios by the recommended semesters



(b) A histogram of course result frequencies by the recommended semesters

Figure 4.6: A comparison of compulsory course completions depending on their recommended semester.

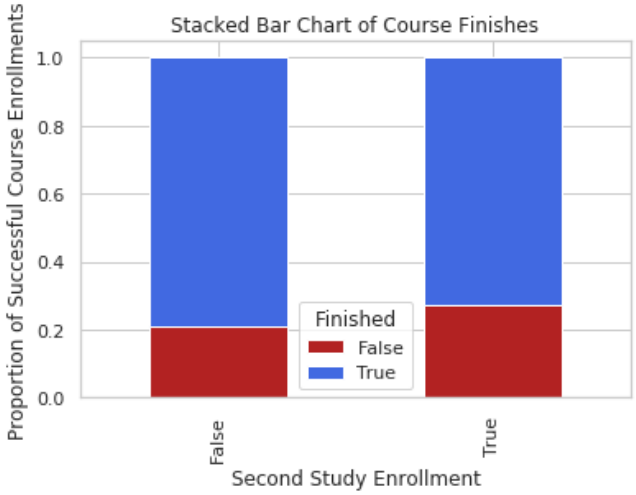


Figure 4.5: The stack bar chart comparing the ratio of successfully finished courses between students of different study plans.

All compulsory subjects have their semester recommendation in a study plan. Students are not bound by these recommendations, but most of them enrol into courses according to them. Therefore, Figure 4.6 can describe which semesters are probably usually the most difficult for students. Most course failures are from the very first semester, but the worst success ratio comes from the second and third semesters, thus these two semesters may be even more significant for the study completion.

There are no differences between the results of both genders. The citizenship remains as the last possible indicator. The countries included in Figure 4.7 represents the citizenships of an absolute majority of the current students. The percentage of the finished courses of people from Kazakhstan is 67%. On the other side are students from the Czech Republic

with a 79% course completion.

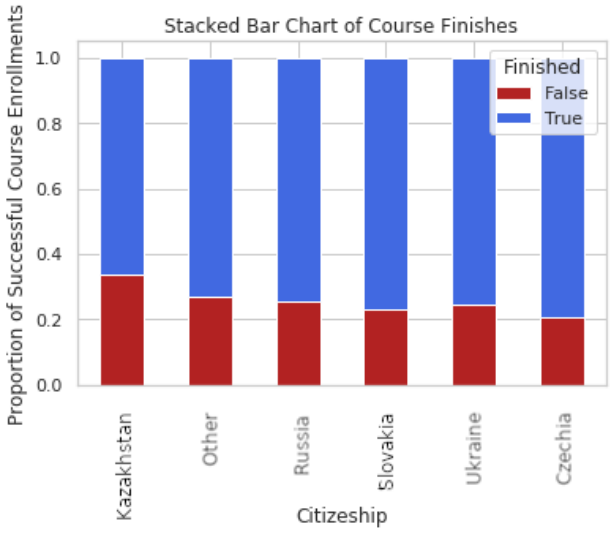


Figure 4.7: A comparison of successfully finished courses between students of different citizenship.

4.4 Summary

There are two major data issues. The first is that student applications' data are not available retrospectively, and therefore, it is not possible to make predictions for the first semester of their study. The second problem is the change in the accreditation of subjects over the years, especially in 2016. The lack of equivalences between old and new subjects makes it difficult to both, prediction of students' dropouts and calculations of the price of studies. However, we were able to add many of these equivalences for the financial demand calculation. This is because there are no longer many students in the new study plans who have enrolled in one of the subjects that have not been taught since 2016. However, there are much more courses in the case of predictions with the missing relations. Therefore, the predictions suffer from a lack of data for classifiers' training and testing. We leave it to future work to create these course equivalence classes, hence gaining more of the historical data and more accurate models.

Chapter 5

The Proposed Approach

In this chapter, we describe the approaches we used to solve the discussed problems. The first section describes how to calculate the minimum financial demand for study plans. The next two parts focus on predicting student success in both, individual subjects and an overall study. The last part presents the data collected by the application, which facilitates the estimation of subjects' occupancy in the following semesters.

5.1 Teaching Demands

5.1.1 Course relations

In section 3.1.2 we formulated the problem where the goal is to find a feasible set of courses for a study plan minimising the financial demand. We also mentioned in 3.1.3 course relations from the scholar system KOS. Course substitution and forbidden enrolment relations 3.3 are problematic for a straightforward problem solution.

The substitution relation allows that a completion of a course **A** causes a completion of a different course **B** in another course group which does not originally contain the course **A**. In such case, we need to raise obtained credits from the both course groups. We solve this problem by modifying groups. If a course **A** substitutes a course **B** we add **A** to **B's** course group. We also ensure that the forbidden enrolment relation exists between the courses in both direction.

As an example, suppose a student must complete at least one course in English from a course group with compulsory subjects of a study programme. In other words, there exists a course group with mandatory courses in English where the minimal number of undertaken courses is equal to one. Their English versions substitutes the Czech versions. Therefore, we can add all English course versions to the other group. Note the substitution is not symmetric. A completion of a compulsory course in Czech language can not raise obtained credits in the group with English subjects. Since the English and Czech versions of the subjects can not be undertaken together, this approach solves the problem with substitution relation.

It is not possible to add the forbidden enrolment relation just by course group modifications. Moreover, it is not obvious why this relation is not symmetrical in the scholar

system. If an enrolment of a course **A** prohibits an enrolment of a course **B**, the enrolment of **B** should do the same. Otherwise, a student would be able to undertake both courses enrolling the course **B** first. Also, the soft-prerequisite relation 3.3 exists because of this reason. Therefore, we take the relation as symmetrical and we suppose a student is forbidden to attend both courses together during his study in any order. We treat the course with forbidden enrolment relations as they are from a same equivalence class (Section 3.1.2.1).

The following Section 5.1.2 proposes an approach with an ILP formulation of the problem.

5.1.2 An ILP Formulation for the MFD

This section proposes an approach for the MFD of a study plan described in 3.1.2 with an ILP formulation. Let $I = \{1, \ldots, n\}$ be a set of all course indices. In the following let the binary variables p_i be set to one if the course i is selected. Let $d_i > 0$ be the financial demands of a respective course i. The ILP formulation of the MFD problem is as follows:

$$\text{minimise} \quad \sum_{i \in I} p_i \cdot (a_i + e_i) \tag{5.1}$$

$$\text{subject to} \quad \sum_{i \in E_t} p_i \leq 1 \qquad \forall k \in \{1, \ldots, r\} \tag{5.2}$$

$$\sum_{i \in I} p_i \cdot c_i \geq Z \tag{5.3}$$

$$\sum_{i \in G_j} p_i \cdot c_i \geq C^j_{min} \qquad \forall j \in \{1, \ldots, m\} \tag{5.4}$$

$$\sum_{i \in G_j} p_i \cdot c_i \leq C^j_{max} \qquad \forall j \in \{1, \ldots, m\} \tag{5.5}$$

$$\sum_{i \in G_j} p_i \geq K_j \qquad \forall j \in \{1, \ldots, m\} \tag{5.6}$$

$$p_i \in \{0, 1\} \qquad \forall i \in I \tag{5.7}$$

The objective function 5.1 minimises the financial demands of a study plan. The first condition ensures that at most one subject can be undertaken from a single equivalence class of courses. The second condition forces an ILP solver to find a solution where the sum of all credits from the result set is greater or equal to Z (180 for bachelor and 120 for master study plans). The following two conditions ensure enough credits from every course group can be obtained and the last condition that enough courses from every course group are selected.

5.2 Course Completion Prediction

This section describes the selected approach for the course completion prediction. Our problem is to decide whether a student will or will not complete a given compulsory course. We chose to use logistic regression which is often used to classify a categorical dependent

variable (for example, to predict whether an email is a spam or not). In logistic regression, we use the so-called *logistic function* to model $p(X)$ where $X = (X_1, \ldots, X_p)$ with p predictors:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \tag{5.8}$$

This function gives us outputs between 0 and 1 for all values of X [15]. The coefficients $\beta_1, \beta_2, \ldots, \beta_p$ are unknown and are chosen using maximum likelihood. In addition to the results p(X), the logistic regression gives us information about the used predictors. This information includes coefficient, z-statistic, p-value and standard error.

Our goal is to estimate the probabilities of students' successful completion of a total of 223 different compulsory courses. The chances are found for students who enrolled in a current semester and students who have to complete the course lately because of their study plan. We describe the approach on a specific subject B0B01MA2 - Mathematical Analysis 2. The same procedure is then used for predictions of all other courses.
We start by a dataset preparation. Data are exported from the proposed application, which we describe in Chapter 4.2. The data include columns with categorical variables like a grade from a subject. We transform these columns into dummy variables. For example variable $grade \in \{A, B, C, D, E, F, /, Z\}$ is transformed into 7 dummy variables $grade\_A, \ldots, grade\_Z$. A dummy variable takes on a value of 1 for students who acquired the specific grade and 0 otherwise. The sum of all grade dummy variables is therefore always equal to 1. The dummy variable approach is also applied on columns of citizenship, gender and type of examination.

After the data transformation, we create two data frames for every subject. The first subject's data frame includes course enrolment records of students who have already undertaken the course. We find all the student's enrolments preceding the enrolment of the course. As we find the subjects, we sum their grades and create one record for the data frame representing the student's history. Also, we compute the attempt variable in the way it represents the number of unsuccessful exam completions. Note that since a student can enrol in a subject twice, the student can occur more than once in the resulting data frame. However, the student's history will always differ, depending on a semester. This data frame will be used as a training data set.

The other subject's data frame is created in the same way. However, we look for enrolments in courses for the current and following semesters. This dataset represents actual histories of students we want to predict.

In the case of B0B01MA2, we have 862 of successful and 328 unsuccessful enrolments. The 72% success rate in the training data set means that we have imbalanced data for the prediction model. Chawla et al. [8] have shown that oversampling the minority class can improve classifier performance. Therefore, we use their Synthetic Minority Oversampling Technique (SMOTE) to oversample the failed enrolments' observations. This algorithm does not copy the instances of the failed enrolments. Instead, it randomly chooses one of the k-nearest-neighbours and uses it to create similar new observations.

```
         Current function value: 0.531194
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                 1198
Model:                          Logit   Df Residuals:                     1189
Method:                           MLE   Df Model:                            8
                                        Pseudo R-squ.:                  0.2336
                                        Log-Likelihood:                -636.37
converged:                       True   LL-Null:                       -830.39
                                        LLR p-value:                6.769e-79
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Repeated       1.4632      0.341      4.293      0.000       0.795       2.131
Attempts      -0.1635      0.045     -3.640      0.000      -0.252      -0.075
Grade_/       -0.3653      0.053     -6.868      0.000      -0.469      -0.261
Grade_A        0.2277      0.039      5.853      0.000       0.151       0.304
Grade_B        0.2980      0.068      4.406      0.000       0.165       0.431
Grade_C        0.2788      0.062      4.499      0.000       0.157       0.400
Grade_D        0.1120      0.051      2.176      0.030       0.011       0.213
Grade_E        0.0110      0.054      0.205      0.837      -0.094       0.116
Grade_F       -0.3557      0.096     -3.713      0.000      -0.544      -0.168
==============================================================================
```

Figure 5.1: Statistics from the prediction model for B0B01MA2.

Figure 5.1 presents data from the prediction model of B0B01MA2. We will describe the terms above using the attempts variable. The **coef** symbolise that with every failed examination attempt from his previous courses, the student will have a 0,1635 unit less chance of successfully finishing the Mathematical Analysis 2 based on the p-value in the table. Large absolute value of **z** indicates evidence against the null hypothesis $H_0 : \beta_i = 0$. In other words, the null hypothesis says the probability of student's success does not depend on the number of his unsuccessful attempts from previous courses. The larger the **z** value gets, the less uncertainty there is. If the **p-value** of a feature is tiny (usually below 0.005) we can reject the null hypothesis [15]. We can say, the unsuccessful course completion depends on the number of previously failed exams. Thus, the attempts may be a good prediction feature. However, as we can see, the chances of student's unsuccessful completion depend more on the number of previously failed courses.

The selected model's accuracy is 73% (tested on randomly selected 30% of the data and trained on the rest with oversampling). However, the accuracy is not a good metric for skewed datasets. In this example, the number of successful enrolments is 72%. We would get 72% by simply labelling all students as successful, disregarding the false positive predictions. Therefore, the classification performance of algorithms in information retrieval is usually measured by precision and recall [8]:

$$precision = \frac{TP}{TP + FP} \tag{5.9}$$

$$recall = \frac{TP}{TP + FN} \tag{5.10}$$

In our case, the precision symbolises the classifier's ability not to mislabel a student as successful. On the other hand, the recall represents the ability to identify all successful students of the course.

	precision	recall	f1-score	support
0	0.53	0.62	0.58	104
1	0.83	0.77	0.80	253
accuracy			0.73	357

Figure 5.2: Precision, recall and accuracy of the B0B01MA2 classifier.

Figure 1 proposes how well the classifier predicts both successful and unsuccessful enrolments for B0B01MA2. The classifier lacks of precision on unsuccessful students. According to the recall 62%, the predictor does not tend to label too many observations as unsuccessful, but the ratio of false negative to true negatives is high. The f-beta score the weighted harmonic mean of the recall and precision. In the case of the f-1 score, we say the precision and recall abilities are equally important. The f-beta score reaches a value between 0 and 1, where the score 1 is considered as the best. The support values display the number of occurrences of a class.

One of the popular tools for comparing classifiers is the receiver operating characteristic (ROC) curve [10]. The advantage of this graph is that it shows the performance of the classifiers at all classification thresholds. The curve plots two parameters: true positive rate (recall) and false positive rate. An ROC curve of a good classifier is as close as possible to the top left corner of the graph. Contrary, the dotted line represents classifier without any information (random classifier). The area under the ROC curve (AUC) then gives the classifier's performance. Figure 5.3 shows an ROC of the B0B01MA2's classifier with an AUC 0.7.

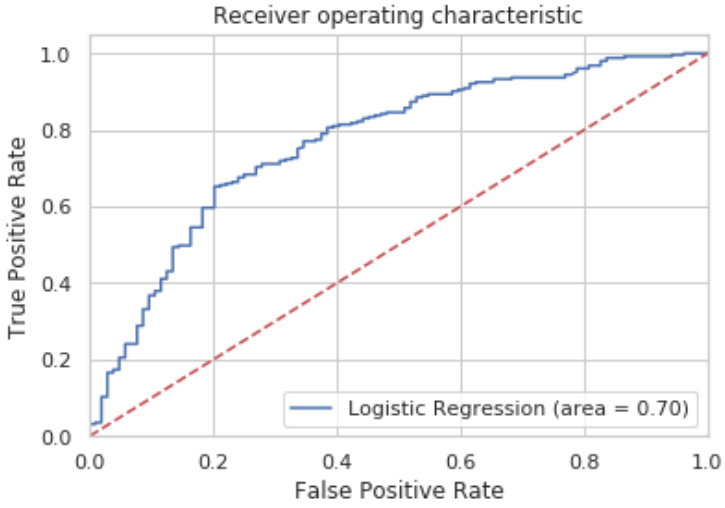


Figure 5.3: An ROC curve of the B0B01MA2 classifier.

A single result of a classifier is usually not so relevant. K-fold cross-validation is commonly used for testing machine learning models since it is simple to implement [16]. During the K-fold testing, the dataset is shuffled and split into k groups. Each dataset is once selected as the test dataset while the remaining are used for training the model. In this way, we obtain k evaluations of the classifier. The mean of the results better estimates the skill of the classifier on unseen data. We use $k = 5$ for our cross-validations. Figure 5.4 shows results of the K-fold Cross-Validation of the Mathematical Analysis 2 classifier. The values represent the means of the evaluations.

```
Accuracy:0.73
AUC:0.72
0: precision 0.51 recall 0.69 support 65.6
1: precision 0.86 recall 0.75 support 172.4
```

Figure 5.4: K-fold Cross-Validation results of the B0B01MA2 classifier. $k = 5$

5.3 Study Completion Prediction

This section describes the approach used for modelling of a classifier for the completion of a study plan. This approach uses a logistic regression as in the previous section. Also, the dataset used for predictions is the same. However, the selected predictor factors differ.

We look for a compulsory study plan subjects whose grades can predict successful study completion. Therefore, we need to evaluate individual grades with a numeric value. We use the following evaluation: $\{"A" : 2.5, "B" : 2, "C" : 1.5, "D" : 1, "E" : 0.5, "F" : -0.5, "/" : -1, "Z" : 0.5\}$. The "$F$" and "/" grades both symbolise the student failed to finish the course. During a course completion study, we have found out that "/" grade obtained when a student does not even take an exam attempt is more significant for classifiers. Thus the absolute value is greater. The *zero* value remains for a student who does not have enrolled in the course yet.

The final vector of variables for a student consists only of grades from his compulsory courses. Histories of all students who have already ended their study are added to the training dataset. Note that successful observations would be only the vectors with completed compulsory subjects. This representation would lead to bad predictions of first-year students. Therefore, we split the records of finished students by semesters, so a record with the student's history is made for each semester.

Since we have no finished students from the new version of programmes, we create a classifier for this study plans together with their predecessors (e.g., the model for OES 2020 is combined with previous OES). Also, we create only one classifier for programmes with more branches such as Open Informatics. Following Figure 5.5 propose the summary for the Open Informatics' classifier. Note the RFE reduced the number of compulsory courses from 38 to 10.

Logit Regression Results

Dep. Variable:	y	No. Observations:	1562
Model:	Logit	Df Residuals:	1552
Method:	MLE	Df Model:	9
		Pseudo R-squ.:	0.3667
		Log-Likelihood:	-685.70
converged:	True	LL-Null:	-1082.7
		LLR p-value:	4.162e-165

	coef	std err	z	P>|z|	[0.025	0.975]
B0B01MA2	1.0651	0.208	5.124	0.000	0.658	1.472
B0B01PST	1.2288	0.189	6.509	0.000	0.859	1.599
B0B36PRP	0.5466	0.067	8.217	0.000	0.416	0.677
B4BPROJ6	1.8092	1.268	1.427	0.154	-0.675	4.294
B0B36PJV	0.3267	0.088	3.728	0.000	0.155	0.498
B4B36PDV	1.1211	0.252	4.445	0.000	0.627	1.615
B4B39VGO	1.0932	0.513	2.131	0.033	0.088	2.099
B0B33OPT	2.5192	0.579	4.351	0.000	1.384	3.654
B6B36TS1	1.2192	0.420	2.901	0.004	0.395	2.043
B0B39PGR	1.0611	0.401	2.646	0.008	0.275	1.847

Figure 5.5: Logit Regression Results for the classifier of Open Informatics' study plans.

We use exactly ten features for the model fitting. By multiple tests, this size of the feature set seems to have the best performance across the different study plans. In the example above 5.5 especially course B4BPROJ6 has a large std error and p-value and could be removed. However, we leave the size of the RFE the same, since the importance of courses may dynamically change each semester, and we do not want to check features of study plans manually every time.

K-Fold Cross-Validation technique is used for testing the classifiers as in the previous section. We would also like to know how well performs the classifiers on students after their first semester. Therefore, we propose another Cross-Validation with different test data frames. We create students' data frames only with a history from their first semester during this K-Fold testing.

```
Accuracy: 0.83
AUC: 0.74
0: precision 0.87 recall 0.91 support 71.6
1: precision 0.68 recall 0.58 support 23.8
```

Figure 5.6: K-Fold test results for Open Informatics' classifier ($k = 5$). The test data set consists of students' histories only from their first semester.

```
Accuracy: 0.83
AUC: 0.82
0: precision 0.84 recall 0.86 support 156.2
1: precision 0.83 recall 0.79 support 127.4
```

Figure 5.7: K-Fold test results for Open Informatics' classifier ($k = 5$). The test data set consists of students' histories from all semesters.

Figures 5.6, 5.7 compare Open Informatics' classifier results between datasets with first-year students and all students. We can see, the classifier has a quite good recall for unsuccessful students after the first semester of their study. The precision of not mislabelling students as unsuccessful is getting better with following semesters. Also note that even the accuracy is the same, the behaviour of the classifier is different.

5.4 Course Occupancy Estimation

We described the problem of occupancy estimation in Section 1. We discussed the estimate would be very inaccurate, and the usage of the results would be limited. Therefore, in the application, we do not estimate any occupancies. We rather collect meaningful data and put them together to save organisers some work and propose a better view of the course.

We find all students with a study plan which includes a given course as compulsory. Furthermore, we will select students who have not yet completed the course, and the recommended semester is the next one, or the students should have already completed the course. In other words, this number of students indicates the number of enrolments according to the occupancy of study plans plus the number of students who should have already finished the course. This value should be used as the minimum for the course occupancy. Next value proposes the average occupancy since 2016 for comparison and the last proposed value represents the predicted success rate of students in the current semester. This may help predict the occupancy if the course repeats every semester and we do not know the number of failed students until the end of the current semester.

In case of the first semester, new students will occur in the predicted occupancy with the first update right after they are added into the KOS database after their successful application.

5.5 Summary

We specified the approach of handling different course relations in the scholar system and formulated the MFD as an ILP task. We also described the classifiers of the successful completion of subjects and the entire study. Logistic regression is used in both predictors. However, we use a histogram of grades to predict subject completion while using a vector of individual subjects' results to predict the study completion.

Finally, we described the data processed by the application, which will help us to estimate the number of students in a course in the next semester. We decided for this solution after

the discussion in Section 3.3 instead of leaving the estimation on the application itself, which could be too inaccurate or erroneously misleading for the teachers.

Results of these approaches are presented in Chapter 6.

Chapter 6

Results

This chapter describes the results of the individual parts of the work. The first part of the chapter compares the minimal and real costs of studying and discusses their reasons. The second part presents the results of the classifiers of successful study completions and individual courses such as recall, precision and AUC. Note that we did not count on the financial demand of students who completed part of their studies abroad and students with an individual study plan, and therefore, they are not included in the results. Likewise, records of study abroad are ignored by individual predictors, because it is not possible to know from the records of the school system whether a student has successfully completed a semester abroad.

6.1 Teaching Demands

6.1.1 Course Financial Demands

This section proposes a view on financial demands of individual courses. We discuss why some subjects have much higher demand per student than others. We also show the *most expensive* and the *cheapest* courses and how the price of courses differs between semesters. The courses with the minimal demand are important for the next section 6.1.2 which presents the minimal demands for study plans, because the *cheapest* subjects are usually to be selected as elective courses of a minimal study plan.

6.1.1.1 Course Demand Comparison

Following Tables 6.1, 6.2 show subjects with minimal and maximal financial demands respectively for accredited hours. The courses are sorted by their credits. There are always listed 3 subjects from each credit category, except the courses for one credit, because A6M33ZPP is the only course for one credit with known demand now. Courses listed in the Table 6.1 have a high number of students in common. On the other side, all the expensive courses except A6M33ZPP were opened for less than 6 students and are usually in English. The accredited hours (ZH) in the tables represent average course ZH from all available semesters.

KOS id	Name	ZH	Credits
A6M33ZPP	Základy první pomoci	4,20	1
B0M36MOOC	Massive Open Online Course	0,93	2
A0B16MPL	Manažerská psychologie	1,24	2
B3B04PSA	Akademické psaní	1,51	2
B6B04PRE	Prezentace	0,73	3
B6B36ZPR	Základy projektového řízení	0,79	3
B6B39ZMT	Základy multimediální tvorby	1,09	3
BD1M15IND	Projekt magisterský	0,20	4
A0M16MPS	Manažerská psychologie	0,62	4
B2B15UEL	Úvod do elektrotechniky	0,62	4
BD1M16EVE	Ekonomika výroby energie	0,40	5
BD1M16ENI	Environmentální inženýrství	0,40	5
B1B01MEK	Matematika pro ekonomii	0,65	5
BD1M16FIM	Finanční management	0,33	6
B3B33KUI	Kybernetika a umělá inteligence	0,59	6
B0B99PRP	Procedurální programování (pro EK a EEM)	0,63	6
B0B01LAGA	Lineární algebra	0,57	7
A8B01MCM	Matematika-vícedimenzionální kalkulus	0,59	7
A8B01DEN	Diferenciální rovnice a numerické metody	0,59	7

Table 6.1: Courses with the minimal average accredited hours per student per one credit.

KOS id	Name	ZH	Credits
A6M33ZPP	Základy první pomoci	4,20	1
BE9M38PRM	Project Management and Marketing	17,28	2
B9M38PRM	Projektové řízení a marketing	7,40	2
A8M32AVL	Laboratoř zpracování audio-video signálů	7,03	2
A6M02FPT	Fyzika pro terapii	8,63	3
B0B02EKE	Ekologie a ekotechnika	5,89	3
A6M33FZG	Fyziologie a anatomie	4,93	3
BE0M02UFL	Introduction to Laser Physics	17,08	4
AE1B37KEL	Communication and Electronics	14,32	4
BEVB14ZVE	Power Electronics	13,50	4
BE1B38EMA	Electrical Measurements and Instrumentation	13,67	5
BE2M37KDK	Coding in digital communications	11,64	5
BE1M14EPT1	Electric Drives and Traction	11,64	5
AE4B38DSP	Distributed Systems and Computer Networks	11,52	6
BE2M37OBFA	Image Photonics	11,51	6
BE2M32DMT	Diagnostics and Measurement in Telecommunications	11,03	6
BE3M38DIT	Diagnostics and Testing	7,90	7
BE5B35ARI	Automatic Control	5,10	7
A8M17RFB	RF funkční bloky	3,81	7

Table 6.2: Courses with the maximal average accredited hours per student per one credit.

6.1.2 Minimal Financial Demands of Study Plans

Table 6.3 shows the minimal accredited hours ZH and the minimal equipment requirement demands KMNP for the listed study plans. Branches of Open Informatics (OI) belongs to study plans with the theoretical lowest financial demands. There are two reasons why Open Informatics study plans have lower financial demands. The first reason is, there are many students which undertake mandatory courses of OI. However, more significant fact is, that (for example for Software branch) only 151 credits come from compulsory or mandatory elective courses. Therefore, almost 30 credits can be filled with the courses with the lowest demands. The opposite case is Applied Electrical Engineering 2018. Only 4 credits are required from elective courses to meet the requirements of the minimum 180 credits. Therefore, even the EEM study plans have a lot of enrolled students, the theoretical minimal demands are much higher.

On the other side, OES and EECS are study plans with the highest financial demands. For both plans, the subjects are usually taught just for a few students. Students of EECS have the compulsory courses in English what makes them even more expensive. However, the difference between these two study plans are the obtained credits from the compulsory subjects. Students of OES do not need any elective courses if they will finish the required minimum from all of their course groups. Students of EECS have to obtain at least 20 credits from some elective subjects. This fact, that they can fill 20 credits with courses with low demands, makes the theoretical minimal financial demand of EECS much lower. Also, note

that students of the English study plan can enrol into elective subjects taught in Czech, although this is unlikely.

KOS id	Name	ZH	KMNP
30020907	OES 2020	428,78	70,76
30019169	EECS (ENG)	303,67	41,57
30018713	EEM - Elektrotechnika a management 2018	230,35	30,40
30018712	EEM - Apl. elektrotechnika 2018	241,19	36,36
30018710	Elektronika a komunikace 2018	237,24	44,87
30018709	OI - Počítačové hry a grafika 2018	149,31	14,71
30018708	OI - Software 2018	149,25	11,37
30018707	OI - Internet věcí 2018	166,00	18,38
30018706	OI - Základy umělé inteligence a počítačových věd 2018	145,55	12,26
30018325	Lékařská elektronika a bioinformatika	191,88	25,20
30013415	OI - Počítačové hry a grafika 2016	149,31	14,71
30013414	OI - Software 2016	150,57	12,55
30013413	OI - Internet věcí 2016	166,00	18,38
30013412	OI - Informatika a počítačové vědy 2016	145,55	12,26
30013411	OI - Kybernetika a Robotika 2016	178,51	25,13
30013410	OI - Elektronika a komunikace 2016	239,01	48,57
30013406	EEM - Elektrotechnika a management 2016	209,73	35,40
30013405	EEM - Aplikovaná elektrotechnika 2016	220,49	41,76
30007918	Softwarové inženýrství a technologie	168,71	17,76
30005719	OES	427,36	80,32

Table 6.3: Theoretical minimal demands per student of the selected study plans.

6.1.3 Real Financial Demands of Study Plans

Table 6.4 shows accredited hours of successfully finished studies of individual study plans. The new study plans are not listed, since there are no finished studies yet. Table 6.4 compares minimal, mean, median and maximal accredited hours of studies to the minimums proposed in Table 6.3. Comparing medians to the theoretical minimums, EECS shows the biggest difference, but demands of only 3 successfully finished studies have been computed this new version of EECS. If we exclude EECS study plan, the biggest difference from the theoretical minimum has Internet of Things 2016 (125,61 ZH). The closest to the minimum are students from Software Engineering and Technology (58,58 ZH).

KOS id	theoretical min	min	avg	median	max	students
30019169	303,67	603,51	638,19	625,96	754,04	3
30013415	149,31	159,09	227,64	223,88	306,78	31
30013414	150,57	205,12	232,98	232,92	315,76	26
30013413	166,00	230,57	293,30	291,61	434,93	6
30013412	145,55	189,11	243,34	238,38	353,15	43
30013411	178,51	214,03	251,05	246,60	371,16	77
30013410	239,01	273,91	330,80	323,55	472,99	53
30013406	209,73	241,45	284,87	275,24	395,76	38
30013405	220,49	266,02	302,31	295,17	407,21	32
30007918	168,71	186,35	241,09	227,29	503,40	97
30005719	427,36	489,03	519,66	500,47	649,49	15

Table 6.4: Accredited hours of finished studies.

Following Table 6.5 compares accredited hours of failed studies to theoretical minimums of study plans. Failed studies are listed only if they represent the student's last study. Study plans of Open Informatics before specialisation are listed separately and the failed studies are not counted after the specialisation again. The medians of financial demands are expectedly lower for study plans which have been added in 2018. The lowest median of accredited hours of older study plans has Electronics and Communications followed by Cybernetics and Robotics. Contrarily, the most expensive failed studies comes from the English study plan EECS.

KOS id	Theoretical min	avg	median	max	students
30020907	428,78	75,64	75,64	75,64	1
30019169	303,67	235,91	199,20	444,08	14
30018713	230,35	82,25	70,97	187,67	10
30018712	241,19	78,92	74,90	116,69	5
30018710	237,24	56,46	47,85	237,28	60
30018709	149,31	95,00	85,49	224,94	13
30018708	149,25	113,22	82,64	283,02	7
30018707	166,00	73,43	72,92	75,03	4
30018706	145,55	85,26	81,55	147,11	10
30018325	191,88	42,70	28,41	102,12	33
30013415	149,31	123,58	116,29	266,40	38
30013414	150,57	132,41	117,03	252,05	22
30013413	166,00	148,45	154,17	241,07	8
30013412	145,55	135,92	137,59	261,97	21
30013411	178,51	54,42	39,08	261,24	155
30013410	239,01	63,76	39,07	335,35	84
30013406	209,73	99,16	74,14	261,97	15
30013405	220,49	79,47	79,47	94,32	2
30007918	168,71	83,96	68,47	284,05	205
30005719	427,36	117,42	65,18	572,34	25
30018705	-	33,28	23,27	173,13	64
30015089	-	33,13	27,17	136,32	74

Table 6.5: Accredited hours of failed studies.

Following histogram 6.1 compares accredited hours of finished and failed studies since 2016. However, if a student enrol in a study plan again after a failure, the study is not displayed. Only failed studies when a student does not continue are listed.

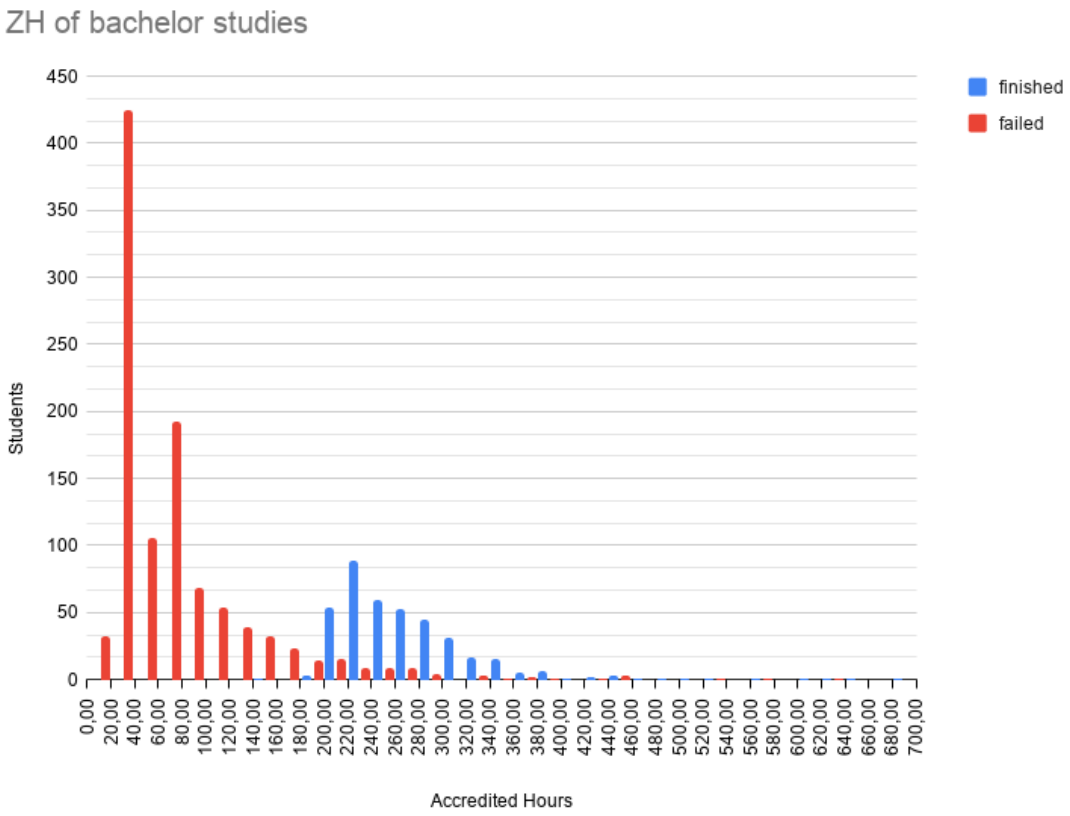


Figure 6.1: The histogram compares accredited hours between failed and successfully finished studies since 2016.

The histogram 6.2 compares demands between 1st and 2nd enrolments for a study. There are shown only studies of students which started at 2016 or later. Therefore, the sample of finished studies with a 2nd study enrolment can be smaller, since some of the students which started at 2016 still study. There are 364 finished studies within a first enrolment and 29 with a repeated study with known financial demands (7%). For comparison, there is 256 out of 1829 currently active studies which are considered as a 2nd enrolment (14%).

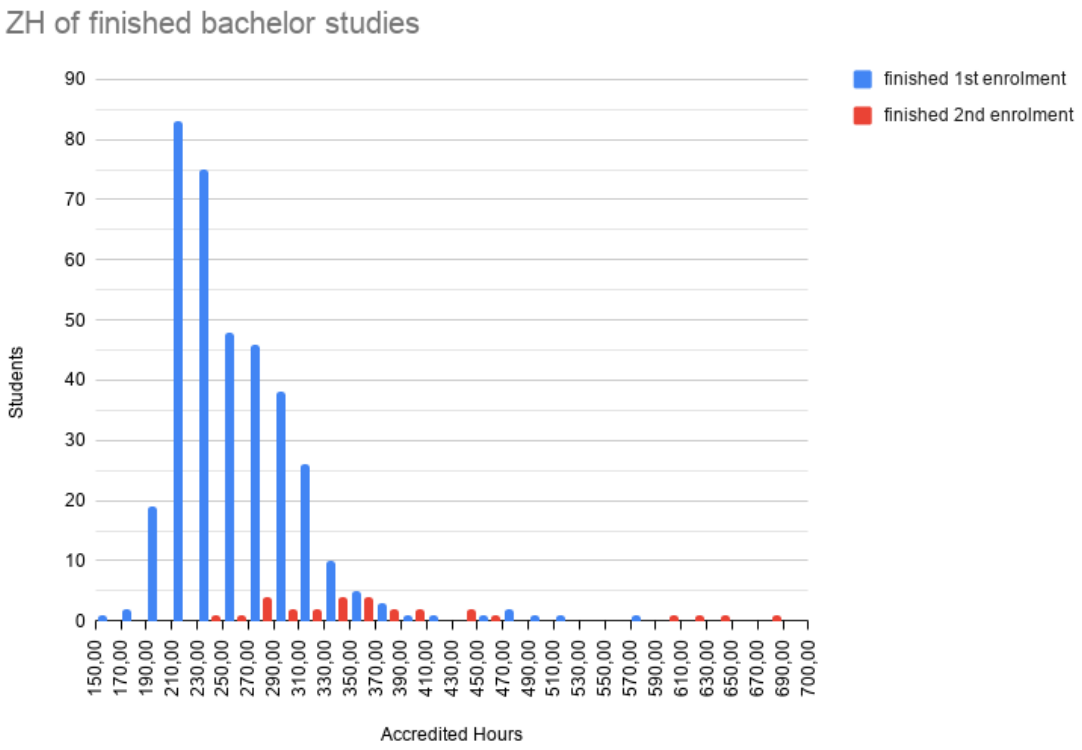


Figure 6.2: The histogram comparing financial demands of finished studies which started in 2016 and later.

Following histograms show occurrence of students' real financial demands, so we can compare demands of successfully finished, failed and currently active studies. Sometimes, a student fails his study, but enrols to a study plan again. In such case, the student's failed study is not displayed in histograms between failed studies. The accredited hours from his failed study are added to his new study instead. Only failed studies, where a student decides not to continue, are displayed as failed.

6.1.3.1 Accredited Hours of SIT

Figure 6.3 compares demands of Software Engineering and Technology study plan. Out of a total of 658 students, the demands were successfully computed for 584 of them (89%).

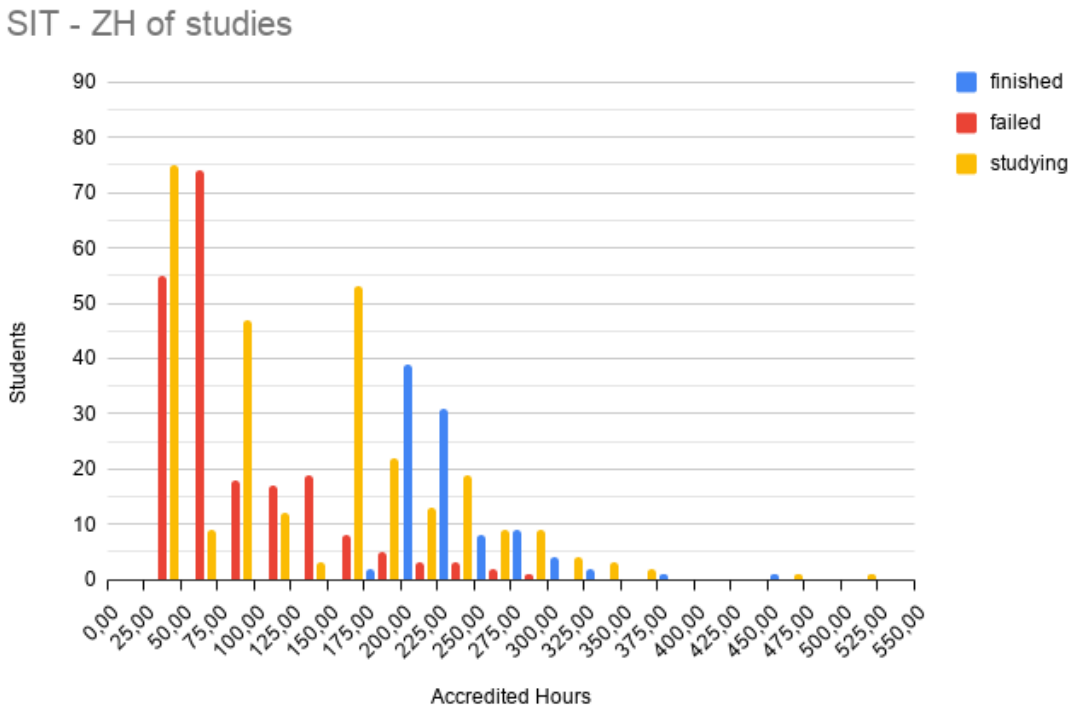


Figure 6.3: The histogram comparing accredited hours of students of Software Engineering and Technology.

6.1.3.2 Accredited Hours of OI

Figure 6.4 below, compares financial demands of students of Open Informatics. New study plans have been introduced in 2018 for Open Informatics programmes and no student have graduated some of them yet. The difference between demands of Open Informatics programmes are according to Table 6.3 very similar to programmes from 2016. Therefore, the histogram combines programmes of Open Informatics accredited from both of these years. Out of a total of 904 students, the demands were successfully computed for 843 of them (93%). Table 6.6 lists Kos ids of the study plans covered by the histogram.

30018709	30018708	30018707	30018706	30018705
30013415	30013414	30013413	30013412	30015089

Table 6.6: Kos ids of Open Informatics study plans covered in histogram 6.4.

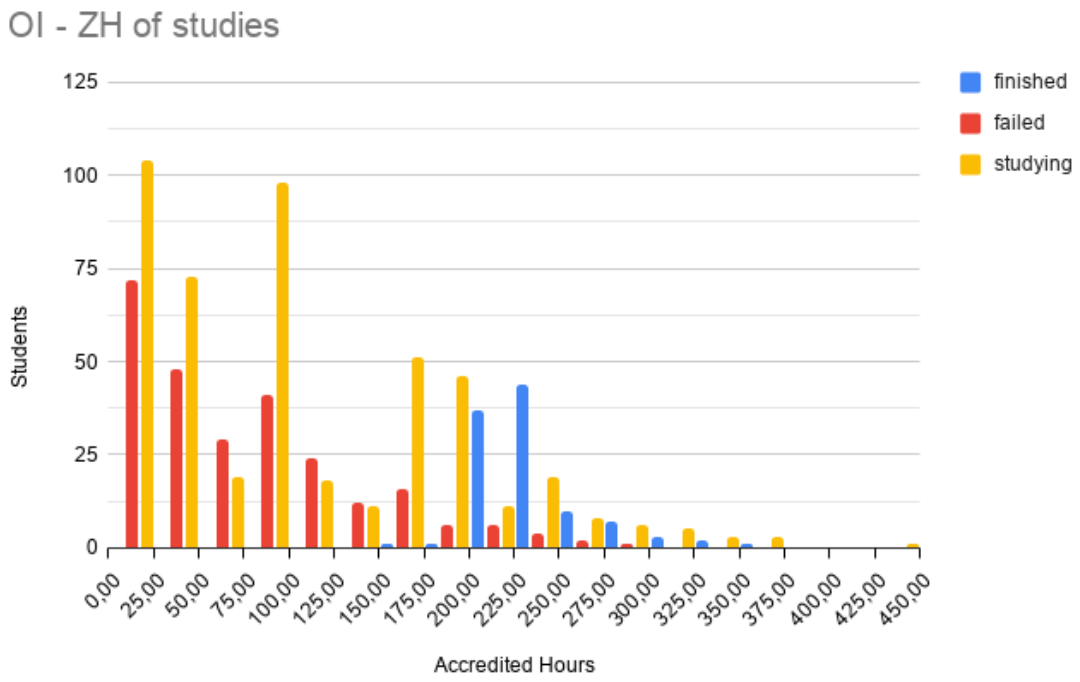


Figure 6.4: The histogram comparing accredited hours of students of Open Informatics.

6.1.3.3 Accredited Hours of KYR

The histogram 6.5 shows financial demands of students of Cybernetics and Robotics. Since this study plan has not changed since 2016, the histogram covers only a study plan with **kos id 30013411**. Successfully computed were 576 of 660 students' demands (87%).

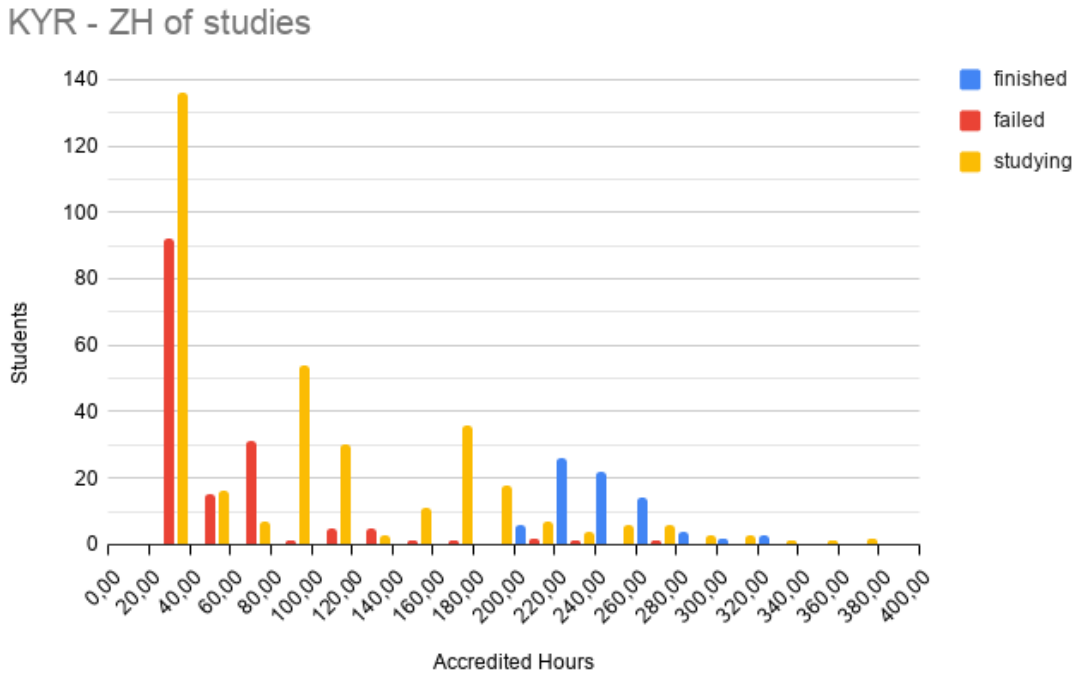


Figure 6.5: The histogram comparing accredited hours of students of Cybernetics and Robotics.

6.1.3.4 Accredited Hours of EECS

Electrical Engineering and Computer Science is the only English bachelor study plan taught at the faculty now. The histogram 6.6 compares demands of 64 students. The financial demands of students of EECS are higher than others, since the language coefficient K_j from KOMETA2 methodology [3] makes the course more expensive. Also, often courses from this study plan are opened just for a few students.

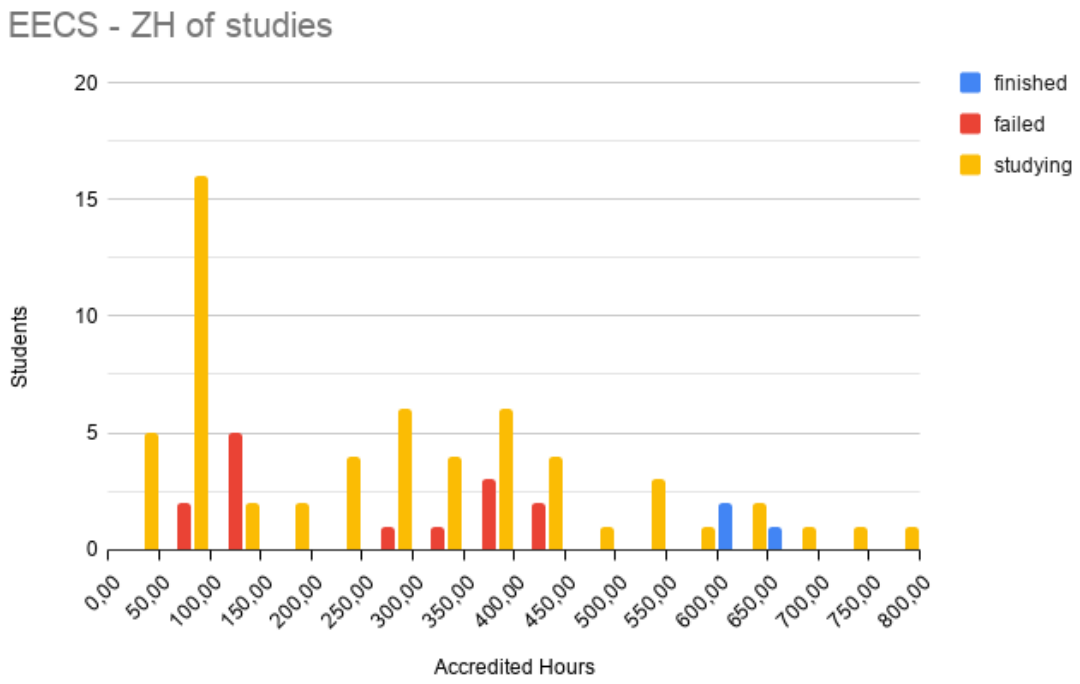


Figure 6.6: The histogram comparing accredited hours of students of Electrical Engineering and Computer Science.

6.1.3.5 Accredited Hours of EEM

Electrical Engineering, Power Engineering and Management programme has two different branches - Applied Electrical Engineering and Electrical Engineering and Management. Currently, there are ending study plans since 2016 and new study plans since 2018. Accredited hours in Figure 6.7 comes from all of the study plans in Table 6.7.

30018713	30018712	30013406	30013405

Table 6.7: Kos ids of Electrical Engineering, Power Engineering and Management study plans covered in histogram 6.7.

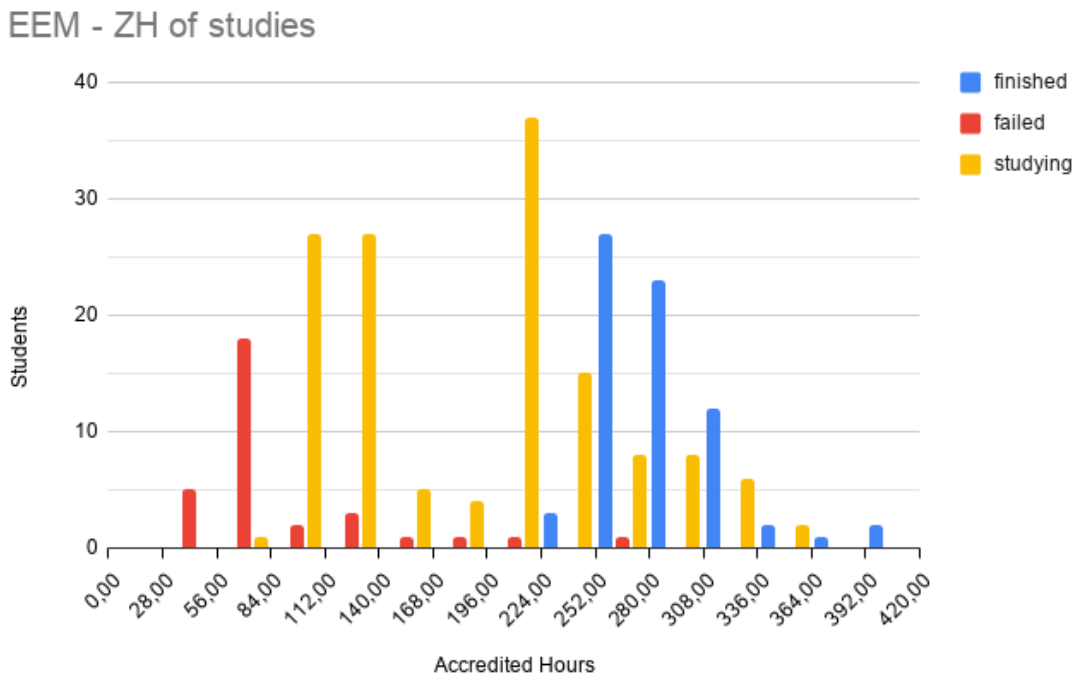


Figure 6.7: The histogram comparing accredited hours of students of Electrical Engineering, Power Engineering and Management.

6.1.3.6 Accredited Hours of EK

Accredited hours have been computed for 410 out of a total 439 students (93%) of Electronics and Communications. This programme has an ending study plan form since 2016 and a new form since 2018. Financial demands of students of study plans in Table 6.8 are combined in Figure 6.8.

30018710	30013410

Table 6.8: Kos ids of Electronics and Communications study plans covered in histogram 6.4.

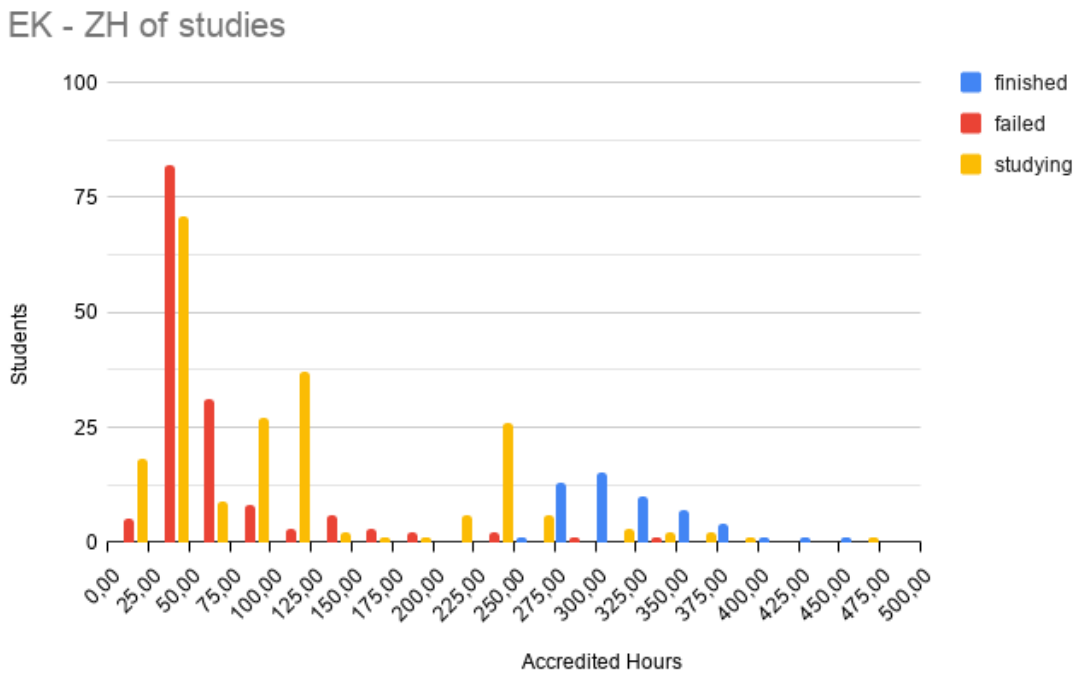


Figure 6.8: The histogram comparing accredited hours of students of Electronics and Communications.

6.1.3.7 Accredited Hours of OES

Open Electronic Systems students study the least. Only 36 students undertake this study plan nowadays. Accreditation of this study plan has been changed recently. However, we combine the new and the old version (Table 6.9) in Figure 6.9 together. Due to the lower number of students, studies of OES belongs between studies with higher financial demands for accredited hours.

30018709	30018708

Table 6.9: Kos ids of Open Electronic Systems study plans covered in histogram 6.4.

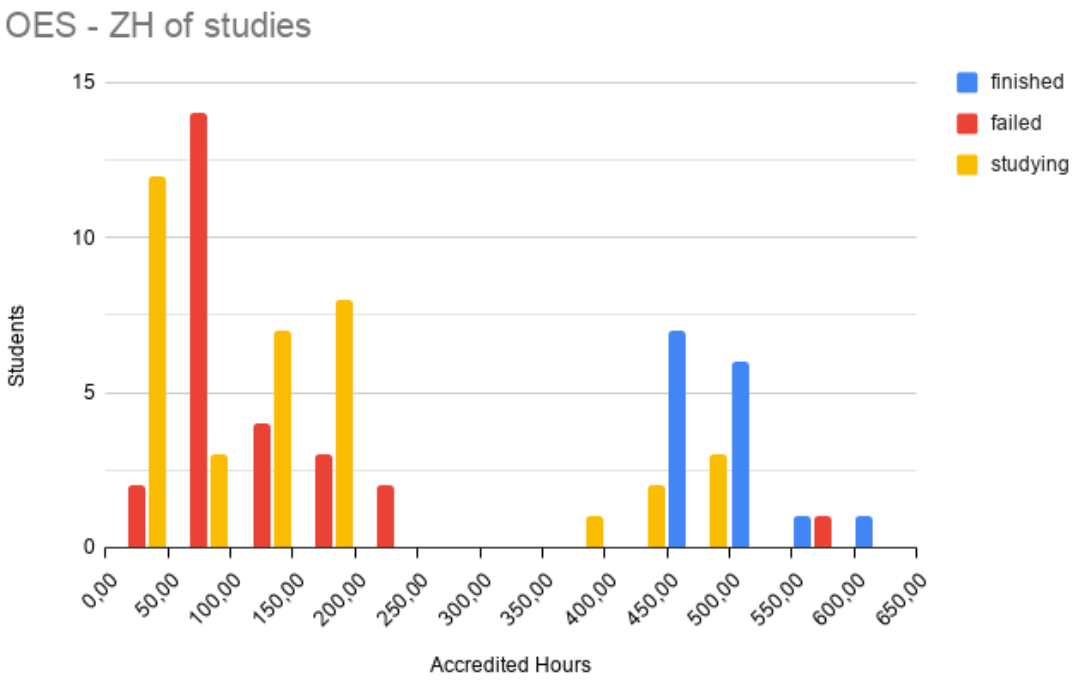


Figure 6.9: The histogram comparing accredited hours of students of Open Electronic Systems.

6.1.3.8 Accredited Hours of Bioinformatics

Medical Electronics and Bioinformatics is a completely new study plan available since 2018. Therefore, there are no finished studies to this date and the figure 6.10 compares only accredited hours of failed and currently active studies.

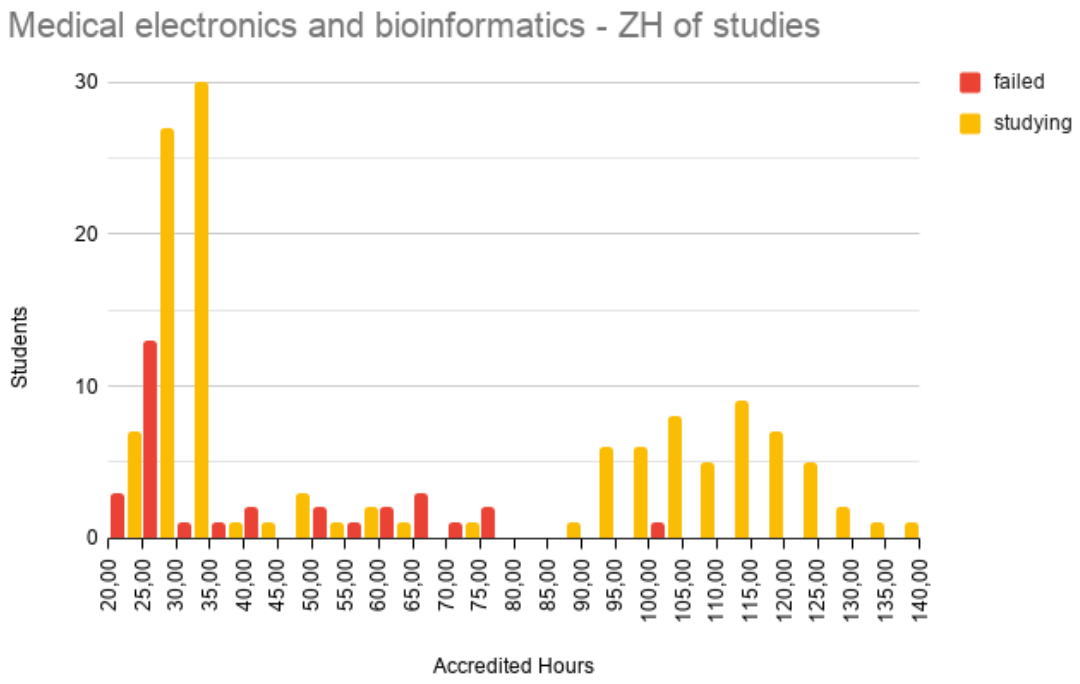


Figure 6.10: The histogram comparing accredited hours of students of Medical Electronics and Bioinformatics.

6.2 Study Completion Prediction

This section describes the results of the study completions' classifiers. As we discussed in Section 5.3, we test the classifiers using the K-Fold Cross-Validation technique. Due to the lack of data, we do not use $k = 10$ as usual, but just $k = 5$ to provide at least a little support for the test results. Students from a study plan are split into five groups, and then each group is used once as the test dataset. Thus, the following test results are the mean of five run-times. Tables 6.10, 6.11 propose the results of the classifiers on the dataset which consists of students' histories from all semesters. Hence every student may occur six times in the test or the training dataset, every time with a different history of course enrolments. However, they can not be in the test and the training dataset at the same time. Tables 6.10, 6.11 provide precision and recall from the unsuccessful and successful completion point of view, respectively. Because of generally low support (number of occurrences) of observations, we also provide the weighted average by the number of observations.

	0				
	precision	recall	support	accuracy	AUC
OES	0,86	0,85	23,40	0,84	0,87
EEM	0,71	0,84	30,80	0,86	0,85
EECS	0,89	0,78	37,40	0,77	0,77
EK	0,91	0,92	60,20	0,91	0,90
KYR	0,83	0,89	83,60	0,89	0,89
OI	0,84	0,86	156,20	0,83	0,82
SIT	0,75	0,83	114,40	0,80	0,80
Weighted Average:	0,82	0,86		0,84	0,84
Average:	0,83	0,85	72,29	0,84	0,84

Table 6.10: Results of classifiers on unsuccessful studies.

	1				
	precision	recall	support	accuracy	AUC
OES	0,79	0,90	20,80	0,84	0,87
EEM	0,94	0,87	92,20	0,86	0,85
EECS	0,58	0,75	14,20	0,77	0,77
EK	0,91	0,89	60,20	0,91	0,90
KYR	0,93	0,89	136,80	0,89	0,89
OI	0,83	0,79	127,40	0,83	0,82
SIT	0,85	0,77	140,40	0,80	0,80
Weighted Average:	0,88	0,83		0,84	0,84
Average:	0,83	0,84	83,34	0,84	0,84

Table 6.11: Results of classifiers on successful studies.

The results of K-Fold Cross-Validation show surprisingly high accuracy and AUC. This may be caused by several reasons. First, the same student occurs in the test dataset many times. If the model successfully predicts the student in one observation, it will more likely successfully predict his success again. However, the student's record is there always with his different history of subjects and therefore, it is not a completely same observation. Also, this fact may work on predictor in the opposite way. Another reason may be that most of the students fail in the first year, especially in the first semester. Thus, we decided to test the prediction models on the datasets that consist of students' histories only from their first semester.

	0				
	precision	recall	support	accuracy	AUC
OES	0,87	0,94	10,20	0,85	0,76
EEM	0,59	0,89	11,00	0,66	0,69
EECS	0,88	0,88	13,60	0,8	0,66
EK	0,92	0,95	35,80	0,90	0,86
KYR	0,84	0,94	47,00	0,84	0,79
OI	0,87	0,91	71,60	0,83	0,74
SIT	0,76	0,86	48,00	0,73	0,67
Weighted Average:	0,84	0,91		0,81	0,74
Average:	0,82	0,91	33,89	0,80	0,74

Table 6.12: Results of classifiers on unsuccessful studies based on results from the first semester.

	1				
	precision	recall	support	accuracy	AUC
OES	0,75	0,59	3,40	0,85	0,76
EEM	0,88	0,49	14,80	0,66	0,69
EECS	0,54	0,43	2,80	0,8	0,66
EK	0,82	0,77	11,60	0,90	0,86
KYR	0,83	0,63	21,80	0,84	0,79
OI	0,68	0,58	23,80	0,83	0,74
SIT	0,64	0,48	25,80	0,73	0,67
Weighted Average:	0,74	0,57		0,81	0,74
Average:	0,73	0,57	14,86	0,80	0,74

Table 6.13: Results of classifiers on successful studies based on results from the first semester.

Tables 6.12 and 6.13 present the results of prediction models tested only on students' histories from their first semester. The average accuracy is still above 80%, but we can see that AUC decreased significantly. Especially recall of successful study completions is much worse after the first semester. At the beginning of the student's study, the predictor tends to underestimate the student and predict more false negatives. Together with a slight overestimation of the student in the last semester, this is expected behaviour. The provided results are quite accurate, but we must not forget that the number of observations in the test datasets is really low, and the results may not be credible.

In Section 5.3 we discussed how a good prediction feature can be selected using the *coef*, *z-statistic* and *p-value*. Accordingly, we selected the three most significant courses for each of the listed study plans. It is useful to look at grades from these subjects whenever we want to predict a successful study completion. Table 6.14 shows the selected courses. This refers to our goal specification of searching variables for predicting unsuccessful students.

Study plan	Course codes		
OES	A8B17EMT	A8B37DIT	A8B34EOD
EEM	B1B14ZPO	B1B17EMP	B1B38EMA
EECS	BE5B33PRG	BE5B02PH1	BE5B01DEN
EK	B0B01LAG	B2B31ZEO	B0B01DRN
KYR	B3B02FY1	B3B31EPO	B0B01LAG
OI	B0B33OPT	B0B01PST	B4B36PDV
SIT	B6B01LAG	B6B16ZPD	B6B36PJC

Table 6.14: Courses which grade is a good predictor feature of study completion.

6.3 Course Completion Prediction

The classifiers of unsuccessful compulsory course completions have been tested in the same way. The prediction model of every compulsory course of the discussed study plans has been tested using the K-Fold Cross-validation with $k = 5$. Figure 6.11 proposes the arithmetical mean of AUCs, precisions and recalls of the classifiers.

```
Accuracy:0.73
AUC:0.73
0: precision 0.48 recall 0.64 support 15.14
1: precision 0.86 recall 0.75 support 48.92
```

Figure 6.11: The average results of the course completion's classifiers.

The classifier based on students' histograms of grades has an average accuracy 72%. However, the predictor is missing precision labelling unsuccessful course completions. Approximately every second prediction marked as unsuccessful is a false negative. Since the individual grades and number of failed attempts on exams are probably not like standard normally distributed data, we tried to scale the variables, so they have zero mean and unit variance. Figure 6.12 proposes the results of the adjusted classifier. We can see that with the scaled features, the precision of unsuccessful completions increases by more than 10%. Nevertheless, the accuracy is still almost the same since we lost the sensitivity of unsuccessful course finishes.

```
Accuracy:0.71
AUC:0.71
0: precision 0.61 recall 0.43 support 15.12
1: precision 0.81 recall 0.83 support 49.36
```

Figure 6.12: The average results of the course completion's classifiers with scaled variables.

None of these classifiers offers convincing results. We cannot simply say whether scaling variables pays off, as it depends on whether we are more concerned with the precision of

determining the future failed course completions or sufficient sensitivity even at the cost of higher false negatives.

Chapter 7

Application

We provide a Spring Boot web application (Surikata) as the prove of concept (POC) of the proposed approaches. The project is managed by Maven and consists of several modules.

- **domain** manages the data access objects (DAO) of the local database
- **fetch** manages updating data from the external sources
- **services** contains business logic of the application
- **commons** goes through all the layers and provides common objects such as exceptions
- **rest** handles the incoming Rest API requests and symbolise the outer interface of Surikata

Figure 7.1 approximates the logical structure of the application.

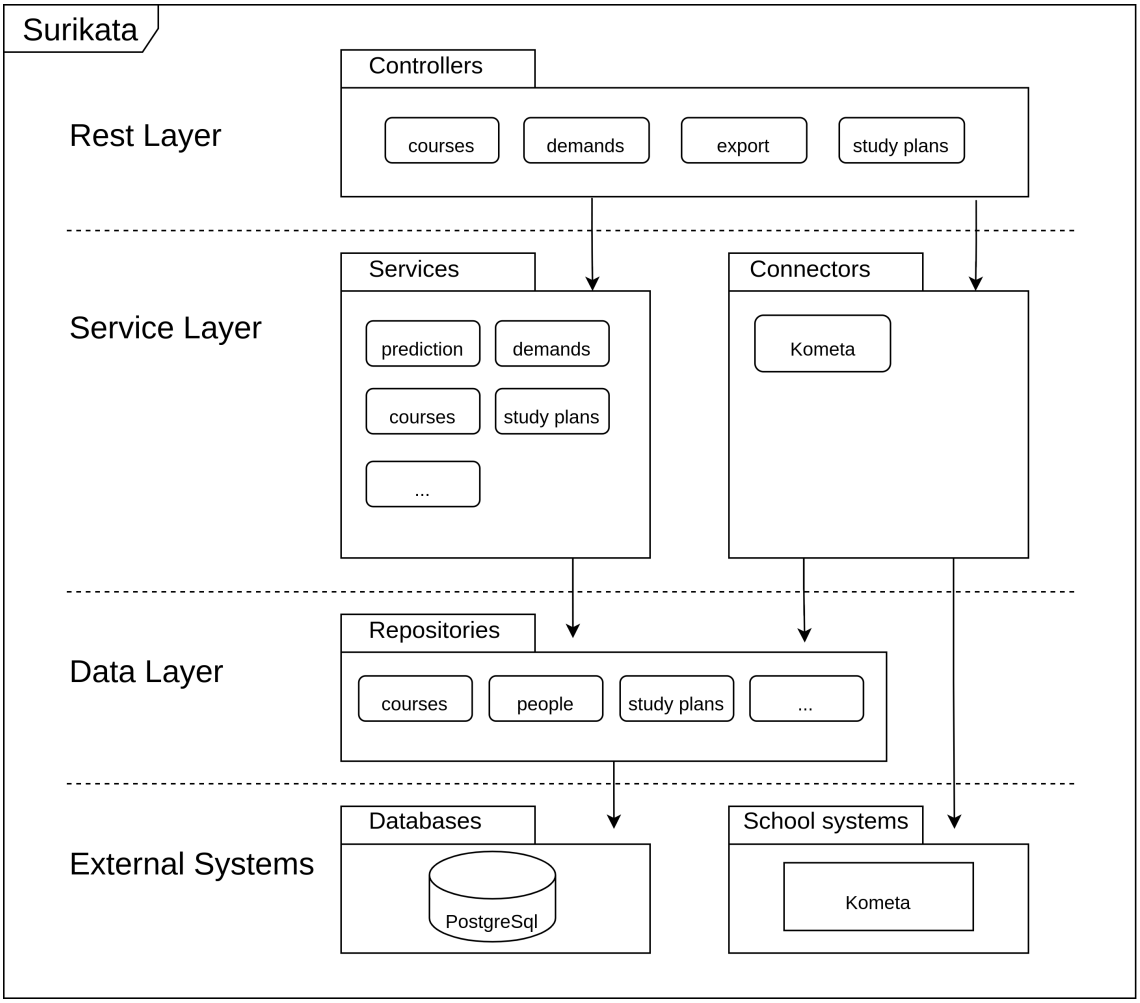


Figure 7.1: Layer diagram of the proposed application.

We use PostgreSQL as the local database. Surikata connects to only a single external system, the school application Kometa, responsible for evaluating courses' accredited hours and material requirements. The application expects and is prepared for regular updates from Kometa using its Rest API with a token authentication (also called bearer authentication). Of course, encrypted communication is expected between these applications.

The scripts for predictions are currently available in Jupyter notebooks. Complete automation of predictions would be time-consuming. Due to the uncertainty of interest, we decided to provide just a simple solution in the POC. Therefore, it is now necessary to manually run the scripts on the server-side after exporting the necessary data using the Rest API. After the script ends, the user has to import the results to the application. The sequential diagram describes the flow of the user's interaction.

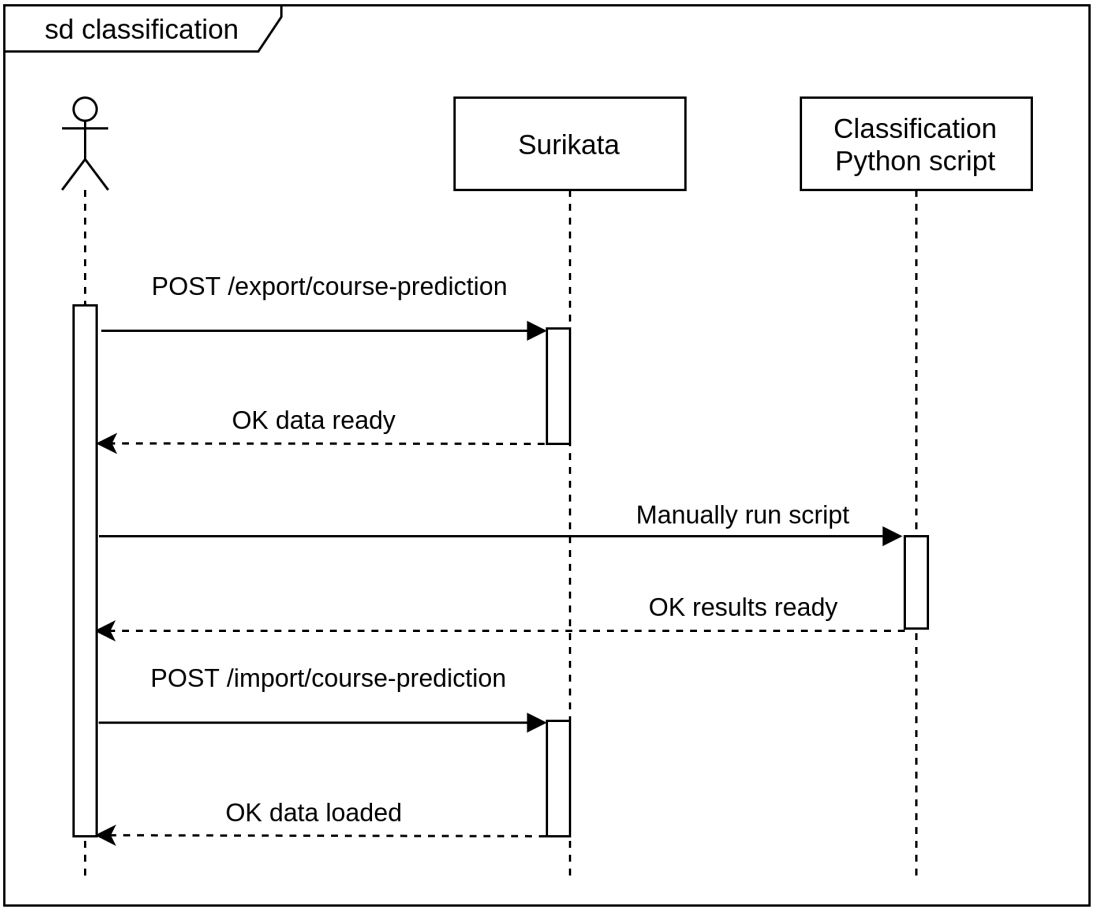


Figure 7.2: Sequence diagram of the classifications.

The ILP solution for the MFD problem is implemented in ILPStudyPlanSolver.class. We use Java Linear Programming Interface SCPSolver proposed by [21] with the GLPK (GNU Linear Programming Kit) package as backend [2]. The CourseService.class is responsible for the occupancy estimations.

7.1 Unit and Integration Testing

We test the application using the unit and integration tests. We use Spring Boot Framework for the testing. Since the application consists of several modules, we define the unit tests as tests within a single module. The tests are isolated from the other layers of the application. To accomplish these unit tests, we use mocking provided by the Spring Boot Test. Thanks to the mocking, we create so-called mocks of instances from other layers and specify their behaviour on specific requests. These mocks then fully substitute the other layers. Hence the correct functionality of the other layers is not necessary anymore for the tests of our module. Also, we do not need any data from a database.

On the other hand, the integration tests help to test the application's behaviour across multiple layers. We provide end-to-end tests to inspect the functionality for the user. In this case, we do not use mocking anymore. In-memory H2 persistence storage is used instead. This in-memory database fully replaces our local PostgreSQL database, so the tests do not affect the present real data.

Chapter 8

Conclusion

This thesis proposes approaches for several independent problems. First, we formulated the problem of searching the minimal financial demand for a study plan using ILP. We implemented the ILP solution within the proposed application and compared the demands with the students' real financial demands. We have shown that the minimal possible demand have study plans of Open Informatics due to the high number of students and the low number of compulsory courses. On the other hand, OES study plan is the most expensive since the subjects are usually opened just for a few students, and they do not even need any elective courses to obtain enough credits for the graduation. According to the real financial demands, we provided comparison of finished, failed and current study demands using histograms for every one of the discussed study plans. Also, we presented the most expansive and the cheapest subjects.

Secondly, we describe approaches for prediction of unsuccessful study and course completions. Since the lack of data from students' high schools and study applications, we use just students' existing results. Thus, the predictions are applicable after the first semester of a student. In the case of course completions, we use logistic regression based on students' histograms of grades. We also use Synthetic Minority Oversampling Technique (SMOTE) for oversampling the imbalance datasets to improve the performance. Based on the K-Fold Cross-Validation we could predict course completions with 73% accuracy on average. On the other hand, the proposed unsuccessful study classifier uses vectors that consist of students' individual compulsory courses' results. With this approach, we could predict an unsuccessful study with 83% precision and 85% recall on average. According to the prediction models' feature coefficients, we select courses of study plans, which results suggest the student's study success chances. The classifiers are implemented in Jupyter notebooks alongside with the application.

Lastly, we provide a tool that may help estimate the expected number of students in a course in advance. After discussion in Section 5.4 we propose a simple tool that sums the number of students which have the compulsory course in their study plan next semester and students who should have already taken the course before. We also propose the average number of students from the last four years.

8.1 Future Work

This section discusses possible future work related to this thesis. In the case of financial demands, the application can also propose results for master's programmes. However, because of some missing equivalency relations between old and new versions of courses, we would have to check the relations to obtain the financial demands of more studies.

We see the next possible future work in analysing the effects of global variables of Kometa2 methodology on the final accredited hours and equipment requirements. How would the MFD of study plans change with a different set of the global variables?

Creation of all equivalency classes between courses would be very valuable. Those relations would allow us to train and test the classifiers also on the already cancelled courses. This may be a very time-consuming task, but results would possibly be more precise with the increased number of newly obtained observations.

Bibliography

[1] Ekvivalence předmětů systému kos. Čvut - fakulta elektrotechnická. https://www.fel.cvut.cz/cz/education/announce/ekvivalentni-predmet.html. (Accessed on 01/04/2021).

[2] Glpk - gnu project - free software foundation (fsf). https://www.gnu.org/software/glpk/#TOCdocumentation. (Accessed on 01/04/2021).

[3] Metodika kometa2. Čvut - fakulta elektrotechnická. https://www.fel.cvut.cz/cz/rozvoj/KOMETA2_2012.pdf. (Accessed on 01/04/2021).

[4] Vztahy předmětů systému kos - Čvut - fakulta elektrotechnická. https://www.fel.cvut.cz/cz/education/announce/0808b.html. (Accessed on 01/04/2021).

[5] A. Alamri, Z. Sun, A. I. Cristea, G. Senthilnathan, L. Shi, and C. Stewart. Is mooc learning different for dropouts? a visually-driven, multi-granularity explanatory ml approach. *Lecture Notes in Computer Science*, page 353–363, 2020.

[6] E. M. Arkin, M. A. Bender, J. S. Mitchell, and S. S. Skiena. The lazy bureaucrat scheduling problem. *Information and Computation*, 184(1):129 – 146, 2003.

[7] A. T. Bolsoni-Silva, R. M. Barbosa, A. S. BrandÃ, and S. R. Loureiro. Prediction of course completion by students of a university in Brazil. *Psico-USF*, 23:425 – 436, 07 2018.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[9] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, 2009.

[10] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine Learning*, 31:1–38, 01 2004.

[11] F. Furini, I. Ljubić, and M. Sinnl. Ilp and cp formulations for the lazy bureaucrat problem. In L. Michel, editor, *Integration of AI and OR Techniques in Constraint Programming*, pages 255–270, Cham, 2015. Springer International Publishing.

[12] F. Furini, I. Ljubić, and M. Sinnl. An effective dynamic programming algorithm for the minimum-cost maximal knapsack packing problem. *European Journal of Operational Research*, 262(2):438 – 448, 2017.

[13] L. Gai and G. Zhang. On lazy bureaucrat scheduling with common deadlines. *Journal of combinatorial optimization*, 15(2):191–199, 2008.

[14] L. Gourvès, J. Monnot, and A. T. Pagourtzis. The lazy bureaucrat problem with common arrivals and deadlines: Approximation and mechanism design. In L. Gąsieniec and F. Wolter, editors, *Fundamentals of Computation Theory*, pages 171–182, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani. *Classification*, pages 127–173. Springer New York, New York, NY, 2013.

[16] G. James, D. Witten, T. Hastie, and R. Tibshirani. Resampling methods. In *An introduction to statistical learning*, pages 175–201. Springer, 2013.

[17] Z. Kovacic. Early prediction of student success: Mining students' enrolment data. 2010.

[18] G. Lassibille and L. Navarro Gómez. Why do higher education students drop out? evidence from spain. *Education Economics*, 16(1):89–105, 2008.

[19] P. Moreno-Marcos, C. Alario-Hoyos, P. Merino, and C. Delgado-Kloos. Prediction in moocs: A review and future research directions. *IEEE Transactions on Learning Technologies*, PP:1–1, 07 2018.

[20] J. Nouri, K. Larsson, and M. Saqr. Bachelor thesis analytics: using machine learning to predict dropout and identify performance factors. *International Journal of Learning Analytics and Artificial Intelligence for Education (iJAI)*, 1(1):116–131, 2019.

[21] H. Planatscher and M. Schober. Scpsolver. http://scpsolver.org/. (Accessed on 01/04/2021).

[22] P. Salvatori. Reliability and validity of admissions tools used to select students for the health professions. *Advances in Health Sciences Education*, 6(2):159–175, 2001.

[23] Č. Štuka, P. Martinková, K. Zvára, and J. Zvárová. The prediction and probability for successful completion in medical study based on tests and pre-admission grades. *New Educational Review*, 28(2):138, 2012.

[24] M. Wati, Haeruddin, and W. Indrawan. Predicting degree-completion time with data mining. In *2017 3rd International Conference on Science in Information Technology (ICSITech)*, pages 732–736, 2017.

Appendix A

List of Abbreviations

ZH Accredited Hours (zapocitatelne hodiny)

KMNP Equipment Requirements (materialova narocnost)

MFD Minimal Financial Demand

ROC Receiver Operating Characteristic

AUC Area Under the ROC Curve

KOS Study Information System of CTU in Prague

MPMKC Maximum-Profit Minimal Knapsack Cover

MCMKP Minimum-Cost Maximal Knapsack Packing

ILP Integer Linear Programming

TP True Positive

FP False Positive

TN True Negative

FN False Negative

CTU Czech Technical University

FEE The Faculty of Electrical Engineering

OI Open Informatics

EEM Electrical Engineering, Power Engineering and Management

EECS Electrical Engineering and Computer Science

EK Electronics and Communications

SIT Software Engineering and Technology

OES Open Electronic Systems

KYR Cybernetics and Robotics

MOOCs Massive Open Online Courses

Appendix B

User Guide

Installation of the application is described in **Help.md**. After running the application, the REST endpoints should be available on port 8081. Documentation of the available end points is in the attachment (exported from swagger).

Appendix C

CD Content

```
.
|– API-documentation – generated Api documentation
| |– index.html
|– DP_Johanides.pdf – master's thesis
|– surikata-master.zip – source code of the proposed application

    1 directory, 3 files
```