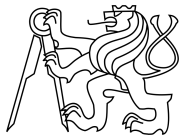
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ



ZADÁNÍ DIPLOMOVÉ PRÁCE

Název:	Predikce studijních výsledků studentů bakalářského programu Informatika FIT ČVUT
Student:	Bc. Magda Friedjungová
Vedoucí:	Ing. Stanislav Kuznetsov
Studijní program:	Informatika
Studijní obor:	Znalostní inženýrství
Katedra:	Katedra teoretické informatiky
Platnost zadání:	Do konce letního semestru 2016/17

Pokyny pro vypracování

1. Získejte data z fakultních systémů, které jsou používány pro zaznamenání studijních výsledků studentů, tzn. EDUX, Progtest a KOS, a proveďte jejich analýzu.
2. Tato data vhodně předzpracujte a zkombinujte s daty z přijímacího řízení a pomocí vhodných metod sestavte prediktivní modely a určete úspěšnost studentů 1. semestru 1. ročníku bakalářského programu Informatika v akademickém roce 2015/2016.
3. Proveďte analýzy výkonů studentů v závislosti na dostupných datech a navrhněte zlepšující doporučení jak pro studenty, tak pro vedení fakulty.
4. Podle reálné úspěšnosti navrženého řešení pro rok 2015/2016 diskutujte podmínky, za jakých by mohla predikce fungovat s vyšší přesností a navrhněte zobecnění modelu pro použití i v následujících letech či mezi jinými semestry.

Seznam odborné literatury

Dodá vedoucí práce.

L.S.

doc. Ing. Jan Janoušek, Ph.D.
vedoucí katedry

prof. Ing. Pavel Tvrdík, CSc.
děkan

V Praze dne 20. listopadu 2015

České vysoké učení technické v Praze
Fakulta informačních technologií
Katedra teoretické informatiky

Diplomová práce

Predikce studijních výsledků studentů bakalářského programu Informatika FIT ČVUT

Bc. Magda Friedjungová

Vedoucí práce: Ing. Stanislav Kuznetsov

4. května 2016

Poděkování

V této práci bych chtěla především velmi poděkovat **Ing. Janu Motlovi** za cennou spolupráci při predektivním modelování a pomoc při nasazování modelů.

Dále bych chtěla poděkovat vedoucímu této práce **Ing. Stanislavu Kuznetsovovi** za duchaplný přístup a nadhled při řešení problematických částí.

Děkuji za nápady, připomínky a hodnou dávku enthusiasmu a odhodlání při nasazování **Ing. Pavlu Kordíkovi, Ph.D.**

Velký dík také patří **Bc. Jakubovi Krejčímu a Bc. Robertu Kotlářovi**, kteří se zasluhují o funkčnost prototypu datového skladu a při realizaci této práci byli nemálo nápomocní.

Ing. Tomáši Kalvodovi, Ph.D., Ing. Karlu Kloudovi, Ph.D. a Ing. Danielu Vašatovi, Ph.D. za velmi produktivní konzultace a cenné rady, tipy a triky sdělené pochopitelným způsobem.

Dále bych chtěla poděkovat **vedení FIT**, které mi umožnilo zpracovat tak zajímavé téma a bez jehož opakované podpory by realizace této práce nebyla vůbec možná. A také děkuji všem zainteresovaným zaměstnancům, kteří mi byli ochotně ku pomoci.

Velké poděkování patří i mému příteli **Bc. Jakubu Průšovi**, který mi byl po celou dobu oporou. Ze stejného důvodu děkuji i mé **rodině**.

Děkuji!

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval(a) samostatně a že jsem uvedl(a) veškeré použité informační zdroje v souladu s Metodickým pokynem o etické přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisů, zejména skutečnost, že České vysoké učení technické v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Praze dne 4. května 2016 .....

Odkaz na tuto práci

Friedjungová, Magda. *Predikce studijních výsledků studentů bakalářského programu Informatika FIT ČVUT.* Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2016.

Abstract

This thesis is about the extraction of data from faculty systems which are being used to record the results of students at the Faculty of Information Technology at CTU. The collected data is preprocessed and predictive models which determine the success rate of students in the first semester of the first year of the bachelor programme Informatics in the academic year 2015/2016 are composed using suitable methods. An analysis of student performance is done based on the results of the prediction and improvements are proposed. The thesis further contains descriptions of the methods and evaluation of the models so that they could be reused in the next academic year.

Keywords Data, educational data mining, preprocessing, predictive modelling, machine learning, data mining, analysis, data warehouse, decision trees.

Abstrakt

Tato práce se zabývá extrakcí dat z fakultních systémů, které jsou používány pro zaznamenávání studijních výsledků studentů na Fakultě informačních technologií ČVUT. Získaná data jsou dále předzpracována a pomocí vhodných metod jsou sestaveny prediktivní modely, které určují úspěšnost studentů v 1. semestru 1. ročníku bakalářského programu Informatika v akademickém roce

2015/2016. Na základě výsledků predikce je provedena analýza výkonů studentů a navržena zlepšující opatření. Práce dále obsahuje popisy postupů a vyhodnocení modelů s cílem jejich znovu použití v nadcházejícím akademickém roce.

Klíčová slova Data, educational data mining, předzpracování, prediktivní modelování, strojové učení, dolování dat, analýza, datový sklad, rozhodovací stromy.

Obsah

Úvod **1**

I Teoretická část **3**

1 Rešerše **5**
1.1 Problematika ve světě 6
1.1.1 Course Signals 6
1.1.2 Student Success System 7
1.1.3 Využití DM metod 8
1.2 Obecné problémy 11
1.3 Vhodná metodika 12

2 Porozumění problematice FIT **15**
2.1 Analýza předmětů 15
2.1.1 Hodnocení předmětů 16
2.1.2 Anomálie 23

3 Porozumění datům **27**
3.1 Datové zdroje 27
3.1.1 KOS 27
3.1.2 Přihláška ČVUT 28
3.1.3 Prototyp datového skladu ČVUT (DWH ČVUT) 28
3.1.4 EDUX 29
3.1.5 Progtest 29
3.1.6 Moodle 29
3.1.7 MARAST 30
3.2 Původ a struktura dat 30
3.2.1 Profil studenta 31
3.2.2 Studijní výsledky 35

II Implementační část **41**

4 Příprava dat **43**
4.1 Předzpracování dat . . . 43
4.1.1 Data z DWH ČVUT . . . 44
4.1.2 Data z Progtestu . . . 48
4.1.3 Data z EDUXu . . . 52
4.2 Problémy s daty . . . 56
4.2.1 Chybějící hodnoty . . . 56
4.2.2 Chybějící metadata . . . 56
4.2.3 Formát dat . . . 57
4.2.4 Povinné sloupce . . . 57
4.3 Normalizace hodnot . . . 57
4.3.1 Min-max normalizace . . . 57
4.3.2 Aplikace v předmětech . . . 59
4.4 Výběr vhodných dat . . . 60
4.4.1 Sestavení datasetu . . . 60

5 Prediktivní modelování **67**
5.1 Volba vhodného modelu . . . 68
5.1.1 Rozhodovací stromy . . . 68
5.1.2 Dvouatributový model . . . 71
5.1.3 Random Forest . . . 72
5.1.4 Kombinování modelů . . . 73
5.2 Získání výsledků . . . 74

6 Využití výsledků **79**
6.1 Oslovení studentů . . . 80
6.1.1 Mailing . . . 80
6.1.2 Dotazník . . . 88

7 Generalizace modelu **91**
7.0.1 Využití v dalších letech . . . 91

Závěr **97**

Literatura **99**

A Přílohy **103**

B Seznam použitých zkratek **107**

C Obsah přiloženého CD **109**

Seznam obrázků

1.1 Diagram CRISP-DM. [18]. 13

2.1 Prostupnost z 1. do 2. semestru BI. 24

4.1 Přesnost predikce v čase. 61
4.2 Transformace pro spojení souborů. 62

5.1 Grafická struktura rozhodovacího stromu CART. Indexy u terminálních uzlů udávají v jakém pořadí došlo k oddělení jednotlivých terminálních uzlů. Prediktoru X_1, X_2 a X_4 jsou spojité, prediktor X_3 je kategoriálná s kategoriemi A, B, C, D. [31] 69
5.2 Deset nejvýznamnějších atributů. 71
5.3 Rozložení bodových zisků z 1. testu BI-MLO a 1. testu BI-PS1 (BI-UOS). Červeně jsou znázorněni studenti, kteří nepostoupili do dalšího semestru, modře pak ti, kteří postoupili. 72
5.4 Distribuce dat v předmětu BI-PS1 v B101-B151. 73
5.5 Kombinování modelů. [29] 74

6.1 Funkce pro převod bodů na kredity. 85

A.1 Ganttův graf pro stanovení termínu sběru dat v B151. 104
A.2 Ukázka rozhodovacího stromu v nástroji BigML. 105

Seznam tabulek

2.1 Klasifikace předmětů 16
2.2 Pravidla hodnocení v BI-PA1 pro B101. 17
2.3 Pravidla hodnocení v BI-PA1 pro B111 a B121. 18
2.4 Pravidla hodnocení v BI-PA1 pro B131 a B141. 18
2.5 Pravidla hodnocení v BI-PA1 pro B151. 19
2.6 Prostupnost daných předmětů v semestrech B101-B151. 23

3.1 Atributy tabulky prih_prihlaska v DWH ČVUT. 35

4.1 Číselníky pro tabulku prih_prihlaska. 46
4.2 Obtížnost domácích úloh v B111. 48
4.3 Obtížnost domácích úloh v B121. 49
4.4 Obtížnost domácích úloh v B131. 50
4.5 Obtížnost domácích úloh v B141. 50
4.6 Znalostní testy v B111 - B141. 51
4.7 Bodové hodnocení v BI-CAO. 53
4.8 Bodové hodnocení v BI-MLO. 53
4.9 Bodové hodnocení v BI-ZMA. 54
4.10 Bodové hodnocení v BI-PS1. 55
4.11 Datová kvalita záznamů. 56
4.12 Popis jednotlivých kroků transformace join_files. 63
4.13 Popis jednotlivých atributů finálního datasetu 65

5.1 Přesnost predikce výsledků 1. testu. 72
5.2 Matice záměn obecně. 75
5.3 Matice záměn pro rozhodovací strom. 76
5.4 Matice záměn pro dvouatributový model. 76
5.5 Matice záměn pro Random Forest. 76
5.6 Přesnost predikce pro jednotlivé třídy a celková. 77

6.1 Velikost obeslaných/neobeslaných skupin. 83

6.2 Průměrná pravděpodobnost dokončení jednotlivých předmětů v B151. 84
6.3 Korelační matice průchodnosti předmětů. 84
6.4 Ukázka doporučení zaměření se na předmět BI-PS1 a BI-MLO na úkor předmětu BI-PA1. 85
6.5 Kontingenční tabulka pro vyhodnocení variant e-mailu. 86
6.6 Výsledky různých kombinací variant. 87

Úvod

Denodenně jsme obklopeni daty, mnohdy prázdnými sděleními, jindy pro nás zásadními informacemi. Data sama o sobě však nemají žádný význam. Význam jim přiřazují až právě lidé, kteří tak proměňují data v informace. Informací však bývá příliš mnoho, lidé jsou často přehlceni, přemíra informací brání efektivitě a může vést k chybnému rozhodování. Schopnost dávat věci do souvislostí (informace s přidanou hodnotou) nazýváme znalost. Znalosti jsou založené na zkušenostech, interpretaci, porozumění a schopnosti umět dát věci do souvislostí. Na základě znalostí je možné se efektivně rozhodovat a jednat. Znalosti jsou proto velmi cenné a na rozdíl od informací jich není nikdy dost.

Strojové učení, dolování dat (en. data mining, dále jen DM), získávání informací a hledání znalostí tak patří mezi jednu z nejpopulárnějších oblastí dnešní doby [1]. Skryté závislosti v datech mohou výrazně pomoci ve všech odvětvích - obchod, průmysl, medicína, strojírenství apod. Mezi jedno z odvětví patří zajisté školství. Dolování dat ve vzdělávacích instuticích, veřejných i soukromých, zažívá v posledních letech velký rozmach, což vedlo ke vzniku nového oboru zvaného Eduactional Data Mining (dále jen EDM). EDM pracuje s daty, která se týkají studentů a jejich studijních výsledků, za účelem zkvalitnění vzdělávání, přizpůsobení nabídky nebo pomoci s palčivými oblastmi studia [2].

Vysoké školství trpí v posledních letech úbytkem studentů ve studijních programech - v roce 2008 byla překročena hranice 50% neúspěšných studentů v bakalářských programech a v prvních ročnících studium neúspěšně ukončila více než třetina studentů [3]. Vhodná aplikace EDM dokáže identifikovat možné příčiny neúspěchu, předejít jim a navrhnout vhodná intervenční opatření, která by pomohla studentům v lepším plnění studijních povinností, čímž by se zvýšila celková studijní úspěšnost a zároveň by neklesla úroveň poskytovaného vzdělání umělým snižováním požadavků na studenty. EDM zahrnuje jak dolování dat, nacházení zajímavých trendů a pravidel, tak prediktivní mo-

delování, na základě kterého jsme schopni předpovědět např. akademickou úspěšnost jednotlivých studentů nebo vývoj celého ročníku. S využitím vysokoškolských dat za tímto účelem se v dnešní době setkáváme po celém světě a můžeme tak pozorovat pozitivní výsledky. Jedná se však o velmi mladé odvětví, takže EDM stále skýtá velký potenciál pro nové poznatky či zlepšení stávajících.

Cíl práce

Fakulta informačních technologií ČVUT (dále jen FIT) v posledních letech usiluje o kvalitní zpracování svých dat, jejich uskladnění a následné analýzy. Mezi jeden z hlavních problémů fakulty patří vysoký úbytek studentů v prvním semestru prvního ročníku v bakalářském programu Informatika (dále jen BI). Tato práce si klade za cíl dostupná data předzpracovat a sestavit prediktivní model, který bude schopen předpovědět, zda daný student projde do druhého semestru prvního ročníku BI. Díky této predikci bude možné identifikovat studenty s nízkou pravděpodobností postupu ve studiu (v této práci označeni jako „ohrožení“), tzn. takové studenty, kteří jsou ve studiu aktivní, ale k získání potřebného minima pro průchod mezi semestry jim vlastní síla z jakýchkoliv důvodů nestačí.

Vedení fakulty vidí pomoc těmto ohroženým studentům jako jedno z hlavních řešení vysoké úmrtnosti mezi semestry. Mezi další řešení pak patří zkvalitnění výuky, doučovací kurzy či podpůrné předměty a v neposlední řadě také správné nastavení podmínek přijímacího řízení. Dále budou pomocí predikce identifikováni studenti, kteří mají velmi nízkou či téměř žádnou pravděpodobnost na postup do druhého semestru. Tato skupina studentů je pro vedení fakulty také velmi důležitá - mohla by pomoci zodpovědět otázku, z jakých důvodů bylo jejich studium neúspěšné. Prostupnost mezi semestry je také důležitým tématem z pohledu financování a plánovaní lidských zdrojů. Na základě stanovení počtu postupivších lze lépe připravit rozvrh na další semestr, finanční zatížení fakulty a mnoho dalších záležitostí.

Tato práce je členěna na teoretickou část, která zahrnuje první tři kapitoly, zabývající se rešerší v oblasti EDM a sestavováním prediktivních modelů, dále analýzou domény (FIT ČVUT) a seznámením se s dostupnými daty. Následuje implementační část, která zahrnuje kapitolu časově nejnáročnější - předzpracování dat, která se zabývá extrakcí, přípravou a řešením problémů týkajících se dat potřebných pro prediktivní modelování. V následující kapitole je popsán návrh a sestavení prediktivních modelů, jejich testování a vyhodnocení. Dále práce popisuje využití výstupů z prediktivních modelů pro oslovení studentů. V poslední kapitole je diskutována možnost opakovaného využití sestavených modelů.

Část I

Teoretická část

Kapitola 1

Rešerše

Prediktivní modelování je hojně využíváno nejen v oblastech obchodu, medicíny apod., ale i ve školství. Mnoho univerzit se postupně učí, jak data o studijních výsledcích využívat jak ku prospěchu vedení, tak ku prospěchu studentů. Jednou z možností je sestavení prektivních modelů, pomocí kterých lze předpovídat úspěšnost studentů a prostřednictvím správných opatření těm méně úspešným ve studiu vhodně pomoci.

Tyto prediktivní modely mohou být využity v informačních systémech, které si kladou za cíl např. včasné upozornění studentů na nedostatečné plnění studijních povinností, upozornění garantů předmětů na studenty s nedostatečným hodnocením nebo mohou poskytovat důležité real-time informace pro vedení fakult a univerzit a významně tak pomáhat v rozhodovacím procesu.

Vizualizace dat pomocí vhodných vizualizačních nástrojů (např. dashboardů) může usnadnit komunikaci a vyhodnocování řady trendů mezi studenty, např. jak studenti postupují ve vztahu k jejich spolužákům, kteří např. sdílejí stejné vstupní podmínky. Vhodné nástroje mohou umožnit rychlé rozeznání vzorů v datech a jejich kombinace s prediktivními modely pak může univerzitě velmi pomoci např. v poznání svých zájemců či v modelování budoucího vývoje [4].

Kombinace dat z různých oblastí (demografické, behaviorální, studijní aj.) poskytuje dostatečně rozmanité atributy ke sledování pokroků studentů a k sestavení předpovědi o jejich budoucím výkonu. Sestavené prediktivní modely by měly být schopny zodpovědět např. tyto otázky:

- Bude student A schopný překonat hranici 15 kreditů, aby postoupil do dalšího semestru, bez pomoci?
- Kolik studentů tento semestr jednoznačně uspěje?
- Kolik studentů vzdalo studium během semestru?

Odpovědi na tyto a mnohé další otázky zajímají i FIT, který by mimo jiné uvítal i nástroj pro sledování průběžných studijních výsledků. Sestavení prediktivního modelu je prvním krokem jak ukázat, že i tento způsob využití dat může mít smysl.

1.1 Problematika ve světě

Na jiných univerzitách, především zahraničích, hojně probíhá vývoj prediktivního modelování a podpůrných informačních systémů pro studenty a vyučující. Univerzitní systémy disponují potenciálem pro dolování informací o výkonech studentů napříč časem.

Největší rozvoj EDM můžeme sledovat v USA, kde jsou ročně utraceny miliardy amerických dolarů pro posílení vzdělávacího systému. Dle statistik [5] více než 35% studentů nikdy nedokončí střední školu (u některých demografických skupin se jedná dokonce o 50-60%), na vysokoškolské úrovni pouze 30% úspěšně absolvuje 2letý program a 50% 3letý program. Snaha dosáhnout lepšího výkonu či úspěšnosti studentů je tedy opodstatněná.

Níže budou popsány různé systémy a výzkumy, které se snaží zvýšit akademický výkon studentů pomocí EDM.

1.1.1 Course Signals

Course Signals (dále jen CS) pochází z univerzity v Purdue [6] a byl zřejmě úplně prvním projektem, který prokázal, že studentovy e-learningové aktivity mohou mít prediktivní schopnost a mohou tak být využity pro pozitivní ovlivnění studijních výkonů.

Prediktivní algoritmus studentova úspěchu (en. student success algorithm, dále jen SSA) je aplikován na žádost vyučujícícho. CS doluje data z různých univerzitních systémů, jedním z nich je Blackboard Vista, portál, který univerzita využívá pro řízení výuky (en. learning management system, dále jen LMS). SSA se skládá ze 4 komponent:

- výkonnost - procentuální bodový zisk doposud,
- snaha - studentova interakce s LMS systémem Blackboard Vista v porovnání se spolužáky,
- studijní výsledky před nástupem na univerzitu včetně výsledků z přijímacího řízení,
- údaje o studentovi (např. věk, bydliště aj.).

Studentova šance na úspěch tedy není vyhodnocena na základě jednoho, ale kombinací více faktorů. Každá komponenta je ohodnocena a předána algoritmu. Výstup z SSA je studentům doručen prostřednictvím personalizovaného e-mailu se specifickou barvou, která reprezentuje jeho úspěšnost v daném kurzu. Barvy reflektují standardní semaforová světla se stejným významem. Dále je toto světlo zobrazeno na studentově webové stránce v LMS v daném kurzu.

Projekt tedy identifikuje 3 úrovně rizika:

- červené světlo - vysoká pravděpodobnost, že bude student v kurzu neúspěšný,
- žluté světlo - možný výskyt problémů, tzn. existuje riziko pro nedokončení kurzu,
- zelené světlo - vysoká pravděpodobnost na úspěch v daném kurzu.

Na základě výsledků SSA vyučující může vytvořit intervenčí plán, který většinou obsahuje:

- zobrazení upozorňujícího signálu na studentově domovské stránce,
- personalizované e-maily a upomínky,
- textové zprávy,
- předání studenta studijnímu poradci,
- konzultace vyučujícího se studentem.

V době publikace studie (r. 2012) bylo využití CS na univerzitě v Purdue hodnoceno pozitivně spolu s prokázáním pozitivního vlivu na studijní výsledky ohrožených studentů. Podle údajů zveřejněných univerzitou se hodnocení v 6letém studjiním programu zvedlo o 21,48% [7]. Využití CS se ukázalo jako nejefektivnější v 1. a 2. ročníku studia. Pomoc studentům v tomto období kladně ovlivnila jejich přístup ke studiu i v budoucnu. Celkem bylo CS ovlivněno přes 23 000 studentů ve více než 100 kurzech a využilo ho přes 140 vyučujících [6]. CS pořád skýtá potenciál pro rozvoj a vylepšení SSA algoritmu. Projekt bohužel neposkytuje žádný pohled na příčiny (diagnostické informace) svých výsledků.

1.1.2 Student Success System

Student Success System (dále jen S3) byl vyvinut společností Desire2Learn Inc. za účelem poskytnout komplexní analytický pohled na studentův akademický výkon. Narozdíl od CS tento projekt nabízí rozšíření o diagnostické

informace k daným výsledkům [5]. Jádrem S3 je flexibilní modelovací nástroj, který využívá strojové inteligence a statistických metod s cílem určit ohrožené studenty. Dále nabízí možnost pochopení, proč jsou v ohrožení, navrhuje zásahy ke zmírnění ohrožení a poskytuje zpětnou vazbu sledováním účinnosti aplikovaného zásahu. S3 se zaměřuje na poskytnutí zobecněné modelovací strategie, která se hodí pro podporu rozsáhlých potřeb vzdělávacích institucí a pro plné využití prediktivní analýzy.

S3 je založeno na metodách skládání modelů (en. ensemble methods) [5]. Tyto metody jsou navrženy tak, aby zvýšily prediktivní zobecnitelnost sloučením předpovědí více modelů. Stohování (en. stacking) pořízených dat, adaptivní posílení a jiné související ensemblovací techniky jsou úspěšně používány v mnoha oblastech ke zvýšení přesnosti predikce nad úroveň získanou jakýmkoliv jedním modelem [8].

S3 kombinuje skládání modelů s dekompozicí na sémantické jednotky, kde každá jendotka má význam v učícím se procesu. Ukázalo se, že dekompozice je velice flexibilní metodou pro případnou generalizaci modelu napříč rozdílnými kontexty. Tato dekompozice spojená s vizualizací dat a vhodným nástrojem pro koordinaci spolupráce (en. case management) umožňuje rozšíření predikcí o tvorbu personalizovaných zásahů [5]. S3 je převratné v překonání propasti mezi prediktivním modelováním a následnou akcí.

1.1.3 Využití DM metod

Jako ukázka konkrétního využití data miningových metod dobře poslouží studie provedená v Turks & Caicos Islands Community College [9] , která se zabývala porovnáním výkonů studentů mezi semestry.

Pro prediktivní modelování bylo k dispozici celkem 2215 záznamů, ze kterých po pečlivém zvážení důsledků neznámých a chybějících dat zbylo pouze 1369 (61,82%) záznamů vhodných pro modelování. Chybějící a nekompletní data byla stále zahrnuta v datasetu kvůli důležitosti učení se z neúplných údajů. Operační data se skládala ze studentových demografických informací a výsledků plnění 5 odlišných kurzů během 4 semestrů. Na základě analýzy všech studentových činností byl sestaven jeden prediktivní model.

Testovány byly následující metody: ANN[1], CART[2], rozhodovací strom (algoritmus C5.0) a CHAID[3]. Tedy techniky učení s učitelem. Při porovnání použitých algoritmů vycházela nejlépe technika rozhodovacího stromu (C5.0) s průměrnou přesností 97,3% a nejhůře CHAID s 57,8% přesností.

[1] Artificial Neural Network
[2] Classification And Regression Tree
[3] Chi-squared Automatic Interaction Detector

Pomocí prediktivního modelování byli studenti rozřazeni do následujících skupin: propadající, na hraně, dostateční, dobří, vynikající, s chybějící klasifikací a skupina studentů bez predikce. Bylo zjištěno, že k poklesu výkonu docházelo především ve druhém semestru, ve třetím a čtvrtém semestru byl výkon studentů již stabilizován.

Dalším příkladem může být výzkum provedený na univerzitě v Mexiku (UNAM) [10], který měl za cíl sestavit prediktivní model studijní úspěšnosti studentů prvního ročníku technického zaměření. Dataset obsahoval záznamy z prvního semestru studia napříč generacemi studentů. Dataset byl rozdělen na skupiny podle počtu splněných předmětů - žádný nebo až 2 předměty (slabší studenti), tři nebo čtyři předměty (průměrní studenti) a více než 5 splněných předmětů (výborní studenti). Využitím techniky pro dolování dat, konkrétně naivní bayesův klasifikátor, byl sestaven model s 60% přesností. Tento model byl použit pro predikci studijních výsledků a po následné validaci byla zjištěna 50% úspěšnost správné klasifikace. Nicméně u některých skupin model dosahoval přesnosti vyšší i než 70%.

Trénovací dataset obsahoval 70% dat a testovací 30% dat. Optimalizací byl získán model s 60% přesností. Celkem byl k dispozici dataset o velikosti 6584 záznamů. Na první pohled se zdá přesnost modelu nízká. Pokud se ovšem na tuto hodnotu podíváme z více úhlů, zjistíme, že je oprávněná a ve své podstatě velmi dobrá vzhledem k řešenému problému rozřazení studentů dle predikovaných výkonů.

Na Can Tho University v Thajsku [11] se zabývali porovnáním přesnosti rozhodovacího stromu a bayesovské sítě při predikci studijních výsledků vysokoškolských a postgraduálních studentů na dvou velmi rozdílných institucích: Can Tho University (dále jen CTU), což je velká státní univerzita ve Vietnamu, a Asian Institute of Technology (dále jen AIT), což je malý postgraduální institut v Thajsku, který zahrnuje studenty z 86 zemí. Vzhledem k rozdílnosti obou skupin bylo porovnání schopnosti algoritmů dosáhnout stejné úrovně přesnosti zajímavou výzvou. V rámci studie na CTU šlo o predikci výkonu ve 3. ročníku na konci 2. ročníku, v rámci AIT se jednalo o predikci výkonu na konci prvního ročníku na základě údajů z přihlášky.

Sestavené modely dosáhly 70-90% přesnosti. Byly používány pro identifikaci a pomoc méně schopným studentům (64% přesnost) nebo pro výběr velmi dobrých studentů, kteří by mohli mít nárok na prospěchové stipendium (82% přesnost). Pro vyhodnocení přesnosti predikce byla využita křížová validace. Výsledky této studie prokázaly, že algoritmus rozhodovacího stromu je pro danou problematiku přesnější až o 12% než algoritmus bayesovské sítě.

V Belgii byla provedena analýza studijních výsledků prvního ročníku belgických frankofonních vysokých škol [12], díky které bylo zjištěno, že cca 60% studentů bylo ve svém studiu neúspěšných. Už v roce 2001 v zemi pozorovali neměnící se trend ve studijní úspěšnosti v posledních 10 letech.

V rámci této analýzy byl vytvořen seznam faktorů, které by mohly být příčinami úspěchu či neúspěchu ve studiu. Vznikly 3 sady faktorů:

1. Faktory vztahující se k osobní historii studenta - jeho totožnost, sociální minulost, akademická minulost apod.

2. Faktory popisující důsledky studentova chování - účast na volitelných aktivitách, konzultace s vyučujícími, zpětná vazba při kontrole úkolů apod.

3. Faktory týkající se studentova subjektivního vnímání akademického kontextu, vyučujících, kurzů apod.

Během studie byl vytvořen dotazník, který poskytl velké množství informací o určitém počtu studentů. Na základě tohoto dotazníku byla navržena databáze, která obsahuje profily studentů. Predikce akademického úspěch pak probíhá na základě dolování dat z této databáze.

Pro prediktivní modelování byly použity tyto metody: rozhodovací strom, Random Forest, neuronové sítě a lineární diskriminační analýza. Dosažené přesnosti modelů se pohybovaly kolem 50%. Takto malá přesnost byla pravděpodobně zapříčiněna nedostatečnou velikostí datasetu a sběrem dat ze 3 různých institucí.

Nicméně zabývání se faktory, které mají přímý vliv na studentovu akademickou úspěšnost, je velmi důležité z důvodu pochopení a hledání příčin výsledků prediktivního modelování. Jedna z mnoha studií [13] jasně poukazuje na velký vliv socio-ekonomického zajištění studenta - zda dotyčný pochází z města či vesnice, zda se univerzita nachází ve velkém městě, z jaké společenské třídy student pochází a jaký má vztah se svými vyučujícími. Tyto faktory údajně patří mezi jedny z nejvlivnějších. Zároveň se ukázalo, že socio-ekonomické podmínky mají větší vliv na ženy než na muže.

Například socio-ekonomický stav studenta na začátku studia hraje velmi důležitou roli - pokud je úroveň nízká, má student menší šanci na úspěch oproti studentovi s vysokou úrovní. Avšak tato úroveň neříká nic o studentových schopnostech, a tak se student s velmi nízkou úrovní může během semestru projevit jako velmi nadaný, a naopak.

Na základě těchto studií stojí za zvážení, jaké atributy máme k dispozici, jaké informace můžeme z dostupných dat dále vydolovat, tedy vytvořit nové atributy, viz [14], a zda jsme schopni od studentů získat zcela nové informace, např. pomocí dotazníku, a o jaký typ informací by se mělo jednat, např. socio-ekonomické, spokojenost s výukou apod.

Závěrem lze říct, že obdobné výzkumy probíhají po celém světě a snaží se zodpovědět stejné otázky jako v této práci. Mezi nejvíce využívanou metodu patří rozhodovací stromy, které dosahují uspokojivých výsledků, které lze jednoduše interpretovat. Rostoucí oblibě se těší neuronové sítě, avšak interpretace jejich výsledků není tak jednoznačná a pro tento problém se zdají řešením předimenzovaným. Nicméně existují doslova tisíce možností a s každým rokem přibývají. Častým problémem je tedy volba správného učícího algoritmu. Všechny algoritmy jsou víceméně kombinací následujících tří komponent:

- reprezentace,
- vyhodnocení,
- optimalizace.

V počátcích strojového učení měl každý svůj oblíbený přístup k učení spolu s důvody, proč zrovna ten jeho je nejlepší. Časem se začalo od metody hledání nejlepšího řešení upouštět a přiklánět se ke kombinaci několika modelů, která dává mnohem lepší výsledky než jeden model [1]. Volba vhodné metody záleží vždy na analýze domény, dostupných datech a účelu.

1.2 Obecné problémy

Současné přístupy v prediktivním modelování v rámci EDM se potýkají se dvěma vážnými omezeními. Za prvé jsou prediktivní modely jednorázové a nelze je snadno přenést z jednoho prostředí do jiného. Nemůžeme tedy jednoduše předpokládat, že prediktivní model vyvinutý pro konkrétní kurz na konkrétní univerzitě bude fungovat i pro další předměty. Otázkou tedy je, zda a jak můžeme navrhnout flexibilní a škálovatelnou metodiku pro tvorbu prediktivních modelů, které pojmou značnou variabilitu přes různé kurzy, semestry nebo instituce. A za druhé stávající modelovací přístupy, přestože generují platné předpovědi, neposkytují dostatek informací pro tvorbu smysluplných personalizovaných zásahů [5]. Častým problémem je také nezohlednění lidského pohledu na věc, který by mohl model ladit v případě potřeby.

Macfadyen a Dawson diskutovali omezení generalizovaného přístupu (např. u Course Signals) [15] - především jeho celkovou zobecnitelnost a interpretaci. Zejména zobecnitelnost takových modelů je omezena vzorky dat použí-

vaných pro modelování. Toto omezení jednoznačně brání rozsáhlému nasazení objeveného modelu ve vzdělávacích institucích smysluplným způsobem. Dále omezuje potenciální výhody, které by instituce mohly čerpat ze svých dat rozvojem prediktivních analýz.

Výzkumy dokazující stimulaci a urychlení učení cílenou interakcí se studentem jsou teprve v počátcích. Zajímavé výsledky mohou přinést i analýzy sociálních sítích, které mohou poskytnout vhled do studijní komunity a odhalit zajímavé vzory vzájemných interakcí.

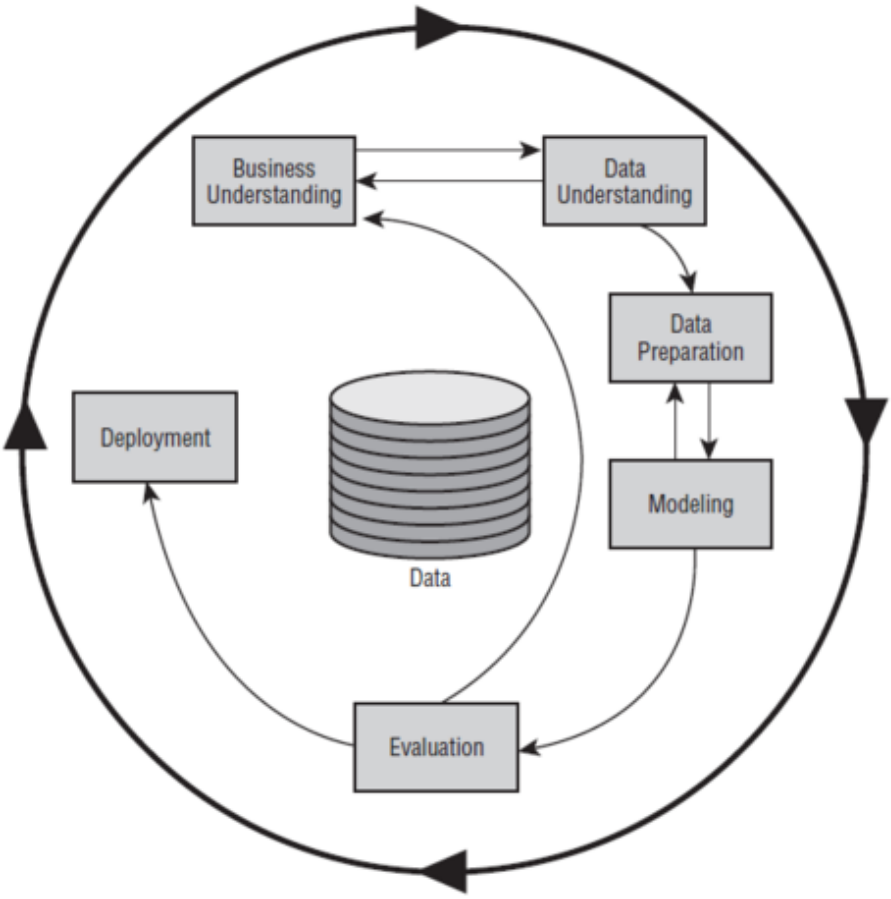
1.3 Vhodná metodika

Obecně se v dolování dat používá ustálený postup (např. [16]), jak znalosti získávat, který se skládá z těchto bodů:

- seznámení se s datovými zdroji,
- původ a struktura dat,
- výběr vhodných dat,
- předzpracování dat,
- sestavení datasetu,
- výběr vhodné metody,
- prediktivní modelování,
- dolování znalostí,
- interpretace a vyhodnocení výsledků,
- využití výsledků.

Přístupů a metodik v oblasti data miningu existuje celá řada. Velmi populárním přístupem je CRISP-DM [4] metodika [17], která se skládá z 6 částí:

- porozumění problematice (en. Business Understanding),
- porozumění datům (en. Data Understanding),
- příprava dat (en. Data Preparation),
- volba vhodného modelu (en. Modeling),
- získání výsledků (en. Evaluation),
- využití výsledků (en. Deployment).



Obrázek 1.1: Diagram CRISP-DM. [18].

Posloupnost jednotlivých částí je vidět na obrázku 1.1.

V CRISP-DM metodice jsou zahrnuty i body z obecného postupu pro dolování dat [19] jako jednotlivé úkoly. V této práci bude CRISP-DM hlavní použitou metodikou, jelikož nejlépe reflektuje postup nutný při řešení zadaného problému a ve světě dolování dat je již jakýmsi standardem. S jejím použitím jsme se také setkali v některých studiích [20] [21]. Tato metodika je také vyučována na FIT [22].

[4]CRoss-Industry Standard Process for Data Mining

Kapitola 2

Porozumění problematice FIT

Akademický rok je na ČVUT dělen na dvě části - zimní (první) a letní (druhý) semestr. Řádná délka bakalářského studia programu Informatika na FIT je stanovena na tři roky s možností prodloužení na čtyři roky. V prvním semestru BI absolvují všichni studenti stejný blok povinných předmětů, který je společný pro všechny obory. Na každého studenta jsou tedy během semestru kladeny stejné požadavky jako na jeho spolužáky. FIT, stejně jako ostatní fakulty ČVUT, podléhá Evropskému systému přenosu a akumulace kreditů (en. European Credit Transfer and Accumulation System, dále jen ECTS) [23], což znamená, že jednotlivé předměty jsou ohodnoceny různým počtem kreditů, kdy jeden kredit v průměru odpovídá 30 hodinám studijní zátěže průměrného studenta. Za studium bakalářského programu by měl student absolvovat předměty v celkové hodnotě 180 kreditů.

Každému předmětu je tedy přidělen kreditový zisk podle jeho náročnosti. Tyto kredity student získá za úspěšné absolvování předmětu (= splnění podmínek hodnocení v předmětu). V prvním semestru BI jsou studentům automaticky zapsány předměty s celkovým ziskem 30 kreditů. Pro úspěšný postup do druhého semestru musí student získat alespoň 15 z těchto kreditů. Pokud student získá méně kreditů, je jeho studium po prvním semestru ukončeno a prohlášeno za neúspěšné.

2.1 Analýza předmětů

Fakulta informačních technologiích vznikla v roce 2009, je tedy relativně mladou fakultou. V roce 2015 bylo nutné znovu akreditovat studijní obory BI. Tato reakreditace se nijak zásadně nedotkla předmětů vyučovaných v prvním semestru BI - předměty v tomto semestru po obsahové stránce zůstávají víceméně stejné od roku 2009. Bohužel často dochází ke změně vnitřní struktury předmětů - napříč roky se mění počet testů v semestru, bodové hodnocení,

podmínky získání zápočtu apod.

Jako první krok je proto nevyhnutelné seznámit se s podmínkami hodnocení v předmětech v každém roce, porovnat je s předchozími běhy a zvážit použití vhodných normalizačních metod kvůli schopnosti srovnání různých běhů.

Jak již bylo řečeno výše, v prvním semestru BI absolvují všichni studenti stejný blok povinných předmětů, který je společný pro všechny obory. Níže jsou stručně popsány jednotlivé předměty s podmínkami hodnocení pro konkrétní akademické roky, v našem případě od akademického roku 2010/2011 po rok 2015/2016 (aktuální). Přestože fakulta zahájila provoz akademickým rokem 2009/2010, nemáme z tohoto období k dispozici potřebná data z určitých zdrojových systémů, jak bude zmíněno v následujících kapitolách. Podmínky hodnocení v předmětu z aktuálního akademického roku lze nalézt na portále EDUX [24] vždy v sekci Hodnocení v konkrétním předmětu. Pro předměty z jiných akademických let lze tyto informace nalézt v archivu portálu EDUX [25]. Vzhledem k tomu, že se v pravidlech vyskytují různé anomálie, byla tato pravidla ještě diskutována s jednotlivými garanty předmětů, kteří za celý předmět, včetně jeho hodnocení, ručí.

2.1.1 Hodnocení předmětů

Každý předmět může být zakončen těmito způsoby: zápočtem, klasifikovaným zápočtem, zkouškou a nebo zápočtem a zkouškou. Zápočet a klasifikovaný zápočet jsou studentovi uděleny za jeho aktivitu během semestru (plnění domácích úkolů, docházka apod.), zkouška (podmíněná či nepodmíněná zápočtem) se uděluje ve zkouškovém období podle výkonu, který student podal přímo na zkouškovém termínu. Výsledná známka (klasifikace) je studentovi udělana podle celkového bodového zisku v rámci předmětu dle tabulky 2.1.

	A	**B**	**C**	**D**	**E**	**F**
Bodové rozmezí	≥ 90	89–80	79–70	69–60	59–50	< 50
Známka	1	1,5	2	2,5	3	4
Slovně	výborně	velmi dobře	dobře	uspokojivě	dostatečně	nedostatečně

Tabulka 2.1: Klasifikace předmětů [26]

V případě předpovědi, zda student postoupí do dalšího semestru, během probíhajícího semestru, např. v 6 týdnu, nemáme k dispozici informaci o studentově úspěchu či neúspěchu u zkoušky, která proběhne až ve zkouškovém období. Máme tedy k dispozici pouze údaje do 6. týdne studentova studia,

a tak je bezpředmětné se v analýze zabývat pravidly hodnocení u zkoušky. Analýza pravidel hodnocení zkoušky by byla důležitá, pokud by cílem byla např. předpověď konkrétních známek v jednotlivých předmětech. Naším cílem je prediktivní model určující, zda student postoupí do dalšího semestru, a pro jeho sestavení je nutné pracovat s údaji známými v době jeho nasazení.

Hlavním cílem analýzy předmětů je pochopení pravidel hodnocení studentů v jednotlivých předmětech během semestru.

Pro označení jednotlivých semestrů bude použit následující ustálený formát zkratek:

- B101 - zimní semestr v akademickém roce 2010/2011,
- B102 - letní semestr v akademickém roce 2010/2011,
- B111 - zimní semestr v akademickém roce 2011/2012,

a analogicky dále.

2.1.1.1 Programování a algoritmizace 1 (BI-PA1)

ECTS kreditů: 6
Týdenní rozsah výuky: 2 přednášky + 2 proseminář + 2 cvičení

Předmět je zakončen zápočtem a zkouškou. Pro úspěšné absolvování předmětu je nutné získat obojí. Zápočet student získává za práci během semestru. Pokud student nezíská v semestru alespoň stanovené minimum bodů (v každém semestru se liší) nebo z dané aktivity nedosáhne minimálního počtu bodů (je-li minimum stanoveno), nemá nárok na zápočet a musí si předmět BI-PA1 zopakovat. Pokud student dosáhne alespoň minima bodů, získá zápočet a je připuštěn k absolvování zkoušky.

V semestru B101 byl minimální bodový zisk pro udělení zápočtu stanoven na 30 bodů za aktivity zmíněné v tabulce 2.2.

Aktivita	**Počet aktivit * max. počet bodů**
Domácí úloha	7 * 5 = 35
Soutěžní úloha	1 * 15 = 15
Test u počítače v semestru	1 * 25 = 25
Semestrální úloha	1 * 20 = 20

Tabulka 2.2: Pravidla hodnocení v BI-PA1 pro B101.

V semestru B111 byla zvýšena hranice minimálního bodového zisku na 40 bodů a také je vidět změna v možných aktivitách v tabulce 2.3 - oproti B101 přibyla 1 domácí úloha, snížil se bodový zisk z testu u počítače a semestrální úloha byla nahrazena 4 znalostními testy u počítače. Tato pravidla zůstala stejná i pro semestr B121.

Aktivita	**Počet aktivit * max. počet bodů**
Domácí úloha	8 * 5 = 40
Soutěžní úloha	1 * 15 = 15
Test u počítače v semestru	1 * 20 = 20
Znalostní test u počítače	4 * 5 = 20

Tabulka 2.3: Pravidla hodnocení v BI-PA1 pro B111 a B121.

Velmi důležitou změnou, která byla provedena v semestru B111 a která přetrvala až do semestru B151 včetně, bylo přidání lehké a těžké varianty ke každé domácí úloze, viz 2.4. Úloh v hodnocení je však stále 8, protože se započítává pouze ta varianta úlohy, ve které bylo dosaženo vyššího bodového zisku.

Pro semestr B131 a B141 archiv portálu EDUX neobsahuje žádná pravidla hodnocení. Po dohodě s garantem předmětu jsou však následující: Položka s testem u počítače v semestru byla stanovena jako součást zkouškového zadání a nemá vliv na zisk zápočtu. Dále byl o polovinu snížen bodový zisk ze znalostních testů u počítače. Tyto změny jsou viditelné v tabulce 2.4. V důsledku těchto změn musel student získat minimálně 25 bodů pro udělení zápočtu.

Aktivita	**Počet aktivit * max. počet bodů**
Domácí úloha 1-4	4 * 5 = 20
Domácí úloha 5-8	4 * 5 = 20
Soutěžní úloha	1 * 15 = 15
Znalostní test u počítače	4 * 2,5 = 10

Tabulka 2.4: Pravidla hodnocení v BI-PA1 pro B131 a B141.

V semestru B151 oproti B141 přibyla pouze jedna domácí úloha č. 0, tzv. zahřívací, s 1 bodovým ohodnocením a bez variant (lehká/těžká). Zbytek aktivit

zůstal netknutý, stejně tak i minimum pro udělení zápočtu.

Aktivita	Počet aktivit * max. počet bodů
Domácí úloha 0	1 * 1 = 1
Domácí úloha 1-4	4 * 5 = 20
Domácí úloha 5-8	4 * 5 = 20
Soutěžní úloha	1 * 15 = 15
Znalostní test u počítače	4 * 2,5 = 10

Tabulka 2.5: Pravidla hodnocení v BI-PA1 pro B151.

Dále bylo v předmětu možné získávat body za aktivitu během cvičení, které však nejsou průběžně evidovány v žádném systému a jsou přičítány až k hodnocení na konci semestru. Z tohoto důvodu tyto body nebudeme brát v potaz.

Shrnutí
V předmětu BI-PA1 byla v semestrech B111-B151 téměř konstantní struktura, co se týče počtu hodnocených aktivit. Změny v počtu bodů za danou aktivitu lze ošetřit vhodnou normalizací. Přidání 0. domácí úlohy v semestru B151 lze zanedbat vzhledem k tomu, že oproti ostatním rokům přebývá a nemáme k ní tedy žádné informace, které by se daly pro prediktivní model využít. Semestr B101 se od ostatních liší výrazně - chybějící znalostní testy, nerozdělení domácích úloh na lehkou/těžkou variantu, o jednu domácí úlohu méně a existence semestrální úlohy.

2.1.1.2 Základy matematické analýzy (BI-ZMA)

ECTS kreditů: 6
Týdenní rozsah výuky: 3 přednášky + 2 cvičení

Předmět je zakončen zápočtem a zkouškou, pro úspěšné absolvování předmětu je nutné získat obojí. Během semestru se na cvičeních píší zápočtové písemky a student může získat body za aktivitu u tabule. Na cvičení je povinná docházka, resp. je povolena z pravidla 1-2 absence, každá další je penalizována -3 body. Pokud student získá požadované minimum bodů ze semestru, obdrží zápočet a je připuštěn ke zkoušce.

V semestru B101 se psaly 4 zápočtové písemky, každá za 10 bodů. Aktivita

u tabule byla hodnocena 1 bodem a student takových bodů mohl získat maximálně 10. Pro udělení zápočtu musel student získat minimálně 20 bodů.

Semestr B111 prošel následujícími změnami: Během semestru se uskutečnily 3 zápočtové písemky, z toho první byla za maximálně 10 bodů, druhá a třetí za 15 bodů každá. Během semestru byly zadány 3 nepovinné domácí úlohy, každá za 3 body. Hodnocení aktivity u tabule a minimální bodový zisk pro zápočet zůstaly beze změny.

V semestru B121 se psaly pouze 2 zápočtové písemky, každá za 20 bodů. Domácí úlohy byly zrušeny, body za aktivitu a spodní hranice pro zápočet zůstávají stejné. Přibyla možnost opravné písemky pro ty studenty, kteří získali ke konci semestru minimálně 10 bodů, ale méně než 20 bodů - pokud student z tohoto testu získal minimálně 10 bodů z 20, byl mu udělen zápočet.

V B131 a B141 zůstávají podmínky stejné jako v B121, pouze se snížil počet bodů za aktivitu z 10 na 5.

Semestr B151 prodělal několik zásadních změn. Předmět obohatila nová aktivita - on-line kvízy, které byly celkem 3 a studenti na vypracování každého měli cca týden. Za každý splněný kvíz pak mohli získat až 8 bodů, za vyplněnou polovinu 4 body. Během cvičení se psali 3 zápočtové písemky, každá za maximálně 12 bodů. Studenti mohli dále získat 5 bodů za aktivitu u tabule a 5 bodů za excelentní řešení kvízu. Pro udělení zápočtu bylo nutné získat minimálně 30 bodů, z toho alespoň 12 ze zápočtových písemek.

Shrnutí
Za nejvíce podobné či téměř stejné se dají považovat semestry B121, B131 a B141. Vzhledem k tomu, že se v ostatních bězích předmětu mění nejen maximální celkový bodový zisk ze semestru, ale také počet testů, bude nutné vhodnou normalizaci probrat s garantem předmětu. Po obsahové stránce se předmět nijak nelišil.

2.1.1.3 Programování v shellu 1 (BI-PS1)

ECTS kreditů: 5
Týdenní rozsah výuky: 2 přednášky + 2 cvičení

Tento předmět je zakončen klasifikovaným zápočtem, což znamená, že pro jeho úspěšné absolvování je nutné získat minimálně 50 bodů během semestru. BI-PS1 se dříve jmenovalo Úvod do operačních systémů (BI-UOS), v B141 došlo k přejmenování na BI-PS1, nicméně průběh předmětu je totožný s BI-UOS. V textu bude používána pouze aktuální zkratka BI-PS1.

Během semestru B101 se psali 3 menší testy po 15 bodech a jeden zápočtový (závěrečný) test za 55 bodů. Aby mohl student psát zápočtový test, musel z malých testů získat minimálně 20 bodů.

V semestru B111 došlo k úpravě bodového rozložení na 2 testy po 15 bodech, 1 test za 20 bodů a zápočtový test za 50 bodů. Podmínka 20 bodů pro psaní zápočtového testu zůstala stejná.

Semestr B121 také prošel přeskupením - 1 test za 15 bodů, 2 testy po 20 bodech a za zápočtový test bylo možné získat 45 bodů. Podmínka 20 bodů pro psaní zápočtového testu opět zůstala.

V semestru B131 zůstalo bodové hodnocení testů stejné jako v B121. Pouze podmínka pro psaní zápočtového testu se snížila na minimum 5 bodů.

V B141 došlo opět k úpravě bodového rozložení - studenti měli možnost psát 3 testy po 20 bodech a zápočtový test za 40 bodů. O podmínce pro psaní zápočtového testu není v tomto semestru řečeno nic.

Hodnocení v semestru B151 se skládalo ze 4 testů po 25 bodech. Po skončení semestru měli ti studenti, kteří z předchozích testů nedosáhli v součtu 50 bodů, možnost absolvovat další test za 25 bodů, který nahradil nejhorší výsledek testu ze semestru. Během semestru měli studenti také možnost získat 5 bodů za vyřešení domácích úloh zadaných na cvičeních nebo přednáškách.

Shrnutí
Napříč běhy se neměnil celkový bodový zisk ze semestru, pouze jednotlivé testy se liší o +-5 bodů, ale jejich počet zůstává stejný.

2.1.1.4 Matematická logika (BI-MLO)

ECTS kreditů: 5
Týdenní rozsah výuky: 2 přednášky + 1.5 cvičení

Předmět je zakončen zápočtem a zkouškou, pro úspěšné absolvování je nutné získat obojí. Zápočet student získává za svoji aktivitu během semestru, která je popsána níže.

V semestru B101 psali studenti na cvičeních celkem 5 malých testů, každý za 6 bodů. Každý student mohl získat body za aktivitu, vyučující na cvičení mohl během celého semestru rozdat celkem $2*n$ bodů, kdy n je počet studentů zapsaných na dané cvičení. Kromě testů a bodů za aktivitu mohl student získat maximálně 10 bodů za odevzdání správných řešení kontrolních úloh, které

byly zadány na přednáškách. Pro udělení zápočtu bylo nutné získat minimálně 15 bodů ze cvičení.

V semestru B111 byly zrušeny body za aktivitu a maximum bodů za odevzdání kontrolních úloh. Zbytek pravidel zůstal stejný jako v B101. Tato pravidla zůstala stejná i pro B121.

V B131 byla provedena redukce počtu testů na 3, každý za 10 bodů. Zbytek hodnocení zůstal stejný jako v B121.

V semestru B141 se psali pouze 2 testy, každý za 15 bodů. Bonusové body nebyly udělovány žádné, minimum 15 bodů pro zápočet zůstalo.

Počet testů zůstal stejný i pro semestr B151, ale opět se vrátila možnost získání až 10 bodů za kontrolní úlohy zadané na přednášce.

Shrnutí
V tomto předmětu se neliší celkový bodový zisk ze semestru, napříč semestry se liší pouze počet testů a bodový zisk z nich. Probíraná látka však zůstává stejná, neměli bychom tedy při sjednocení narazit na problémy.

2.1.1.5 Číslicové a analogové obvody (BI-CAO)

ECTS kreditů: 5
Týdenní rozsah výuky: 2 přednášky + 2 cvičení

Předmět je zakončen zápočtem i zkouškou, nicméně zkouška je nepovinná pro ty studenty, kteří získají nadprůměrný počet bodů během semestru (z 50 možných bodů získají 40 a více). Zápočet je studentovi udělen po zisku minimálně 25 bodů.

BI-CAO napříč semestry nezměnilo své hodnocení. V každém z analyzovaných semestrů se tedy psal jeden test za 10 bodů a 2 testy po 20 bodech. Během cvičení mohl student získat až 10 bodů za aktivitu, excelentní řešení zadaných úloh apod.

Shrnutí
Co se pravidel hodnocení v tomto předmětu napříč běhy týče, je BI-CAO ukázkovým předmětem, s jehož daty lze bez problémů pracovat a porovnávat je mezi sebou, protože struktura hodnocení zůstává každý semestr stejná.

2.1.1.6 Právo a informatika (BI-PAI)

ECTS kreditů: 3
Týdenní rozsah výuky: 2 přednášky

Tento předmět je zakončen pouze zkouškou, která není podmíněna žádnou aktivitou studenta během semestru.

Shrnutí
Vzhledem k neexistenci hodnocení či jiných údajů (např. docházka na přednášky) během semestru nebudeme s tímto předmětem pracovat.

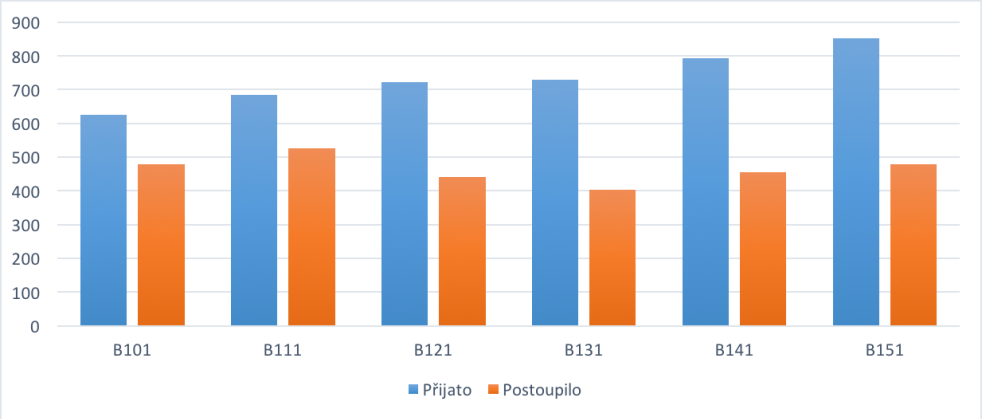
2.1.2 Anomálie

Vzhledem k tomu, že klesající prostupnost mezi prvním a druhým semestrem BI je velmi tížívým problémem (2.1), přistoupilo vedení fakulty v posledních dvou letech k možnosti zavedení podpůrných volitelných předmětů. Tyto předměty většinou slouží jako doporučená prerekvizita nebo doučování k náročnějším předmětům - ty s nejnižší prostupností, viz 2.6 (pro zjednodušení zobrazení není použit prefix BI-).

Předmět	**CAO**	**MLO**	**PA1**	**PAI**	**PS1**	**ZMA**
B101	84%	82%	55%	85%	56%	63%
B111	82%	76%	71%	81%	53%	60%
B121	87%	64%	56%	81%	38%	45%
B131	70%	50%	48%	71%	37%	42%
B141	77%	53%	44%	75%	32%	48%
B151	72%	44%	35%	70%	29%	38%

Tabulka 2.6: Prostupnost daných předmětů v semestrech B101-B151.

Za úspěšné absolvování těchto podpůrných předmětů získávají studenti kredity, které se jim započítávají do prvního semestru BI a pomáhají tak studentům dosáhnout alespoň minimální hranice 15 kreditů pro postup. Tyto předměty jsou však jednorázovou akcí, proto nejsou jejich data při sestavování prediktivního modelu použita, avšak je důležité mít tyto události na paměti při dalších analýzách.



Obrázek 2.1: Prostupnost z 1. do 2. semestru BI.

2.1.2.1 Úvod do vyšší matematiky (BI-UVM)

Tento předmět byl spuštěn pouze jednorázově a to v B141. S blížícím se koncem semestru začínal problém s prostupností mezi semestry sílit - nejvíce studentů, kterým z nějakých důvodů utíkala šance na zápočet (a tedy i možnost úspěšného absolvování zkoušky a zakončení předmětu), se vyskytovalo v předmětu BI-ZMA. Vedení fakulty se rozhodlo vypsat předmět BI-UVM, který byl doporučen primárně studentům, kteří měli málo bodů z prvního testu v BI-ZMA, přestože aktivně navštěvovali cvičení. Předmět BI-UVM si studenti mohli zapsat pouze po zrušení zápisu předmětu BI-ZMA, jelikož se jednalo o jednoduší alternativu.

Tento předmět byl vyučován on-line pomocí systému Marast během zkouškového období s nepovinnou docházkou na cvičení, kde byly zadávány a odevzdávány domácí úkoly. Celkem obsahoval 8 výukových okruhů, po každém následoval on-line kvíz. Kurz byl zakončen zápočtem, pro který musel student absolvovat 6 z 8 on-line kvízů a odevzdat 6 z 8 domácích úloh. Za úspěšné absolvování BI-UVM student získal 5 kreditů, tzn. rovnocenný kreditový zisk s BI-ZMA. Předmět si zapsalo celkem 161 studentů a úspěšně jej absolvovalo 111 studentů. Údajně cca 80 studentům předmět pomohl v postupu do dalšího semestru. Další rok se už tento předmět nevyučoval.

2.1.2.2 Přípravný kurz matematiky (BI-PKM)

Během léta před zahájením akademického roku 2015/2016 byl otevřen předmět BI-PKM pro nové studenty BI. Tento předmět měl studenty připravit na pro mnohé problematické BI-ZMA. Předmět probíhal ve stylu e-learningového kurzu v systému Marast obdobně jako předmět BI-UVM s tím rozdílem, že

obsahoval pouze 7 výukových okruhů a celá výuka probíhala on-line - student po každém výukovém okruhu musel absolvovat on-line kvíz, ve kterém bylo nutné zodpovědět minimálně 6 otázek správně. Za 3 špatné odpovědi v pořadí následovala penalizace ve formě znepřístupnění kvízu na 120 minut. Předmět byl zakončen absolvováním písemného testu v 1. výukovém týdnu prvního semestru a bylo možné za něj získat 4 kredity. Studentů, kteří zahajovali semestr B141 s těmito 4 kredity bylo 321.

2.1.2.3 Úvod do Linuxu (BI-ULI)

BI-ULI sloužil jako podpůrný předmět pro studenty, kteří nezvládají BI-PS1. Předmět si bylo možné zapsat v B151, podmínkou zápisu tohoto předmětu byl souběžný zápis předmětu BI-PS1. Předmět byl nepovinný a náplň se z velké části překrývala s obsahem předmětu BI-PS1. Informace byly však podávány jinou formou a měly přispět k pochopení souvislostí, získání nadhledu nad problematikou a zvýšit tak pravděpodobnost úspěšného absolvování povinného předmětu BI-PS1. Předmět byl bez kontaktní výuky, studenti se s jednotlivými tématy seznamovali samostatně pomocí e-learningového systému. Závěrečný zápočtový test se konal v prostorách FIT. Předmět si zapsalo 389 studentů, z toho 314 jej úspěšně absolvovalo a získalo tak 2 kredity.

2.1.2.4 Shell Minimum (BI-SM)

Začátkem zkouškového období zimního semestru B151 byl vypsán volitelný předmět určený těm studentům, kteří v tomto semestru neuspěli v předmětu BI-PS1, tzn. absolvovali ho se známkou F (neprospěl). Z tohoto důvodu byl zápis možný až během zkouškového období. Předmět bylo nutné absolvovat do konce zimního zkouškového období a skládal se pouze z jedné konzultace a jednoho přípravného soustředění. Předmět byl zakončen testem z omezené podmnožiny témat, která byla probrána na soustředění. Za úspěšné absolvování BI-SM student získal 2 kredity, takových studentů bylo 123 ze 187 zapsaných.

Kapitola 3

Porozumění datům

Pro porozumění datům je důležité pochopit, z jakého zdrojového systému data pocházejí a jakým způsobem se získávají. Dále je nutné porozumět struktuře, ve které jsou data uložena, a významu, jaký data nabývají, především v případě chybějící dokumentace.

3.1 Datové zdroje

ČVUT využívá přes 600 různých subsystémů, které jsou zaměřené na různé oblasti a spravované různými fakultami. V této práci je aktuální pouze znalost systémů využívaných FIT za účelem evidence studijních výsledků.

3.1.1 KOS

Systém KOS (Komponenta studia ČVUT) slouží jako celouniverzitní systém k evidenci výsledné klasifikace všech studentů. Najdeme zde všechny informace o vypsaných a zapsaných předmětech, rozvrhy, zkouškové termíny, počet zkouškových pokusů, jednotlivé známky či jednorázové akce. Právě v tomto systému je uveden výsledný stav studentova studia, kolik získal kreditů a zda splnil podmínky pro postup do dalšího semestru. Také jsou zde informace o zvoleném oboru či nároku na sociální stipendium a kontaktní informace o každém studentovi.

KOS je reprezentován databází Oracle, do které má FIT omezený přístup pomocí účtu FitAgent, kterému jsou zobrazeny pouze určité tabulky či pohledy. Stejný přístup je využíván i KOSapi[5]. Setkáváme se zde s omezeným přístupem ke zdrojovému systému a systém jako celek nám zůstává neznámý.

[5]Poskytuje aplikační rozhraní (API) v podobě RESTful webových služeb, které zprostředkovává přístup k vybrané části dat v databázi KOS.

3.1.2 Přihláška ČVUT

Systém KOS také uchovává informace z celouniverzitní aplikace Přihláška ČVUT, která slouží k podávání elektronických přihlášek uchazečů o studium. Aplikace se nachází na adrese prihlaska.cvut.cz a její vyplnění provádí přímo uchazeči. Po přijetí takového uchazeče jsou jeho údaje z přihlášky uchovávány právě v KOSu, v případě nepřijetí uchazeče jsou z důvodu právních předpisů data smazána. V KOSu je pak zaznamenána i informace o výsledku přijímacího řízení (např. bodový zisk z přijímací zkoušky).

K získání osobních údajů a informací ohledně předchozích studií studentů poslouží jako zdroj dat právě Přihláška, nicméně data nebudou získávána pomocí systému KOS, ale z prototypu datového skladu ČVUT (dále jen DWH ČVUT).

3.1.3 Prototyp datového skladu ČVUT (DWH ČVUT)

Prototyp datového skladu ČVUT byl realizován v rámci projektu „Datová čistota a datový audit v doménách studium a hodnocení kvality výuky na ČVUT“ (IP 2015, DÚ č. 20) pod vedením Ing. Michala Valenty, Ph.D., jehož realizačního týmu jsem součástí. Navržené řešení prozatím integruje doménu Studium v rámci FIT. Hlavním datovým zdrojem je KOS, dále pak Anketa ČVUT a nyní se připravuje integrace fakultních systémů jako jsou EDUX, Progtest, Moodle aj. Souběžně s návrhem DWH ČVUT probíhá vývoj informačního portálu jménem EBIE (Extended Business Intelligence Encyclopedia), který si klade za cíl poskytování reportingu, business slovníku, přehledu procesů aj. služeb, které budou podporovat datový audit, proces rozhodování na základě důvěryhodných dat a zvyšovat kvalitu informovanosti.

Vize DWH ČVUT je integrovat všechny dostupné systémy a poskytnout tak ucelené informace na jednom místě. V budoucnu proto již nebude nutné popisovat jednotlivé datové zdroje, neboť veškerá data pro analýzy budou čerpána z datového skladu, jehož tým data smyslupně sjednocuje. Velkou výhodou DWH ČVUT oproti KOS je schopnost historizace, kterou KOS nedisponuje a nelze tak evidovat změny záznamů. Vývoj datového skladu je velmi časově náročný, neboť se jedná o podnikový návrh. Předchozí fakultní datový sklad [27] není svojí architekturou a obsahem vyhovující stávajícím potřebám, proto není v této práci zmíněn a je postupně nahrazen DWH ČVUT.

DWH ČVUT je aktuálně realizován jako PostgreSQL databáze s využitím nástrojů pro plnění a transformace dat od společnosti Pentaho. Obojí je open-source řešení.

3.1.4 EDUX

Systém EDUX je na rozdíl od KOSu používán pouze v rámci FIT a slouží pro zaznamenávání průběžných studijních výsledků každého studenta. Používání EDUXu je pro všechny vyučující ve všech předmětech povinné. V tomto systému můžeme zjistit, jak si student během semestru vede, kolik získal bodů z testů nebo domácích úloh a zda má nárok na udělení zápočtu. Systém tedy slouží jako podpůrná prerekvizita pro KOS.

EDUX je velmi heterogenní systém, lze do něj zaznamenávat jak výsledky testů a zkoušek, tak domácích úkolů, aktivitu během výuky, docházku nebo zde studenti mohou nahrávat řešení svých domácích úloh do osobního prostoru. Dále zde najdeme ke každému předmětu potřebné studijní materiály, anotaci a podmínky hodnocení v předmětu.

EDUX je reprezentován formou open-source wiki, která využívá jednoduchou syntaxi DokuWiki. Je to velmi použitelná forma především pro vytváření dokumentace a podporuje spolupráci v týmu více lidí. Všechna data jsou uložena v textových souborech, zaniká tedy potřeba databáze. EDUX měl sloužit pouze jako přechodný nástroj po založení fakulty vzhledem k tomu, že se nejedná o systém vhodný pro podporu výuky (LMS).

3.1.5 Progtest

Systém Progtest slouží jako podpůrný nástroj pro testování studentů. Původně byl využíván pouze v předmětech zaměřených na programování (především BI-PA1, BI-PA2 apod.), ale v poslední době se těší oblibě i v předmětech jiných v rámci FIT. V tomto systému najdeme především testy a domácí úlohy - jejich zadání, jednotlivé pokusy studentů o odevzdání, bodové ohodnocení, penalizace, nápovědy, dále pak bodový zisk ze znalostních testů a zkoušková zadání a jejich výsledky apod. Progtest je vyvíjen na půdě FIT a obsahuje velmi užitečné informace jako je počet pokusů o odevzdání úlohy s přesnými datumy a bodovým ziskem, funkcionalitu pro odhalování opisování apod.

3.1.6 Moodle

Systém Moodle je celouniverzitní systém, ale na FIT není hojně využíván. Setkáme se s ním jen v několika málo předmětech. Systém slouží nejen k zaznamenávání studijních výsledků, docházky a sdílení studijních materiálů jako je tomu u EDUXu, ale také umožňuje sledování pohybu studentů (zda se zaregistroval, jak často portál navštěvuje, které materiály si zobrazuje apod.) a možnost on-line testování studentů. Tento systém by tedy mohl poskytnout mnoho užitečných informací ohledně chování a výsledků studentů během se-

mestru, ale v 1. semestru BI není využíván žádným předmětem. Dále se s tímto systémem setkáme u přijímací zkoušky do magisterského studia FIT, která je tvořena testem s uzavřenými otázkami.

Moodle je bezplatným a open-source řešením pro formu e-learningu (LMS), jehož placeným konkurentem je řešení Blackboard, které využívá např. zmíněná univerzita v Purdue v rámci systému Course Signals.

3.1.7 MARAST

Projekt MAtematika RAdoSTně (zkráceně MARAST) je webový portál na adrese marast2.fit.cvut.cz, který vznikl pro podporu výuky matematických předmětů na KAM FIT [6]. Cílem je vytvořit on-line cvičebnici, kterou vyučující průběžně plní různými příklady. Z těchto příkladů jsou generovány i zápočtové písemky a portál tak pomáhá vyřešit velký problém s vytvořením dostatečného počtu stejně obtížných písemek. Během posledního roku portál také podporoval on-line materiály pro výuku dvou volitelných bakalářských předmětů - BI-UVM a BI-PKM.

3.2 Původ a struktura dat

Kritickou součástí strojového učení jsou data - strojové učení funguje jen na zadání, pro která máme dostatek dat. Bez dostatku tréninkových dat, obsahujících ty správné informace, nebude strojové učení fungovat. Proto je jednou z nejdůležitějších a nejnáročnějších částí této práce právě získání dat ze zdrojových systémů a jejich předzpracování.

Dlouhodobým problémem FIT je získávání dat ze zdrojových systémů, které není nijak automatizované a exporty jsou prováděny na žádost přímo správci systémů. Tento proces je neefektivní, jelikož jsou správci už tak přehlceni různými požadavky a zároveň tento postup neumožňuje získávání real-time exportů dle potřeby.

Momentálně je k dispozici KOS API a účet Fitagent, prostřednictvím kterého lze automatizovaně získávat data pro DWH ČVUT. Bylo by ideální, kdyby obdobné API nabízely i ostatní systémy. Kromě usnadnění získávání dat pro analýzy je také nespornou výhodou možnost integrace datového zdroje do datového skladu - v budoucnu by odpadla velká část předzpracování dat a sestavování datasetu, což by výrazně urychlilo celý proces prediktivního modelování.

[6] Katedra aplikované matematiky FIT

3.2.1 Profil studenta

Profil studenta je tvořen informacemi, které student v roli zájemce o studium uvede v přihlášce ke studiu pomocí aplikace Přihláška ČVUT. Tato data jsou pak uložena v systému KOS, ze kterého jsou propagována do DWH ČVUT a odtud čerpána pro sestavení datasetu. Další informace, než-li z přihlášky, o studentovi nemáme. Během jeho studia přibývají informace o jeho studijních výsledcích.

Data z přihlášky jsou v DWH ČVUT integrována v jedné tabulce prih_prihlaska. Tato tabulka je dále napojena na tabulky s osobními údaji o studentovi osob_osoba a koud_adresa. Databázové schéma [7] celého DWH ČVUT je součástí přiloženého CD a bylo vytvořeno v nástroji Visual Paradigm. Přestože se jedná o jeho 129. revizi, prochází schéma neustálým vývojem v závislosti na dostupných datech a rozšiřování zpracovávané domény.

V rámci seznámení se s daty bylo nutné popsat význam jednotlivých atributů pomocí obchodních názvů (en. business metadata) a zjistit, zda jsou atributy využívány. Pro porozumění jednotlivým atributům posloužil dokument KOSpec [8], který vytvořil Ing. Jakub Jirůtka, a diplomová práce Ing. Elišky Hrubé [28], která se mj. zabývala integrací přihlášky do fakultního datového skladu, který byl nahrazen DWH ČVUT. Nutné podotknout, že data v DWH ČVUT nekopírují zcela přesně atributy z KOSu, některé atributy byly při implemetaci vyřazeny a jiné zase přidány. Tabulka 3.1 popisuje obsah aktuální tabulky prih_prihlaska pro BI, který bude dále zpracováván v rámci sestavení datasetu. Tabulka obsahuje pouze hodnoty využívané na FIT, tedy význam jednotlivých atributů je přizpůsoben právě tomuto prostředí.

Atribut	**Popis**	**Vyplněno**
id_prihlasky_bk	identifikátor přihlášky	100%
cislo_prihlasky	pořadové číslo přihlášky	100%
kodprihl	kód přihlášky	19%
druhst_kod	kód druhu studia	0%
oborst1_kod	kód oboru 1, na který se uchazeč hlásí	0%
oborst2_kod	kód oboru 2, na který se uchazeč hlásí	0%
oborst3_kod	kód oboru 3, na který se uchazeč hlásí	0%
odkud_skola_kod	kód odkud se uchazeč hlásí	99,90%
obor_st_sk	kód oboru střední školy	83%

[7] Schéma nezahrnuje úpravy implementované v rámci této práce z důvodu neposkytnutí licencí pro rok 2016 VIC ČVUT. Úpravy budou do schématu implementovány dodatečně, nicméně do DWH ČVUT jsou implementovány souběžně s touto prací.

[8] Specifikace KOS API dostupná na adrese gitlab.fit.cvut.cz/rozvoj/kospec.

Atribut	Popis	Vyplněno
predmet1	předmět 1 ze střední školy, u kterého se sledují známky	30%
predmet2	předmět 2 ze střední školy, u kterého se sledují známky	0,04%
predmet3	předmět 3 ze střední školy, u kterého se sledují známky	0%
predmet4	předmět 4 ze střední školy, u kterého se sledují známky	0%
predmet5	předmět 5 ze střední školy, u kterého se sledují známky	0%
znamky1	známky z předmětu 1	20%
znamky2	známky z předmětu 2	0,02%
znamky3	známky z předmětu 3	0%
znamky4	známky z předmětu 4	0%
znamky5	známky z předmětu 5	0%
prumer1	průměr z předmětu 1	0,3%
prumer2	průměr z předmětu 2	0,3%
prumer3	průměr z předmětu 3	0,3%
prumer4	průměr z předmětu 4	0,26%
prumer5	průměr z předmětu 5	0%
zamestnavatel	zaměstnavatel uchazeče	0%
predch_st_kde	předchozí studium	20%
skolitel	ID učitele, který bude školitelem pro doktorské studium	0%
tema	téma závěrečné práce	0%
rozhodnuti	kód rozhodnutí (způsob přijetí)	99,9%
poznamka	poznámka	0%
jazyky	jazykové znalosti	0%
organizacni_jednotka_id	identifikátor střediska	0%
typ_prihlasky	typ přihlášky	19%
matur_prum	průměr známek z maturity	4,70%
cismist_zkousky	číslo místnosti řízení	0%
datum_zkousky	datum a čas přijímací zkoušky	0%
hodnoceni_celkem	hodnocení přijímací zkoušky	84%
promin_zkousky	příznak prominutí přijímací zkoušky	22%
hodnoceni1	varianta přijímací zkoušky	49%
hodnoceni2	výsledek přijímací zkoušky (počet bodů)	43%

Atribut	Popis	Vyplněno
hodnoceni3	zda je odevzdáno maturitní vysvědčení	98%
hodnoceni4	odevzdáno potvrzení o SZZ	3,4%
hodnoceni5	zkouška z češtiny (B2)	56%
hodnoceni6	absolvované olympiády	40%
hodnoceni7	percentil ze Scio testů	44%
hodnoceni8	průměr z bakalářského studia FIT	0%
hodnoceni9	hodnocení 9	0%
hodnoceni10	hodnocení 10	0%
zapsal	kód, zda se student dostavil k zápisu	84%
stipendium	žádost o stipendium	0%
stipcizi	žádost o cizí stipendium	0%
kolej	žádost o kolej	0%
konzultant	identifikátor učitele, který bude konzultantem	0%
skupina	číslo studijní skupiny	0,1%
typss	typ střední školy	99,9%
rok_mat	rok maturity	100%
komise	identifikátor komise	0%
studijni_program	kód studijního programu	100%
nove_prijaty	příznak, zda je uchazeč nově přijatý	100%
navazujici_sp	příznak, zda se jedná o navazující studium	0%
financovani	způsob financování	100%
stupen_predchoziho _vzdelani	stupeň předchozího vzdělání	100%
typ_prog	typ studijního programu	100%
forma_studia	forma studia	100%
t1_datum	datum 1. přijímací zkoušky	0%
t1_cas	čas 1. přijímací zkoušky	0%
t1_misto	místo 1. přijímací zkoušky	0%
t2_datum	datum 2. přijímací zkoušky	0%
t2_cas	čas 2. přijímací zkoušky	0%
t2_misto	místo 2. přijímací zkoušky	0%
t3_datum	datum 3. přijímací zkoušky	0%
t3_cas	čas 3. přijímací zkoušky	0%
t3_misto	místo 3. přijímací zkoušky	0%
prumerpp	průměr známek ze střední školy	20,7%
program1_id	identifikátor studijního programu 1, na který se uchazeč hlásí	100%

Atribut	Popis	Vyplněno
program2_id	identifikátor studijního programu 2, na který se uchazeč hlásí	0%
zamereni1_id	ID zaměření 1, na které se uchazeč hlásí	0%
zamereni2_id	ID zaměření 2, na které se uchazeč hlásí	0%
zamereni3_id	ID zaměření 3, na které se uchazeč hlásí	0%
zamereni4_id	ID zaměření 4, na které se uchazeč hlásí	0%
zamereni5_id	ID zaměření 5, na které se uchazeč hlásí	0%
zamereni6_id	ID zaměření 6, na které se uchazeč hlásí	0%
zamereni7_id	ID zaměření 7, na které se uchazeč hlásí	0%
zamereni8_id	ID zaměření 8, na které se uchazeč hlásí	0%
zamereni9_id	ID zaměření 9, na které se uchazeč hlásí	0%
obor1_id	ID oboru 1, na který se uchazeč hlásí	0%
obor2_id	ID oboru 2, na který se uchazeč hlásí	0%
obor3_id	ID oboru 3, na který se uchazeč hlásí	0%
obor4_id	ID oboru 4, na který se uchazeč hlásí	0%
obor5_id	ID oboru 5, na který se uchazeč hlásí	0%
obor6_id	ID oboru 6, na který se uchazeč hlásí	0%
datum_reg	datum registrace přihlášky	19%
pistest	příznak, zda uchazeč bude psát písemný test	0%
pred_vs_kod	kód předchozí VŠ	5%
pred_fak_kod	kód předchozí fakulty	0%
cislo_uchazece	identifikátor uchazeče	19%
scio_test	příznak, zda uchazeč úspěšně absolvoval Scio test	8%
olympiady	seznam olympiád, kterých se uchazeč zúčastnil	1%
postizeni_kod	kód postižení	0,20%
postizeni_poznamka	poznámka k postižení	0,02%
uspesne_bc_cvut	příznak absolvování bc. programu na ČVUT	19%
uspesne_bc_fakulta	příznak absolvování bc. programu na fakultě, na kterou se hlásí	19%
stpl_studijni _programprogram_id	ID studijního programu	100%
osob_osobaperidno	osobní číslo uchazeče	100%
stud_studiumid _studia	ID studia uchazeče (přiděleno při nastoupení studia)	100%

Atribut	Popis	Vyplněno
str_skola	kód střední školy (IZO)	99,9%
technical_key	technický klíč pro DWH	100%
version	verze nahrání záznamu do DWH	100%
date_from	datum platnosti záznamu od v DWH	100%
date_to	datum platnosti záznamu do v DWH	100%
matur	známky z maturity	5%
rodstav	rodinný stav	100%
obcanstvi	kód občanství	100%

Tabulka 3.1: Atributy tabulky prih_prihlaska v DWH ČVUT.

V kapitole zabývající se předzpracováním dat bude blíže vysvětleno, které atributy budou použity.

3.2.2 Studijní výsledky

Pro vytvoření datasetu potřebujeme znát studijní výsledky studentů. V analýze předmětů byly zmíněny potřebné předměty, data získáme z těchto zdrojových systémů:

- BI-PA1 - Progtest,
- BI-ZMA, BI-PS1, BI-MLO, BI-CAO - EDUX,
- Informace o postupu do letního semestru - DWH ČVUT.

3.2.2.1 Progtest

Data z předmětu BI-PA1 jsou z Progtestu získána formou jednorázového XML exportu (co semestr, to soubor), který na žádost provedl Ing. Ladislav Vagner, Ph.D., tvůrce systému Progtest. Každý export obsahuje stejnou a velmi dobře čitelnou strukturu, kterou lze po dohodě s Ing. Ladislavem Vagnerem, Ph.D. definovat.

Struktura exportu ve formátu xml je následující:

```
<!– Hlavička předmětu –>
<Course id="identifikátor kurzu"name="název předmětu"subject="kód
předmětu"year="název akademického roku»
<!– Seznam zadaných úloh –>
<Tasks>
<Task id="identifikátor úlohy"name="název úlohy"openDate="datum ote-
vření úlohy"deadlineDate="datum uzavření úlohy"/>
</Tasks>
<!– Seznam zadaných testů –>
<Quizes>
<Quiz id="identifikátor testu"name="jméno testu"/>
</Quizes>
<!– Výsledky studenta –>
<Student name="username studenta»
<!– Seznam všech odevzdaných úloh s jednotlivými pokusy –>
<Task id="identifikátor úlohy»
<Submit date="datum odevzdání řešení úlohy"result="bodový zisk
z úlohy"/>
<Submit date="datum odevzdání řešení úlohy"result="bodový zisk
z úlohy"/>
<Submit date="datum odevzdání řešení úlohy"result="bodový zisk
z úlohy"/>
</Task>
<!– Seznam všech výsledků z testů –>
<Quiz id="identifikátor testu"result="bodový zisk z testu"/>
</Course>
```

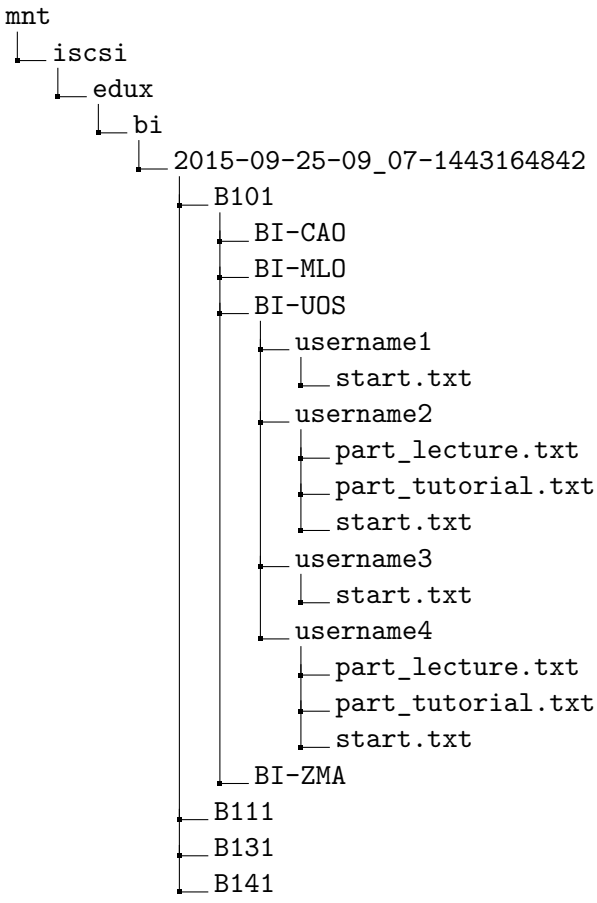
3.2.2.2 EDUX

Na začátku semestru je v EDUXu v oddíle Klasifikace studenta (přístupný pouze vyučujícím) vygenerován seznam studentů zapsaných na předmět na základě dat, která poskytuje KOS a jeho rozvrhové lístky. Tento seznam lze v EDUXu dále manuálně spravovat, avšak pokud jsou během semestru v KOSu provedeny nějaké změny, např. si student zruší zápis předmětu, do EDUXu se tyto změny již automaticky nepromítají.

Struktura klasifikace určuje, jaké údaje budou o studentovi zaznamenávány (docházka, aktivita, testy, domácí úlohy apod.) a umožňuje definovat výrazy, kterými je možné vyhodnotit např. studentův nárok na zápočet. Strukturu klasifikace si v každém předmětu definuje vyučující (garant předmětu) sám, stejně tak i položky hodnocení pro daný běh předmětu. Detailní popis možností tvorby klasifikace není pro tuto práci nijak zajímavý, vyučujícím jsou na EDUXu k dispozici návody, kde je vše srozumitelně popsáno.

Data z předmětů BI-ZMA, BI-PS1, BI-MLO, BI-CAO byla z EDUX poskytnuta jako jednorázový manuální export Ing. Tomášem Kadlecem, který zodpovídá za správu systému EDUX a je vedoucím ICT oddělení FIT. Tento export byl později zautomatizován do té míry, že data jsou k dispozici každý den v 6:00 na určité adrese dostupné po autorizaci.

Struktura manuálního exportu je následující:



- bi - bakalářský program Informatika
- 2015-09-25-09_07-1443164842 - datum a identifikátor pořízení exportu
- B101 - kód semestru
- BI-CAO - název předmětu
- username1 - username studenta zapsaného na předmět v daném semestru
- start.txt - textový soubor obsahující hodnocení studenta

Každá složka studenta obsahuje 1-5 textových souborů podle toho, jak je předmět koncipován a zda si jeho zápis student nezrušil:

1. start.txt - informace o studiu,
2. part_tutorial.txt - hodnocení cvičení,
3. part_lecture.txt - hodnocení přednášky, zkoušky,
4. seminar.txt - hodnocení proseminář,
5. lab.txt - hodnocení laboratoře.

V požadovaném exportu jsem se setkala pouze s textovými soubory typu 1, 2 a 3. Soubory typu 4 a 5 nebudou v této práci popsány, jelikož k nim nemám přístup a netýkají se analyzovaných předmětů. Struktura dat bude demonstrována např. na předmětu BI-UOS v B101, což je pro její pochopení postačující. Nesmíme však zapomenout, že se tato struktura semestr od semestru a předmět od předmětu liší a je tedy důležité se seznámit s daty napříč všemi zpracovávanými semestry, přestože se týkají jednoho předmětu, který měl pro všechny běhy stejné podmínky.

Velkou nevýhodou jsou chybějící metadata napříč celou klasifikací v EDUXu. Přesná rekonstrukce významu jednotlivých položek by šla pouze na základě analýzy s konkrétním vyučujícím, který danou klasifikaci vytvořil, což je vzhledem ke stáří dat (až 6 let) a počtu předmětů nespolehlivé a často i nemožné. Proto se omezíme pouze na popis, který zcela postačí k práci s daty při jejich výběru a předzpracování.

start.txt
Obsahuje základní informace o studentovi, který si daný předmět zapsal. Tento soubor je vygenerovaný pro všechny studenty v daném běhu, pokud se stane, že tento soubor chybí, nastala zřejmě chyba při exportu a nebo EDUX nedisponuje daty (data nebyla po skončení semestru uložena a jsou nenávratně ztracena).

Struktura souboru:

```
=======Příjmení Jméno=======
* namespace: [[:student:username:start]]
~~classification:username:jazyk studia:forma studia::~~
```

- namespace - odkazuje na osobní prostor studenta na EDUXu
- jazyk studia - nabývá hodnot cs/en a udává informaci, zda student studuje v českém nebo anglickém jazyce
- forma studia - může být fulltime/partime a říká, zda se jedná o studenta prezenční nebo kombinované formy studia

part_tutorial.txt
V tomto souboru jsou zapsány studentovy výsledky z definovaných aktivit na cvičení. Tento soubor je vygenerován pouze těm studentům, u kterých nějaká aktivita existuje, tzn. že u studenta, který po celý semestr neprojeví vůbec žádnou aktivitu (neodevzdá úkol, nezapíše se do docházky apod.), tento soubor neexistuje. Aktivity (položky hodnocení) mohou být různé podle toho, jak je nadefinuje vyučující.

Struktura souboru:

~~classification:username:jazyk studia:forma studia:typ souboru:a:22:{s:3:"
název položky hodnocení";s:1:"získané hodnocení";s:3:"d02";s:1:"0";s:3:"d03";
s:1:"0";s:3:"d04";s:1:"0";s:3:"d05";s:1:"0";s:3:"d06";s:1:"0";s:3:"d07";s:1:"0";s:3:"
d08";s:1:"0";s:3:"d09";s:1:"0";s:3:"d10";s:1:"0";s:3:"d11";s:1:"0";s:3:"d12";s:1:"0";
s:3:"d13";s:1:"0";s:3:"mt1";s:1:"5";s:3:"mt2";s:2:"12";s:3:"mt3";s:2:"11";s:2:"vt";s:
2:"30";s:2:"_d";i:0;s:3:"_mt";i:28;s:4:"_mtu";s:1:"1";s:1:"c";i:58;s:2:"kz";s:1:"E"
;}~~

- typ souboru - nabývá hodnot tutorial/lecture aj., podle toho, o jaký soubor se jedná
- { } - ohraničuji položky hodnocení a získané výsledky
- s:číslo: - např. s:3: uvádí název položky hodnocení, který následuje v uvozovkách

  - písmena a čísla mohou být různá a mít různé významy, vše záleží na definici vyučujícím

  - každá taková dvojice písmeno:číslo:""je zakončena středníkem

part_lecture.txt
V tomto souboru by se mělo nacházet hodnocení přednášky, které však většinou jako takové neexistuje, nebo zkoušky. V reálu se zde nacházejí položky hodnocení, které jsou potřeba ke klasifikaci studenta, tzn. pro udělení zápočtu, zkoušky a výsledné známky. Nachází se zde např. součty bodů za aktivitu na cvičeních či za domácí úlohy nebo jen celkový součet získaných bodů pro udělení zápočtu. Může se stát, že některé informace soubor obsahuje duplicitně se souborem part_tutorial.txt.

Struktura souboru:

~~classification:username:jazyk studia:forma studia:typ souboru:a:11:{s
:8:"aktivita";s:1:"5";s:7:"absence";i:0;s:6:"oprava";s:0:"";s:5:"testy";i:21;s:7:"
cviceni";i:26;s:7:"pisemka";s:4:"31.5";s:4:"suma";d:57.5;s:6:"znamka";s:1:"E";
s:8:"pisemka1";s:4:"21.5";s:8:"pisemka2";s:4:"31.5";s:8:"pisemka3";s:0:"";}~~

Výsledné hodnocení (zda byl udělen zápočet, absolvována zkouška apod.) by se mělo nacházet v souboru part_lecture, nicméně tomu nemusí tak vždy být a hodnocení se může objevit i v part_tutorial. Stejně tak se oba soubory nemusejí vyskytovat společně - některé předměty obsahují pouze part_tutorial.txt, jiné pouze part_lecture.txt. Další anomálií jsou studenti - pro některé je vygenerován jak part_tutorial.txt, tak i part_lecture.txt, pro jiné zase pouze jeden z nich bez žádného zřejmého pravidla.

3.2.2.3 DWH ČVUT

Z datového skladu získáme pomocí SQL dotazu informaci, kolik studenti prvního ročníku získali na konci zimního semestru kreditů a zda splnili podmínku minimálně 15 kreditů pro postup do letního semestru.

Daný SQL skript je součástí přiloženého CD.

Část II

Implementační část

Kapitola 4

Příprava dat

Příprava dat spočívá ve sběru všech potřebných dat, resp. jejich exportů, jejichž struktura byla popsána v předchozí kapitole, následné úpravě hodnot, tzv. předzpracování, v selekci potřebných hodnot a sestavení datasetu, který bude výchozím vstupem pro prediktivní modelování. Velká část této kapitoly bude věnována právě předzpracování dat, jelikož se jedná o nejdůležitější a časově nejnáročnější část celého data miningu. Pokud budou k prediktivnímu modelování použita nekvalitní data, pak výsledek bude také nekvalitní.

4.1 Předzpracování dat

Předzpracování dat spočívá ve vytvoření datasetu, který bude zpracován jednotlivými analytickými metodami. Data by měla obsahovat pouze relevantní údaje a být ve tvaru, který vyžaduje příslušná analytická metoda. Při předzpracování byl využit primárně nástroj Microsoft Excel.

Mezi používané kroky patří:

- selekce dat,
- čištění,
- validace dat,
- tranformace dat (typové konverze, diskretizace, binomizace apod.),
- doplnění chybějících údajů,
- integrování dat,
- redukce dat.

Připravovaný dataset bude obsahovat data ze semestrů B101 - B141, kterým se budeme během předzpracování věnovat se zohledněním výsledků analýzy předmětů. Data ze semestru B151 budou předzpracována stejným způsobem,

ale až po výběru vhodných dat, vzhledem k tomu, že tato data se dynamicky vytvářejí během semestru a v době vzniku této práce ještě nebyla k dispozici.

4.1.1 Data z DWH ČVUT

Vzhledem k tomu, že tato data prošla určitým předzpracováním již při nahrávání do datového skladu, budeme pracovat především se selekcí (mnoho atributů je duplicitních nebo nepotřebných) a vytvářením nových atributů. Data získáme pomocí SQL dotazu v open-source nástroji PgAdmin z PostgreSQL databáze DWH ČVUT (DWH2_TARGET).

Filtry
Zadání této práce je vymezené na studenty 1. ročníku bakalářského studia progmau Informatika v českém jazyce, z čehož vyplývají následující filtry:

- studijní program bakalářská Informatika: studijni_program = BI
- bakalářský typ programu: typ_prog = B
- prezenční forma studia: forma_studia = P

Další omezení se vztahuje na použité semestry B101 - B151. Z dat je nutné vyselektovat přihlášky studentů, kteří se hlásili do B091. Tabulka prih_prihlaska neobsahuje údaj o semestru a ani žádný tomu podobný (např. datum zápisu nebo datum podání přihlášky), protože tento údaj historicky chybí ve zdrojovém systému. Přihlášky pro semestr B151 již obsahují datum registrace. Pro ostatní semestry toto bylo vyřešeno napojením na tabulku stud_studium a získání datumu, kdy bylo studium zahájeno. Díky této informaci byli vyselektováni studenti se zápisem v B091.

```
SELECT
  semestry_studentu.studiumid_studia,
  semestry_studentu.arok_semestrsemestr_id,
  ROW_NUMBER()
  OVER (PARTITION BY semestry_studentu.studiumid_studia
    ORDER BY semestry_studentu.arok_semestrsemestr_id ASC) AS
        semestr_studia
FROM
  (SELECT DISTINCT
     zap.studiumid_studia,
     beh.arok_semestrsemestr_id
   FROM zapi_predmet zap LEFT JOIN pred_beh_predmetu beh ON zap.
       pred_beh_predmetubeh_predmet_id = beh.beh_predmet_id_bk
   WHERE
     zap.technical_key > 1 AND zap.date_to = '2199-12-31
         00:00:00.000000' AND beh.date_to = '2199-12-31
         00:00:00.000000'
```

```
    AND zap.studiumid_studia IN (SELECT studiumid_studia
                                 FROM stud_studium
                                 WHERE organizacni_jednotka_id = 10)
  ORDER BY zap.studiumid_studia, beh.arok_semestrsemestr_id)
     semestry_studentu;
```

Po tomto napojení přes ID studia (atribut stud_id) se ukázalo, že pro 75 záznamů v tabulce stud_studium DWH ČVUT chybí ID studia a datum zahájení studia. Po ověření těchto záznamů přímo v KOSu bylo zjištěno, že obsahují příznak smazání (DEL), tedy již dále nejsou z nějakého důvodu aktivní, a proto se tyto záznamy do DWH ČVUT nepropsaly. Nicméně data z přihlášek těchto osob DWH ČVUT uchovává, jedním z možných vysvětlení je, že se jedná o bývalé studenty, kteří přestoupili na jinou fakultu v rámci ČVUT a vzhledem k tomu, že máme přístup pouze k datům FIT, tak už nevidíme o těchto záznamech žádné další údaje, kterými jsou i studijní výsledky. Proto bylo těchto 75 záznamů smazáno a uvádím je zde pouze jako příklad potřeby důslednosti při validaci dat a dotazů.

Po aplikaci těchto podmínek zůstalo k dispozici 4299 záznamů z původních 6985. Export dat provedeme do formátu xls (Microsoft Excel) pro pozdější potřeby předzpracování.

Redukce
Na základě seznámení se s daty a jejich strukturou byla provedena redukce atributů, které:

1. Neobsahují žádná data, tzn. stav vyplnění je 0%.
2. Neobsahují data s informační hodnotou, tzn. slouží pouze k organizačním účelům nebo všechny vyplněné záznamy obsahují stejnou hodnotu.
3. Vysoce korelují s jinými atributy, např. PSČ a název města. Pokud bychom si nebyli korelací jistí, je potřeba použít metodu Feature ranking pro ohodnocení přínosu jednotlivých atributů.

Čištění dat
Během procesu nahrávání dat do DWH ČVUT neprobíhá kontrola datové kvality či validace dat ve smyslu jejich sjednocení. Data jsou do skladu nahrávána v originálním znění a jejich validace a úpravy probíhají až na sémantické vrstvě. Při sestavení datasetu proběhla kontrola datové čistoty a tedy i validace nastavených procesů na sémantické vrstvě.

Transformace dat
Některé atributy obsahují kódy, které by správně měly vést na číselníky s textovým popisem dané hodnoty (tzv. rozklíčování), pokud v tabulce neexistuje takový atribut. V implementaci DWH ČVUT tyto číselníky chyběly vzhledem

k tomu, že tato práce je prvním ostrým využitím dat, která DWH ČVUT může poskytnout, a proto byly v jejím rámci doplněny (4.1). Nahrazení kódů textovými hodnotami je v některých případech výhodnější a také usnadňuje interpretaci výsledků.

Název atributu	**Název číselníku**	**Přiřazení přes atribut**
odkud_skola_kod	cis_uchazec_odkud	odkud_kod
obor_st_sk	cis_obor_st_sk	obor_st_sk_kod
rozhodnuti	cis_rozhodnuti	rozhodnuti_kod
typss	cis_typ_ss	typ_ss_kod
stupen_predchoziho_vzdelani	cis_st_predch_vzdelani	st_predch_vzdelani_kod
str_skola	cis_ss	ss_izo
pred_vs_kod	cis_vs_kody	vs_kod
rodstav	cis_rodstav	rodstav_kod
obcanstvi	cis_obcanstvi	obcanstvi_kod

Tabulka 4.1: Číselníky pro tabulku prih_prihlaska.

Rozšíření dat

V rámci předzpracování byla data z přihlášky rozšířena o atributy, u kterých se domnívám, že by mohly mít vliv na výsledky predikce:

rok_narozeni

Atribut obsahující rok narození byl sestaven z datumu narození následujícím SQL příkazem:

```
EXTRACT(YEAR FROM osob.datum_narozeni) AS rok_narozeni
```

vek_pri_nastupu

Atribut obsahující věk uchazeče v době podání přihlášky, vypočítaný na základě začátku zahájení studia a datumu narození:

```
EXTRACT(DAYS FROM pocatecni_semestr.datum_zacatku_studia -
    datum_narozeni) / 365) :: INT AS vek_pri_nastupu
```

puvodem_praha
Obsahuje ano/ne odpověď na otázku, zda uchazeč pochází z Prahy na základě atributu koud_adresa.misto, typ=trvale.

typ_prijeti
Tento atribut obsahuje informaci o tom, jakým způsobem byl student přijat a může nabývat následujících hodnot:

- zkouska - pokud atribut hodnoceni2 obsahuje body z přijímací zkoušky pořádané FIT, nabyde atribut typ_prijeti hodnotu *zkouska*,
- scio - pokud atribut hodnoceni7 obsahuje percentil ze Scio testu, nabyde atribut typ_prijeti hodnoty *scio*,
- oboji - pokud jsou vyplněny atributy hodnoceni2 a hodnoceni7, je do atributu typ_prijeti zaznamenána hodnota *oboji* namísto hodnot zkouska nebo scio,
- soutez - pokud je vyplněn atribut hodoceni6, nabyde atribut typ_prijeti hodnoty *soutez*
- prominuti - pokud atribut promin obsahuje hodnotu 1, což znamená prominutí přijímací zkoušky, nabyde atribut typ_prijeti hodnoty *prominuti*

U tohoto atributu bylo ověřeno, že nedochází ke kolizím jednotlivých hodnot.

semestr
Semestr obsahuje kód semestru, do kterého se uchazeč hlásí. Získáme jej z roku uvedeného v počátku studia (atribut datum_zah v tabulce stud_studium, který porovnáme s rokem z datumu začátku semestru (atribut zacatek z tabulky arok_semestr):

```
prih_prihlaska prihlaska LEFT JOIN (SELECT
    id_studia_bk,
    semestr_id_bk AS semestr,
    datum_zah    AS datum_zacatku_studia
        FROM stud_studium stud
        JOIN arok_semestr seme
        ON EXTRACT(YEAR FROM seme.zacatek) = EXTRACT(YEAR FROM stud.
            datum_zah) AND
        semestr_id_bk LIKE '%1'
AND seme.date_to = '2199-12-31 23:59:59.999'
        WHERE stud.date_to = '2199-12-31 23:59:59.999')
            pocatecni_semestr
```

4.1.2 Data z Progtestu

Na základě analýzy byly z předmětu BI-PA1 zvoleny k předzpracování tyto položky:

- domácí úlohy (B111-B141)
- znalostní testy (B111-141)

Přestože je export z Progtestu velmi přehledný, je potřeba daná data upravit na strukturu, ve které všechny studijní výsledky jednoho studenta budou tvořit jeden záznam (řádek), který bude začínat usernamem studenta a končit poslední úlohou/testem v semestru zadanou. Jedná se tedy o vytvoření sloupců z řádků, které obsahují výsledky z jednotlivých úloh či testů. Tato úprava musí dodržovat časovou posloupnost zadávání úloh a také by bylo dobré odlišit, zda se jednalo o splnění úlohy lehké nebo těžké. Pro tyto účely vznikly tabulky 4.2, 4.3, 4.4 a 4.5.

Pořadí	**Původní ID**	**Obtížnost**	**Datum začátku**	**Datum konce**
1	42	lehká	07/10/11 22:00	23/10/11 21:59
1	634	těžká	07/10/11 22:00	23/10/11 21:59
2	569	lehká	14/10/11 22:00	30/10/11 22:59
2	636	těžká	14/10/11 22:00	30/10/11 22:59
3	28	lehká	21/10/11 22:00	06/11/11 22:59
3	637	těžká	21/10/11 22:00	06/11/11 22:59
4	640	lehká	28/10/11 22:00	13/11/11 22:59
4	740	těžká	28/10/11 22:00	13/11/11 22:59
5	742	lehká	04/11/11 23:00	20/11/11 22:59
5	743	těžká	04/11/11 23:00	20/11/11 22:59
6	745	lehká	11/11/11 23:00	27/11/11 22:59
6	746	těžká	11/11/11 23:00	27/11/11 22:59
7	203	lehká	18/11/11 23:00	05/12/11 22:59
7	748	těžká	18/11/11 23:00	05/12/11 22:59
8	749	lehká	25/11/11 23:00	18/12/11 22:59
8	750	těžká	25/11/11 23:00	18/12/11 22:59
9	744	soutěžní	11/11/11 23:00	31/12/11 22:59

Tabulka 4.2: Obtížnost domácích úloh v B111.

Pořadí	Původní ID	Obtížnost	Datum začátku	Datum konce
1	831	lehká	12/10/12 22:00	28/10/12 22:59
1	832	těžká	12/10/12 22:00	28/10/12 22:59
2	833	lehká	19/10/12 22:00	04/11/12 22:59
2	834	těžká	19/10/12 22:00	04/11/12 22:59
3	841	lehká	26/10/12 22:00	11/11/12 22:59
3	859	těžká	26/10/12 22:00	11/11/12 22:59
4	838	lehká	02/11/12 23:00	18/11/12 22:59
4	840	těžká	02/11/12 23:00	18/11/12 22:59
5	862	lehká	09/11/12 23:00	25/11/12 22:59
5	864	těžká	09/11/12 23:00	25/11/12 22:59
6	869	lehká	16/11/12 23:00	02/12/12 22:59
6	871	těžká	16/11/12 23:00	02/12/12 22:59
7	203	lehká	23/11/12 23:00	09/12/12 22:59
7	874	těžká	23/11/12 23:00	09/12/12 22:59
8	877	lehká	30/11/12 23:00	16/12/12 22:59
8	878	těžká	30/11/12 23:00	16/12/12 22:59
9	846	soutěžní	09/11/12 23:00	31/12/12 22:59

Tabulka 4.3: Obtížnost domácích úloh v B121.

Pořadí	Původní ID	Obtížnost	Datum začátku	Datum konce
1	969	těžká	18/10/13 22:00	03/11/13 22:59
1	972	lehká	18/10/13 22:00	03/11/13 22:59
2	975	těžká	25/10/13 22:00	10/11/13 22:59
2	976	lehká	25/10/13 22:00	10/11/13 22:59
3	977	těžká	01/11/13 23:00	17/11/13 22:59
3	978	lehká	01/11/13 23:00	17/11/13 22:59
4	980	těžká	08/11/13 23:00	24/11/13 22:59
4	981	lehká	08/11/13 23:00	24/11/13 22:59
5	982	těžká	15/11/13 23:00	01/12/13 22:59
5	983	lehká	15/11/13 23:00	01/12/13 22:59
6	989	lehká	22/11/13 23:00	08/12/13 22:59
6	990	těžká	22/11/13 23:00	08/12/13 22:59
7	991	lehká	29/11/13 23:00	15/12/13 22:59
7	992	těžká	29/11/13 23:00	15/12/13 22:59

Pořadí	Původní ID	Obtížnost	Datum začátku	Datum konce
8	877	lehká	06/12/13 23:00	22/12/13 22:59
8	993	těžká	06/12/13 23:00	22/12/13 22:59
9	967	soutěžní	08/11/13 23:00	31/12/13 22:59

Tabulka 4.4: Obtížnost domácích úloh v B131.

Pořadí	Původní ID	Obtížnost	Datum začátku	Datum konce
1	1091	těžká	17/10/14 22:00	02/11/14 22:59
1	1095	lehká	17/10/14 22:00	02/11/14 22:59
2	1092	těžká	24/10/14 22:00	09/11/14 22:59
2	1093	lehká	24/10/14 22:00	09/11/14 22:59
3	1096	těžká	31/10/14 23:00	16/11/14 22:59
3	1097	lehká	31/10/14 23:00	16/11/14 22:59
4	1099	těžká	07/11/14 23:00	23/11/14 22:59
4	1102	lehká	07/11/14 23:00	23/11/14 22:59
5	1104	těžká	14/11/14 23:00	30/11/14 22:59
5	1105	lehká	14/11/14 23:00	30/11/14 22:59
6	1106	těžká	21/11/14 23:00	07/12/14 22:59
6	1107	lehká	21/11/14 23:00	07/12/14 22:59
7	1114	těžká	28/11/14 23:00	14/12/14 22:59
7	1115	lehká	28/11/14 23:00	14/12/14 22:59
8	750	těžká	05/12/14 23:00	21/12/14 22:59
8	1118	lehká	05/12/14 23:00	21/12/14 22:59
9	1098	soutěžní	07/11/14 23:00	31/12/14 22:59

Tabulka 4.5: Obtížnost domácích úloh v B141.

Stejný proces je potřeba zopakovat i u znalostních testů, viz 4.6:

Pořadí	B111 Původní ID	B121	B131	B141
1	3	57	141	341
2	4	59	143	343

Pořadí	B111 Původní ID	B121	B131	B141
3	5	61	145	345
4	6	63	147	347

Tabulka 4.6: Znalostní testy v B111 - B141.

Všechna data již prošla normalizací (hodnoty se pohybují v rozmezí 0 až 1, případně 1.2, kde jsou zahrnuty bonusové body např. za včasné odevzdání úlohy) a nepotýkáme se zde s nečistotou nebo chybějícími daty.

Rozšíření dat
Kromě bodového zisku máme k dispozici také veškeré pokusy, které student uskutečnil, a jejich úspěšnost. Můžeme tak dataset obohatit o nové atributy, které by mohly mít prediktivní schopnost.

dny_od_zacatku
Atribut obsahuje počet dnů od spuštění úlohy, kdy se student poprvé pokusil úlohu odevzdat. Můžeme tak zjistit, jak velký vliv při prediktivním modelování má svědomitý přístup k řešení úloh.

dny_do_konce
Obsahuje počet dnů do uzavření úlohy ode dne, kdy student odevzdal svůj poslední pokus. Na základě tohoto údaje je možné sledovat, zda se více studentů pokouší o odevzdání před vypršením časového limitu apod.

pocet_pokusu
Z počtu záznamů lze spočítat, kolikrát student danou úlohu odevzdával.

nejvice_bodu
Atribut nejvice_bodu obsahuje nejvyšší bodový zisk, kterého student v dané úloze dosáhl a který je mu vždy jako jediný ze všech započítán do koncového hodnocení.
Každý tento atribut je uvozen prefixem, který se skládá z pořadového čísla úlohy a počátečního písmena obtížnosti, např. 1L_dny_do_zacatku, 4H_pocet _pokusu.

Transformace vybraných řádků na sloupce a přidání atributů bylo realizováno pomocí skriptu napsaného v programovacím jazyce Python.

4.1.3 Data z EDUXu

Časově a obsahově nejnáročnější bylo předzpracování dat ze systému EDUX, kdy prvotní export obsahoval 45 380 textových souborů. Nejdříve bylo nutné upravit strukturu exportu, kdy z analýzy struktury dat vyplynulo, že žádané hodnoty se nacházejí mezi složenými závorkami a jsou většinou odděleny středníky, kdy první hodnota je název sloupce a druhá je pak samotná hodnota. Z těchto hodnot je opět nutné sestavit dataset, ve kterém bude každý student mít jeden záznam, který bude obsahovat jeho username a studijní výsledky z daného předmětu. Pro toto zpracování byl napsán parser v programovacím jazyce Python a byl spuštěn vždy v každém předmětu v daném semestru na všechny soubory part_lecture.txt a part_tutoril.txt zvlášť. Username byl získán z názvů složek a obsah souboru start.txt nebyl nijak zohledněn. Předzpracování tak zahrnuje i studenty kombinované formy studia, kterých je malé množství a ve finálním datasetu nebudou zahrnuti, jelikož studijní výsledky budou připojeny k předzpracovaným záznamům z přihlášky, kde tato selekce již proběhla.

Takto upravené soubory pak byly převedeny do XLS formátu pro přehlednější práci např. s filtry apod. V tomto formátu byly prozkoumány odlišnosti a stav souborů part_tutorial.txt a part_lecture.txt. Pro předzpracování byl použit vždy jen jeden z nich v závislosti na analýze daného předmětu. Předzpracování pak probíhalo po jednotlivých předmětech napříč semestry. Data z každého předmětu byla obohacena o informaci, kolik bodů celkem student získal, zda mu byl udělen zápočet, zda úspěšně absolvoval zkoušku a s jakou známkou předmět zakončil. Tyto informace nebudou nijak využity ve finálním datasetu, nicméně mohou být užitečné pro další analýzy.

Nutné zmínit, že po bližším prozkoumání export z EDUXu za semestr B101 nedopatřením obsahuje i některé studenty zapsané v semestru B091, pro který data nejsou údajně vůbec k dispozici vzhledem k nefunční záloze z tohoto roku. Jelikož data ze semestru B091 jsou nekompletní a nejsou k dispozici pro všechny předměty, byla sice předzpracována z důvodu pozdější integrace do DWH ČVUT, ale nebyla zahrnuta do finálního datasetu.

Všechny exporty byly zpracovány po semestrech, jelikož se jednotlivé běhy předmětů od sebe liší. Jejich sloučení bude možné po provedení normalizace hodnot. Každý předmět také obsahuje tabulku, ve které jsou zaznamenány bodové zisky z aktivit na základě analýzy předmětů.

BI-CAO
Z analýzy předmětu vyplynulo, že pravidla hodnocení v BI-CAO byla v každém semestru (B101-B151) vždy stejná, viz 4.7. Pro předzpracování byly použity soubory part_tutorial.txt a vzhledem k relativně dobré kvalitě byly ze

všech dostupných dat (1. test, 2. test, 3. test a aktivita) stanoveny následující atributy:

- username,
- cao_test1 - cao_test3,
- cao_cv1 - cao_cv13,

kdy cao_test1-3 obsahuje bodový zisk z jednotlivých testů a cao_cv1-13 pak počet bodů udělených za aktivitu v jednotlivých cvičeních. Výsledná data nijak nezohledňují opravný test, vzhledem k tomu, že jej nelze v datech identifikovat.

Položka	**B101**	**B111**	**B121**	**B131**	**B141**	**B151**
1. test	10	10	10	10	10	10
2. test	20	20	20	20	20	20
3. test	20	20	20	20	20	20
Aktivita	10	10	10	10	10	10

Tabulka 4.7: Bodové hodnocení v BI-CAO.

BI-MLO

V předmětu BI-MLO byly brány v potaz pouze testy. Kromě testů jsou v předmětu udělovány také body za aktivitu, které však nejsou zaznamenávány do EDUXu, proto se v datasetu nevyskytují. Pro předzpracování byl použity soubory part_tutorial.txt, který obsahuje potřebné informace pro udělení zápočtu. Soubor part_lecture pak obsahuje informace potřebné pro udělení známky.

Položka	**B101**	**B111**	**B121**	**B131**	**B141**	**B151**
1. test	6	6	6	10	15	15
2. test	6	6	6	10	15	15
3. test	6	6	6	10	-	-
4. test	6	6	6	-	-	-
5. test	6	6	6	-	-	-
6. test	6	6	6	-	-	-
Aktivita	2 x n	-	-	-	-	-

Tabulka 4.8: Bodové hodnocení v BI-MLO.

BI-ZMA
V předmětu BI-ZMA byly předzpracovány pouze testy a on-line kvízy, přestože některé běhy obsahovaly i docházku, která ale byla čitelná pouze v semestru B101, kdy atribut pro docházku obsahoval "x"pro zaznamenání studentovy přítomnosti a "a"pro evidenci absence. V ostatních semestrech nebylo vyplňování těchto polí jednotné, formát se lišil vždy cvičící od cvičícího a někteří tato pole využívala jako poznámku, která např. obsahovala, zda byl student u tabule, že nerozumí limitám, přestože je měl v domácím úkolu správně apod. Některé záznamy dokonce obsahovaly vzorce pro výpočet bodového zisku z dané hodiny. V semestru B131 všechny záznamy u atributů pro docházku obsahovaly hodnotu 0. Předzpracování takových hodnot by bylo neúměrně náročné s nejistým výsledkem vzhledem k celkovému objemu dat.

Dále exporty obsahují sloupec, ve kterém je zaznamenán celkový bodový zisk za aktivitu a docházku na cvičení. Tento atribut, ač byl předzpracován, nebude zahrnut ve finálním datasetu, protože obsahuje hodnotu nabytou na konci semestru a neříká nám žádnou informaci o konkrétním průběhu v jednotlivých týdnech.

Položka	**B101**	**B111**	**B121**	**B131**	**B141**	**B151**
1. test	10	10	20	20	20	12
2. test	10	15	20	20	20	12
3. test	10	15	-	-	-	12
4. test	10	-	-	-	-	-
Aktivita	10	10	10	5	5	10
Domácí úlohy	-	3 x 3	-	-	-	-
On-line kvízy	-	-	-	-	-	3 x 8

Tabulka 4.9: Bodové hodnocení v BI-ZMA.

BI-PS1
V předmětu BI-PS1 byly pro předzpracování použity soubory part_lecture.txt, soubory part_tutorial.txt obsahují i bodové zisky z jednotlivých sekcí testů. Tato informace by byla užitečná v případě sestavení předpovědi úspěšnosti studentů v jednotlivých předmětech, nicméně pro rozsah této práce se jedná o informaci nadbytečnou. V tomto předmětu se nijak nezaznamenává docházka, výsledkem předzpracování jsou tedy bodové zisky z testů.

Položka	B101	B111	B121	B131	B141	B151
1. test	15	15	15	15	20	25
2. test	15	15	20	20	20	25
3. test	15	20	20	20	20	25
4. test	55	50	45	45	40	25
Opravný test	-	-	-	-	-	25
Aktivita	-	-	-	-	-	5

Tabulka 4.10: Bodové hodnocení v BI-PS1.

4.1.3.1 Shrnutí

V semestru B101 je nutné zohlednit přítomnost některých záznamů z B091, které po předzpracování vypadly především díky přiřazení infromace, zda student v B101 předmět úspěšně zakončil. V BI-ZMA v semestru B111 a B121 došlo k velké ztrátě záznamů, což bylo způsobeno nejednotným procesem vyplňování hodnot, kdy napsaný parser nedokázal zareagovat a vhodně data separovat. Nutno podotknout, že se nejedná vyloženě o chybu parseru, ale o zanesený šum při vyplňování hodnot. Např. pokud někdo vyplní do pole středník, stane se tak textový soubor nepoužitelným. Další problém může nastat při používání odlišného počtu atributů cvičícími. Obdobný problém je zaznamenán i v předmětu BI-MLO v B131. Kromě lidského faktoru při vyplňování EDUXu může také nastat chyba při poskytování exportu. Nejednou se stalo, že některé předměty byly chybně (obsahovaly špatné počty studentů, víz B101) nebo dokonce nebyly vůbec exportovány (obsahovaly prázdné složky). Některé exporty mohou být neúplné i z důvodu již vymazaných/ztracených dat z EDUXu, vzhledem k tomu, že wiki není vhodným nástrojem pro uchovávání historie. Dále byly odstraněny prázdné záznamy.

Tabulka 4.11 ukazuje, kolik procent záznamů je po předzpracování použitelných.

Semestr	BI-CAO	BI-MLO	BI-ZMA	BI-PS1 (BI-UOS)
B101	41%	49%	38%	77%
B111	84%	99%	18%	79%
B121	82%	91%	21%	83%
B131	78%	26%	80%	90%
B141	89%	86%	87%	81%

Semestr	BI-CAO	BI-MLO	BI-ZMA	BI-PS1 (BI-UOS)

Tabulka 4.11: Datová kvalita záznamů.

4.2 Problémy s daty

S každými daty přichází určité problémy, obzvlášť pokud jsou hodnoty vyplňovány manuálně uživateli. Problémy, které se vyskytovaly v přihlášce, byly shrnuty v [28] a v rámci implementace DWH ČVUT byly některé napraveny. Proto zde nebudou již zmiňovány. V této sekci se zaměříme především na data ze systému EDUX.

4.2.1 Chybějící hodnoty

Chybějící hodnota může znamenat, že žádná z možných není vhodná, že hodnota při sběru nebyla zaznamenána (vyučující zapomněl hodnotu zapsat), že student získal z dané aktivity 0 a vyučující ji nezapsal nebo že se student dané aktivity vůbec nezúčastnil.

Specifický problém nastává s nulou - nemáme žádnou informaci o tom, zda student, který v hodnocení obdrží 0, se aktivity účastnil s výsledkem 0 a nebo na aktivitě chyběl, a tak automaticky obdržel 0 bodů. Výsledek je pořád stejný, ale každý s sebou nese jinou informaci, která by při předpovědi dalšího vývoje mohla být důležitá a mohla by zodpovědět např. otázky: Jedná se o studenta, který se neúčastní testů? Nebo se jedná o studenta, který se nějakým způsobem snaží, ale jeho výkon je mizerný?

Z tohoto důvodu byly chybějící hodnoty nahrazeny řetězcem *null*, aby nedošlo k záměně s 0 a zanesení ještě většího šumu.

Pokud by bylo chybějících hodnot v dané instanci neúměrně mnoho, např. bychom disponovali pouze daty z přihlášky, ale chyběly by nám studijní výsledky, je mnohem výhodnější takové instance odstranit.

Mezi další postupy nakládání s chybějícími daty patří např. nahrazení průměrem/modem atributu, čímž by došlo ke zfalšování výsledků a tak zanesení významového šumu v predikci úspěšnosti studentů.

4.2.2 Chybějící metadata

Exporty neposkytují žádná metadata, která by blíže specifikovala jednotlivé atributy a obsahující hodnoty.

Pro identifikaci obsahujících hodnot je nyní nutné projít sekci Hodnocení na portále EDUX, obsahující pravidla hodnocení, která si v ideálním případě nechat potvrdit garantem daného předmětu, jelikož občas dochází ke změnám, které na EDUXu uvedené nejsou, např. vynásobení všech výsledků z posledního testu konstantou, aby se navýšila průchodnost v předmětu BI-UOS aj. akce. Další problém může nastat u chybějící zálohy stránky v archivu, takže pravidla mohou být nedohledatelná.

Tento způsob s sebou tedy nese velkou režii s nejistým výsledkem, protože EDUX nám nedává žádnou informaci, ve kterém konkrétním sloupci se dané aktivity vyskytují a jaký je požadovaný formát dat.

4.2.3 Formát dat

Z exportů ani EDUXu není nijak jasný formát dat, resp. zda jednotlivé atributy mají omezení na datové typy. Formát dat byl diskutován se správcem EDUXu Ing. Tomášem Kadlecem a jeho neomezení je údajně z poskytnutí volnosti garantům předmětů. Toto řešení je v některých případech nešťastné, viz textové poznámky v docházce v BI-ZMA, nicméně se do budoucna nebude měnit.

4.2.4 Povinné sloupce

EDUX momentálně nedisponuje funkcí, která by vyžadovala povinnost vyplnění některých sloupců. Mohou tedy nastat případy, kdy vyučující nebude do EDUXu zapisovat žádné studijní výsledky a studenty o nich bude informovat pomocí jiného nástroje, např. Google Documents, Moodle apod. Přestože existuje vnitřní nařízení o povinném používání EDUXu, existují předměty, které toto nařízení ignorují a na EDUXu o nich nenajdeme téměř žádné informace.

4.3 Normalizace hodnot

Jedním z kroků předzpracování dat je jejich normalizace, která řeší problém různých rozsahů vzdáleností. V našem případě se jedná o porovnávání hodnot mezi semestry, kdy je každá z jinak velkého intervalu a je nutné informaci o velikosti intervalu zohlednit. Možnosti normalizace byly zkonzultovány s Ing. Tomášem Kalvodou, Ph.D. (zástupce vedoucího KAM FIT). Pro tento problém se nejlépe hodí tzv. min-max normalizace [29].

4.3.1 Min-max normalizace

Při min-max normalizaci se hodnoty lineárně transformují do nového oboru hodnot, nejčastěji v intervalu [0, 1] nebo [-1, 1]. V našem případě se bude

jednat o normalizaci pouze v kladném intervalu [0, 1], abychom se vyhnuli případným potížím se zápornými hodnotami. Podmínkou je znát minimální a maximální hodnotu původního intervalu. Distribuce hodnot po aplikaci zůstává stejná.

Případně můžeme narazit na problém s odlehlými hodnotami (en. outliers). Takové případy nám pomohou odhalit distribuční grafy, které budou zároveň sloužit k validaci hodnot.

Použitý vzorec pro min-max normalizaci:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

kde $x = (x_1, ..., x_n)$ a z_i je i-tá aktuálně normalizovaná hodnota.

Při použití min-max normalizace existují dvě varianty maxima:

1. Maximum bodů ze semestru
   Porovnání získaných bodů z dané aktivity relativně k maximu bodů z celého semestru s sebou nese informaci, jak moc se student danou aktivitou přibližuje k zápočtu (většinou polovina maxima bodů ze semestru).

2. Maximum bodů z aktivity
   Porovnání získaných bodů z aktivity ku možnému maximu z dané aktivity nám dává informaci, jak si student vede při jednotlivých aktivitách bez informace, jak si vede vůči celému semestru.

Data byla upravena pro každou variantu a postoupena rozhodovacímu stromu pro testování výsledků. Varianta s maximem bodů ze semestru vykazovala až o 8% menší prediktivní přesnost, což může souviset i s nevhodným uplatněním pro rozhodovací stromy - je možné, že se model naučí např. na hodnotách v intervalu [0, 0.3] , protože nebude možné se daným testem více než třetinou přiblížit zápočtu v daném semestru. Avšak při aplikaci v jiném semestru dojde ke změně pravidel a hodnoty se budou pohybovat v intervalu [0, 0.5]. Naučený model takové hodnoty neočekává, a tak lze jeho výsledky změnou intervalu zcela znehodnotit.

Maximum bodů z aktivity je v našem případě efektivnější, např. pokud porovnáváme testy z různých semestrů mezi sebou a nemáme informaci o celkovém počtu testů. Tuto informaci většinou máme díky analýze předmětů, ale nemáme ji jak předat modelu. V úvahu připadá pouze přiřazení vah, ale po konzultaci s Ing. Tomášem Kalvodou, Ph. D. byla tato možnost zavrhnuta, vzhledem k velké odlišnosti napříč semestry a nemožnosti přiřazení smysluplných vah.

4.3.2 Aplikace v předmětech

Při každé aplikaci normalizace byl pro daný atribut sestaven distribuční graf, který pomohl ke snazší identifikaci odlehlých hodnot. Pokud nalezená hodnota byla větší než maximum bodů z testu, byla snížena na maximum, pokud byla nižší než minimum, byla taková hodnota zvýšena na minimum. Jednalo se např. o záporné body při absenci v BI-ZMA nebo bonusové body za excelentní řešení. Takových hodnot bylo minimum.

BI-CAO
V tomto předmětu nebyla data znormalizována, protože se napříč semestry pravidla hodnocení nezměnila.

BI-PA1
Hodnoty v BI-PA1 byly již z normalizovány, data se pohybují v intervalu [0, 1], místy 1.2 díky bonusovým bodům např. za včasné odevzdání úlohy. Přístup k původním hodnotám nemáme.

BI-PS1
Absolvování předmětu BI-PS1 každý semestr zahrnovalo vždy 4 testy, které se v bodovém hodnocení lišily +- 5 body, nicméně celkový součet zůstal vždy roven 100 bodům. Aby byla tato data mezi sebou porovnatelná, jednotlivé bodové zisky z testů byly vždy vyděleny maximálním možným ziskem z daného testu. Takto upravené hodnoty se tedy vždy pohybují v intevalu 0 až 1 a udávají studentovu úspěšnost vůči danému testu, kdy 0 je vždy minimum a 1 maximum, tzn. plný počet. Jedná se tedy o aplikaci min-max normalizace.

BI-ZMA
BI-ZMA se svojí strukturou liší oproti ostatním předmětů nejzásadněji. Pro možnost porovnání hodnot byla zvolena také min-max normalizace, kdy maximem (1) je maximální bodový zisk z daného testu a minimum (0) představuje minimální bodový zisk. V B151 je po konzultaci hodnota testu tvořena vždy součtem bodů z n-tého testu a n-tého kvízu a tato hodnota je následně znormalizována.

BI-MLO
V předmětu BI-MLO byl zvolen jiný způsob normalizace vzhledem ke konstantnímu celkovému zisku za semestr. Sjednocení hodnot proběhlo následujícím způsobem:

- B101, B111, B121: 1. + 2. + $\frac{1}{2}$ 3. testu = **mlo_test1**, $\frac{1}{2}$ 3. + 4. + 5. test = **mlo_test2**
- B131: - 1. + 1/2 2. testu = **mlo_test1**, 1/2 2. + 3. test = **mlo_test2**

- B141 je stejný jako B151 - ponechány původní hodnoty

4.4 Výběr vhodných dat

Součástí práce je také stanovení vhodného termínu pro aplikaci prediktivního modelu. Výsledky této predikce by měly v semestru B151posloužit k identifikaci skupiny ohrožených studentů, která bude oslovena s nějakou formou pomoci. Proto je velmi důležité stanovit termín aplikace tak, aby měl model dostatek atributů pro rozhodování a zároveň nebylo příliš pozdě na pomoc studentům.

Pro názornost byl sestaven Ganttův graf, který je hojně využíván při řízení projektů pro znázornění činností v čase. Jednotlivé činnosti na sebe nijak nenavazují a každá činnost pokrývá vždy celý týden, protože každý předmět je vyučován v několika paralelkách v různé časy a dny. Garanti všech zainteresovaných předmětů se zaručili, že všechny testy budou opraveny nejpozději do 1 týdne od jejich napsání. S tímto faktem je potřeba počítat.

Z Ganttova grafu, viz příloha A.1, vyplývá, že vhodným týdnem pro sběr dat je začátek 8. výukového týdne, kdy budou k dispozici tato data:

- **BI-ZMA:** 1. kvíz a 1. test
- **BI-PA1:** 0. úkol, 1. úkol, 2. úkol, 1. a 2. znalostní test
- **BI-CAO:** 1. test
- **BI-MLO:** 1. test
- **BI-PS1:** 1. test

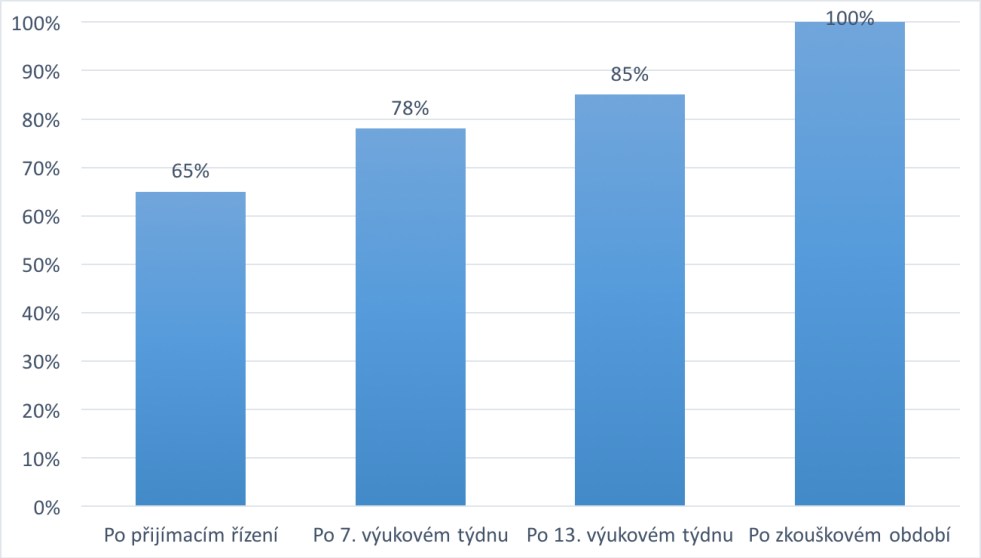
Toto rozhodnutí bylo podpořeno testováním přesnosti predikce v čase na prediktivním modelu rozhodovacího stromu, kterému byly předány všechny dostupné atributy pro B101-B141, které byly postupně odebírány.

Data ze semestru B151 byla získáná a předzpracována stejným způsobem, jak bylo popsáno pro semestry B101 až B141.

4.4.1 Sestavení datasetu

Na základě stanovení vhodých dat byl sestaven finální dataset.

Pro jeho sestavení bylo nutné spojit všechny předzpracované datasety s daty z přihlášky. Exporty obsahují pouze username studenta, který není jednoznačným identifikátorem studia. Student může být na univerzitě zapsán do více studií na více fakultách, některá studia mít aktivní, jiná ukončená, ale po celou

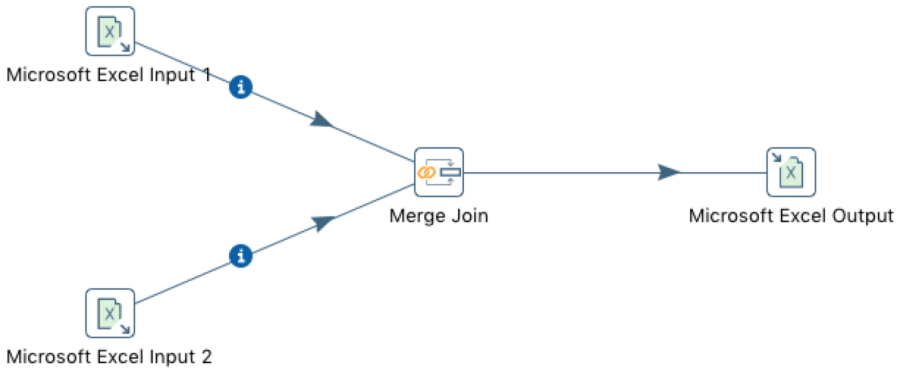


Obrázek 4.1: Přesnost predikce v čase.

dobu mu zůstává totožný username. Např. student mohl nastoupit studium v B101, v B102 studium ukončit a v B111 nastoupit nové studium, což znamená, že má stejný username, ale dvě odlišná ID studia a dvě odlišné přihlášky.

Abychom tedy ke studentovi přiřadili správnou přihlášku a tak i studium, je nutné do exportů z EDUXu doplnit ID studia, která získáme z DWH ČVUT pomocí SQL dotazu. Spojování bude probíhat přes username, abychom se vyhnuli duplicitám a přiřadili k username správné studium, bude provedeno po jednotlivých semestrech. Z DWH ČVUT tedy získáme username a stud_id všech studentů zapsaných na konkrétní běh předmětu.

```
SELECT DISTINCT
  zapsane_predmety."kod_semestru",
  zapsane_predmety.stud_id,
  osoba.username,
  predmet.kod AS "kod_predmetu"
FROM (
      SELECT
        zap.studiumid_studia     AS stud_id,
        beh.arok_semestrsemestr_id AS "kod_semestru",
        beh.pred_predmetpredmety_id
      FROM
        public.zapi_predmet zap LEFT JOIN pred_beh_predmetu beh
          ON zap.pred_beh_predmetubeh_predmet_id = beh.
              beh_predmet_id_bk
```



Obrázek 4.2: Transformace pro spojení souborů.

```
      WHERE beh.arok_semestrsemestr_id = 'B101') zapsane_predmety
 -- misto B101 lze zadat libovolny semestr
  LEFT JOIN pred_predmet predmet ON zapsane_predmety.
      pred_predmetpredmety_id = predmet.predmet_id_bk
  LEFT JOIN stud_studium studium ON studium.id_studia_bk =
      zapsane_predmety.stud_id
  LEFT JOIN osob_osoba osoba ON studium.osob_osobaperidno = osoba.
      peridno_bk
WHERE predmet.kod = 'BI-ZMA';
 -- misto BI-ZMA lze zadat libovolny predmet
```

Po získání těchto dat je nutné je spojit s předzpracovanými daty z EDUXu a Progtestu, která také obsahují pouze username. Výsledné soubory pak můžeme připojit k připraveným datům z přihlášky. Pro spojení dat byl použit open-source nástoj Pentaho Kettle (součást Pentaho Data Integration), který je primárně určen pro ETL/ELT operace, které jsou využívány při nahrávání dat do datového skladu, nicméně se dá také použí při předzpracování dat. Tento nástroj jsem zvolila z důvodu jeho znalosti v rámci implementace DWH ČVUT.

V Pentaho Kettle byla vytvořena jedna transformace join_files.ktr, u které jsou vždy obměňovány vstupní soubory a název výstupního, viz 4.2.

Název kroku	**Popis**
Microsoft Excel Input	Nahrání vybraného XLS souboru

Název kroku	Popis
Merge Join	Sloučení vstupních souborů podle klíčové hodnoty
Microsoft Excel Output	Uložení zpracovaných dat do zvoleného XLS souboru

Tabulka 4.12: Popis jednotlivých kroků transformace join_files.

V kroku Merge Join lze nastavit druh sloučení (Join Type) na inner, left outer, right outer nebo full outer. V našem případě byl zvolen left outer, jelikož Microsoft Excel Input 1 obsahuje vždy požadovaná data a Microsoft Excel Input 2 obsahuje jejich rozšíření (např. stud_id). Dále je nutné nastavit klíčovou hodnotu (Key field), tzn. atribut, podle kterého budou soubory sloučeny. V případě rozšiřování dat z EDUXu a Progtestu bude klíčovým atributem username, v případě napojování dat k příhlášce bude klíčovým atributem ID studia. Před spuštěním transformace je nutné, aby byla data stejně seřazená (kvůli kroku Merge Join).

Výsledný dataset obsahuje 52 atributů, z nichž stud_id slouží pouze pro identifikaci konkrétních studentů a postup označuje cílový sloupec predikce.

Název atributu	Popis
stud_id	identifikátor studia
cislo_prihlasky	pořadové číslo přihlášky
semestr	kód semestru, kterým student zahájil studium
pohlavi	pohlaví studenta
vek_pri_nastupu	věk v době zápisu do studia
rok_narozeni	rok narození
rodinny_stav	rodinný stav
obcanstvi	občanství
misto_narozeni	místo narození
bydliste_prechodne	město přechodného bydliště
bydliste_trvale	město trvalého bydliště
puvodem_praha	příznak, zda má trvalé bydliště v Praze
odkud_skola	odkud se uchazeč hlásí (např. zaměstnání, SŠ apod.)
predch_studium	předchozí studium
matur_prum	průměr známek z maturity
matur_znamky	známky z maturity

Název atributu	Popis
rok_matur	rok maturity
typ_ss	typ střední školy
st_predch_vzd	stupeň předchozího vzdělání
rozhodnuti_prijeti	zda student absolvoval přijímací řízení
typ_prijeti	typ přijetí
hodnoceni_zk	hodnocení přijímací zkoušky
cislo_zk	číslo zadání přijímací zkoušky
pocet_bodu_zk	počet bodů z přijímací zkoušky
cestina_zk	příznak, zda student absolvoval zkoušku z češtiny
olympiady	příznak, zda student absolvoval olympiády
scio	výsledek ze SCIO testu
cao_cv1	body za aktivitu na 1. cvičení BI-CAO
cao_cv2	body za aktivitu na 2. cvičení BI-CAO
cao_cv3	body za aktivitu na 3. cvičení BI-CAO
cao_cv4	body za aktivitu na 4. cvičení BI-CAO
cao_cv5	body za aktivitu na 5. cvičení BI-CAO
cao_cv6	body za aktivitu na 6. cvičení BI-CAO
cao_test1	body z 1. testu BI-CAO
mlo_test1	body z 1. testu BI-MLO
zma_test1	body z 1. testu BI-ZMA
ps1_test1	body z 1. testu BI-PS1
1L_dny_od_zacatku	počet dnů do deadline 1. lehké úlohy ode dne 1. odevzdání
1L_nejvice_bodu	počet nejvíce dosažených bodů z 1. lehké úlohy
1L_pocet_pokusu	počet pokusů o devzdání 1. lehké úlohy
1T_dny_od_zacatku	počet dnů do deadline 1. těžké úlohy ode dne 1. odevzdání
1T_nejvice_bodu	počet dnů do deadline 1. těžké úlohy ode dne 1. odevzdání
1T_pocet_pokusu	počet pokusů o devzdání 1. těžké úlohy
2L_dny_od_zacatku	počet dnů do deadline 2. lehké úlohy ode dne 1. odevzdání
2L_nejvice_bodu	počet nejvíce dosažených bodů z 2. lehké úlohy
2L_pocet_pokusu	počet pokusů o devzdání 2. lehké úlohy
2T_dny_od_zacatku	počet dnů do deadline 2. těžké úlohy ode dne 1. odevzdání
2T_nejvice_bodu	počet dnů do deadline 2. těžké úlohy ode dne 1. odevzdání

Název atributu	**Popis**
2T_pocet_pokusu	počet pokusů o devzdání 2. těžké úlohy
pa1_test1	body z 1. znalostního testu BI-PA1
pa1_test2	body z 2. znalostního testu BI-PA2
postup	příznak, zda student postoupil do dalšího semestru

Tabulka 4.13: Popis jednotlivých atributů finálního datasetu

Kapitola 5

Prediktivní modelování

Prediktivní modelování [16] je jednou z metod strojového učení, jehož počátky spadají až do 50. let 20. století. Jedna z definic strojového učení zní: *„Učení je proces, kdy systém zvyšuje svůj výkon na základě zkušeností.“* Mnoho algoritmů strojového učení spočívá ve vybudování prediktivního modelu pro vytvoření rozhodnutí a předpovědí na základě dostupných dat.

Rozlišujeme 3 základní typy úloh strojového učení:

1. **Regrese** - odhaduje číselnou hodnotu výstupu na základě vstupních dat.

2. **Klasifikace** - rozděluje vstupní data do dvou a více tříd.

3. **Shlukování** - objekty s podobnými vlastnostmi zařazuje do skupin.

Dále lze algoritmy strojového učení rozdělit podle způsobu učení:

1. **Učení s učitelem** (en. supervised learning) - pro vstupní data je definován správný výstup.

2. **Učení bez učitele** (en. unsupervised learning) - pro vstupní data není znám výstup.

3. **Kombinace předchozích** (en. semi-supervised learning) - pro některá data známe správný výstup, pro zbývající ne.

Sestavení prediktivního modelu pro předpověď postupu studentů mezi semestry lze zařadit jako úlohu klasifikace vhodnou pro algoritmy schopné učení s učitelem. Výsledkem modelu je klasifikace záznamů do dvou skupin - postoupil, nepostoupil, a protože máme k dispozici záznamy se správnými výsledky předchozích let, jedná se právě o učení s učitelem.

5.1 Volba vhodného modelu

Jak ukazuje rešerše, i ostatní vzdělávací instituce se potýkají se stejným klasifikačním problémem a disponují víceméně stejnými daty jako my nyní. Mezi nejpoužívanější metody jednoznačně patří rozhodovací stromy. Náš problém a kvalita dostupných dat přímo směřuje k použití rozhodovacích stromů či podobných metod. V rámci této práce budou vysvětleny 3 použité algoritmy, které při procesu sestavování a testování podaly nejlepší výsleky.

Při prediktivním modelování jsem spolupracovala s Ing. Janem Motlem, studentem doktorského studia FIT pod vedením Ing. Pavla Kordíka, Ph.D. Výsledky tohoto modelování budou vhodným způsobem aplikovány na studentech a mohou mít kladný či záporný dopad na jejich studijní chování. Tato spolupráce byla tedy nezbytná a velmi přínosná.

5.1.1 Rozhodovací stromy

Rozhodovací stromy jsou velmi oblíbené díky své přehlednosti a jednoduché interpretovatelnosti výsledků. Při tvorbě rozhodovacího stromu se postupuje metodou rozděl a panuj (en. divide and concquer) [30]. Data jsou postupně rozdělena na stále se zmenšující podmnožiny, tzv. uzly stromu. V těchto podmnožinách převládají příklady jedné třídy. Na konci tohoto procesu tedy máme podmnožiny tvořené příklady stejné třídy. Cílem je nalézt takový strom, který bude konzistentní s daty.

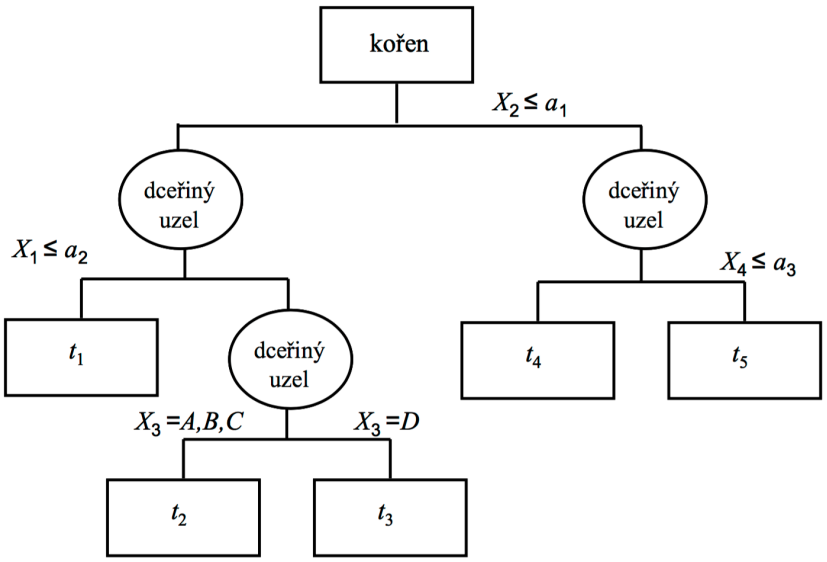
Stěžejní otázkou při konstrukci rozhodovacího stromu je výběr vhodného atributu pro větvení stromu, tzn. vybrat takový atribut, který od sebe nejlépe odliší příklady různých tříd. Vhodnými charakteristikami atributu mohou být např. entropie, informační zisk, poměrný informační zisk, Chí kvadrát nebo Giniho míra diverzity.

Druhým důležitým problémem je stanovení podmínky zastavení dělení při konstrukci rozhodovacího stromu. Pokud bychom nepoužili žádnou podmínku, dospěli bychom k úplnému stromu, který by sice trénovací data klasifikoval se 100% úspěšností, ale měl by tendenci k přeučení a byl by pravděpodobně příliš velký. Pro stanovení zastavení se nejčastěji používají tyto metody:

- Zastavovací pravidlo - Proces dělení se zastaví v případě, kdy neexistuje statisticky signifikantní rozdělení.
- Prořezávání (en. pruning) - Strom je zkonstruován do maximální šíře a při zpětném průchodu odstraníme listy a větve, které podle vhodně zvoleného statistického kritéria nelze považovat za významné.

5.1.1.1 Implementace

V dnešní době existuje celá řada nástrojů, která umožňuje použití různých algoritmů pro rozhodovací stromy, nicméně po zvážení všech důsledků této práce jsem se rozhodla pro použití nástroje BigML a to především z důvodu jeho dostupnosti (webová aplikace dostupná na www.bigml.com), jednoduché interpretaci (velmi srozumitelné grafické zpracování), efektivitě a možnosti rychlého opakovaného použití bez hlubší znalosti i v budoucnosti. BigML je dostupné ve vývojářském módu při kapacitě do 16MB úloh (např. nahrávání dat, evaluace, predikce apod.) zcela zdarma, což je v našem případě dostačující. BigML také psokytuje celou řadu metod klasifikace, regrese a shlukování, disponuje prostředím pro detekci anomálií v datasetech či hledání asociačních pravidel.



Obrázek 5.1: Grafická struktura rozhodovacího stromu CART. Indexy u terminálních uzlů udávají v jakém pořadí došlo k oddělení jednotlivých terminálních uzlů. Prediktoru X_1, X_2 a X_4 jsou spojité, prediktor X_3 je kategoriálná s kategoriemi A, B, C, D. [31]

Po seznámení se s veškerou funkcionalitou BigML a po testování různých modelů a nastavení byl zvolen model, který využívá binární rozhodovací strom inspirovaný CART (Classification and Regression Trees) pro úlohu klasifikace [32]. CART patří mezi jeden z nejnámějších a nejpoužvanějších algoritmů pro vytváření binárních stromů a je jejich základním představitelem. Ostatní binární stromy lze získat modifikací právě CART [31], obdobně jako je tomu

u BigML, které CART rozšiřuje o další pravidla.

Pro větvení bylo využito poměrného informačního zisku [30], který mění informační zisk tak, aby netíhl k výběru atributů s mnoha hodnotami, což je jeho hlavní problém. Informační zisk i poměrný informační zisk jsou míry odvozené z entropie. Entropii spočteme jako

$$H(S) = -\sum_{i=1}^{n} P(s_i) log_2 P(s_i)$$

kde $P(s_i)$ je pravděpodobnost, že libvovolný příklad S je typu s_i. Informační zisk se pak spočte jako rozdíl entropie pro celá data (cílový atribut) a pro uvažovaný atribut. Informační zisk tak měří redukci entropie způsobenou volbou atributu S:

$$Informační\ zisk(S) = H(C) - H(S)$$

V případě entropie se snažíme najít atribut s minimální hodnotou, v případě informačního zisku hledáme atribut s maximální hodnotou. Ani jedno kritérium nebere v potaz počet hodnot zvoleného atributu, proto se používá poměrný informační zisk:

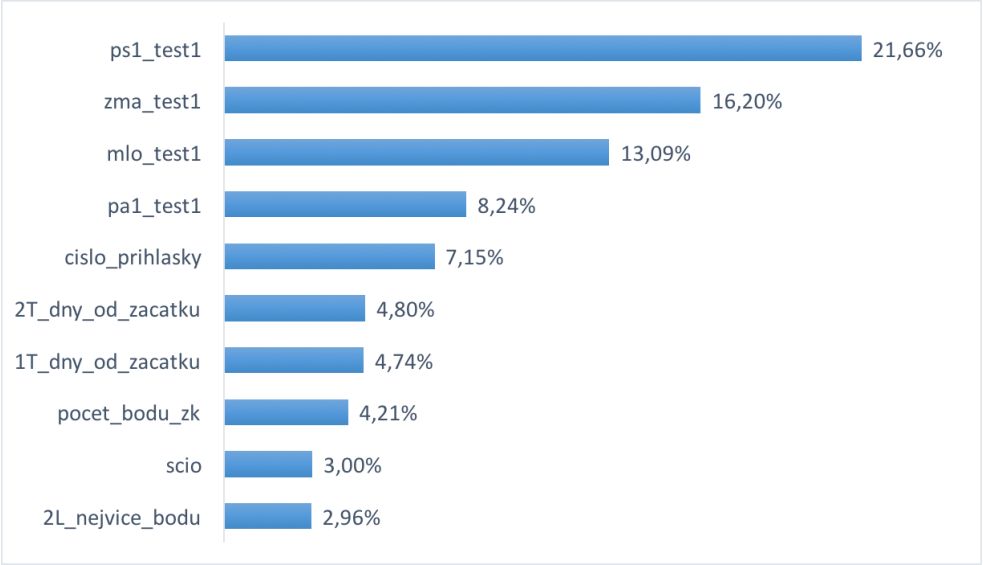
$$Poměrný\ informační\ zisk\ (S) = \frac{Informační\ zisk\ (S)}{Větvení\ (S)}$$

Větvení(S) je entropie dat vzhledem k hodnotám atributu S, resp. entropie distribuce instancí do větví.

Dále zvolený model využívá jako pravidlo pro zastavení růstu statistické prořezávání (en. statistical pruning), které se ukázalo jako velmi efektivní řešení při zabránění přeučení (en. overfitting) modelu. Prořezávání je založeno na odhadech spolehlivosti.

V grafu 5.2 je zobrazeno pořadí prvních 10 atributů s největším vlivem při prediktivním modelování. Je vidět, že mezi nejvlivnější atributy patří 1. test z BI-PS1, který tvoří kořen rozhodovacího stromu, dále pak 1. test z BI-ZMA a BI-MLO. Z přihlášky má znatelný vliv pouze její pořadové číslo, počet bodů z přijímací zkoušky organizované FIT a počet bodů ze Scio testů. Zajímavé je, že předmět BI-CAO má při predikci minimální vliv stejně jako socio-demografické údaje o studentovi.

Tento sestavený rozhodovací strom dosáhl při testování 75% přesnosti klasifikace.



Obrázek 5.2: Deset nejvýznamnějších atributů.

5.1.2 Dvouatributový model

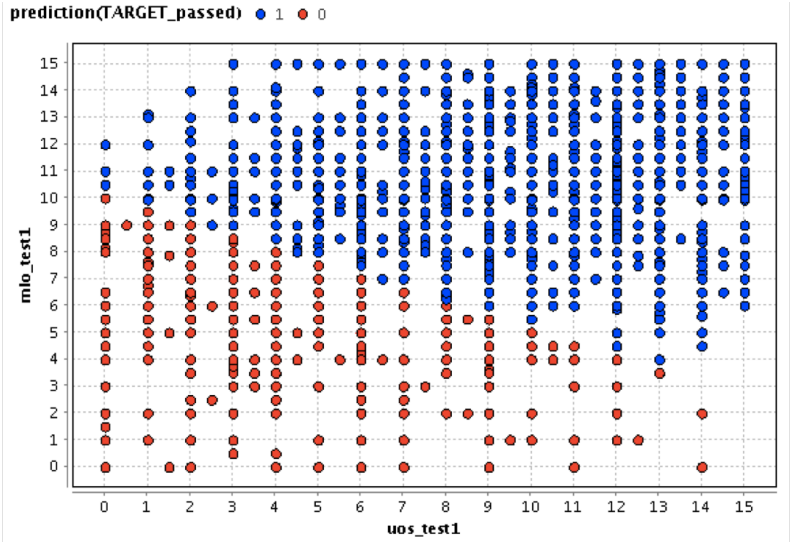
Tento model byl navržen Ing. Janem Motlem na základě dodaného finálního datasetu. Predikce je založena pouze na dvou atributech:

- výsledky prvního testu z BI-MLO - atribut mlo_test1,
- výsledky prvního testu z BI-PS1 - atribut ps1_test1.

Ostatní atributy byly vyhodnoceny s nulovým přínosem nové informace.

Na obrázku 5.3 vidíme, že téměř všichni studenti, kteří získali z 1. testu v BI-MLO 10 a více bodů, postoupili do dalšího semestru. Z výsledků z 1. testu v BI-PS1 nic takového usuzovat nelze. Pomocí lineární regrese [33] bylo ukázáno, že není nutné při predikci přihlížet na otatní předměty. V tabulce 5.1 lze vidět, s jakou přesností je možné na základě prvních testů v BI-MLO a BI-PS1 (BI-UOS) předpovídat výsledky testů z ostatních předmětů.
Takto vytvořený model dosáhl při testování 77% přesnosti správné klasifikace.

Pro sestavení tohoto modelu byl použit open-source nástroj Rapid Miner Studio, který je používán v předmětech zabývajících se data miningem na FITu a zároveň se v dnešní době jedná o jeden z nejpopulárnějších nástrojů v komunitě zabývající se strojovým učením.



Obrázek 5.3: Rozložení bodových zisků z 1. testu BI-MLO a 1. testu BI-PS1 (BI-UOS). Červeně jsou znázorněni studenti, kteří nepostoupili do dalšího semestru, modře pak ti, kteří postoupili.

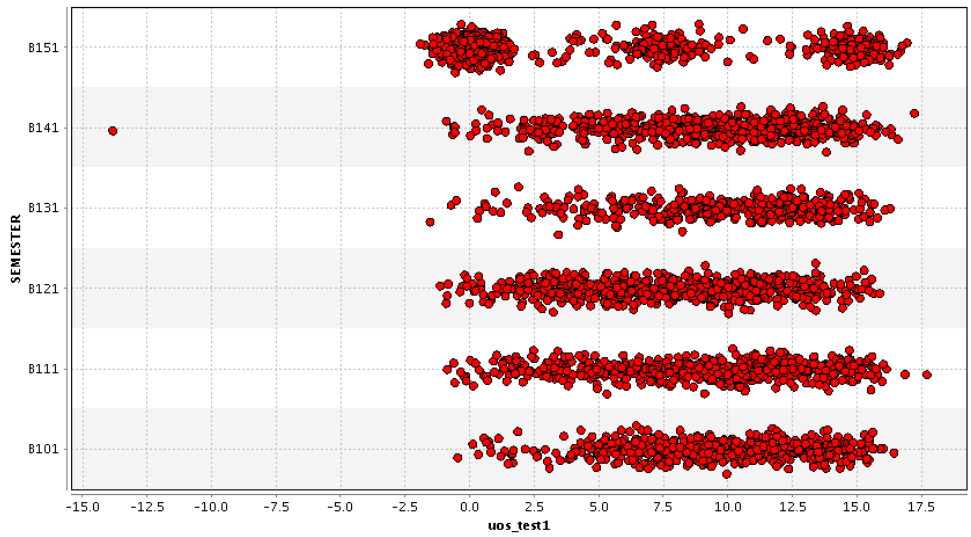
Předmět	MLO koeficient	PS1 koeficient	Přesnost - 1. test
BI-CAO	100%	0%	**98%**
BI-MLO	100%	0%	**100%**
BI-PA1	69%	31%	**96%**
BI-PAI	-	-	**-**
BI-PS1	0%	100%	**100%**
BI-ZMA	100%	0%	**98%**

Tabulka 5.1: Přesnost predikce výsledků 1. testu.

5.1.3 Random Forest

Po obdržení dat z aktuálního semestru B151 byla zjištěna změna v distribuci dat v 1. testu z BI-PS1. Tato změna lze vidět na distribučním grafu 5.4. Tato změna byla způsobena nasazením systému Progtest v BI-PS1 v B151, který provádí automatické vyhodnocení testů, tedy bez zásahů vyučujících. Vzhledem k tomu, že předchozí dvě techniky jsou citlivé na distribuční změny, byl ve spolupráci s Ing. Janem Motlem vytvořen nový model - Random Forest (náhodné lesy).

Metoda Random Forest spočívá ve vytvoření několika rozhodovacích stromů a následně v jejich složení (en. ensemble), díky čemuž se metoda lépe vyrovnává



Obrázek 5.4: Distribuce dat v předmětu BI-PS1 v B101-B151.

se změnami v distribuci dat. Metoda kombinuje tzv. bagging, který spočívá v náhodném výběru učící podmnožiny, a tak zachovává různorodost jednotlivých modelů, s náhodně vybranou podmnožinu vstupních atributů. Výstup je pak tvořen nejčastější hodnotou (tzv. modus náhodné veličiny) tříd vrácených jednotlivými stromy.

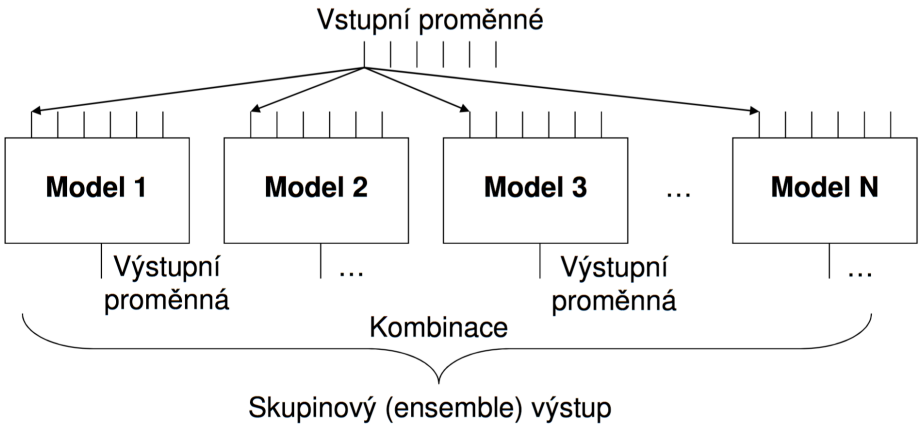
Tento model má o něco nižší přesnost, než modely předchozí, a to 62%.

Pro modelování byl použit finální dataset se všemi atributy a opět nástroj Rapid Miner Studio.

5.1.4 Kombinování modelů

Přestože byly postaveny modely poměrně výkonné, každý z modelů se chová lépe za odlišných podmínek. Z tohoto důvodu bylo přistoupeno k metodě kombinování modelů (en. ensemble methods) [30], která má za úkol ověřit prediktivní sílu modelů a utvrdit nás v přesnosti predikovaných výsledků.

Princip je velmi jednoduchý - skupina modelů je naučena na stejném úkolu. Výstupy takto naučených modelů jsou zkombinovány, jak je vidět na obrázku 5.5. Kombinované modely by měly být co nejvíce různorodé, což může být zajištěno např. různými množinami datasetů nebo různými metodami konstrukce jednotlivých modelů. Při takovém kombinování dochází k redukci rozptylu (tzv. variance) a zaujetí (tzv. bias). Jedná se tedy o možnost zlepšení



Obrázek 5.5: Kombinování modelů. [29]

výsledků, kterých dosáhly jednotlivé modely.

Mezi základní metody kombinování patří:

- Bagging (Bootstrap Aggregating) - modely jsou naučeny nezávisle, kombinován je až jejich výstup.
- Boosting - modely jsou učeny sekvenčně, trénovací data jsou závislá na chybách předcházejících modelů.
- Stacking - modely jsou naučeny nezávisle, kombinují se naučením tzv. meta modelu pomocí jejich výstupů použitých jako trénovací data.

V našem případě disponujeme třemi modely, které byly postaveny různými metodami nad datasety obsahujícími různé atributy. Není tedy pochyb o zajištění jejich různorodosti a vzhledem k tomu, že i učení modelů probíhalo nezávisle na sobě, nabízí se využití metody bagging. Důležité je zmínit, že všechny modely mají rovnocennou váhu.

Takto získaný model dosáhl při testování přesnosti 75%, což není výrazně lepší než modely předchozí, nicméně lze tyto výsledky považovat za nejstabilnější ze všech. Proto byl tento model použit při predikci skupiny ohrožených studentů v B151.

5.2 Získání výsledků

Klasifikace jako taková se skládá ze dvou fází:

		Skutečnost	
		ANO	NE
Klasifikace	ANO	TP	FP
	NE	FN	TN

Tabulka 5.2: Matice záměn obecně.

- učící,
- vybavovací.

Cílem prediktivního modelování je vytvoření klasifikátoru s co nejlepší vybavovací fází. Abychom mohli hodnotit chybovost této fáze, náhodně jsme rozdělili dataset na trénovací a testovací data v poměru 80 : 20. Na trénovacích datech naučíme model, na testovacích pak spočítáme chybu klasifikace. Tento poměr rozdělení dat je běžný, jelikož zmenšením trénovací množiny by hrozilo nekvalitní naučení modelu a tím zvětšení chyby klasifikátoru a naopak malá množina testovacích dat by snížila přesnost zjištění chyby.

Dalším způsobem stanovení chybovosti, je hodnocení algoritmu, který vytváří prediktivní model pomocí křížové validace (en. cross validation) [33], kdy jsou data rozdělena na n stejně velkých částí a následně n-krát opakujeme použití $n - 1$ částí pro naučení modelu a zbývající část pro zjištění chyby klasifikátoru. Výsledná chyba je pak průměrem z n dílčích chyb. Tato metoda je výhodná z důvodu použití všech dat pro učení.

V této práci byly v rámci testování použity obě výše zmíněné metody.

Pro vyhodnocení úspěšnosti klasifikace použijeme matici záměn (en. confusion matrix) [30], která předpokládá binární klasifikaci a říká, kolik prvků bylo klasifikováno správně a kolik špatně. V našem případě uvažujme dvě třídy, ano a ne, v matici jsou pak uvedeny všechny možné typy klasifikace:

- TP - true positive - klasifikován správně jako ANO
- TN - true negative - klasifikován správně jako NE
- FP - false positive - klasifikován jako ANO, ale ve skutečnosti je NE
- FN - false negative - klasifikován jako NE, ale ve skutečnosti je ANO

Matice záměn pro jednotlivé modely jsou ukázány na datech z B151. Vyhodnocení bylo realizováno až po skončení zkouškového období zimního semestru B151, kdy jsme měli k dispozici informaci o průchodu studentů do letního

Rozhodovací strom		**Skutečnost**	
		ANO	NE
Klasifikace	ANO	257	48
	NE	213	305

Tabulka 5.3: Matice záměn pro rozhodovací strom.

Dvouatributový model		**Skutečnost**	
		ANO	NE
Klasifikace	ANO	311	75
	NE	159	278

Tabulka 5.4: Matice záměn pro dvouatributový model.

Random Forest		**Skutečnost**	
		ANO	NE
Klasifikace	ANO	435	181
	NE	35	172

Tabulka 5.5: Matice záměn pro Random Forest.

semestru. Studenti, kteří postoupili dále, jsou skryti pod hodnotou *ANO*, studenti, kteří nepostoupili, jsou označeni jako *NE*. Pro kombinovaný model matice záměn sestavena nebyla z důvodu transformace jeho výsledků pro definici skupiny ohrožených studentů.

Jako míru pro ohodnocení klasifikátoru budeme používat celkovou správnost (en. overall accuracy), která se spočítá jako:

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

a přesnost pro jednotlivé třídy:

$$Acc_{ANO} = \frac{TP}{TP + FP}$$

a

$$Acc_{NE} = \frac{TN}{TN + FP}$$

.

Klasifikátor	**ANO**	**NE**	**Celek**
Rozhodovací strom	54,7 %	86,4 %	68,3 %
Dvouatributový model	66,2 %	57,8 %	63,8 %

Klasifikátor	**ANO**	**NE**	**Celek**
Random Forest	92,6 %	48,7 %	73,8 %
Kombinování modelů	75,1 %	73,7 %	74,5 %

Tabulka 5.6: Přesnost predikce pro jednotlivé třídy a celková.

V tabulce 5.6 vidíme, že některé modely byly úspěšnější při predikci úspěšných studentů a jiné byly úspěšnější při predikci neúspěšných studentů. Při kombinování modelů můžeme pozorovat konstantní přesnost nezávisle na třídě, což potvrzuje stabilitu a důvěryhodnost této metody pro naše zadání.

Kapitola 6

Využití výsledků

Výše popsané prediktivní modely byly použity pro predikci postupujících a nepostupujících studentů 1. ročníku BI do letního semestru akademického roku 2015/2016. Výstupy jednotlivých modelů byly zkombinovány pomocí metody bagging. Výsledkem této predikce je také skupina ohrožených studentů, tzn. studentů, kteří byli zařazeni do tříd postupujících nebo nepostupujích s průměrnou správností klasifikace na základě kombinace výstupů modelů.

Celkem z 823 studentů (původní dataset obsahoval 825, ale dva studenti změnili formu studia na kombinovanou) byly definovány tyto skupiny studentů:

- studenti s vysokou pravděpodobností postupu,
- studenti s nízkou pravděpodobností postupu,
- zcela neaktivní studenti.

Skupina studentů s vysokou pravděpodobností průchodu zahrnuje 444 studentů. Jedná se o studenty, kteří by měli být schopni postoupit do dalšího semestru bez větších komplikací. Jakékoliv upozornění ze strany fakulty by na tyto studenty mohlo mít negativní vliv, proto je stavíme stranou našeho zájmu.

Pro vedení fakulty jsou nejtíživější studenti s nízkou pravděpodobností postupu,tzv. ohrožení, jelikož je u nich největší potenciál pro nabídnutí pomocné ruky a následné zajištění jejich průchodu do dalšího semestru. Tato skupina čítá 243 studentů.

Třetí skupina je tvořena 121 neaktivními studenty, kteří do 6. výukového týdne neprojevili žádnou aktivitu během svého studia. Tito studenti s nejvyšší pravděpodobností nemají o studium na FIT zájem. Studium neukončili např. z důvodu ponechání statusu studenta nebo nevědomosti. Tato skupina studentů je pro fakultu také bolestivá, především z finančních a režijních důvodů jako je např. kapacita rozvrhu.

6.1 Oslovení studentů

Vedení fakulty zvažovalo mnoho možností, jakým způsobem studentům s nízkou pravděpodobností průchodu nabídnout pomoc a jaký formát by taková pomoc měla mít. Na základě dostupných dat a provedených analýz vyplynulo, že studenti mají největší problémy v předmětu BI-ZMA, případně v BI-MLO. Nicméně předmět BI-ZMA v B151 prošel razantními změnami a předcházel mu nově předmět BI-PKM, viz sekce Analýza předmětů. Předmět BI-MLO oproti předešlým běhům nevykazuje žádné změny a je nadrámec možností vyučujících nyní nějaké provádět, např. v podobě zavedení podpůrného předmětu.

Po dlouhém zvažování různých možností dospělo vedení fakulty k možnosti obeslání těchto studentů upozorňujícím e-mailem. Zcela jasně se nejedná o nejefektivnější formát nabídnutí pomoci, ale prediktivní model za účelem pomoci studentů s reálným nasazením byl sestaven poprvé a je potřeba jej otestovat a v případě mýlky nijak významně nenarušit studijní motivaci.

Upozornění prostřednictvím e-mailu využívá i Course Signals, jedná se však o automatizovaný a ověřený proces, který upozorňuje jak vyučujícího, tak souběžně i žáky v jednotlivých předmětech.

Velkým plusem narvhovaného mailingu jsou personalizované informace a také viditelná akce ze strany fakulty napříč předměty, což může vést k motivaci.

6.1.1 Mailing

Ve spolupráci s vedením bylo připraveno několik variant e-mailu, které byly zaslány náhodně vybraným skupinám studentů s nízkou pravděpodobností postupu. Tyto varianty obsahují celkem tři informační bloky, které byly mezi sebou různě kombinovány:

- informace o pravděpodobnosti postoupení do dalšího semestru,
- informace o pravděpodobnosti dokončení jednotlivých předmětů,
- doporučení předmětu, který zvýší šanci na postup.

Konečné znění e-mailu, které bylo schváleno vedením a které obsahuje všechny informační bloky:

Předmět e-mailu Upozornění na dosavadní studijní výsledky

Vážený [oslovení křestním jménem - 5. pád],

pro mnohé studenty naší fakulty je první semestr studia kritický. V minulých letech řada studentů nezískala na konci 1. semestru potřebné minimum 15 kreditů pro postup do dalšího studia, a to i přesto, že měli pro dokončení studia předpoklady.

Tento semestr jsme v 8. výukovém týdnu pilotně vyhodnotili data o dosavadních studijních výsledcích, která máme k dispozici a pomocí metod prediktivního modelování jsme identifikovali studenty, u kterých mohou podobná rizika nastat. To, že jste adresátem tohoto e-mailu ještě neznamená, že problém skutečně máte nebo ho určitě mít budete, protože naše prediktivní modely nemají vždy pravdu. Nicméně věříme, že naši iniciativu uvítáte, a v případě, že jsme se ve Vašem případě dopustili „falešného poplachu“, se na nás nebudete zlobit.

V naší analýze vycházíme zjednodušeně řečeno z faktu, že Váš profil a studijní výsledky se podobají studentům, kterým se v minulých letech nepodařilo získat alespoň 15 kreditů nutných pro postup do druhého semestru.

Informace o pravděpodobnosti postoupení

Abychom byli konkrétní, na základě Vašeho profilu, dosavadích studijních výsledků a dat z minulých let je pravděpodobnost toho, že úspěšně postoupíte do dalšího semestru [doplnit pravděpodobnost]%.

Informace o jednotlivých předmětech

Dle Vašeho profilu a výsledků do 8. týdne náš prediktivní model předpovídá šance na úspěšné dokončení předmětů takto:

- BI-CAO: [doplnit pravděpodobnost]%
- BI-MLO: [doplnit pravděpodobnost]%
- BI-PA1: [doplnit pravděpodobnost]%
- BI-ZMA: [doplnit pravděpodobnost]%
- BI-PS1: [doplnit pravděpodobnost]%
- BI-PAI: [doplnit pravděpodobnost]%

Protože nám velmi záleží na tom, aby Vaše studium zdárně pokračovalo, máme následující návrhy, které Vám mohou pomoci.

Pokud některé látce nerozumíte, je nejvyšší čas, abyste vyhledal/a svého vyučujícího, případně garanta daného předmětu, a požádal/a ho o pomoc s danou látkou. Nebojte se konzultovat svou situaci s cvičícím. Pokud projevíte snahu, jistě Vám poradí a pomohou.

Zaměřte se na klíčové předměty, které máte šanci úspěšně dokončit, a předměty s nízkou pravděpodobností dokončení si nechte na příští rok - nezapomeňte, že potřebujete alespoň 15 kreditů pro postup do dalšího semestru. Mnoho Vašich kolegů si uvědomilo svůj studijní problém příliš pozdě a špatně rozložilo úsilí a strategii, na co se soustředit a co vypustit.

Doporučení předmětu, který zvýší šanci na úspěch

Pokusili jsme se nasimulovat různé scénáře budoucího vývoje a náš model mj. ukazuje, že úspěšné zvládnutí předmětů [název doporučeného předmětu č. 1] a [název doporučeného předmětu č. 2] výrazně zvýší Vaši šanci na úspěch.

Samozřejmě jsme si vědomi, že naše prediktivní modely nemají dost informací, aby mohly Vaši situaci komplexně posoudit, a jen Vy nejlépe víte, jak na tom jste a co musíte udělat.

Pokud budete mít chuť vyjádřit se k relevanci našich predikcí a doporučení, napište na friedmag@fit.cvut.cz.

Věříme, že svou situaci zvládnete a přejeme Vám mnoho úspěchů v dalším studiu.

S pozdravem
Bc. Magda Friedjungová
Oddělení pro rozvoj, referent

6.1.1.1 Rozesílání

Kurzívou jsou uvedeny pouze poznámky pro čtenáře, hranaté závorky budou nahrazeny informací, kterou požadují. E-mail byl zaslán celkem 243 studentům na školní e-mailovou adresu s rozlišením ženského a mužského rodu. Pro skloňování křestních jmen byl využit kód, který používá pro takové oslovení portál pro Spolupráci s průmyslem (ssp.fit.cvut.cz, dále jen SSP). Takto vyskloňovaná jména byla dále manuálně zvalidována, aby nedošlo k chybě a byla zároveň poskytnuta zpětná vazba tvůrcům skloňování v SSP.

Při tvorbě e-mailu byla stanovena velikost obeslaných skupin jednotlivými variantami:

Název bloku	Zaslat	Nezaslat
Pravděpodobnost postupu do dalšího semestru	20%	80%
Pravděpodobnost dokončení jednotlivých předmětů	80%	20%
Doporučení předmětu se šancí na úspěch	80%	20%

Tabulka 6.1: Velikost obeslaných/neobeslaných skupin.

Celkem existuje 2^3 variant e-mailu, nicméně varianta neobsahující ani jeden informační blok nebyla zaslána žádným studentům z důvodů etických a morálních, přestože by skupina neobeslaných studentů byla důležitá pro vyhodnocování úspěšnosti jednotlivých variant. Do budoucna je vize rozeslání pouze jedné, té nejúspěšnější varianty.

Pro rozesílání byla sestavena binární matice o velikosti 7*3, kdy 1 označovala zaslání daného informačního bloku a 0 nezaslání. Každému studentovi pak byla náhodně přidělen jeden řádek matice současně se zachováním velikosti obeslaných skupin.

6.1.1.2 Informační bloky

Obsahem e-mailu jsou také informace, které prediktivní model neposkytuje. Tyto informace byly zajištěny ve spolupráci s Ing. Janem Motlem, níže je stručně popsán postup jejich získávání.

Pravděpodobnost postupu do dalšího semestru
Tuto informaci nám poskytuje prediktivní model, jedná se o procentuální přesnost predikce.

Pravděpodobnost dokončení jednotlivých předmětů
Pro pravděpodobnost dokončení jednotlivých předmětů byl sestaven prediktivní model metodou Random Forest v Rapid Miner Studiu. V tomto případě jsme využili informaci o zakončení předmětu, která byla při předzpracování součástí každého datasetu z běhu předmětu.

Pravděpodobnosti dokončení předmětů (viz 6.2 prošly optimalizačními úpravami. Pravděpodobnosti pro BI-ZMA a BI-PS1 se pohybují v rozsahu [0,3 - 0,7]. Pokud u těchto předmětů chybí hodnoty (test pravděpodobně nepsali), byla pravděpodobnost stanovena na 0,1 z důvodu možnosti dopsání testu.

BI-CAO	BI-MLO	BI-PA1	BI-PAI	BI-PS1	BI-ZMA
0,837	0,695	0,609	0,795	0,49	0,388

Tabulka 6.2: Průměrná pravděpodobnost dokončení jednotlivých předmětů v B151.

Pro vysvětlení dokončení jednotlivých předmětů byla sestavena korelační matice jejich průchodnosti 6.3 . Pro zjednodušení zobrazení není v tabulce používán prefix BI-.

Atribut	CAO	MLO	ZMA	PS1	PA1	PAI
CAO	1	0,530	-0,374	0,276	0,445	0,743
MLO	0,530	1	-0,463	0,435	0,577	0,568
ZMA	-0,374	-0,463	1	-0,376	-0,430	-0,384
PS1	0,276	0,435	-0,376	1	0,427	0,320
PA1	0,445	0,577	-0,430	0,427	1	0,477
PAI	0,743	0,568	-0,384	0,320	0,477	1

Tabulka 6.3: Korelační matice průchodnosti předmětů.

Doporučení předmětu se šancí na úspěch
Pomocí optimalizačního kritéria bylo napočítáno doporučení, které reflektuje fakt, jestliže se student zlepší v nějakém předmětu, jaký to bude mít dopad na postup do dalšího semestru. Pro modelování byl použit předpoklad mírného zlepšení o 10% (neboli bodů při stupnici 0 - 100 bodů), abychom předešli závěrům, že se má student zlepšit ve všech předmětech.

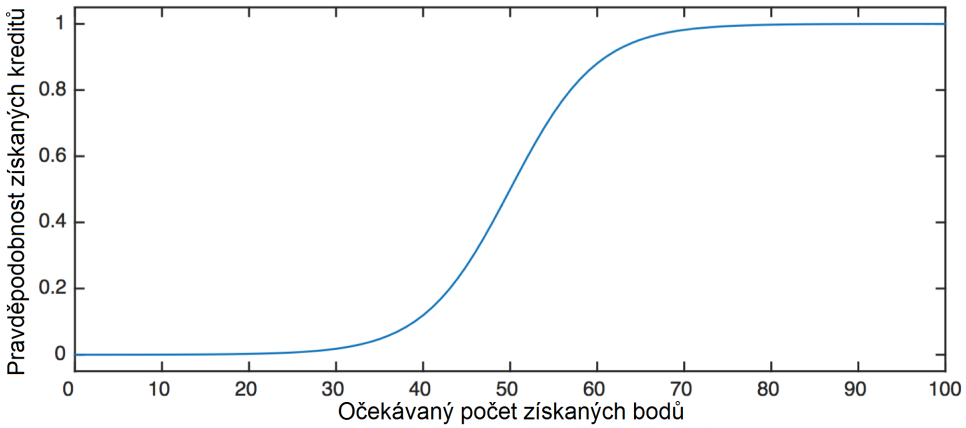
Doporučení bylo definováno jako optimalizace zisku kreditů napříč všemi předměty. Jelikož je každý předmět ohodnocen jiným počtem kreditů, jedná se o vážený součet pravděpodobností, že student daný předmět úspěšně absolvuje:

$$ocekavany\_pocet\_kreditu = \sum_{i=1}^{n} w_i * p_i$$

Na vstupu je vektor očekávaných bodových zisků pro každý z předmětů. Protože nás zajímá zisk kreditů, nikoliv bodů:

$$ocekavany\_pocet\_kreditu = \sum_{i=1}^{n} w_i * f(ocekavany\_pocet\_bodu_i)$$

Převod bodů na kredity je definován zkušebním řádem [26]. Vzhledem k tomu, že Heavisideova funkce [34] není vhodná k optimalizaci, byla nahrazena sig-



Obrázek 6.1: Funkce pro převod bodů na kredity.

Předmět	**CAO**	**MLO**	**PA1**	**PAI**	**PS1**	**ZMA**
Očekávaný bodový zisk	83	49	26	86	49	39
Doporučení	0	12	-26	0	12	0

Tabulka 6.4: Ukázka doporučení zaměření se na předmět BI-PS1 a BI-MLO na úkor předmětu BI-PA1.

moidou, viz 6.1.

Výsledkem je doporučení 1-2 předmětů, jejichž úspěšné zakončení by mělo mít pozitivní vliv na studentův postup do dalšího semestru. Zároveň zlepšení se v těchto předmětech by nemělo být pro studenta nemožné, tzn. že v předmětu již vykázal nějakou aktivitu, která však doposud nesměřovala ke zdárnému zakončení bez vynaložení dalšího úsilí. Zároveň byl při doporučování zohledněn zápis BI-ULI, který může ovlivnit absolvování BI-PS1. Ukázka doporučení ve výsledném souboru viz tabulka 6.4. Pro zjednodušení zobrazení není v tabulce používán prefix BI-.

Optimalizační metoda byla také rozšířena o schopnost doporučení předmětů, kterým by se měl student přestat věnovat. U předmětů, kde je předpokládaný bodový zisk nižší než 30 bodů z 50 možných, byly body odčerpány a s určitou ztrátou přeskupeny do jiných předmětů. Tuto možnost doporučení zanechání předmětu výsledný e-mail nezahrnoval.

Varianty	0		1	
	Počet	**Počet variant**	**Počet**	**Počet variant**
1 (100)	33,3%	2	66,7%	4
2 (010)	62,5%	20	37,5%	12
3 (110)	80,0%	8	20,0%	2
4 (001)	55,2%	16	44,8%	13
5 (101)	84,6%	11	15,4%	2
6 (011)	69,3%	88	30,3%	39
7 (111)	69,2%	18	30,8%	8

Tabulka 6.5: Kontingenční tabulka pro vyhodnocení variant e-mailu.

6.1.1.3 Vyhodnocení mailingu

Vyhodnocení jednotlivých variant e-mailů lze provést několika metodami, je však nutné zvážit naše předpoklady.

Na základě sestavené binární matice pro obesílání a percentuální velikosti využití variant, viz tabulka 6.1, bylo obesláno 7 různě velkých skupin studentů. Nestejná velikost obeslaných skupin pro vyhodnocení vylučuje použití některých statistických metod, které vyžadují vyvážený vzor. Stejně tak některé metody vyžadují spojitou odezvu, která je v našem případě binární. Důležitým parametrem pro vyhodnocení je také fakt, že jednotlivé skupiny studentů nemají žádné společné rysy a jejich rozdělení bylo provedeno zcela náhodně.

Vhodná metoda pro vyhodnocení vlivu jednotlivých variant na postup studentů do dalšího semestru byla konzultována s Ing. Danielem Vašatou, Ph.D. z KAM FIT. Na základě výše popsaných předpokladů byla sestavena kontingenční tabulka 6.5, která je typická pro posouzení, zda je postup studentů je závislý na obdržené variantě e-mailu.

Na základě výsledků v tabulce 6.5 vidíme, že nejlepší odezvy dosahuje varianta 1 (100) a to 66,7%, nicméně tato varianta byla zaslána pouze celkem 6 studentům (2 nepostupivší + 4 postupivší), což je pro vyslovení závěru velmi malý vzorek. Vyhodnocení bylo provedeno pomocí testu homogenity multinomických rozdělení [35], který vychází z dvojrozměrné kontingenční tabulky, obsahující empirické četnosti n_{ij}, a matice pravděpodobností p_{ij}.

Test homogenity byl zvolen kvůli předem stanoveným marginálním řádkovým četnostem. Jeho provedení je totožné jako provedení testu nezávislosti, který pracuje s náhodnými veličinami.

Nejčastější úlohou je právě provedení testu hypotézy, kdy veličiny Y (vari-

anty emailů) a Z (četnosti) jsou na sobě nezávislé. Tuto hypotézu nezávislosti H_0 můžeme zapsat jako:

$$H_0 : p_{ij} = p_{i.}p_{.j}, \qquad i = 1, ..., r; j = 1, ..., c,$$

kdy $p_{i.}$ a $p_{.j}$ označují marginální pravděpodobnosti, r označuje počet řádků (veličina Y) a c počet sloupců (veličina Z) v kontigenční tabulce 6.5. V našem případě $r = 7$ a $c = 2$.

Uveďme si pouze stěžejní vzorec pro rozdělení χ^2, pro postup jeho odvození viz [35]:

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}^2}{n_{i.}n_{.j}} - n$$

s počtem stupňů volnosti $rc - (r + c - 2) - 1 = (r - 1)(c - 1)$.

Ke shodě s limitním rozdělením je vyžadováno, aby všechny teoretické četnosti $n_{i.}n_{.j}/n$ byly větší než 5, což v našem případě dodrženo není. Proto byly spojeny sobě nejbližší hodnoty (řádky), tzn. varianta 1 a 4 a varianta 3 a 5. Výsledky různých kombinací lze porovnat v tabulce 6.6. V rámci ověřování byly spojeny i varianty 1 a 3 s vědomím, že poskytují opačné výsledky.

Počet variant	**Stupeň volnosti**	α	χ^2	$\chi^2_{stupeň\ volnosti}(\alpha)$	**p-value**
7	6	0,1	8,162	**10,645**	0,226
6 (spojení 1 a 3)	5	0,1	4,464	**9,237**	0,485
5 (spojení 1 a 4, 3 a 5)	4	0,1	7,034	**7,779**	0,134

Tabulka 6.6: Výsledky různých kombinací variant.

K zamítnutí hypotézy H_0 o nezávislosti veličin Y a Z dochází, pokud $\chi^2 \geq \chi^2_{(r-1)(c-1)}(\alpha)$. Na základě výsledků v tabulce 6.6 nemůžeme hypotézu H_0 ani v jednom případě zamítnout.

Pro srovnání bylo provedeno vyhodnocení mailingu pomocí logistické regrese [30] [34] Ing. Janem Motlem, Ph.D., který dospěl ke stejnému závěru.

Shrnutí
Ze statistického hlediska nemají jednotlivé varianty e-mailu prokazatelný vliv na studentův postup do dalšího semestru. Bohužel za znatelnou komplikaci vyhodnocení lze považovat rozeslání variant nerovnoměrně velkým skupinám studentů (6.1).

Jestliže má být v budoucnu zvolena pouze jedna varianta e-mailu, jednalo by se o variantu 1 (100 - pravděpodobnost postupu do dalšího semestru) v kombinaci s variantou 4 (001 - doporučení předmětu se šancí na úspěch).

Z celkem 243 oslovených studentů postoupilo do dalšího semestru pouze 80.

Nutné zmínit, že rozesílání tohoto e-mailu neproběhlo v plánovaném 7. týdnu, přestože byla dle plánu definována skupina studentů s nízkou pravděpodobností postupu. E-mail byl rozeslán až během 10. výukového týdne, především z důvodů organizačních. Obsah e-mailu tedy zcela přesně nereflektoval studentovu aktuální situaci. V některých předmětech již studenti neměli téměř žádnou šanci na nápravu svých výsledků. Vedení fakulty si toto nešťastné načasování uvědomuje a v budoucnu by měl být zásah aplikován v doporučeném týdnu.

6.1.2 Dotazník

Data z prediktivního modelování byla dále využita pro kontaktování skupiny neaktivních studentů. Pro tuto skupinu byl ve spolupráci s Ing. Lucií Kolomazníkovou, referentkou Oddělení pro rozvoj FIT, sestaven on-line dotazník, který měl za úkol zjistit, z jakého důvodu studenti zanechávají studia.

Dotazník obsahoval celkem 14 otázek, některé s předem definovanými odpověďmi a jiné s možností pro odpověď volnou.

Společně s průvodním textem byl dotazník zaslán e-mailem na soukromé e-mailové adresy 120 neaktivních studentů 1. ročníku BI. Soukromé e-mailové adresy byly zvoleny z důvodu již možného nepoužívání školního e-mailu vzhledem k jejich neaktivitě při studiu. Tyto adresy byly získány z atributu emailova_adresa tabulky koud_emaily z DWH ČVUT.

Znění dotazníku a průvodního e-mailu je součástí přiloženého CD.

6.1.2.1 Vyhodnocení

Dotazníku se zúšastnilo celkem 26 neaktivních studentů, což je 21,5% z celkového počtu oslovených. Toto číslo odpovídá standardní očekávané vyplněnosti dotazníků. Jeho velikost je také opodstatněná skupinou oslovovaných, což jsou neaktivní studenti vůči FIT.

Přes 53% respondentů uvedlo, že aktivně navštěvovali výuku a účastnili se testů, přestože jejich výsledky byly nulové. Neméně překvapivé je jejich hodnocení obtížnosti předmětů, které by pravděpodobně korelovalo s odpověďmi

aktivních studentů. Jako nejvíce obtížný předmět 50% respondentů zvolilo BI-PS1. Za zajímavé lze považovat, že 69% respondentů nezanechalo studia z vlastní vůle, ale kvůli nesplnění požadavků podle studijního a zkušebního řádu, 53,8% respondentů uvedlo, že si nezvolilo špatnou fakultu a 57% zcela nesouhlasí s tvrzením, že by zjistili, že je informatika a informační technologie nebaví. U hodnocení, zda FIT splnil představy a očekávání respondentů nebo zda pro ně bylo studium příliš náročné, je rozložení kladných a záporných odpovědí vyrovnané.

Za poněkud matoucí můžeme považovat odpovědi, které se vztahovaly k předmětům 2. či 3. semestru nebo vyučujícím, kteří vyučují předměty určené studentům od 2. semestru BI. Nejedná se však o chybu v datech, ale o studenty, kteří již FIT v minulosti studovali, ale z nějakého důvodu studia zanechali a v roce 2015 si znovu podali přihlášku ke studiu na FIT a úspěšně absolvovali přijímací řízení, čímž se oficiálně stali studenty 1. ročníku, a tak i součástí datasetu. Takoví studenti byli zohledněni pouze při zařazení do skupiny s nízkou pravděpodobností postupu a byl jim přizpůsoben obsah e-mailu. U skupiny neaktivních studentů toto zohledněno nebylo, jelikož se jednalo o minimum studentů.

Celé znění odpovědí a jejich grafický souhrn je k dispozici na přiloženém CD, za velmi důležité zjištění lze považovat fakt, že více než polovina respondentů měla o studium zřejmě vážný zájem.

Kapitola 7

Generalizace modelu

Zadání této práce mimo jiné obsahuje požadavek na zhodnocení možnosti využití sestavených modelů i v příštím akademickém roce nebo dokonce mezi semestry. Generalizace modelů patří mezi jeden ze dvou největších problémů, se kterými se prediktivní modelování potýká. V našem případě tomu není bohužel jinak.

7.0.1 Využití v dalších letech

Modely v této práci sestavené lze použít i v příštích zimních semestrech, nicméně za stávajících podmínek je nevyhnutelný následující postup:

- Získání dat ze zdrojových systémů - vzhledem k chybějící automatizaci exportů a nemožné integraci do datavého skladu bude nezbytná součinnost správců zdrojových systémů, aby byla poskytnuta nová data.
- Analýza pravidel hodnocení - jelikož v některých předmětech dochází každý semestr ke změnám v pravidlech hodnocení, je potřeba tato pravidla na portále EDUX s garanty předmětů před každým zpracováním vždy zkontrolovat a zanalyzovat jejich změny.
- Předzpracování dat - pokud budou exporty dat dodávány ve formátech jako doposud, je potřeba nová data předzpracovat způsoby uvedenými v kapitole zabývající se předzpracováním. Stejně tak bude nutné zvolit vhodnou normalizaci nových hodnot.
- Kontrola distribuce dat - díky různým změnám v předmětu může dojít ke změně distribuce dat, jako tomu bylo např. v BI-PS1 v B151. Proto je nezbytné před každým spuštěním modelu na nových datech zkontrolovat jejich distribuci a případně zvolit vhodnou nápravu.
- Zhodnocení modelu - je možné, že v pravidlech hodnocení předmětů nebo distribuci dat dojde ke změnám, které budou mít zásadní dopad

na funkčnost sestavených modelů, proto je nutné vyhodnotit, zda jsou modely pro daný semestr stále použitelné či funkční. Bohužel jsou tyto změny nepředvídatelné, takže na ně nelze modely nijak připravit.

Tato práce by měla sloužit jako dostatečný návod k replikování všech potřebných postupů a k pochopení jednotlivých kroků a modelů. Na tomto základě by tedy nemělo být nemožné prediktivní modelování i mezi semestrem letním a zimním - lze využít zde zmíněné postupy, nicméně bude nutné provést analýzu předmětů v daném semestru, získat veškerá, tzn. aktuální a historická, data a provést jejich předzpracování a normalizaci. Současně při predikci úspěšnosti v postupu z letního do zimního semestru lze využít studijních výsledků z předchozího zimního semestru a je otázkou, zda používat i údaje z přihlášky, tzn. bude nutné vyhodnotit vliv všech možných atributů a zvážit jejich použití ve finálním datasetu. Dále je nutné zohlednit možnost zápisu volitelných předmětů, kterého studenti od 2. semestru hojně využívají a dochází tak k nekonzistencím. Při stávajícím postupu by tak mohlo dojít k sestavení datasetu v podobě řídké matice. Pro modelování pak lze velmi pravděpodobně využít rozhodovací strom a Random Forest, nicméně bude nutné modely znovu sestavit, přestože se jedná o stejnou doménu, ale jiná data, která ještě nejsou zpracována.

Sestavení prediktivního modelu pro postup mezi letním a zimním semestrem z výše zmíněných důvodů přesahuje rozsah této práce a lze očekávát, že vyřešení tohoto úkolu může být stejného, ne-li většího rozsahu než je tato práce. Stejně tak při modelování postupu mezi ročníky nebo úspěšnosti v různých předmětech napříč studiem.

7.0.1.1 Doporučení

Sestavené prediktivní modely dosáhly uspokojivých výsledků, nicméně během práce jsme se potýkali s několika problémy, které by bylo vhodné do budoucna napravit.

Automatizace exportů
Jako jedno z hlavních doporučení lze zmínit automatizované poskytovaní exportů dat pro jejich automatizované nahrání do datového skladu, ze kterého by byla data v jednotných formátech poskytována pro analýzy.

Automatizace je dlouhodobým tématem na FIT, pomocí API lze již přistupovat k datům ze systému Progtest, ale k datům ze systému EDUX žádný takový přístup prozatím není. Automatické a pravidelné poskytování exportů by usnadnilo mnoho času a zkvalitnilo zpracování dat.

Integrace dat
Všechna získaná data ze zdrojových systémů by bylo ideální integrovat do datového skladu DWH ČVUT. Pro uskladnění je nezbytné navrhnout jednotnou strukturu pro všechny předměty, tedy je nutné pokrýt všechny situace, které v hodnocení předmětů mohou nastat. Důležité je zmínit, že na FIT je vypsáno více než 200 předmětů a při manuální analýze, která byla provedená v této práci na 5 předmětech v 6 bězích, je to úkol téměř nereálný. Analýzu by měla nahradit metadata, která však v systému EDUX chybí.

Metadata
Metadata představují strukturovaný popis dat jako takových. Pomáhají nám se v datech lépe orientovat a porozumět jim. Právě metadata by nám měla poskytnout informaci o významu a obsahu jednotlivých atributů a pomoci nám tak jednoduše analyzovat jednotlivé exporty. Chybějící metadata lze považovat za jeden z nejdůležitějších problémů exportů z EDUXu. Bez nich lze totiž obsah a význam jednotlivých atributů pouze hádat a analýza takového obsahu zabírá neúměrné množství času. Zároveň díky neexistenci metadat nelze jednotlivé exporty předmětů efektivně integrovat do datového skladu a to z důvodu právě chybějící informace o struktuře každého exportu, které by byl přizpůsoben návrh tabulek v databázi datového skladu.

V rámci této práce byla navržena struktura metadat hodnocení předmětů, která by mohla být implementována v rámci vývoje nového informačního systému (dále jen IS) pro zaznamenávání výsledků hodnocení během semestru. Na tomto systému pracuje ICT oddělení FIT a měl by mít velmi podobnou struktura jako stávající EDUX. Rozšíření o metadata by mohlo pomoci ve zpracování poskytovaných dat.

Navržený koncept požaduje z nového IS dva výstupy:

- pravidla hodnocení v předmětu,
- studijní výsledky studenta.

Pravidla hodnocení v předmětu by obsahovala následující položky:
ID pravidla; název; kód předmětu; ID semestru; ID položky; číslo výukového týdne; pořadí; minimum bodů z položky; maximum bodů z položky; příznak, zda je vyplnění položky povinné; příznak, zda položku v editaci zobrazovat; poznámka.

ID položky by vedlo na číselník, ve kterém by byly definovány všechny existující položky v předmětech na FIT.

Studijní výsledky studenta by obsahovaly:

ID hodnocení; username; ID studia; kód předmětu; ID semestru; ID vyučujícího; ID pravidla; hodnota; timestamp.

Tato struktura exportů by umožnila plynulou integraci dat do DWH ČVUT, která momentálně z důvodu chybějících metadat není možná.

Rozšíření dat
Ve studiích provedených na jiných univerzitách lze pozorovat aplikaci prediktivního modelování i na odlišná data, než kterými disponuje FIT. Jedná se především o informace:

- o rodinném zázemí - vzdělání, plat a zaměstnání rodičů, počet sourozenců, zda jsou starší nebo mladší než uchazeč, apod.,
- o střední škole - absolvované předměty, průměr z těchto předmětů, vztah k předmětům, dojezdová vzdálenost vzdělávací instituce od bydliště, výsledky z maturitní zkoušky, apod.,
- o zájmových kurzech - zda a jaké uchazeč navštěvoval/navštěvuje, jak dlouho, jaké úrovně znalosti dosáhl v dané oblasti apod.,
- o sociálních dovednostech - subjektivní hodnocení, jak se uchazeč vnímá ve společnosti, jaký má vztah ke kolektivu, zda se cítí být oblíbený nebo spíše outsider apod.,
- o sportu - uchazečův vztah ke sportovním aktivitám,
- o využívání komunikačních prostředků - jak často a k jakým účelům používá mobilní telefon, zda vlastní smartphone, zda je aktivní na sociálních sítích a jakých apod.,
- o pracovních zkušenostech - informace, zda uchazeč absolvoval nějakou brigádu (jakou a po jak dlouhou dobu) nebo zeměstnání před nástupem na VŠ, v jakém oboru apod.

Tyto a mnohé další ukazatele lze získat např. prostřednictvím on-line formuláře, který by mohl být studentům zprostředkován v rámci přijímacího řízení (např. v portálu Přiříz [9]) nebo by mohl být zaslán na jejich e-mailové adresy po nástupu do studia. Dále je možné přihlášku jako takovou rozšířit a povinná pole, např. co se týče výsledků z maturity a dalších informací o studiu na střední škole.

Kromě rozšíření ukazatelů lze také sledovat studentův pohyb na studijních portálech, tzn. zda a jak dlouho si zobrazil určitý studijní materiál, zda využívá studentského portálu www.fit-wiki.cz, s jakou frekvencí, zda projevil

[9] Aplikace používaná na FIT pro správu přijímacího řízení.

zájem o spolupráci s průmyslem prostřednistvím SSP [10] apod.

Také tyto údaje rozšířit o informace ze sociálních sítí jako je např. Facebook nebo Twitter, kde lze také sledovat studentovu aktivitu, ze které lze získat informaci, zda se student o obor zabývá i ve svém volném čase (navštěvuje přednášky, konference, soutěže apod.). Dalším zdrojem informací může být profesní síť LinkedIn, kde můžeme získat informace o profesních zkušenostech našich studentů, kterými např. SSP nedisponuje.

Veškeré tyto informace mohou být v rámci EDM velmi zajímavé a užitečné. Stejně jako např. průběžné monitorování spokojenosti studentů a mnohé další aktivity, které může fakulta zajistit. Nicméně získání alespoň některých dat a jejich využití může být během na dlouhou trať a celý proces je velmi závislý na podpoře a součinnosti vedení FIT.

Procesy
Predikce postupu studentů 1. ročníku BI z 1. do 2. semestru a aplikce těchto výsledků za účelem pozitivního ovlivnění výsledků studentů byla na FIT novinkou. Není proto nijak překvapivé, že jsme se po celou dobu potýkali s nastavováním různých procesů, které mimo jiné způsobovalo časové prodlevy v realizaci úkolu.

Považuji za důležité poukázat na některé komplikace:

1. Poskytnutí historických dat.
   Při sběru dat ze zdrojových systémů jsme se nejednou setkali s neochotou správců systémů spolupracovat a poskytnout nám potřebná data, přestože se jedná o zadání přímo od vedení fakulty. Situace si proto vyžádala vystavení příkazu děkana, který již byl dotyčnými správci uposlechnut.

2. Chybějící jednoznačný identifikátor studenta.
   V EDUXu slouží jako identifikátor studenta jeho username, který není v souvislosti s jednotlivými studii unikátní. Během předzpracování dat bylo ukázáno, že tato nejednoznačnost komplikuje práci a data musí být obohacena o ID studia. Toto ID studia lze z KOSu získat stejnou cestou jako studentův username, neměl by tedy být problém tuto položku v EDUXu přidat.

3. Definice pravidel hodnocení v předmětech v aktuálních běžích.
   Týden před zahájením semestru jsou garanti předmětů povinni zveřejnit pravidla pro hodnocení předmětu na portále EDUX. Tento termín však nebyl vždy dodržován a některá pravidla byla upravena/doplněna v 1.

[10] Portál Spolupráce s průmyslem provozovaný FIT ČVUT.

výukovém týdnu semestru. Pokud došlo ke změnám pravidel až po provedení jejich analýzy, přestože se garant předmětu za pravidla zaručil, neexistuje žádné automatické upozornění na tyto změny.

4. Zajištění včasného zápisu dat do EDUXu.
   Při definici týdne, ve kterém proběhne sběr dat v B151, se garanti předmětů zaručili, že veškeré testy a úkoly budou vždy do 1 výukového týdne opraveny a jejich hodnoty budou zapsány do EDUXu. Opět toto nelze nijak automaticky zkontrolovat, nicméně od studentů víme, že některé testy v některých paralelkách byly opraveny mnohem později než bylo domluveno.

5. Nekonzistence nabídky předmětů.
   Během zkouškového období v semestru B151 byl vytvořen předmět BI-SM, o kterém jsme se, jakožto nestudenti 1. ročníku BI, dozvěděli jen díky sociální síti Facebook, na které o tomto předmětu studenti diskutovali. Pravidla hodnocení v tomto předmětu byla vytvořena jako součást předmětu BI-PS1 na EDUXu a po skončení zkouškového období byla tato pravidla z EDUXu smazána. Takovéto jednání považujeme za nešťastné, jelikož má zásadní dopad na postup do dalšího semestru pro desítky studentů a neexistuje žádné oficiální médium, které by o těchto aktivitách informovalo.

6. Nedodržení vhodného termínu pro akci.
   Při realizaci této práce bylo domluveno spuštění prediktivního modelu v 8. týdnu následně s předáním výsledků vedení fakulty pro možnou akci, pro kterou byl tento týden vyhodnocen jako nejefektivnější. Kvůli přípravě vhodné akce a jejímu schvalování proběhlo využití výsledků až v během 10. výukového týdne, tedy výsledky prediktivního modelování zcela přesně nereflektovaly situaci.

Práce jako taková nemá za cíl nastavení procesů, nicméně došlo k prvnímu pokusu a nastínění postupu, který by mohl být využit i při příštím prediktivním modelování. Za důležité lze považovat, že práce navzdory různým překážkám byla úspěšně realizována v rozsahu přesahujícím původní zadání.

Závěr

Cílem této práce bylo sestavení prediktivního modelu pro předpověď úspěšnosti studentů 1. semestru 1. ročníku BI na FIT ČVUT a vhodné využití těchto výsledků.

V první fázi práce jsem se zabývala získáním potřebných dat z fakultních a univerzitních systému, jejich analýzou a vhodným předzpracování. V této fázi práce jsem se potýkala s nějvětšími problémy a také byla časově nejnáročnější. Data poskytnutá ze systému EDUX nebyla příliš dobře strukturovaná a bez jakéholiv popisu. Provedla jsem tedy důkladnou analýzu jejich možného obsahu a zvolila jsem vhodné metody pro předzpracování. Dále jsem u většiny hodnot provedla vhodnou normalizaci. Připravila jsem celkem 3 441 použitelných záznamů, z let 2010 - 2014.

Pro tato data jsem navrhla několik prediktivních modelů. Model s nejlepšími výsledky a vlastnostmi při testování (75% přesnost) jsem aplikovala na data dostupná do 7. výukovém týdnu letošního roku [11], která jsem před aplikací podrobila předzpracování (825 záznamů). Ná základě výsledků predikce jsem vybrala skupinu studentů, kteří dosáhli nízké pravděpodobnosti průchodu do letního semestru. Těmto studentům jsem ve spolupráci s vedením fakulty zaslala personalizovaný e-mail upozorňující na dosavadní studijní výsledky.

V rámci testování nejlepší varianty e-mailu jsem studenty obeslala celkem 7 variantami. Na základě jejich vyhodnocení se ukázalo, že ani jedna z variant neměla prokazatelný vliv na výkon studentů. To mohlo být způsobeno podmínkami pro zaslání tohoto e-mailu.

Hlavním přínosem této práce je řešení pro opakovanou identifikace ohrožených studentů, návrh a testování možné intervence, její vyhodnocení a doporuču-

[11] Akademický rok 2015/2016

jící opatření pro budoucí aplikaci řešení. Dalším přínosem je zpracování dat pro jejich integraci do datového skladu. Navzdory několika problémům bylo splněno zadání práce a dosaženo uspokojivých a použitelných výsledků.

Za podstatnou komplikaci v realizaci lze považovat měnící se pravidla hodnocení v předmětech napříč semestry a vytváření nových nepovinných předmětů s životností 1 semestr (např. BI-UVM, BI-SM apod.). Výzvou zůstává sestavení prediktivního modelu se schopností automatické generalizace, aby mohl být využit ve všech semestrech BI i MI. Návrh a implementace takového algoritmu bohužel přesahuje rozsah diplomové práce a představuje úkol, který by mohl být řešen v rámci práce dizertační, např. po vzoru projektu Course Signals. Schopnost generalizace takových modelů je stále nevyřešeným problémem.

Literatura

[1] Domingos, P.: A Few Useful Things to Know about Machine Learning. [online], 2012, [Cited 2014-02-05]. Dostupné z: https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

[2] International Educational Data Mining Society. [online], 2016, [Cited 2014-02-03]. Dostupné z: http://www.educationaldatamining.org

[3] Drbohlav, J.: Vývoj studijní úspěšnosti na českých VVŠ mezi lety 2003-2014. [online], 2015, [Cited 2014-01-30]. Dostupné z: http://www.msmt.cz/uploads/odbor_30/Jakub/Studijni_uspesnost_na_ceskych_VVS_2003_2014_web.pdf

[4] Murphy, D.: Visualizing Student Progress to Provide Actionable Information. [online], 2013, [Cited 2014-02-05]. Dostupné z: http://tinyurl.com/gtpmggy

[5] Essa A., Ayad H.: Improving student success using predictive models and data visualisations. [online], 2012, [Cited 2014-02-05]. Dostupné z: http://www.researchinlearningtechnology.net/index.php/rlt/article/view/19191

[6] Arnold K. E., Pistilli M. D.: Course Signals at Purdue: Using Learning Analytics to Increase Student Success. [online], 2012, [Cited 2014-02-02]. Dostupné z: http://goo.gl/VRmzjY

[7] Dormehl, L.: This Algorithm Can Predict Your Success At University. [online], 2013, [Cited 2014-02-02]. Dostupné z: http://www.fastcompany.com/3020036/this-algorithm-can-predict-your-success-at-university

[8] Oza N. C., Tumer K.: Classifier ensembles: select real-world applications. [online], 2008, [Cited 2014-02-05]. Dostupné z: http://www.sciencedirect.com/science/article/pii/S1566253507000620

[9] Ogor E. N.: Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques. [online], 2007, [Cited 2014-02-05]. Dostupné z: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4367712&tag=1

[10] Pathros E., Garci'a I., Mora P. M.: Model Prediction of Academic Performance for First Year Students. [online], 2011, [Cited 2014-02-05]. Dostupné z: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6118995&tag=1

[11] Nghe N. T., Janecek P., Haddawy P.: A Comparative Analysis of Techniques for Predicting Academic Performance. [online], 2007, [Cited 2014-02-05]. Dostupné z: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4417993

[12] Superby J. F., Vandame J. P., Meskens N.: Determination of factors influencing the achievement of the first-year university students using data mining methods. [online], 2006, [Cited 2014-02-05]. Dostupné z: http://tinyurl.com/gsbg3pq

[13] Atalay V., Üstün S., Bülbül S.: The Determination Of Socio-Economic Factors Affecting Student Success By Data Mining Methods. [online], 2013, [Cited 2014-02-05]. Dostupné z: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=\&arnumber=6786167

[14] Márquez-Vera C., Cano A., Romero C., Ventura S.: Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. [online], 2013, [Cited 2014-02-05]. Dostupné z: http://link.springer.com/article/10.1007%2Fs10489-012-0374-8#/page-1

[15] Macfadyen L. P., Dawson S.: Mining LMS data to develop an 'early warning system' for educators: A proof of concept. [online], 2010, [Cited 2014-02-05]. Dostupné z: http://www.sciencedirect.com/science/article/pii/S0360131509002486

[16] Berka, P.: *Dobývání znalostí z databází*. Academia, 2011, ISBN 80-200-1062-9.

[17] Contributors of Wikipedia: Cross Industry Standard Process for Data Mining. [online], 2016, [Cited 2014-02-05]. Dostupné z: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

[18] Contributors of Talent Analytics: Predictive Approach. [online], 2016, [Cited 2014-02-05]. Dostupné z: http://www.talentanalytics.com/technology/methodology/

[19] Marbán Ô., Mariscal G., Segovia J.: A Data Mining & Knowledge Discovery Process Model. [online], 2016, [Cited 2014-02-05]. Dostupné z: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

[20] Bunkar K., Bunkar R., Singh U. K., Pandya B.: Data Mining: Prediction for Performance Improvement of Graduate Students using Classification. [online], 2011, [Cited 2014-02-05]. Dostupné z: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6335530

[21] Al-Radaideh Q. A., Al-Shawakfa E. M., I. Al-Najjar M.: Mining Student Data using Decision Trees. [online], 2006, [Cited 2014-02-06]. Dostupné z: https://www.researchgate.net/publication/242076740_Mining_Student_Data_Using_Decision_Trees

[22] Kordík P., Jiřina M.: Lecture 1: Introduction, CRISP-DM, DM software. [online], 2015, [Cited 2014-02-05]. Dostupné z: https://edux.fit.cvut.cz/courses/MI-PDD/_media/lectures/01/pdd_01.pdf

[23] Contributors of European Commission's portal: European Credit Transfer and Accumulation System (ECTS). [online], 2016, [Cited 2014-02-08]. Dostupné z: http://ec.europa.eu/education/ects/ects_en.htm

[24] EDUX FIT ČVUT. [online], 2016, [Cited 2014-02-08]. Dostupné z: https://edux.fit.cvut.cz

[25] EDUX FIT ČVUT - archiv. [online], 2016, [Cited 2014-02-08]. Dostupné z: https://edux.fit.cvut.cz/archive/

[26] Studijní a zkušební řád pro studenty Českého vysokého učení technického v Praze ze dne 8. července 2015. [online], 2015, [Cited 2014-02-06]. Dostupné z: https://www.cvut.cz/sites/default/files/content/7e72349e-3ea5-4693-9853-5147f1238481/en/20160122-studijni-a-zkusebni-rad-pro-studenty-cvut-ze-dne-8-7-2015.pdf

[27] Kuznetsov, S.: *Datový sklad fakulty*. Diplomová práce, České vysoké učení technické v Praze, Fakulta informačních technologií, 2013, [Cited 2014-02-05]. Dostupné z: https://dip.felk.cvut.cz/browse/pdfcache/kuznesta_2013dipl.pdf

[28] Hrubá, E.: *Analýza výsledků absolventů středních škol na VŠ*. Diplomová práce, České vysoké učení technické v Praze, Fakulta informačních technologií, 2014, [Cited 2014-02-05]. Dostupné z: https://dip.felk.cvut.cz/browse/pdfcache/hrubaeli_2014dipl.pdf

[29] Kordík P., Motl J.: Přednášky z předmětu BI-VZD. [online], 2011, [Cited 2014-02-15]. Dostupné z: https://edux.fit.cvut.cz/oppa/BI-VZD/prednasky

[30] Witten I. H., Frank E., Hall M. A.: *Data Mining : practical Machine Learning Tools and Techniques*. Elsevier, třetí vydání, 2011, ISBN 978-0-12-374856-0.

[31] Komprdová, K.: Rozhodovací stromy a lesy. [online], 2012, [Cited 2014-02-19]. Dostupné z: https://www.iba.muni.cz/res/file/ucebnice/komprdova-rozhodovaci-stromy-lesy.pdf

[32] BigML Support. [online], 2016, [Cited 2014-02-19]. Dostupné z: https://support.bigml.com/hc/en-us/articles/206616279-What-kind-of-algorithm-does-BigML-use-to-build-the-decision-trees-and-how-does-it-work-

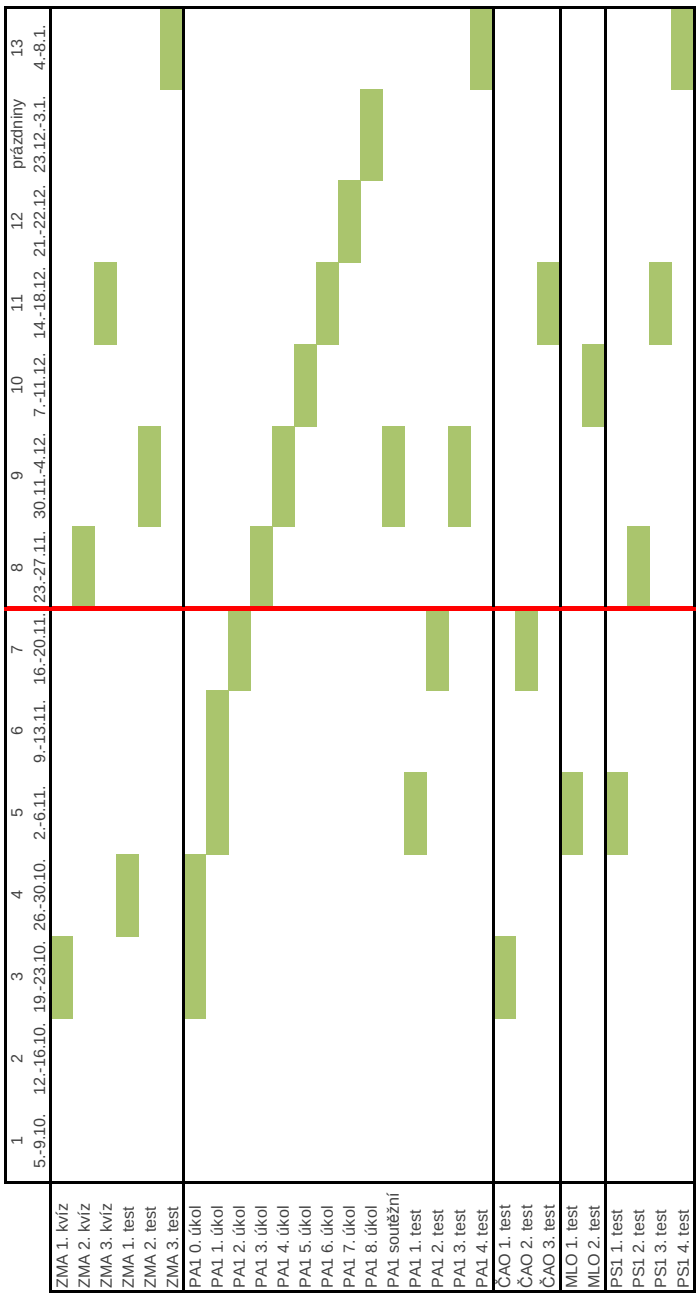
[33] Hastie T., Tibshirani R., Friedman J.: *The Elements of Statistical Learning*. Springer, druhé vydání, 2011, ISBN 80-200-1062-9.

[34] Bishop Ch. M.: *Pattern Recognition and Machine Learning*. Springer Science + Business Media, 2006, ISBN 978-0387-31073-2.

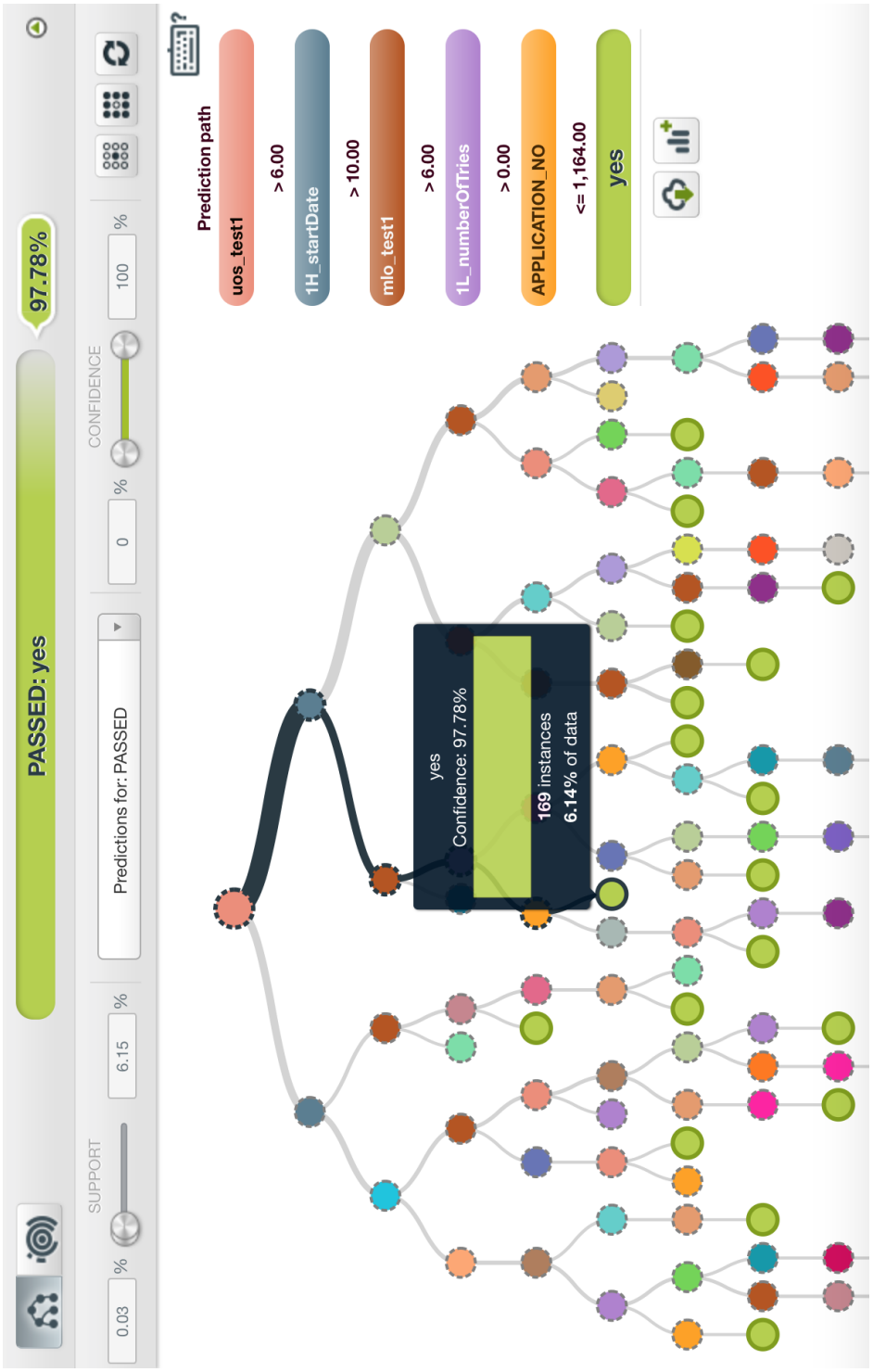
[35] Anděl, J.: *Základy matematické statistiky*. Matfyzpress, 2011, ISBN 978-80-7378-162-0.

Příloha A

Přílohy



Obrázek A.1: Ganttův graf pro stanovení termínu sběru dat v B151.



Obrázek A.2: Ukázka rozhodovacího stromu v nástroji BigML.

Příloha B

Seznam použitých zkratek

AIT Asian Institute of Technology

ANN Artificial Neural Network

BI Bakalářský studijní program Informatika

CART Classification And Regression Tree

CS Course Signals

CRISP-DM Cross Industry Standard Process for Data Mining

CTU Can Tho University

ČVUT České vysoké učení technické v Praze

EBIE Extended Business Intelligence Encyclopedia

ECTS European Credit Transfer System

EDM Educational Data Mining

ELT Extract, Load, Transform

ETL Extract, Transform, Load

FIT Fakulta informačních technologií

DM Data Mining

DWH ČVUT Prototyp datového skladu ČVUT

CHAID Chi-squared Automatic Interaction Detector

IZO Identifikační znak organizace

KAM Katedra aplikované matematiky

LMS Learning Management System

MARAST Matematika Radostně

MI Magisterský studijní program Informatika

PDI Pentaho Data Integration

SSA Student Succcess Algorithm

SSP Portál Spolupráce s průmyslem

3S Student Success System

UNAM Universidad Nacional Autónoma de México

XML Extensible Markup Language

Příloha C

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu CD
- `src` ..... adresář obsahující podklady pro jednotlivé části práce
  - `analyzy` ..... podklady pro analýzy
  - `datovy_sklad` ..... rozšíření datového skladu
  - `doporuceni` ..... podklady pro jednotlivá doporučení
  - `dotaznik` ..... podklady a vyhodnocení dotazníku
  - `mailing` ..... podklady a vyhodnocení mailingu
  - `prediktivni_modely` ..... podklady pro prediktivní modelování
  - `predzpracovani` ..... podklady pro předzpracování dat
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `DP_Friedjungova_Magda_2016.pdf` ..... text práce ve formátu PDF