FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Zadání bakalářské práce

Název:	Predikce výkonů studentů
Student:	Anna Kapitánová
Vedoucí:	Ing. Magda Friedjungová, Ph.D.
Studijní program:	Informatika
Obor / specializace:	Znalostní inženýrství
Katedra:	Katedra aplikované matematiky
Platnost zadání:	do konce letního semestru 2022/2023

Pokyny pro vypracování

FIT disponuje daty o studijních výkonech studentů za posledních 11 let. Tato data jsou uložena v datovém skladu ČVUT. Cílem této práce je data efektivně využít pro usnadnění rozhodovacích procesů.

1) Seznamte se s problematikou "educational data miningu" (EDM) a proveďte rešerši aplikace metod EDM na studijní výsledky studentů. Seznamte se s daty dostupnými v datovém skladu ČVUT.
2) Data analyzujte a nad vybranými navrhněte prediktivní modely inspirované rešerší predikující např. postup studentů do dalších ročníků, studijní úspěšnost jednotlivců, pravděpodobnost dokončení studia apod. (konkrétní úkoly budou specifikovány po diskuzi s vedoucím).
3) Modely patřičně vyhodnoťte a získané výsledky interpretujte.
4) V případě uspokojivých výsledků řešení automatizujte a výsledky zpřístupněte konkrétní skupině uživatelů prostřednictvím webového portálu.
5) Diskutujte znovu použití/zobecnění modelů i v následujících letech při zvážení nové akreditace BI.

Elektronicky schválil/a Ing. Karel Klouda, Ph.D. dne 29. září 2021 v Praze.

Bakalářská práce

PREDIKCE VÝKONŮ STUDENTŮ

Anna Kapitánová

Fakulta informačních technologií
Katedra aplikované matematiky
Vedoucí: Ing. Magda Friedjungová, Ph.D.
10. května 2022

České vysoké učení technické v Praze
Fakulta informačních technologií
© 2022 Anna Kapitánová. Odkaz na tuto práci.
Tato práce vznikla jako školní dílo na Českém vysokém učení technickém v Praze, Fakultě informačních technologií. Práce je chráněna právními předpisy a mezinárodními úmluvami o právu autorském a právech souvisejících s právem autorským. K jejímu užití, s výjimkou bezúplatných zákonných licencí a nad rámec oprávnění uvedených v Prohlášení na předchozí straně, je nezbytný souhlas autora.

Odkaz na tuto práci: Kapitánová Anna. *Predikce výkonů studentů.* Bakalářská práce. České vysoké učení technické v Praze, Fakulta informačních technologií, 2022.

Obsah

Seznam obrázků

Seznam tabulek

Chtěla bych moc poděkovat Ing. Magdě Friedjungové, Ph.D. za čas, který věnovala vedení mé bakalářské práce. Děkuji za všechny cenné rady, vstřícnost a ochotu při vedení práce. Děkuji také pracovníkům datového skladu, zvláště pak Adamovi Marhefkovi, za ochotu a pomoc s problémy s daty. Dále bych moc ráda poděkovala všem mým blízkým, kteří mě v průběhu mého studia podporovali a bez nichž bych se nikdy nedostala až do tohoto bodu. Speciální dík patří mé rodině a příteli Kubovi Jančičkovi, který mi byl v průběhu studia vždy oporou. Děkuji.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracovala samostatně a že jsem uvedla veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při 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. V souladu s ust. § 2373 odst. 2 zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisů, tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programů, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo"), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli způsobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelům). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu) licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným způsobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 10. května 2022 .....

Abstrakt

Dlouhodobým jevem českých vysokých technických škol je nízká úspěšnost studentů v dokončení studia. Školy se proto snaží poskytnout včasnou pomoc studentům, kteří mají problémy se studiem. K rozpoznání takových studentů může posloužit datová analýza a vytěžování znalostí ve vzdělávání. Tato práce se zabývá komplexní analýzou dat týkajících se studentů Fakulty informačních technologií Českého vysokého učení v Praze a záznamů jejich studijních výsledků. Použitá data z let 2009 až 2021 jsou získána z datového skladu ČVUT. Na předzpracovaných datech jsou vytvořeny prediktivní modely za pomocí metod strojového učení. Úkolem vytvořených modelů je získání předpovědi úspěšného dokončení studia na fakultě, rovněž předpověď průchodnosti studentů mezi jednotlivými semestry a také predikce výsledných známek studentů z vybraných fakultních předmětů. Současně je výstupem práce automatizace procesu predikování a diskuse použitelnosti výsledků v budoucích letech.

Klíčová slova vytěžování znalostí ve vzdělávání, vytěžování znalostí, datový sklad ČVUT, předzpracování dat, strojové učení

Abstract

The low study programme completion rate remains a long-term phenomenon among Czech technical universities. Schools therefore seek to provide timely assistance to students who may have difficulties studying. Data analysis and educational data mining may present a way to identify such students. This thesis presents a comprehensive analysis of study records and other associated data of the students of the Faculty of Information Technology, Czech Technical University in Prague. The data warehouse of CTU was used to obtain the data, with the records ranging from the years 2009 to 2021. Machine learning methods were utilised to create the predictive models from the pre-processed data. The created models are tasked with predicting the successful completion of the study programme by individual students and the overall ratio of students successfully completing the current semester. For chosen faculty courses, the students' final grades are also predicted. On top of that, the discussion of the results' applicability in the future and the automation of the prediction process are among the outputs of this thesis.

Keywords educational data mining, data mining, CTU data warehouse, data preprocessing, machine learning

Seznam zkratek

AdaBoost	Adaptive Boosting
BICS	Bakalářský studijní program informatika, prezenční forma v českém jazyce
CIPS	Centrum informačních a poradenských služeb
CRISP-DM	Cross-industry standard process for data mining
CTU	Czech technical university in Prague
ČVUT	České vysoké učení technické v Praze
DM	Data mining
DWH ČVUT	Prototyp datového skladu ČVUT
ECTS	European Credit Transfer System
EDM	Educational data mining
FIT	Fakulta informačních technologií
IZO	Identifikační znak organizace
kNN	K-Nearest Neighbor
KOS	Komponenta studia ČVUT
MAE	Mean absolute error
MICS	Magisterský studijní program informatika, prezenční forma v českém jazyce
ML	Machine learning
MLP	Multilayer perceptron
MSE	Mean Squared Error
MŠMT	Ministerstvo školství, mládeže a tělovýchovy
PP předměty	Předměty povinné programu
REDIZO	Resortní identifikátor právnické osoby
RMSE	Root Mean Square Error
SVM	Support vector machines
VŠ	Vysoká škola / vysokoškolský
XGBoost	Extreme Gradient Boosting

Kapitola 1
Úvod

S tím, jak jde vývoj stále dopředu, se zároveň klade stále vyšší důraz na kvalitní vzdělání. Univerzity po celém světě se zabývají otázkou, jak zlepšit vzdělání a služby poskytované svým studentům. Jak docílit toho, aby byli jejich absolventi dobře uplatnitelní na trhu práce a šířili tak dál dobré jméno školy. Důležitým faktorem, který by chtěli vysoké školy zlepšit, je počet studentů, kteří náročné studium dokončí.

Dle dat dostupných na MŠMT [1, 2] na české vysoké školy ročně nastoupí kolem 50 000 poprvé zapsaných studentů do bakalářských programů. Ale pouze zhruba 50 % z nich nakonec úspěšně dostuduje. Kolem 40 % studentů končí studium již v prvním ročníku. Ať už z důvodu přestupu na jinou školu, tak kvůli nezvládnutí studia. Na českých technických školách je tato statistika ještě smutnější. Na naší fakultě dokončí bakalářské studium jen zhruba 30 % studentů. Téměř 60 % studentů nedokončí první rok studia.

Pro studenty ČVUT přitom existuje řada podpůrných prostředků, které se snaží pomoci studentům při studiu. Jednou z cest, kterou se ČVUT snaží podpořit své studenty, je Centrum informačních a poradenských služeb ČVUT[1], zkráceně CIPS. CIPS nabízí podporu studentům již od roku 2003. ČVUT se tak inspirovalo poradenskými centry na zahraničních univerzitách, které jsou zde běžné. Dále existuje Středisko pro podporu studentů se specifickými potřebami ELSA[2]. Centrum ELSA nabízí celou řadu podpůrných prostředků pro studenty se specifickými poruchami učení. Neméně důležitou podporou studentům jsou samozřejmě samotní kantoři. Ti v rámci svých předmětů vypisují konzultační hodiny a z mé zkušenosti mohu potvrdit, že byli vždy ochotni podat studentům pomocnou ruku při studiu.

Proč tedy takové množství studentů přece jen nedokončí studium úspěšně? Velice důležitá je včasná pomoc. Bohužel ve většině případů ji začnou studenti vyhledávat, až když je již pozdě a řádná pomoc je již nedokáže před případným neúspěchem zachránit. Pro školy je tedy nesmírně důležitá schopnost včas rozpoznat studenty, kteří budou potřebovat v budoucnu pomoci se studiem. K tomuto nelehkému úkolu může pomoci právě datová analýza a vytěžování znalostí z dat, čímž se tato práce zabývá.

Fakulta informačních technologií existuje od roku 2009. Od roku založení studovalo na fakultě přes 16 000 studentů a studentek. K nim existuje zhruba 350 000 záznamů klasifikace. Od doby jejího vzniku se mnohé změnilo. V pravidelných intervalech dochází ke změnám akreditace, kdy dochází ke slučování starých předmětů do nových nebo naopak rozšíření učiva ve starých předmětech, a proto rozdělení předmětů do nově vzniklých. Vznikají nové obory, mění se jejich uspořádání. Roste počet studentů a mění se jejich složení. S přibývajícími lety a nárůstem nových absolventů fakulty roste i množství dat, které fakulta o svých studentech udržuje.

Náplní této práce je průzkum těchto dat a hledání souvislostí v datech. Tak abychom pomocí

[1] https://www.cips.cvut.cz/
[2] https://www.elsa.cvut.cz/

získaných znalostí byli schopni předpovídat úspěšnost našich studentů a poskytnout jim včasnou podporu k jejich studiu.

1.1 Cíle práce

Cílem teoretické části práce je důkladná analýza problematiky *educational data miningu*. Dále je důležité seznámit se s daty z datového skladu ČVUT a správně je interpretovat. Na základě znalostí získaných analýzou dat vybrat vhodná data a obeznámit se s fungováním studia na Fakultě informačních technologií Českého vysokého učení v Praze od doby jejího vzniku.

Cílem praktické části práce je kvalitně předzpracovat vybraná data a následně navrhnout vhodné prediktivní modely inspirované rešerší. Po diskuzi s vedoucí bakalářské práce bylo vybráno predikování úspěšného dokončení studia studentů bakalářského programu Informatika a analýza predikování u navazujícího magisterského programu Informatika, predikování výsledné známky z povinných předmětů bakalářského programu Informatika a predikování úspěšného dokončení jednotlivých semestrů studentů bakalářského studijního programu Informatika. Následným krokem je zhodnocení a interpretace výsledků predikce. V případě uspokojivých výsledků je úkolem automatizovat řešení a zpřístupnit výsledky konkrétní skupině uživatelů, stejně tak jako diskuse znovupoužitelnosti vytvořených modelů a jejich využití v budoucích letech při zvážení nové akreditace.

Výsledky bakalářské práce mohou posloužit pro vedení fakulty jako vhled do problematiky neúspěšnosti studentů a zlepšit tak včasnou pomoc studentům fakulty.

Část I

Teoretická část

Kapitola 2
Rešerše

Zpracování velkého množství dat, jejich analýza, hledání vzorů a struktur nebo třeba detekce anomálií patří do oboru, který se nazývá vytěžování dat (anglicky *data mining*) [3]. Oblast vytěžování dat, která se zaměřuje na školství a vzdělávání, se nazývá vytěžování znalostí ve vzdělávání (anglicky *educational data mining*) - dále jen EDM. Právě EDM se tato práce zabývá.

2.1 Co je to EDM

Jak již bylo zmíněno výše, EDM je podoblast vytěžování dat, která je zaměřená na vzdělávání. Jedná se o obor s velmi širokým polem působnosti. Práce [4] pěkně shrnuje nejčastěji zkoumané oblasti EDM:

Predikce (Prediction) Predikováním se obecně rozumí vývoj modelů, které jsou schopny odvodit jeden aspekt dat (predikovaná proměnná, někdy také nazývaná jako výsledná proměnná) z nějaké kombinace jiných aspektů dat (příznaky). Pro ilustraci uveďme příklad: Na základě informací (věk, pohlaví, známky ze střední školy, . . .) o studentovi budeme predikovat postup studenta prvním ročníkem vysokoškolského studia.

Shlukování (Clustering) Shlukování seskupuje data podle podobných vzorců v nich nalezených. Jednotlivé shluky (*clustery*) potom obsahují sobě podobná data. To, co mají jednotlivé shluky představovat, nemusí a většinou ani není předem známé [5]. Pro ukázku uveďme článek [6]. V něm se autoři zabývají problematikou rozdělení 612 kurzů na *Korean higher education institute*. Jednotlivé shluky vytváří na základě logovacích souborů z tamějšího systémů pro správu vzdělávání. Výsledkem bylo rozdělení kurzů do 4 shluků, podle typického chování pro účastníky kurzu.

Hledání vztahů mezi proměnnými (Relationship Mining) Touto oblastí se rozumí hledání pravidel a vztahů mezi příznaky v datasetu, který jich obvykle obsahuje velké množství. Zmiňme alespoň nejběžnější způsob hledání vztahů mezi proměnnými, kterým jsou asociační metody (*association rule mining*). Asociačními metodami se zabývá článek [7]. Autoři využívají asociačních pravidel pro profilování studentů podle jejich hodnocení. Na datech studentů Fakulty informačních technologií ČVUT v Praze využívala asociačních metod ve své diplomové práci i Ing. Eliška Hrubá [8].

Extrakce dat pro posouzení člověkem Sem patří různé metody vizualizace dat, které umožňují lidem lepší vhled do zkoumané problematiky. Vizualizovat můžeme jak výsledek modelování, tak i samotná data. Vizualizaci dat můžeme využít i v rámci předzpracování dat pro

rychlou detekci anomálií a odlehlých hodnot[1].

Další dělení EDM může být podle stupně granularity zkoumané oblasti, jak zmiňují autoři práce [10]:

- **Na úrovni vyučování** – Do této úrovně spadají úkoly jako: predikce úspěchu v testech nebo co nejlepší rozdělení studentů do studijních skupin podle jejich znalostí, případně včasná identifikace studentů, kterým je potřeba pomoci k úspěšnému zvládnutí kurzu. Na této úrovni EDM se často využívá dat z různých studijních portálů a fór, výsledků z testů či informací o aktivitě v hodinách.

- **Na úrovni kurzu** – Jedná se o úroveň zaměřující se na celý kurz jako na celek. Můžeme sem zařadit predikce, zda student kurz dokončí a případné výsledné známky či predikování průchodnosti kurzu na základě výsledků ostatních kurzů, rozdělování kurzů podle jejich náročnosti, ... Pro predikce na této úrovni mohou být použita data ze školních informačních systémů jako např. známky, průchodnosti předmětů, informace o studentech, kteří kurzy absolvovali, aj.

- **Na úrovni celého studia** – Jedná se o nejvyšší úroveň, která může využívat dat a výsledků obou předchozích oblastí. Zabývá se otázkami typu: Jaká je šance, že student studium dokončí úspěšně? Jak dopředu dokážeme predikovat, jestli bude mít v budoucnu student problémy s dokončením studia?

2.1.1 Problémy EDM

EDM jako takové se potýká s celou řadou problémů, kterým musí datový analytici čelit. V první řadě je predikování v EDM velmi vázané na danou instituci, pro kterou je problém řešen. Jedná se často o velice specifické prostředí, čemuž musí analytik věnovat patřičnou pozornost. Velmi podstatné je mít dostatečnou orientaci a to jak v současném prostředí školy, tak i v minulosti, poněvadž školství je často velmi proměnlivé. Příkladem mohou být obsahy předmětů, jejich uspořádání, proměnlivost jednotlivých ročníků, či vnější vlivy (COVID, změna vedení školy, ...). Z toho také vyplývá velmi obtížná přenositelnost výsledných modelů na jiné instituce.

Dalším specifikem EDM oproti jiným oblastem vytěžování dat je vysoká závislost na lidském faktoru. V rámci EDM se často analytici zaměřují na podobnosti jednotlivých studentů, kdy na základě získaných znalostí aplikují vzorce chování na sobě podobné studenty. Chování a výkon studentů je však často velmi nekonzistentní.

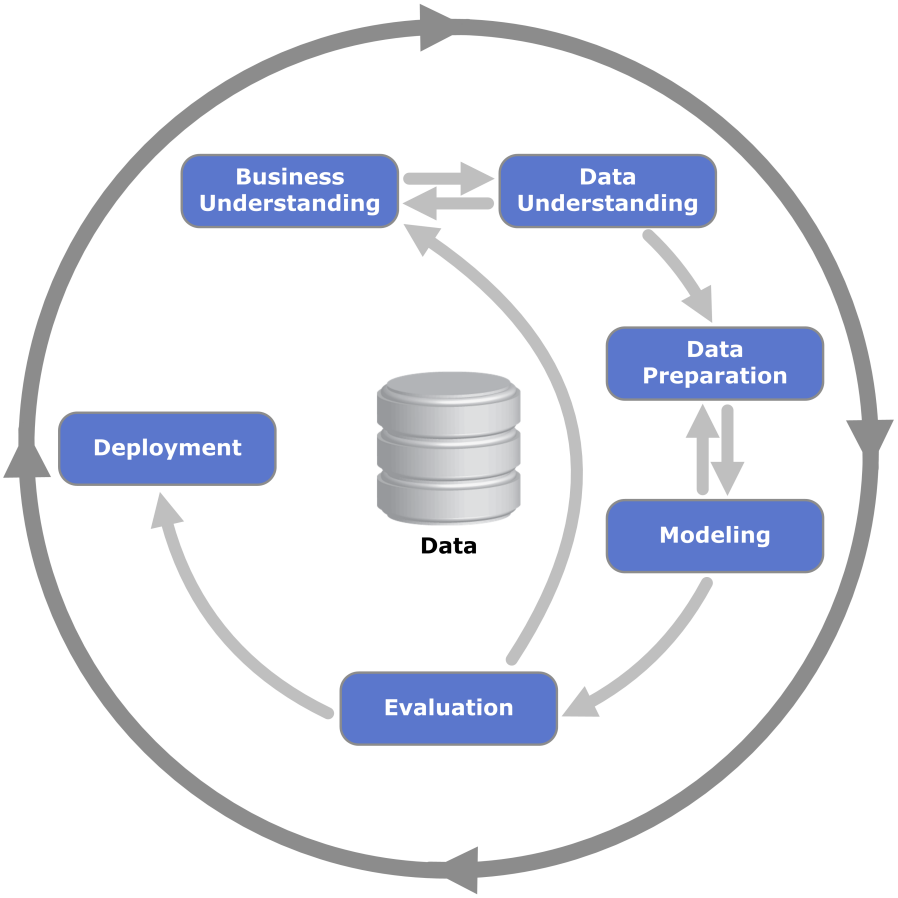
Neméně problematická je také nutnost pracovat s dostatkem kvalitně předzpracovaných dat pro modelování. Množství dat je specifické pro každou instituci. Pro kvalitní predikce je potřeba mít nejlépe zásobu dat z několika ročníků, abychom model nepřeučili na specifikách jednoho akademického roku. Na druhou stranu slepé braní dat z velkého množství ročníků může způsobit naučení modelu na již neaktuálních datech. Hledání správné rovnováhy je tedy pro vytváření výsledné predikce zásadní.

Na závěr bych ještě ráda zdůraznila důležitost včasné predikce. K tomu je zapotřebí mít ideálně celou škálu informací o studentech, ne pouze výsledné známky v předmětech – jak je vidět v sekci 2.5. Je vhodné mít k dispozici jak známky studentů, tak i bodové hodnocení studenta v semestru, jeho chování na hodinách, připravenost, aktivitu v hodinách, informace o mimoškolních aktivitách, sociodemografické informace, data o předchozím vzdělání studenta, jeho ekonomické zázemí a další neméně důležité informace. Zisk těchto informací je ovšem často velmi složitý a nespolehlivý.

[1] Jedná se o datový bod (záznam v datasetu), který se významným způsobem odlišuje od ostatních [9].

2.2 Vybraná metodika

Jedna z nejrozšířenějších metodik v oblasti vytěžování dat je metodika CRISP-DM, plným názvem *CRoss Industry Standard Process for Data Mining* [11]. Metodika je uplatněna ku příkladu v práci [12], ve které se autoři zabývají speciálním rozšířením CRISP-DM o fázi získávání dat v jejich oboru. CRISP-DM se skládá z šesti mezi sebou propojených fází, které na sebe navazují, ale přeskakovaní z navazujících fází na předešlé je v praxi běžné a ve většině situací i žádoucí, jak je pěkně vidět na diagramu 2.1.



Obrázek 2.1 Diagram CRISP-DM (Zdroj: [11]).

Jednotlivé fáze jsou:

- porozumění problematice (*Business understanding*),
- porozumění datům (*Data understanding*),
- příprava dat (*Data preparation*),
- modelování (*Modeling*),
- vyhodnocení výsledků (*Evaluation*),
- využití výsledků (*Deployment*).

Na bázi metodiky CRISP-DM je vypracovaná i tato práce.

2.3 Předzpracování dat

Předzpracování dat (*Data Preprocessing*) je klíčovou součástí vytěžování dat. Kvalitní předzpracování by mělo být stabilním pilířem, na kterém se následně staví veškerá další práce. Běžně se v odborné literatuře můžeme dočíst, že se jedná o časově nejnáročnější úkol vytěžování dat, odhadem zabírá kolem 80 % času [13].

Připomeňme nyní v praxi používané části předzpracování dat, tak jak je uvádí článek [14]:

Čištění dat Hlavním cílem čištění dat je snaha upravit data tak, aby nad nimi mohly pracovat vybrané techniky modelování. To zahrnuje odstranění nerelevantních a vadných dat. Stejně tak i doplnění chybějících hodnot. Mezi nejběžnější způsoby doplnění chybějících hodnot patří: odhad pomocí modelování (kNN), doplnění zvolenou výchozí hodnotou (např. 0, konstantou reprezentující neznámou hodnotu, ...) a doplnění nejčastější hodnotou či průměrem. V případě potřeby může dojít k vybrání podmnožiny z již vyčištěných dat.

Integrace dat Pod pojmem integrace dat si můžeme představit slučování dat z různých zdrojů nebo také propojování datasetů. K tomu je potřebná důkladná analýza dat, aby došlo k propojení pouze hodnot, které náleží stejné entitě.

Redukce dat V rámci zrychlení celého procesu je redukce dat klíčovým aspektem. Je potřeba rozhodnout, jaká data budou dále využita. Redukce se může týkat jak jednotlivých záznamů, tak i redukcí příznaků (*Feature selection*) jednotlivých datasetů. Redukce příznaků je mimo jiné jednou z cest boje pro tzv. prokletí dimenzionality[2] (*Curse of dimensionality*). Úkolem datového analytika je správné protřídění dat tak, aby výsledná přesnost modelování nad vyčištěným datasetem byla stejná nebo lepší než přesnost před čištěním.

Transformace dat Úkolem transformace dat je změna formátu, popř. struktury dat. Patří sem například normalizace dat (tj. přeškálování hodnot), diskretizace dat (roztřídění dat do diskrétních intervalů), agregace dat (agregování více příznaků do jednoho, ku příkladu součtem hodnot jednotlivých příznaků).

2.4 Teorie

Práce je koncipována pro čtenáře se základními znalostmi v oblastech strojového učení. V této sekci proto podrobněji rozebereme pouze použité metody strojového učení, které jsou nad rámec základních znalostí.

2.4.1 XGBoost

XGBoost neboli *eXtreme Gradient Boosting* je v současné době velmi populární metodou strojového učení. Jedná se o *boosting ensamble* metodu, jež využívá gradientních rozhodovacích stromů. Zároveň pro zrychlení celého procesu a zlepšení výkonu využívá paralelizace, zpětného prořezávání stromů a různých algoritmických vylepšení jako regularizace – dochází k penalizaci složitějších modelů, což mimo jiné napomáhá proti přeučení modelu [16].

2.4.2 Stratified k-fold cross-validation

K-násobná křížová validace (anglicky *k-fold cross-validation*) je jedna z nejrozšířenějších způsobů ladění hyperparametrů. Spočívá ve rozdělení datasetu do k stejně obsáhlých množit záznamů, na

[2] Množství dat pro trénování modelů musí pro zachování stejné přesnosti predikce růst exponenciálně ku velikosti dimenze (počtu příznaků). Z toho vyplývá, že s počtem dimenzí roste exponenciálně obtížnost učení modelů [15].

$k-1$ takto rozdělených množinách je natrénován model, který je následně validován na zbývající množině záznamů. Celý proces se k-krát opakuje, tak aby každá z množit dat byla použita jako testovací právě jednou. Výsledná chyba je spočtena jako průměr ze všech chyb naměřených na jednotlivých množinách. Tento postup je opakován pro všechny kombinace hyperparametrů. Vybrána je kombinace s nejnižší průměrnou chybou.

Speciálním případem k-násobné křížové validace, který je užíván v této práci, je tzv. *stratified k-fold cross-validation*[17]. Ta spočívá v rozdělování trénovacích a testovacích datasetů, tak aby poměr zastoupení možných hodnot výsledné proměnné byl v obou datasetech stejný.

2.4.3 Vyhodnocení úspěšnosti modelů

Způsob vyhodnocení úspěšnosti modelů se liší u klasifikačních a regresních úloh. V této sekci se podrobněji seznámíme se způsoby vyhodnocování úspěšnosti modelů, které jsou využity v rámci této práce.

Klasifikace

U klasifikačních úloh se zavádí pojem Matice záměn (*Confusion Matrix*). Matice záměn má jinou podobu pro binární a vícetřídní klasifikaci [18]. Neboť úkolem klasifikačních úloh v této práci je binární klasifikace (snažíme se rozdělit na studenty na úspěšné a neúspěšné v rámci celého studia nebo při průchodu semestrem), zaměříme se na matici záměn pro binární klasifikaci. Při binární klasifikaci predikujeme pouze dva typy hodnot 0 (Ne – *Negative*) a 1 (Ano – *Positive*). Zaměřme se nyní na tabulku 2.1, která představuje obecné uspořádání matice záměn pro binární klasifikaci. Jak je vidět matice udržuje počet čtyř typů ohodnocení výsledků predikce:

- **TP** – ***True Positive*** – Správně klasifikované záznamy jako Ano,
- **FN** – ***False Negative*** – Špatně klasifikované záznamy jako Ne,
- **FP** – ***False Positive*** – Špatně klasifikované záznamy jako Ano,
- **TN** – ***True Negative*** – Správně klasifikované záznamy jako Ne.

Pro měření přesnosti klasifikátoru existují různé míry přesnosti. Vzhledem k této práci se zaměříme na dvě a to Klasifikační přesnost (*Overall accuracy* nebo také pouze *Accuracy*) a F1 skóre (*F1 score*) [19].

Tabulka 2.1 Matice záměn

		Predikovaná třída	
		Positive	*Negative*
Skutečná třída	*Positive*	TP – *True Positive*	FN – *False Negative*
	Negative	FP – *False Positive*	TN – *True Negative*

Klasifikační přesnost – bývá často považována jako nejintuitivnější míra přesnosti. Vypočítá se jako poměr správně predikovaných výsledků ku celkovému počtu záznamů.

$$Klasifikační\ přesnost = \frac{TP + TN}{TP + FP + FN + TN}.$$

F1 skóre – před zavedením samotného F1 skóre je nutné zadefinovat vzorce pro *precision* a *recall*. *Precision* je míra procentuální úspěšnosti predikce pro *positive prediction*. Platí, že maximalizací *precision* se minimalizuje počet záznamů predikovaných jako FP.

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

Naopak maximalizací *recall*, někdy nazývaného také jako *sensitivity*, se minimalizuje množství záznamů predikovaných jako FN.

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

F1 skóre potom právě *precision* a *recall* kombinuje, jedná se o jejich harmonický průměr. Využití nachází F1 skóre zejména, pokud je zastoupení hodnot výsledné proměnné nestejné a převažují záznamy patřící do jedné třídy.

$F1\ Score = \frac{2 \cdot recall \cdot precision}{recall + precision}$.

Regrese

U regresních metod nejde měřit přesnost modelů tak jako u klasifikace. Výkonost modelů se měří pomocí metrik, které určují chybu predikce. Cílem je tedy, aby výsledný prediktivní model měl co nejmenší naměřenou chybu. Zajímáme se o to, jak moc výsledky predikování blíží ke skutečným hodnotám.

RMSE – než zadefinujeme samotné RMSE, je důležité představit, co je to MSE. *Mean Squared Error* zkráceně MSE je jedna z nejpoužívanějších metrik měření chyby u regresních úloh [20]. Je často využívána jako tzv. ztrátová funkce při řešení optimalizačních úloh. Vzorec pro výpočet MSE je následující: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$, kde n je počet záznamů, y značí skutečnou hodnotu výsledné proměnné a $\tilde{y}$ predikovanou hodnotu výsledné proměnné.

RMSE neboli *Root Mean Squared Error* je vypočítáno jako odmocnina z MSE.

2.5 Problematika ve světě

Napříč celým světem vznikají práce, které se zabývají problematikou EDM ve všech možných směrech. Sbírání dat o studentech mnohdy začíná již v předškolním věku, kdy se sleduje jejich chování ve školce, a jednotlivci jsou pak sledováni až do jejich dospělosti napříč celým školním systémem. Sleduje se prospěch studentů, docházka, ale třeba i sociální zázemí, vzdělání jejich rodičů, socioekonomická situace v rodině a další aspekty, které mohou mít vliv na případný akademický úspěch či neúspěch.

Uveďme tedy několik příkladů odlišných přístupů v EDM:

- Studie [21] zkoumá vliv sociodemografických dat studentů (jestli studenti již pracují při studiu a s tím související ekonomická situace v rodině, vzdělání rodičů aj.) pro včasnou identifikaci „zranitelných“ studentů, kteří mají vyšší pravděpodobnost tíhnout k neúspěchu. Výsledky prokazují vliv těchto příznaků.
- Další práce [22] zase prozkoumává rozdíl mezi prvorozenými studenty a studentkami oproti ostatním v akademické sféře. Stejně tak jako vliv pohlaví či příjmu rodičů. Výsledky naznačují vliv těchto příznaků.
- Experiment [23] se zaměřuje na dopad množství času, který studenti stráví na svých telefonech, na jejich výkon ve škole. Skupina dobrovolníků z řad studentů si na svoje mobilní zařízení nainstalovala sledovací aplikace, které snímaly čas strávený na telefonech. Byla zaznamenána korelace mezi takto stráveným časem a studijními výsledky. Ti, kteří trávili na mobilu větší množství času, měli častěji horší známky a naopak.

Pojďme se nyní blíže seznámit s několika zajímavými pracemi, které využívají technik EDM.

2.5.1 Brazilian Public University

Autoři studie [24] se zabývali tím, jak využít efektivně metod vytěžování dat pro včasné odhalení studentů, kteří s vyšší pravděpodobností nedokončí jeden ze dvou nezávislých úvodních programovacích předmětů na *Brazilian Public University*. Jeden kurz byl vyučován distanční formou a druhý prezenční.

Pro distanční kurz vytvořili dataset, který obsahoval záznamy o 262 studentech. Použili jak příznaky obsahující osobní informace o studentech (věk, pohlaví, rodinný stav, město, příjem studenta, registrace, období, třída, semestr, kampus), tak příznaky týkající se aktivity studentů v distančním kurzu – jak často se student přihlašoval, účast v diskuzním fóru, počet doručených a prohlédnutých souborů, jak používají učební pomůcky systému (blog, slovník, kvízy, wiki, zprávy), rok absolvování kurzu, status disciplíny a úspěchy studentů v týdenních aktivitách a kurzech.

Dataset vytvořený pro prezenční kurz obsahoval 161 studentů, kteří se kurzu účastnili v roce 2014 během 16 týdnů. Každý týden byla vyhodnocována jejich aktivita v kurzu a navíc kurz obsahoval 4 testy v 4., 8., 12. a 16. týdnu. Byly použity příznaky: věk, pohlaví, rodinný stav, město, příjem, registrace, období, třída, semestr, kampus, rok zapsání kurzu, status disciplíny, počet vypracovaných cvičení, počet správně vypracovaných cvičení, úspěch studenta v týdenních aktivitách a testech.

Autoři využívají metodu podpůrných vektorů (SVM) a úspěšnost predikce vyhodnocují pomocí F1 skóre. SVM dokázala předpovědět s přesností 92 % u distančního kurzu a 83 % u prezenčního kurzu, zda daný student neuspěje, pokud absolvoval alespoň 50 % kurzů. Dále ve studii využívali: Naivní bayesův klasifikátor (přesnost 86 % a 78 %), rozhodovací strom (87 % a 78 %) a neuronovou síť (88 % a 74 %).

2.5.2 EDM framework

Článek [25] navrhuje komplexní EDM framework, který predikuje na základě analýzy úspěšnost studentů a zároveň je schopen poskytnout rozhodnutí, která k těmto závěrům vedla.

Jejich dataset byl tvořen daty dvou středních škol v Portugalsku. Autoři použili celkem 33 příznaků, které se týkaly známek studentů, sociodemografických dat a příznaky související s danou školou.

Ve studii byla použita logistická regrese, metoda podpůrných vektorů a neuronová síť.

Autoři se snažili predikovat rozdělení studentů do tří kategorií, které reprezentovaly, jak byli daní studenti úspěšní: *high*, *medium*, *poor*. Tato predikce se jim povedla s přesností 79 %, kterou změřili pomocí klasifikační přesnosti.

2.5.3 Analýza nedostudovaných studentů

Cílem studie [26] byla analýza dat studentů čtyřletého bakalářského studia IT a následné poskytnutí informací o výkonech těchto studentů osobám, které jim jsou schopny pomoci.

Autoři se zaměřili se na dva směry:

1. Predikovali úspěch studentů na konci celého čtyřletého programu, a to na základě výsledné známky celého bakalářského programu. Tato známka se vypočítá jako 10 % průměrné známky z prvního ročníku, 20 % z druhého ročníku, 30 % z třetího ročníku a 40 % ze čtvrtého ročníku.
2. Studovali typické průběhy studia a jejich souvislost s výsledkem predikce.

Dataset se skládal z dat 210 studentů, kteří se zapsali v letech 2007/2008, 2008/2009 (dva ročníky). Obsahoval jak počet bodů získaných na střední škole, tak body ze všech předmětů v celém čtyřletém univerzitním programu. Jako trénovací dataset byla použita data z ročníku 2007/2008 a jako testovací data z ročníku 2008/2009.

Pro evaluaci výsledků na testovacím datasetu byly použity metriky Klasifikační přesnost a *Cohenovo kappa*.

Tabulka 2.2 Analýza nedostudovaných studentů – výsledky modelů

Model	Klas. přesnost	Cohenovo kappa
Rozhodovací strom měřen Gini indexem	68,27 %	0,493
Rozhodovací strom měřen Informačním ziskem	69,23 %	0,498
Rozhodovací strom měřen Klasifikační přesností	60.58 %	0,325
Indukci pravidel měřenou Informačním ziskem	55,77 %	0,352
k-nejbližších sousedů, s k = 1	74,04 %	0,583
Naivní Bayesův klasifikátor	83,65 %	0,727
Neuronové sítě	62,50 %	0,447
Náhodný les měřený Gini indexem	71,15 %	0,543
Náhodný les měřený Informačním ziskem	69,23 %	0.426
Náhodný les měřený Klasifikační přesností	62,50 %	0,269

Z tabulky 2.2 je vidět, že nejlepší přesnost naměřili u Naivního Bayesova klasifikátoru. Ačkoliv se Naivní Bayesův klasifikátor ukázal být nejlepším modelem, rozhodli se pro použití jejich vlastního modelu, který se kvůli lepší čitelnosti výsledků skládal z rozhodovacího stromu vylepšeného jejich vytvořenou selekční technikou příznaků. Tato selekční technika příznaků zlepšila výkon téměř poloviny klasifikátorů a zejména pak rozhodovacích stromů.

K vytvoření rozhodovacího stromu pro predikci použili pouze těchto 5 příznaků:

- HS205/206 – *Islamic Studies or Ethical Behaviour*,
- MS-121 – *Applied Physics*,
- CS-251 – *Logic Design and Switching Theory*,
- HS-207 – *Financial Accounting and Management*,
- CT-255 – *Assembly Language Programming*.

Jedná o známky z kurzů z prvních dvou ročníků – další ročníky nebyly brány v potaz, neboť účel studie byl včasné predikování výsledků studenta, jak je popsáno výše. Tyto příznaky vybrali na rozšířeném datasetu o další dva ročníky (blíže není specifikováno). V podstatě šlo o to, že autoři zkoušeli různé modely rozhodovacích stromů a v prvním kroku vybrali takové příznaky, které se při větvení stromů objevovaly alespoň v polovině všech rozhodovacích stromů. Ve druhém kroku se z množiny příznaků z prvního kroku vybraly příznaky takové, které navíc vedly ke dvěma typům uzlů – „čistým uzlům“ a „nečistým uzlům“.

Pro identifikaci výjimečně dobrých studentů rozdělovali výzkumníci uzly následovně:

- „čisté uzly“, které obsahovaly pouze data studentů s výslednou známkou B,
- „nečisté uzly“, které obsahovaly navíc data studentů s výslednou známkou A.

Pro identifikaci špatných studentů rozdělovali uzly následovně:

- „čisté uzly“, které obsahovaly pouze data studentů se známkou D nebo E,
- „nečisté uzly“, které obsahovaly i známku C.

Výsledky naznačují, že pokud se zaměříme na malé množství předmětů, u kterých je velké množství studentů se zvláště dobrými nebo naopak špatnými výsledky, tak je možné poskytnout včasné varování a podporu studentům se špatnými výsledky a zároveň poskytnout rady či příležitosti studentům s dobrými výsledky.

Konkrétně autoři dospěli k těmto třem poznatkům:

1. Pokud mají v prvním ročníku studenti v předmětu *MS-121* počet bodů ≤ 63, mají vyšší pravděpodobnost, že budou na konci studia ohodnoceni známkou D nebo E.
2. Pokud mají v druhém ročníku studenti v předmětu *CS-251* počet bodů < 43 nebo v předmětu *HS-207* < 60, mají vyšší pravděpodobnost, že budou na konci studia ohodnoceni známkou D nebo E.
3. Pokud mají v druhém ročníku ≥ 80 bodů v předmětu *HS-207* nebo mají ≥ 86 bodů v předmětu *CT-255*, mají vyšší pravděpodobnost, že budou na konci studia ohodnoceni známkou A nebo B.

2.5.4 Predikce známky v kurzech na FIT, Masarykova Univerzita v Brně

Zajímavý přístup zvolili také výzkumníci ve studii [27] na Fakultě informačních technologií na Masarykově Univerzitě v Brně. Autoři studie chtěli poskytnout studentům na začátku každého semestru pomocí predikcí lepší vhled do toho, jak si rozvrhnout během semestru pracovní zátěž. Dále může být predikce využita pro doporučování zápisu předmětů. Na začátku kurzu je predikce nejtěžší, chybí totiž data o tom, jak se student v kurzu snaží atd. Postupem času, když se začnou objevovat data jako např. známky z testů, se predikce zpřesňuje.

Ve studii se zaměřili na dva hlavní úkoly:

1. predikce známky v kurzech,
2. predikce úspěšného dokončení studia.

K tomu bylo využito dvou přístupů:

1. Je založen na klasifikačních a regresních algoritmech, které predikují jejich studijní výkon.
2. Je založen na technikách kolaborativního filtrování. Konečné známky studentů jsou předpovídány na základě známek podobných studentů a informací o předmětech.

Oba přístupy byly validovány na datech 3584 studentů univerzity, kteří absolvovali 138 kurzů v letech 2010 až 2013. Trénovací dataset obsahoval záznamy z dat z let 2010–2012 (37 005 záznamů). Testovací dataset obsahoval data z roku 2013 a měl 11 026 záznamů.

Studie zadefinovává 4 hypotézy:

- **H1** – Hypotéza 1 předpokládá, že sociální vazby a chování korelují s výkonem studentů.
- **H2** – Hypotéza 2 předpokládá, že známky z jednotlivých předmětů můžou charakterizovat znalosti studentů. Pokud by platila, k predikci průchodu konkrétního studenta daným předmětem by mohlo stačit vybrat pouze studenty s podobnými zájmy a znalostmi.
- **H3** – Hypotéza 3 předpokládá, že podobné předměty vyžadují podobné dovednosti k jejich úspěšnému absolvování. Pokud by platila, k predikci výsledných známek by mohlo stačit méně předmětů aniž by se nějak signifikantně zhoršila přesnost predikce, a tím by se dala snížit výpočetní náročnost predikce.
- **H4** – Předpokládá, že každý ze dvou přístupů, které jsou popsané výše, se hodí na jiné předměty.

Všechny tyto hypotézy jsou následně v práci prokázány. Pojďme se nyní blíže seznámit s tím, jak autoři ve své práci postupovali.

Klasifikace a regrese

Pro účely prvního přístupu založeném na klasifikačních a regresních algoritmech rozdělili autoři studie data na dvě skupiny:

1. Osobní data studenta, tedy pohlaví, datum narození, rok přijetí, počet získaných kreditů a průměr známek.
2. Data popisující jak chování studenta, tak vzájemnou kooperaci studentů na diskusních fórech. Tato data obsahovala příspěvky a komentáře na diskusních fórech, statistiky emailů a publikování či sdílení souborů. Z těchto dat byl poté vytvořen sociogram, pomocí kterého bylo možné odvodit např. vážený průměr přátel daného studenta, počet přátel studenta, míru důležitosti studenta v síti nebo počet návštěv daného kurzu studentem.

Na datech z první skupiny tvůrci studie predikovali jak úspěšné dokončení studia, tak predikci známky. Pro predikci úspěšně dokončeného studia využili metod klasifikace – výsledky jednotlivých metod jsou vidět v tabulce 2.3. Z tabulky vyplývá, že pouze osobní data studenta nestačí k dobré predikci úspěšného dokončení studia. Nejlepší výsledky měla metoda podpůrných vektorů, kde přesnost měřená pomocí F1 skóre vychází na 55,9 %. Model Baseline je pouze kontrolní model, který predikuje vždy neúspěch.

Tabulka 2.3 Predikce známky v kurzech na FIT, Masarykova Univerzita v Brně – výsledky modelů, predikujících úspěšné dokončení studia, nad osobními daty studentů

Model	F1 skóre	MAE	Recall
Metoda podpůrných vektorů	0,559	0,161	0,444
Naivní Bayesův klasifikátor	0,554	0,251	0,467
J48	0,552	0,182	0,397
Náhodný les	0,550	0,173	0,362
Part	0,543	0,202	0,417
IB1	0,536	0,216	0,436
OneR	0,508	0,183	0,321
Baseline	0,326	0,822	1

K predikci známky bylo využito regresních metod, jak je vidět z tabulky 2.4. Baseline tentokrát představuje model, který vždy predikuje průměrnou známku. Nejlepší model je opět tvořen metodou podpůrných vektorů.

Tabulka 2.4 Predikce známky v kurzech na FIT, Masarykova Univerzita v Brně – výsledky modelů, predikujících známku, nad osobními daty studentů

Model	MAE	Recall
Metoda podpůrných vektorů reg.	0,605	0,196
Lineární regrese	0,615	0,152
Aditivní regrese	0,634	0,165
RepTree	0,643	0,184
Náhodný les	0,668	0,216
IBk	0,767	0,294
Baseline	0,806	0

Pro výslednou predikci známky tedy byla zvolena metoda podpůrných vektorů. Pokud model predikující úspěšné dokončení studia predikoval neúspěch nebo pokud byla predikována známka F, výsledný model předpovídal neúspěch studenta. Úspěšnost výsledné predikce je vidět v tabulce 2.5.

Tabulka 2.5 Predikce známky v kurzech na FIT, Masarykova Univerzita v Brně – výsledky modelů

Dataset	MAE	Recall
Trénovací dataset	0,701	0,524
Testovací dataset	0,744	0,414

Tabulka 2.6 Predikce známky v kurzech na FIT, Masarykova Univerzita v Brně – výsledky modelů, osobní data i data z fóra

Dataset	Atributy	MAE	Recall
Trénovací dataset	Osobní data	0,701	0,524
	Osobní data + data z fóra	0,629	0,528
Testovací dataset	Osobní data	0,744	0,414
	Osobní data + data z fóra	0,688	0,427

Po obohacení datasetu o data popisující chování studentů na diskusních fórech se predikce zlepšila (viz tabulka 2.6), čímž byla hypotéza 1 (tedy, že sociální vazby korelují s výkonem studenta) potvrzena.

Kolaborativní filtrování

Jak již bylo zmíněno výše, druhý přístup byl založen na metodách kolaborativního filtrování. Pro predikci známek zavedli autoři pojem tzv. *baseline* studenta. Konkrétně zadefinovali dva typy *baseline* studentů:

1. **Average student** – student, který má průměrnou známku ze všech předmětů.
2. **Uniform student** – student, jehož průměr známek za všechny předměty je označen jako průměrný (jako průměrná známka byla vypočítána známka D, ta odpovídala hodnotě 2,5 v číselném označení).

Následně vytvořili dataset v podobě matice, jejíž řádky tvořily záznamy studentů a sloupce představovaly všechny předměty, u nichž se predikovala známka. Buňky matice obsahovaly záznamy klasifikace studentů v patřičných předmětech, pokud student daný předmět neabsolvoval, zůstalo pole nevyplněné. Nad takto sestavenou maticí následně autoři vypočítali sobě nejpodobnější studenty. K tomu využili čtyři různé míry podobnosti – *Mean absolute difference* (MAD), *Root mean squared difference* (RMSD), *Cosine similarity* (COS) a *Pearson's correlation coefficient* (PC). Vytvořili tak tzv. Podobnostní matici (anglicky *Similarity matrix*)[3].

Po vytvoření Podobnostní matice museli autoři (vzhledem k velkému množství předmětů) rozdělit předměty podle jejich kapacit na malé (do 30 studentů), střední (mezi 30 a 70 studenty) a velké (předměty s kapacitou vyšší než 70 studentů). Následně hledali, jak velké okolí, tj. velikosti shluků, mají vyhledávat u jednotlivých typů předmětů. U malých předmětů se rozhodli pro maximálně 10 studentů v jednom shluku, u středních pro 15 studentů a pro velké až 30 studentů. Výsledné shluky následně obsahovaly takové studenty, kteří si byli vzájemně více podobní, nežli byli podobní *baseline* studentům. Konečné známky vyšetřovaného studenta byly odhadnuty na základě známek studentů patřících s ním do stejného shluku. K tomu bylo využito metod jako průměr, max, median, ale i pokročilejší metody, které využívaly vážení významnosti. Přesnost predikce byla měřena pomocí MAE – 0,650 a *recall* – 0,267.

Autoři se také zaměřili na hledání podobností mezi jednotlivými předměty. K tomu využili jak informace o průchodnosti jednotlivých předmětů, jejich kapacity, tak i další rozšiřující informace jako jsou prerekvizity (množina kurzů, kterými je potřeba projít před zapsáním kurzu), literatura

[3]Matice jejíž sloupce a řádky tvoří jednotlivé datové body (zde představují studenty) a pole matice jsou vyplněny naměřenou podobností příslušných dvou datových bodů [28].

(autoři doporučené literatury), vyučující a jiné. Následně s pomocí těchto charakteristik provedli shlukování předmětů. Shluky obsahovaly 2 až 22 předmětů, v průměru jeden shluk obsahoval tři předměty. Výzkumníci hledali způsoby, jak co nejefektivněji shlukovat předměty podle jejich podobnosti. Shlukování předmětů výzkumníci využili při predikci známek studentů, kdy se pro výslednou predikci předmětu stačilo zaměřit pouze na podobné předměty zkoumaného předmětu. Výsledky měřili opět pomocí MAE (0,661 na trénovacím datasetu a 0,685 na testovacím) a *recall* (0,470 na trénovacím datasetu a 0,418 na testovacím).

Výsledné srovnání obou přístupů

Ukázalo se, že oba přístupy měly podobné průměrné výsledky. Oba přístupy byly dostatečně kvalitní, aby mohly být využity k předpovědi finálních známek v kurzech. První přístup, predikce známek pomocí regrese, vykazoval lepší výsledky pro kurzy s menším počtem zapsaných studentů. Druhý přístup využívající kolaborativní filtrování byl přesnější ve vyhodnocování matematických kurzů. Přestože pro většinu kurzů dokázali predikovat výsledky spolehlivě, existovalo pár kurzů, u kterých výsledky nebyly uspokojivé.

Kapitola 3
Situace na FIT ČVUT v Praze

Dne 1. července 2009 došlo k oficiálnímu založení Fakulty informačních technologií ČVUT v Praze [29]. V říjnu téhož roku tak na fakultě započalo své studium téměř 500 prvních studentů bakalářského programu Informatika a následujícího roku se otevřel navazující magisterský program Informatika.

Standardní délka bakalářského studia jsou tři roky s možným bezplatným prodloužením studia na čtyři roky. U magisterského studia délka standardního studia odpovídá dvěma letům s možností bezplatného prodloužení na tři roky. Oba programy Informatika se dále dělí na obory a specializace. Programy akreditované před novelou VŠ zákona se dělí na obory, programy akreditované po září 2016 se dělí na specializace. Některé obory/specializace se dále ještě dělí na jednotlivá zaměření [30].

Akademický rok se skládá ze dvou částí – zimního a letního semestru. Během prvního ročníku nemají studenti ještě zpravidla zapsán žádný obor či specializaci, ty si vybírají až v průběhu studia – avšak nejpozději při výběru zadání závěrečné práce.

V této kapitole se hlouběji seznámíme s fungováním fakulty.

3.1 Předměty na FIT

Studenti si zapisují předměty dle studijních plánů jejich oborů/specializací. Doporučené (avšak nikoliv povinné) průchody studijním plánem jsou zveřejněny děkanem fakulty v elektronické Bílé knize[1] [30].

Na konci studia musí ještě studenti bakalářského i magisterského programu stvrdit své znalosti nabrané studiem při státní závěrečné zkoušce a odevzdáním závěrečné práce (bakalářská a diplomová práce).

3.1.1 Typy předmětů

Předměty se na naší fakultě dělí na předměty povinné a volitelné. Povinné předměty se ještě dále rozlišují na:

- **PO** – Povinné předměty oboru,
- **PP** – Povinné předměty programu,
- **PE** – Povinné předměty ekonomické,

[1] https://bk.fit.cvut.cz/ a https://bilakniha.cvut.cz/

- **PZ** – Povinné předměty zaměření,
- **PJ** – Povinná zkouška z angličtiny,
- **PS** – Povinné předměty specializace,
- **PT** – Povinná tělesná výchova, sportovní kurzy,
- **PV** – Povinně volitelné předměty.

Volitelné předměty se dále dělí na:

- **V** – Volitelné předměty,
- **VE** – Povinně volitelné ekonomicko-manažerské,
- **VH** – Povinně volitelné humanitní,
- **VO** – Volitelné předměty oboru/specializace.

Platí, že pro úspěšné dokončení bakalářského studia musí studenti absolvovat všechny povinné předměty programu (jedná se o předměty, které mají všichni studenti společné), všechny povinné předměty oboru (nově specializace) či zaměření, jeden předmět povinně ekonomický (BI-EMP), jeden předmět povinně volitelný ekonomicko-manažerský (dle svého výběru), jeden předmět povinně volitelně humanitní (na výběru studenta), dvakrát tělesnou výchovu, povinnou zkoušku z angličtiny a nespecifikovaný počet volitelných předmětů (dle své volby), tak aby měl na konci studia získáno minimálně 180 kreditů.

Studenti magisterského programu musí pro úspěšné dokončení studia absolvovat všechny povinné předměty programu, jeden předmět povinně volitelně ekonomicko-manažerský, jeden předmět povinně volitelně humanitní a patřičný počet předmětů volitelných a volitelných oborových, tak aby počet získaných kreditů dával na konci studia v součtu alespoň 120 kreditů.

3.1.2 Kódy předmětů

Každý předmět na FIT má svůj unikátní kód pro rychlejší a přehlednější orientaci v předmětech. Předměty patřící do bakalářského programu Informatika v prezenční formě v českém jazyce začínají předponou `BI`. Předměty magisterského programu Informatika Informatika v prezenční formě v českém jazyce akreditované před nejnovější magisterskou akreditací z roku 2020 mají předponu `MI`, předměty z nové akreditace mají předponu `NI`. Zbytek kódu je tvořen zkratkou názvu předmětu, která obsahuje tři znaky. Některé předměty obsahují ještě číselnou koncovku oddělenou tečkou, která specifikuje verzi předmětu – například předmět BI-SI1.2.

3.1.3 Povinné společné předměty programu

Jedná se o předměty, které mají všichni studenti společné a musí je splnit, aby úspěšně dokončili studium. Právě s těmito předměty tato práce nadále kalkuluje, je pro ně používáno značení PP předměty. Samozřejmě pro účely práce by bylo zajímavé se zaměřit i na předměty povinné oborové, bohužel jsem ale neměla k dispozici informace o tom, do jakých oborů studenti patří, a případné roztřídění studentů podle oborů tedy nebylo možné.

Bakalářský program

V bakalářském programu Informatika v prezenční formě v českém jazyce musí student splnit povinné předměty programu a BI-EMP, jelikož každý student musí absolvovat jeden předmět povinně ekonomický, přičemž fakulta nabízí pouze předmět BI-EMP. Pojďme se nyní podrobněji seznámit s jednotlivými PP předměty, jejich vývojem v průběhu let a také návazností na novou akreditaci z roku 2021 (bližší detaily k akreditaci 2021 jsou popsány v sekci 3.3).

BI-PA1 – Programování a algoritmizace 1

Jedná se o předmět vyučovaný od založení fakulty. Zároveň je jeho zapsání doporučováno po celou dobu jeho výuky na první semestr.

BI-PAI – Právo a informatika

Předmět, který je rovněž vyučován po celou dobu existence fakulty a doporučený zápis je též v prvním semestru po celou dobu jeho vyučování. V nové akreditaci existuje jako předmět BI-PAI.21 a jeho doporučený zápis je v druhém či čtvrtém semestru dle zvolené specializace.

BI-CAO – Číslicové a analogové obvody

Předmět, který má doporučený zápis v prvním semestru a existuje od doby vzniku fakulty. V nové akreditaci byl předmět nahrazen novým předmětem BI-TZP.21 (Technologické základy počítačů), oba předměty jsou vzájemně uznatelné a jejich osnovy jsou srovnatelné.

BI-PS1 – Programování v shellu 1

Jedná se o předmět, který v rámci akreditace v roce 2015 nahradil předmět BI-UOS (Úvod do operačních systémů), osnova BI-PS1 téměř kopírovala předchozí předmět. V současné době byl v nové akreditaci nahrazen předmětem Unixové operační systémy a navrátil se tak ke původní zkratce pouze s koncovkou .21, tedy BI-UOS.21. Všechny verze předmětu byly vzájemně uznatelné. V doporučeném průchodu studiem je po celou dobu výuky všech jeho verzí uváděn v prvním semestru a momentálně je vyučován pouze v zimním semestru, až do roku 2015 (včetně) byl nicméně vyučován i v letním semestru.

BI-MLO – Matematická logika

Matematická logika je předmět existující od vzniku fakulty s doporučeným zápisem v prvním semestru. V nové akreditaci došlo nicméně ke spojení předmětů BI-MLO a BI-ZDM do nového předmětu BI-DML.21 (Diskrétní matematika a logika), který oba předměty nahradil, bližší informace o uznatelnosti předmětů staré a nové akreditace se můžete dočíst v sekci 3.3.

BI-ZMA – Základy matematické analýzy

Další z předmětů, který byl vyučován od doby vzniku fakulty a jeho doporučený zápis byl v prvním semestru. V nové akreditaci došlo, k rozšíření stávajícího učiva a rozložení předmětu do dvou nových – BI-MA1.21 a BI-MA2.21. Oba zmíněné předměty jsou pro studenty bakalářského programu Informatika povinné a bližší informaci o uznatelnosti mezi starým a novými předměty se můžete dočíst v sekci 3.3.

BI-PA2 – Programování a algoritmizace 2

Jedná se o předmět, který navazuje na předmět BI-PA1. Existuje od vzniku fakulty a jeho doporučený zápis je po celou dobu v druhém semestru. V nové akreditaci existuje ve verzi BI-PA2.21.

BI-DBS – Databázové systémy

Předmět, který funguje od doby vzniku fakulty. Od roku 2009 až do roku 2015, byl doporučený zápis předmětu stanoven na třetí semestr. Od roku 2015 je doporučen zápis již ve druhém semestru. V nové akreditaci byl nahrazen předmětem BI-DBS.21 a oba předměty jsou vzájemně uznatelné.

BI-SAP – Struktura a architektura počítačů

Opět se jedná o předmět, který je vyučován od doby založení fakulty a jeho doporučený zápis je po celou dobu ve druhém semestru. V současné době byl v nové akreditaci nahrazen předmětem BI-SAP.21, který obsahuje identické učivo a jsou mezi sebou vzájemně uznatelné.

BI-LIN – Lineární algebra

Předmět Lineární algebra existoval od vzniku fakulty a měl stanoven doporučený zápis na druhý semestr. V nové akreditaci došlo k rozšíření stávajícího učiva a rozdělení předmětu na dva – BI-LA1.21 a BI-LA2.21. Pro všechny studenty je nadále povinný pouze předmět BI-LA1.21. BI-LA2.21 je povinný jen pro vybrané specializace.

BI-AG1 – Algoritmy a grafy 1

Jedná se o předmět, který nahradil předměty BI-EFA (Efektivní algoritmy) a BI-GRA (Grafové algoritmy a základy teorie složitosti). Oba tyto původní předměty nebyly předměty povinné pro všechny studenty, ale pouze pro studenty vybraných oborů. Učivo původních předmětů a učivo předmětu BI-AG1 je poměrně odlišné. V nové akreditaci je předmět vyučován ve verzi BI-AG1.21 a doporučený zápis je ve třetím semestru.

BI-AAG – Automaty a gramatiky

Automaty a gramatiky je předmět, který byl vyučován od doby založení fakulty, a doporučený zápis byl ve třetím semestru. V nové akreditaci je vyučován ve verzi BI-AAG.21.

BI-ZDM – Základy diskrétní matematiky

Předmět, který byl vyučován od doby založení fakulty, doporučený zápis byl ve třetím semestru. Jak již bylo řečeno u předmětu BI-MLO, v nové akreditaci jsou oba tyto předměty sloučeny a nahrazeny předmětem BI-DML.21.

BI-OSY – Operační systémy

Předmět byl vyučován od vzniku fakulty. Od roku 2009 do roku 2011 s doporučeným zápisem ve druhém semestru, od roku 2012 byl doporučený semestr změněn na čtvrtý. V nové akreditaci je vyučován ve verzi BI-OSY.21.

BI-PSI – Počítačové sítě

Předmět vyučovaný od začátku fakulty, doporučený semestr byl vždy čtvrtý semestr. V nové akreditaci je vyučován ve verzi BI-PSI.21.

BI-BEZ – Bezpečnost

Další předmět, který byl vyučován od doby vzniku fakulty. V letech 2009 až 2010 měl doporučený zápis v pátém semestru. Od roku 2011 měl doporučený semestr absolvování stanoven na čtvrtý semestr. V nové akreditaci byl předmět nahrazen předmětem BI-KAB.21 (Kryptografie a bezpečnost), oba předměty jsou vzájemně uznatelné.

BI-PST – Pravděpodobnost a statistika

I Pravděpodobnost a statistika byla vyučována od založení fakulty, v letech 2009 až 2010 byl doporučen k absolvování čtvrtý semestr, od roku 2011 potom pátý semestr. V nové akreditaci je vyučován ve verzi BI-PST.21.

BI-DPR – Dokumentace, prezentace, rétorika

Od roku 2009 do roku 2014 existoval předmět BI-PPR (Projekt, prezentace a rétorika), jehož doporučený semestr byl stanoven na pátý semestr. Tento předmět byl od akreditace roku 2015 nahrazen nynějším předmětem BI-DPR, jehož doporučený zápis připadá na semestr šestý. Oba předměty měly podobné probírané učivo a byly vzájemně uznatelné. V nové akreditaci nahradil předmět BI-DPR nový předmět BI-TDP.21 (Tvorba dokumentace a prezentace), který je rovněž se svou starou verzí vzájemně uznatelný.

BI-SI1.2 – Softwarové inženýrství I

Předmět vyučovaný od založení fakulty. Mezi roky 2009 až 2012 existoval ve verzi BI-SI1, od roku 2013 byl předmět změněn na nynější předmět BI-SI1.2, platila vzájemná uznatelnost mezi předměty. Zajímavostí je oficiální doporučený semestr uváděný v Bílé knize, ten je totiž stanoven na sedmý semestr[2] (standardní délka studia je šest semestrů). Sedmý semestr je vybrán, neboť není jeden doporučený semestr a výběr, kdy si předmět zapsat, je plně v kompetenci studenta, ač v rámci jednotlivých oborů jsou stanoveny doporučené semestry absolvování. Zároveň bylo v minulosti absolvování předmětu odpuštěno, pokud student absolvoval předmět BI-ZSI (Základy softwarového inženýrství). V nové akreditaci byl předmět nahrazen předmětem BI-SWI.21 (Softwarové inženýrství) a již nepatří mezi PP předměty pro všechny studenty, předmět je povinný jen pro vybrané specializace.

BI-EMP – Ekonomické a manažerské principy

Již od založení fakulty platilo, že pro úspěšné dokončení studia musí student absolvovat jeden z povinně ekonomických předmětů. Mezi lety 2009 a 2012 existovali předměty BI-EPD (Ekonomika podnikání) a BI-EKP (Ekonomika podniku). V roce 2013 došlo ke sloučení těchto předmětů do nového předmětu BI-EPD.2 a zároveň byl navýšen počet kreditů z původních 4 kreditů (BI-EPD) na 5 kreditů (BI-EKP měl taktéž 5 kreditů). V akreditaci roku 2015 byl předmět BI-EPD.2 nahrazen nynějším předmětem BI-EMP a počet kreditů byl opět snížen na 4 kredity. Doporučený semestr byl u všech vyjmenovaných předmětů stanoven (ze stejných důvodů, jako u předmětu BI-SI1.2) na sedmý semestr. V nové akreditaci již předmět BI-EMP nepatří mezi PP předměty.

V minulosti ještě existoval a patřil mezi PP předměty předmět BI-TED (Tvorba elektronické dokumentace) s doporučeným zápisem v druhém semestru. Jeho výuka byla zrušena s akreditací v roce 2015. S tímto předmětem není při vytváření datasetů pro prediktivní modely počítáno právě z důvodu jeho zrušení.

Magisterský program

V magisterském programu Informatika v prezenční formě v českém jazyce musí student absolvovat čtyři níže blíže popsané povinné předměty programu. V minulosti existovali ještě další povinné předměty – MI-TES, MI-TES.2, MI-PRM, MI-IBE a MI-CIO, ty byly ale postupně rušeny jako povinné a od akreditace z roku 2016, které momentálně dobíhá a je nahrazována nejnovější akreditací z roku 2020, nepatřil již ani jeden mezi předměty povinné.

MI-PAA – Problémy a algoritmy

Předmět byl vyučován od roku 2010, mezi lety 2010 až 2015 byl jeho doporučený zápis ve třetím semestru. S akreditací v roce 2016 se jeho doporučený průchod přesunul již do prvního semestru. V nejnovější akreditaci z roku 2020 byl předmět nahrazen předmětem NI-KOP (Kombinatorická optimalizace), předměty jsou vzájemně uznatelné.

MI-MPI – Matematika pro informatiku

Předmět byl taktéž vyučován od roku 2010, mezi lety 2010 až 2011 byl jeho doporučený semestr druhý semestr. Od roku 2012 je jeho doporučeným semestrem semestr první. V nejnovější akreditaci z roku 2020 je předmět nahrazen předmětem NI-MPI, oba předměty jsou vzájemně uznatelné.

MI-PDP.16 – Paralelní a distribuované programování

Roku 2010 byl vyučován ve verzi MI-PAR (Paralelní algoritmy a systémy) a jeho doporučený zápis byl v prvním semestru. V roce 2011 došlo k pozměnění předmětu a jeho přejmenování

[2] https://bk.fit.cvut.cz/cz/pruchody/pr1241646176405.html/

na novější verzi MI-PAR.1. V roce 2013 vznikl odloučením od původního předmětu MI-PAR.1 předmět MI-PPR (Paralelní programování), MI-PAR.1 byl aktualizován a změněn na předmět MI-PAR.2. MI-PPR byl v budoucnu ještě pozměněn a přejmenován na MI-PPR.2. Předmět MI-PPR byl povinný předmět a sloužil jako sesterský předmět MI-PAR.2 až do akreditace roku 2016. V akreditaci roku 2016 došlo ke sloučení obou předmětů do nového předmětu MI-PDP.16 a jeho doporučený semestr byl posunut na druhý semestr. V nejnovější akreditaci z roku 2020 je předmět nahrazen předmětem NI-PDP, předměty jsou vzájemně uznatelné.

MI-SPI.16 – Statistika pro informatiku

V roce 2010 existoval předmět ve verzi MI-SPI a jeho doporučený zápis byl v prvním semestru. Následujícího roku 2011 byl předmět upraven a přejmenován na předmět MI-SPI.1 a doporučován pro první semestr. Od roku 2012 se změnil doporučený semestr na druhý semestr. S akreditací v roce 2016 byl předmět přejmenován na MI-SPI.16. V nejnovější akreditaci z roku 2020 je předmět nahrazen předmětem NI-VSM (Vybrané statistické metody), oba předměty jsou vzájemně uznatelné.

3.2 Hodnocení studentů

Hodnocení studentů na celém ČVUT probíhá prostřednictvím Evropského systému přenosu a akumulace kreditů (ECTS). Tabulka 3.1 zobrazuje systém známkování, který odpovídá standardu ECTS. Jednotlivé známky jsou uvedeny v sestupném pořadí – A odpovídá nejlepší známce, F nejhorší [31].

Tabulka 3.1 Rozdělení známek

	A	B	C	D	E	F
Číselně	1	1,5	2	2,5	3	4
Slovně	výborně	velmi dobře	dobře	uspokojivě	dostatečně	nedostatečně
Počet bodů	100–90	89–80	79–70	69–60	59–50	< 50

Předměty vyučované na FIT mohou být zakončeny následujícími pěti způsoby:

- **Z** – Předměty, které jsou zakončeny pouze zápočtem. Tedy student je ohodnocen pouze binárně – uspěl/neuspěl. Jedná se např. o tělesné výchovy, bakalářskou práci, diplomovou práci či různé zahraniční stáže.
- **Z, ZK** – Předměty, které jsou zakončené zápočtem a zkouškou. Student je ohodnocen výslednou známkou. Do této kategorie spadá většina povinných předmětů.
- **ZK** – Předměty zakončené pouze zkouškou, po které je student ohodnocen výslednou známkou. Z povinných předmětů se jedná například o předmět BI-PAI (Právo a informatika).
- **KZ** – Předměty zakončené kombinovaným zápočtem, student je ohodnocen známkou již za práci v semestru. Příkladem je předmět BI-PS1 (Programování v shellu), nynější BI-UOS (Unixové operační systémy).
- **NIC** – Takto jsou v datamartech označeny tzv. fiktivní předměty, které představují akce, rezervace rozvrhu aj.

Pokud student úspěšně absolvuje předmět, je mu přidělen patřičný počet kreditů. Rozdělení kreditů mezi předměty by mělo odpovídat náročnosti jednotlivých předmětů – uvádí se, že jeden kredit odpovídá zhruba 25–30 hodinám práce v semestru. Na FIT mají povinné předměty nejčastěji okolo 4 až 6 kreditů. Aby student mohl úspěšně dostudovat, je potřeba mít na konci studia odpovídající počet kreditů ke svému studijnímu programu. Zároveň na konci každého roku

(v případě prvního ročníku i po prvním semestru) dochází ke kontrole počtu kreditů. Minimální počet kreditů pro postup ve studiu je uveden v tabulce 3.2.

Tabulka 3.2 Minimální počet kreditů pro postup ve studiu

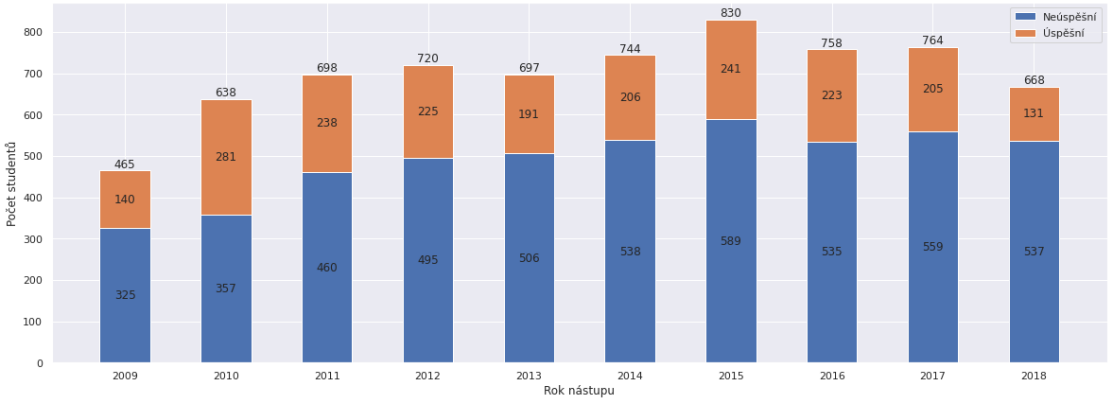
Období kontroly	Bakalářský program	Magisterský program
Po prvním semestru	15	20
Po prvním roce	30	40
Po každém dalším roce	40	40

3.2.1 Průchodnost a průchodnost v průběhu let

Jak již bylo zmíněno, nejvíce neúspěšných bakalářských studentů končí v prvním roce studia, jedná se o zhruba 60 % neúspěšných studentů. S každým dalším ročníkem počet „odpadlých“ studentů výrazně klesá, to je ale samozřejmě ovlivněno i faktem, že v každém dalším roku studia studuje méně a méně studentů.

Nejčastěji bakalářští studenti dostudují úspěšně ve standardní délce studia tří let. Jedná se zhruba o 55 % všech úspěšných studentů, téměř 32 % studentů potom úspěšně dokončí své studium po čtyřech letech studia. U studentů magisterského programu dokončí studium ve standardní délce dvou let 45 % studentů a po třech letech dokončí studium 45,5 % studentů.

Jak je to s průchodností v průběhu let můžeme vidět na grafu 3.1. Graf ukazuje počty úspěšných a neúspěšných bakalářských studentů podle data jejich nástupu na fakultu. Posledním rokem, který zobrazuje, je rok 2018 – data z tohoto roku ale ještě nejsou úplně relevantní, neboť velké procento studentů s tímto rokem nástupu stále studuje.



Obrázek 3.1 Úspěch a neúspěch studentů podle roku nástupu

3.3 Akreditace a návaznost předmětů

Na fakultách českých vysokých škol musí v pravidelných intervalech docházet k obnovám akreditací. Momentálně stanovuje MŠMT šestiletou platnost akreditace vzdělávacích institucí [32].

Pro bakalářský program Informatika jsou zásadní tři akreditace: úplně první akreditace z akademického roku 2009/2010, akreditace z roku 2015/2016, které přinesla výrazné změny do vyučovaných předmětů a nejnovější akreditace, která započala v letošním akademickém roce 2021/2022. V následujícím textu se na tuto akreditaci podíváme více detailně.

Pro magisterský program Informatika jsou zásadní tyto akreditace: první akreditace v roku 2010/2011, akreditace z roku 2016/2017, ta přinesla vůbec největší změny v PP předmětech. Poskládání PP předmětů, se od této akreditace příliš nezměnilo. A samozřejmě nejnovější akreditace z roku 2020/2021, ta přinesla změny hlavně v oborových předmětech.

V akademickém roce 2021/2022 došlo v bakalářském programu Informatika na fakultě k přechodu ze staré akreditace 2015 na novou akreditaci 2021. V rámci akreditací na fakultě se jedná akreditaci, která přináší největší změny výuky. Dosavadní obory jsou nahrazeny specializacemi, došlo k pozměnění rozložení a výběru předmětů v jednotlivých původních oborech. U některých oborů došlo zároveň při konverzi na specializace i k jejich přejmenování. Přistoupilo se rovněž k vytvoření úplně nových specializací bez vazby na předchozí obory, jiné nové specializace zase vznikly rozdělením původních oborů. Celkově obsahuje nová akreditace deset specializací, zatímco stará akreditace měla na výběr pouze z šesti oborů.

Zásadní změny jsou vidět zejména v matematických předmětech [33]. Dosavadní průchod matematickými předměty byl sestaven již při vzniku fakulty roku 2009, nové sestavení předmětů tak reflektuje dlouhodobou zkušenost kantorů i studentů s výukou matematiky. Tabulka 3.3 shrnuje uznatelnost nově vzniklých PP předmětů s předměty staré dobíhající akreditace.

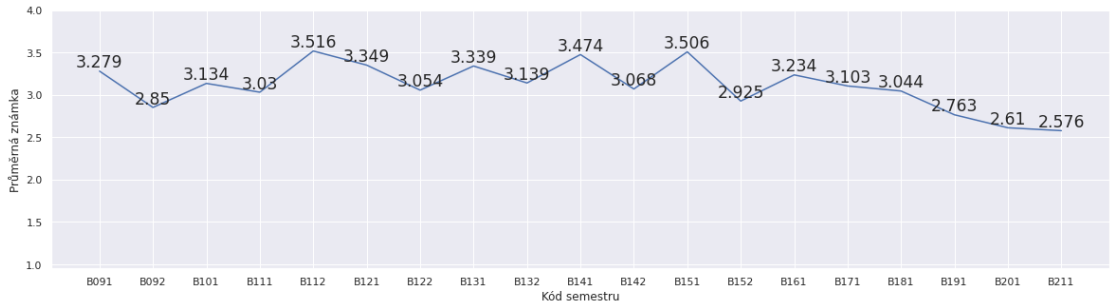
Tabulka 3.3 Vzájemná uznatelnost PP předmětů nové a staré akreditace

Vzájemná uznatelnost		**Uznání starého na nový**		**Uznání nového na starý**	
Starý	Nový	Starý	Nový	Starý	Nový
BI-BEZ	BI-KAB.21	BI-LIN	BI-LA1.21	BI-LIN	BI-LA1.21 + BI-LA2.21
BI-CAO	BI-TZP.21	BI-ZMA	BI-MA1.21	BI-ZMA	BI-MA1.21 + BI-MA2.21
BI-DPR	BI-TDP.21	BI-MLO + BI-ZDM	BI-DML.21	BI-ZDM	BI-DML.21 + BI-MA2.21
BI-EMP	BI-EPP.21			BI-MLO	BI-DML.21 + BI-LOG.21
BI-PS1	BI-UOS.21				
BI-SI1.2	BI-SWI.21				

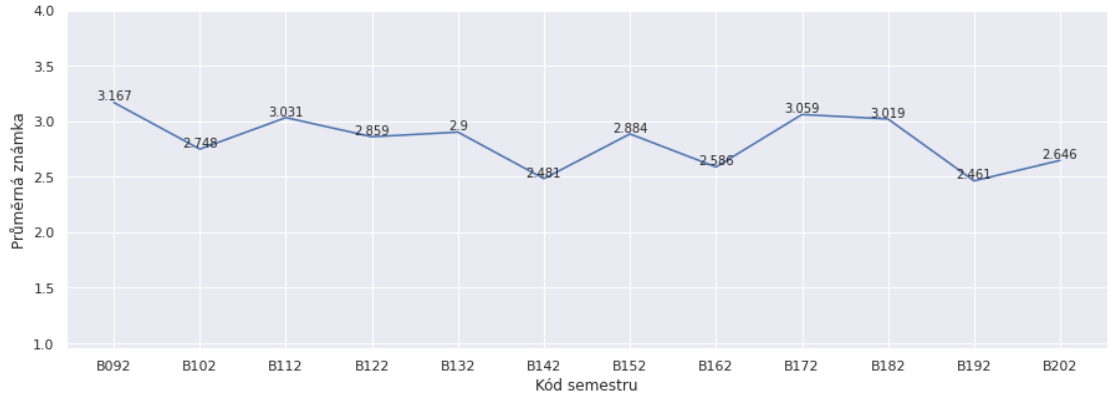
Tak jak dochází ke změnám v předmětech změnou akreditace, dochází i k úpravám osnov učení v předmětech v průběhu jedné akreditace. Garanti vyučovaných předmětů tak v průběhu let reflektují nové poznatky v daném oboru, ale i zpětnou vazbu kantorů a studentů. To jde ovšem ruku v ruce s nekonstantní průchodností a průměrem známek v předmětech v průběhu let. Pěkným příkladem tohoto jevu může být předmět BI-PS1. Na grafu 3.2 si můžeme povšimnout výrazných změn v průměrných známkách předmětu, zejména pak je vidět poměrně strmé zlepšení průměru počínaje semestrem B171. V tomto semestru se začal používat systém *Learnshell* na psaní průběžných malých testů, na kterých si mohli studenti nasbírat cenné body. Další zlepšení je vidět mezi semestry B181 a B191, v semestru B191 došlo k přesunu na platformu *Learnshell* i u zápočtových testů z původní platformy *Progtest*. Až do akademického roku 2016/2017 byl předmět rovněž vyučován také v semestrech letních. Ve většině případů si můžeme povšimnout lepších výsledků v letních semestrech. V letních semestrech často předmět studovali studenti, kteří předmět opakovali ze zimního semestru. Dalším faktorem, který mohl vývoj průměru ovlivnit, byla změna garanta předmětu a jeho jiné představy o obtížnosti předmětu.

Také předmět BI-OSY, jak je vidět na grafu 3.3, se potýká s relativně vysokými výkyvy v průměrných známkách předmětu. Zde si můžeme povšimnout zlepšení průměru v semestru B192, jedná se o semestr, ve kterém došlo k přerušení prezenční výuky kvůli pandemii způsobené virem COVID-19 – dopadu distanční výuky se více věnuje sekce 3.4.

Jak je vidět z výše shrnutých poznatků, proměnlivost vyučovaných předmětů musí být brána v rámci návrhu datasetů pro predikování v potaz a měla by hrát důležitou roli při výběrech



Obrázek 3.2 Vývoj známek v předmětu BI-PS1



Obrázek 3.3 Vývoj známek v předmětu BI-OSY

ročníků, které budou do datasetů zahrnuty.

3.4 COVID a jeho dopad

Dne 10. března 2020 došlo na ČVUT k pozastavení veškeré prezenční výuky kvůli rostoucímu množství lidí nakažených virem COVID-19[3]. ČVUT patřilo nejen mezi první vysoké školy ale i mezi první školy v celé České republice, které učinily tento krok. Znovuzavedení prezenční výuky proběhlo až po téměř roce a půl výuky v online režimu. Studenti, učitelé i vedení školy se prakticky ze dne na den ocitli v naprosto bezprecedentní situaci.

Distanční výuka prošla během covidového období obrovskou změnou. Musíme si uvědomit, že do této doby probíhala veškerá výuka na fakultě. Učební materiály přístupné online byly pouze v podobě přednášek či studijních textů vystavených na fakultních webech. Výuka v prvních měsících byla velice nepředvídatelná, nikdo nevěděl, kdy opatření bránící prezenční výuce skončí. Nejdříve se předpokládalo, že půjde pouze o několik týdnů doma, během kterých budou studenti dostávat pár úkolů, jež budou vypracovávat samostatně a pak se zase situace vrátí do starých kolejí prezenční výuky. Distanční výuka tedy v naprosté většině předmětů prakticky neexistovala a šlo pouze o samostudium.

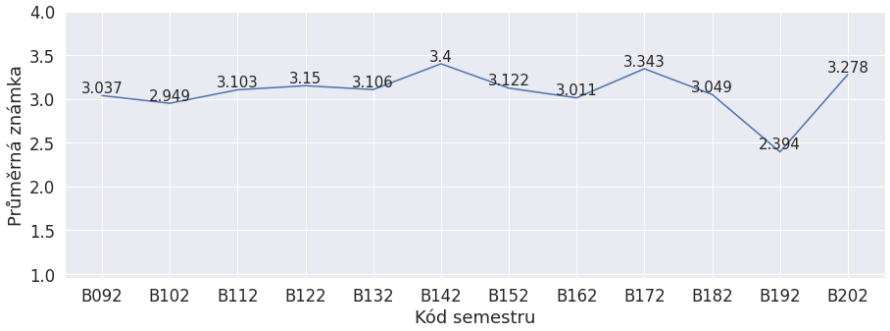
S ubíhajícími týdny stále více a více předmětů přecházelo k alespoň nějaké formě online výuky, která měla podpořit studenty při studiu doma. Některé předměty na tom byly lépe a jejich učitelé

[3]Podle Dodatku č. 1 k příkazu rektora č. 7/2020.
https://www.muvs.cvut.cz/wp-content/uploads/2020/03/PR_2020x07_d1.pdf/

začali poskytovat streamy a videozáznamy k přednáškám a cvičením, podpůrná videa, rozšiřující materiály a zintenzivněli komunikaci se studenty přes studijní portály. Jiné předměty na tom byly podstatně hůře, jejich přerod do online formy byl velmi bolestivý jak pro studenty, tak pro učitele. Celkově se tedy jednalo o velice různorodý přechod, který je velmi těžce uchopitelný. Toto rozporuplné období můžeme datovat zhruba do konce školního roku 2019/2020.

V dalším školním roce již setrvání v distančním studiu nebylo pro všechny takovým šokem. V zimním semestru byla již fakulta schopna zařídit online přenosy přednášek pro většinu předmětů. Samotní kantoři měli také mnohem lepší představu o tom, jak poskládat učivo pro online výuku. A tak můžeme říct, že ač bylo distanční studium pro všechny stále velice náročné, úroveň online výuky se oproti předchozímu semestru podstatně zlepšila.

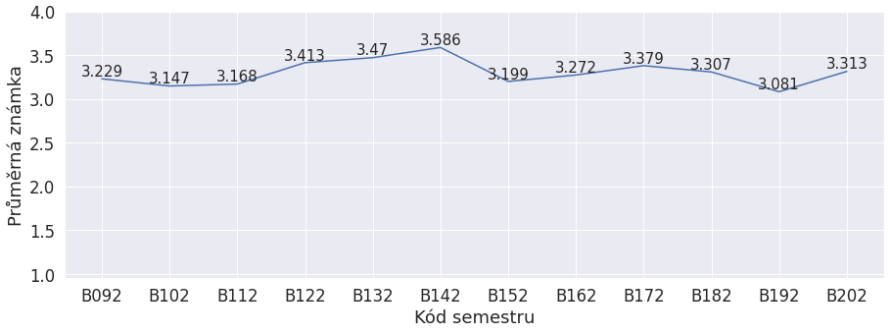
Nabízí se tedy otázka, jak se distanční výuka projevila na výsledcích studentů. Ukažme si tedy pár příkladů. Na grafu 3.4 je pěkně vidět znatelné zlepšení známek z předmětu BI-PA2 v letním semestru B192, jedná se právě o první online semestr v akademickém roce 2019/2020. Průměrná známka se zlepšila ze známky E ($3.0 \leq \mathrm{E} < 4.0$) na známku C ($2.0 \leq \mathrm{C} < 2.5$). V letním semestru 2020/2021 – B202 již vidíme návrat k původní průměrné hodnotě známek. Předmět BI-PA2 byl jeden z prvních předmětů, který přešel na kvalitní plně online výuku. Studenti měli přístup ke profesionálním záznamům veškerých přednášek, proseminářů i cvičení. Za výrazným zlepšením průměru však pravděpodobně také stojí změna způsobu konání a hodnocení zkoušek, které musely být upraveny, aby je studenti mohli skládat ze svých domovů. Garant předmětu BI-PA2 se rozhodl zrušit nepříjemnou závěrečnou programovací zkoušku, kterou nahradily body ze semestru a semestrální práce.



Obrázek 3.4 Vývoj známek v předmětu BI-PA2

Stejná tendence je vidět i u předmětu BI-LIN na grafu 3.5. Zlepšení není tak drastické jako u BI-PA2, ale stále znatelné. V předmětu BI-LIN se přednášky natáčely i před pandemií, a tak k nim měli studenti přístup od začátku uzavření fakulty. Zároveň byly vypisovány i letní termíny zkoušek, kterých mohli studenti, kterým přes online semestr tzv. „ujel vlak“, využít.

Tabulka 3.4 obsahuje průchodnost povinných předmětů programu. Zaměřme se na srovnání průchodnosti předmětů vyučovaných v letním semestru prvního ročníku studia (tedy BI-PA2, BI-DBS, BI-SAP, BI-LIN) v akademických letech 2019/2020 (první semestr v distančním režimu) a 2018/2019 (letní semestr byl ještě v prezenční formě). Vidíme, že až na předmět BI-SAP byla ve všech předmětech vyšší průchodnost. Na první pohled se může tato skutečnost jevit v souvislosti se situací jako nečekaná. Možných vysvětlení je však několik. Vzhledem k nevídané situaci byli kantoři postaveni před rozhodnutí, jak řešit hodnocení v semestru a zakončení předmětu. Garanti některých předmětů (jako tomu bylo ku příkladu v předmětu BI-PA2) se uchýlili k úplnému zrušení zkoušek, jiní zase byli situací nuceni změnit formát zkoušek a zrušit některé jejich části. V průběhu semestru byly často (kvůli prvotnímu úplnému přerušení výuky) odpuštěny či jinak zjednodušeny semestrální testy, práce či jinak běžné úkoly v hodinách. Celkově se kantoři vzhledem k situaci snažili být ke studentům shovívaví. Zároveň došlo ke snížení množství kreditů



Obrázek 3.5 Vývoj známek v předmětu BI-LIN

potřebných k průchodu studiem. Došlo také k posunutí termínu státních závěrečných zkoušek a odevzdání závěrečných prací.

Jak můžeme vidět v tabulce 3.4 následujícího akademického roku 2020/2021, se situace s průchodností v povinných předmětech mění a není již zdaleka tak jednoznačná jako v předchozím letním semestru. Jedná se o rok, který se kompletně odehrál v distančním režimu. Pěkným příkladem může být předmět BI-PA1 – ač můžeme vidět poměrně vysoký výkyv v jeho průchodnosti i mezi roky 2018/2019 a 2019/2020, kdy se ještě v distančním režimu nevyučoval (téměř 8,5 procentních bodů) – propad průchodnosti se v roce 2020/2021 oproti předchozímu roku blíží k 16 procentním bodům. Jedná se předmět, který je vyučován v prvním semestru. Je tedy možné, že byl propad v jeho průchodnosti částečně způsoben tím, že studenti prvního ročníku byli vystaveni velkému tlaku online výuky bez předchozí znalosti vysokoškolského studia. V menší míře je propad v průchodnosti předmětů pro první ročník vyučovaných v zimním semestru také vidět v předmětech BI-ZMA a BI-MLO. Co se týče ostatních předmětů, velký propad vidíme také v předmětech BI-PA2, BI-LIN, BI-AG1, či BI-OSY. Naopak ku příkladu v předmětu BI-PS1 vidíme poměrné vysoké zvýšení průchodnosti (téměř 10 procentních bodů), jak jsme si ale mohli povšimnout v sekci 3.3), předmět BI-PS1 již dlouhodobě zvyšuje svou průchodnost a zlepšuje se průměr známek v předmětu.

Tabulka 3.4 Průchodnost PP předmětů v procentech

Předměty	2018/2019	2019/2020	2020/2021
BI-PA1	43,2	51,6	36,1
BI-PAI	76,9	82,1	85,0
BI-CAO	78,8	80,2	83,9
BI-PS1	61,2	62,9	73,0
BI-MLO	53,6	57,0	52,0
BI-ZMA	38,7	42,8	37,6
BI-PA2	47,9	68,6	36,3
BI-DBS	76,2	80,5	72,5
BI-SAP	71,6	68,3	68,0
BI-LIN	41,4	51,1	39,1
BI-AG1	61,7	54,0	49,4
BI-AAG	63,8	48,7	56,6
BI-ZDM	62,2	52,5	51,1
BI-OSY	67,3	84,2	74,8
BI-PSI	73,0	78,6	72,8
BI-BEZ	74,4	77,8	78,2
BI-PST	88,3	81,6	86,1
BI-DPR	91,9	86,8	82,5
BI-SI1.2	83,4	90,6	86,3
BI-EMP	84,1	88,7	82,6

Část II

Praktická část

Kapitola 4
Porozumění datům

Data, která FIT udržuje o svých studentech, bohužel nedisponují takovým množstvím příznaků, jako tomu bylo ve studiích zmíněných v sekci 2.5. Data, která jsem měla k dispozici, obsahovala pouze základní informace o studentech a zároveň byla velmi nedobré kvality, jak bude podrobně popsáno v této kapitole. Hlavní zdroje, ze kterých bylo čerpáno při zjišťování významu jednotlivých atributů dat, byla diplomová práce Ing. Elišky Hrubé z roku 2014 [8] a diplomová práce Ing. Magdy Friedjungové, Ph.D. z roku 2016 [34].

4.1 Datové zdroje

V rámci celého ČVUT existuje celá řada systémů a aplikací, které udržují univerzitní data – jejich využívání je specifické pro každou konkrétní fakultu. Na půdě Fakulty informačních technologií existuje datový sklad ČVUT. Pro účely bakalářské práce mi byly k dispozici API endpointy datového skladu ČVUT, ze kterých byla data stahována. Primárním zdrojem byla data ze systému KOS. Datový sklad aktualizuje data z KOSu pravidelně každý týden, veškerá další práce s daty je tvořena na datech ke dni 17. 4. 2022.

4.1.1 Datový sklad ČVUT

Datový sklad ČVUT (DWH ČVUT) vznikl pod vedením Ing. Michala Valenty, Ph.D. v rámci rozvojových projektů. Jeho úkolem je integrace dat z několika různých univerzitních systémů, jako hlavní datový zdroj slouží Komponenta studium ČVUT .

4.1.2 KOS

Komponenta studium ČVUT, zkráceně KOS, je celouniverzitní systém, který shromažďuje studijní výsledky studentů, studenti si jeho prostřednictvím zapisují předměty a tvoří rozvrhy, přihlašují se na zkoušky a zápočtové testy, taktéž zde mají přístup k informacím o svých studijních výsledcích včetně záznamů klasifikace. KOS mimo jiné uchovává informace z celouniverzitní aplikace Přihláška ČVUT, ta slouží zájemcům o studium k vyplnění elektronické přihlášky. Tato bakalářská práce je postavena právě na datech dostupných ze systému KOS.

4.2 Co víme o studentech FIT

Data jako taková jsou tvořeny jednotlivými datamarty, které se dále dělí na dimenzionální a faktové tabulky. Dimenzionální tabulky obsahují popisná data, faktové tabulky napočítané metriky [35]. Dimenzionální tabulky jsou v názvu zakončeny koncovkou *dim*, k dispozici pro účely práce byly tyto dimenzionální tabulky:

- **predmet_dim** – udržuje informace o předmětech, které si studenti mohou zapsat,
- **prihlaska_dim** – udržuje informace související se přijímacím procesem studentů,
- **semestr_dim** – obsahuje seznam semestrů,
- **student_dim** – obsahuje osobní informace o studentech,
- **studium_dim** – shromažďuje informace týkající studia studentů,
- **szzk_dim** – obsahuje záznamy státních závěrečných zkoušek,
- **zaverecne_prace_dilci_hodnoceni_dim** – obsahuje dílčí hodnocení závěrečných prací.

Faktové tabulky jsou zakončeny koncovkou *fact*, pro účely této práce byla přístupná data z těchto tabulek:

- **klasifikace_fact** – obsahuje záznamy klasifikace,
- **zaverecne_prace_fact** – obsahuje konkrétnější hodnocení závěrečných prací.

Jako hlavní zdroj dat pak posloužily zejména tabulky: *prihlaska_dim*, *student_dim*, *studium_dim* a *klasifikace_fact*. Blíže se s jejich obsahem seznámíme v následujících sekcích.

4.3 Informace o předmětech

Informace o předmětech, které mohou studenti absolvovat udržuje tabulka *predmet_dim*. V současné chvíli tabulka obsahuje 1 225 unikátních kódů předmětů. Nejedná se pouze o předměty, které jsou vyučovány na naší fakultě, ale obsahuje taktéž předměty, které si mohou studenti zapisovat z ostatních fakult ČVUT.

4.3.1 Problémy s tabulkou predmet_dim

První problém spočíval v absenci sloupce, který uchovává doporučený semestr absolvování PP předmětů. Tento sloupec nebyl v původních datech obsažen a byl přidán až po konzultacích s pracovníky datového skladu v závěru práce. Z tohoto důvodu bylo nutné využít informací o doporučených průchodech studiem z Bílé knihy[1].

Zároveň prvotní formát dat byl velmi nekonzistentní, označení typů předmětů a jejich zakončení nedržel jednotný formát a práce s daty tak byla poměrně obtížná. Také z tohoto důvodu byla veškerá data čerpána z Bílé knihy pomocí metod *web scrapingu* a tabulka *predmet_dim* byla využita pouze k překladu *predmet_id* (číselný identifikátor předmětů), pomocí kterého jsou předměty rozlišovány v záznamech klasifikace, na *kod_predmetu* (kód předmětu popsaný v sekci 3.1.2).

[1] https://bk.fit.cvut.cz/cz/prehled.html/

4.4 Informace z přihlášky

Fakulta informačních technologií udržuje o zájemcích o studium poměrně širokou škálu informací. Tabulka, která integruje data týkající se přihlášek ke studiu a informací o přijímacím řízení se jmenuje *prihlaska_dim*. Tabulka *prihlaska_dim* v sobě udržuje zároveň:

- **Informace z přihlášky** – Data, která pochází z aplikace Přihláška ČVUT. Samotní uchazeči o studium vyplní patřičnou přihlášku, data z ní jsou využity pro účely přijímacího řízení a v případě jejich přijetí jsou informace z přihlášky uchovávány v systému KOS.
- **Informace z KOSu** – Záznamy o předchozích studiích uchazeče na ČVUT, zejména pokud se student rozhodne pokračovat z bakalářského na magisterské studium a chce využít možnosti přijetí na základě průměru známek z bakalářského studia.
- **Informace z průběhu přijímacího řízení** – Mezi takové patří ku příkladu počet bodů, které uchazeč získal u přijímací zkoušky, percentil získaný v případě přijetí na základě SCIO testů a jiné.

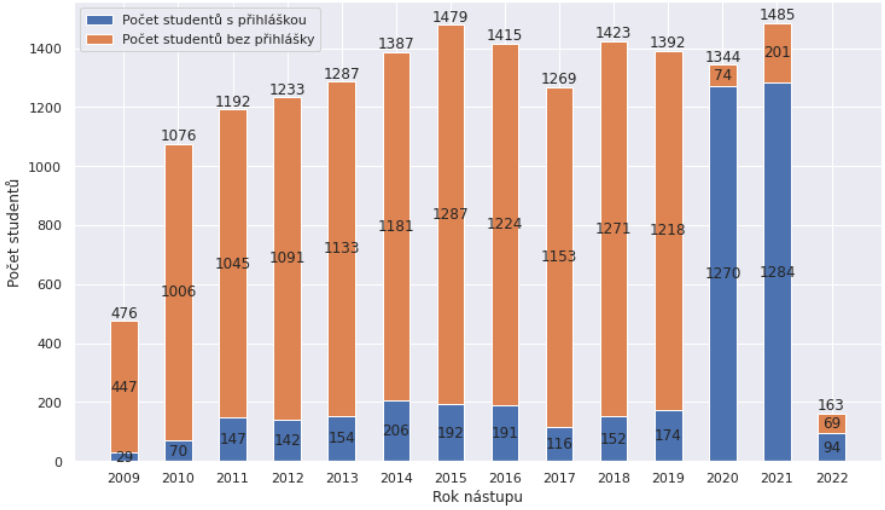
Tabulka obsahující kompletní seznam všech atributů tabulky *prihlaska_dim* a jejich vyplněnost je dostupná v příloze A. Jednotlivé atributy můžeme rozdělit do sedmi kategorií:

- **Osobní údaje uchazeče** – Veškeré atributy, které obsahují osobní informace o uchazeči o studium. Zde můžeme jmenovat ku příkladu informace o trvalém bydlišti uchazeče, tj. město, okres, země, PSČ, zda má uchazeč nárok na stipendium, zda má zažádáno o kolej, informace o jeho případných studijních indispozicích a jiné.
- **Typ střední školy** – Atributy z této skupiny shrnují informace o střední škole, ze které se daný uchazeč hlásí. Mezi takové informace můžeme zařadit IZO a REDIZO kódy střední školy, typ střední školy, či obor střední školy.
- **Prospěch na střední škole** – Atributy obsahující známky uchazeče z vybraných předmětů na střední škole, známky z maturity, typ maturity.
- **Způsob přijetí** – Atributy, které popisují vybraný způsob přijetí uchazeče a jeho výsledky, společně s ostatními doplňujícími údaji (např. datum uskutečnění přijímací zkoušky, místnost a jiné).
- **Předchozí vzdělání** – Atributy, které podávají informace o stupni předchozího vzdělání, v případě, že se uchazeč hlásí z vysoké školy, vyplňuje do přihlášky informace o předchozím vzdělání, či průměr na předchozím bakalářském studiu, pokud žádá o přijetí na základě průměru.
- **Budoucí studium na fakultě** – Patří sem informace o tom, zda-li se uchazeč zapsal ke studiu, studijní skupině, zda je uchazeč nově přijatý, navazující studijní program a jiné.
- **Ostatní** – Obsahuje atributy, které nepatří do žádné výše zmíněné skupiny včetně atributů, které představují identifikátory uchazeče.

4.4.1 Problémy s daty z přihlášky

Bohužel v současné chvíli existuje několik poměrně závažných problémů s daty z přihlášek. V první řadě DWH ČVUT momentálně neeviduje veškeré záznamy přihlášek, k dispozici jsem měla konkrétně 5 081 záznamů – přihláška tedy byla dostupná k pouze 30 % studentů. Na grafu 4.1 si můžeme povšimnout, že jsou evidovány záznamy přihlášek zejména ke studentům s rokem nástupu 2020 a 2021. Problém je, že u studentů s tímto rokem nástupu nejsme schopni (pro

účely predikce) plně rozdělit studenty na studenty úspěšné a neúspěšné (v případě bakalářského programu tito studenti ještě neměli šanci dokončit standardní délku studia 3 roky, v případě magisterského studia máme informace pouze o studentech s rokem nástupu 2020 se standardní dobou studia 2 roky). Na tomto místě by bylo dobré zmínit, že počet přihlášek, které datový sklad udržuje se v průběhu vypracovávání mé bakalářské práce několikrát měnil. Na začátku práce jsem měla k dispozici 3 599 záznamů přihlášek, v průběhu práce většina těchto záznamů zmizela a dostupných bylo pouze 323 záznamů. Po konzultaci s pracovníky datového skladu jsem obdržela finální počet 5 081 záznamů.



Obrázek 4.1 Rozložení záznamů přihlášek podle roku nástupu studentů

Dalším problémem je fakt, že aplikace Přihláška ČVUT neudržuje po celou dobu své existence konstantní množství atributů (sloupců). Jak ve své diplomové práci uvádí Ing. Eliška Hrubá [8], ku příkladu v prvních dvou letech existence Fakulty informačních technologií, tedy v letech 2009/2010 a 2010/2011, bylo o uchazečích o studium udržováno 177 atributů. V akademickém roku 2011/2012 počet atributů stoupl na 207. O počtu atributů v akademickém roce 2012/2013 kvůli problémům v datech (nebylo zde vyplněno rodné číslo uchazečů, které sloužilo jako identifikátor) nemáme dostupné informace. Následujícího roku 2013/2014 se o uchazečích uchovávalo již 224 atributů. O konkrétním počtu atributů z dalších let bohužel nemám dostupné informace, neboť přístup k datům, který mi byl poskytnut obsahuje pouze sadu dat v aktuálním formátu. Momentálně *prihlaska_dim* obsahuje 97 atributů. Tak jak se měnil v průběhu let počet udržovaných atributů, měnil se i formát dat a pojmenování jednotlivých atributů. Stejně tak i metodika vyplňování patřičných atributů nebyla v průběhu let jednotná.

Velice těžko řešitelným problémem s daty z *prihlaska_dim* je fakt, že tím, jak se v průběhu let měnily podmínky pro přijetí uchazečů, měnila se i množina údajů, které musely být povinně vyplněny. Dobrým příkladem zde může být povinnost vyplnění známek ze střední školy. V prvních letech bylo povinné doložit známky z matematiky, studijní průměr a známky z maturity, to se ale změnilo, a tak počet záznamů s vyplněnými sloupci, které se týkají známek ze střední školy, je prakticky zanedbatelný. Konkrétně se o uchazeči můžou udržovat známky až z pěti předmětů. Nejčastěji jsou vyplněny známky maximálně ze dvou předmětů (to o jaké předměty se jedná se různí u každého uchazeče). Z celkem 5 081 záznamů jsou známky ze dvou předmětů udržovány u 61 studentů, známky z jednoho předmětu u 65 uchazečů, známky ze tří a čtyř předmětů má pouze jeden záznam, známky z pěti předmětů nemá ani jeden záznam.

Zmiňme ještě v krátkosti problematiku nejednoznačného a špatně interpretovatelného vyplňování atributů tabulky. Za prvé u řady atributů aplikace přihláška nekontroluje správně formát

vyplněných dat. To se dotýká zejména atributů o střední škole. V případě, že se uchazeč hlásí do bakalářského programu, je vyzván k tomu, aby vyplnil IZO své střední školy. Neděje se tak ale prostřednictvím nabídky, ze které uchazeč vybere patřičný IZO kód, ale kód musí být zadán ručně uchazečem. Zároveň správnost kódu není nijak ověřována, takže velice často je místo IZO kódu vyplněno např. REDIZO střední školy, IČO či jiné hodnoty. Zejména se problém týká uchazečů, kteří nepochází z České republiky, zde jsou často místo IZO kódů vyplněny výchozí hodnoty a to i přesto, že střední školy v zemi původu uchazeče IZO kódy evidují (ku příkladu Slovensko). Přitom právě IZO kód je hlavním zdrojem informace o typu střední školy, neboť je také obsažen v tabulce *studium_dim*, která má mnohem lepší poměr vyplněnosti svých atributů nežli *prihlaska_dim*, jak píši v sekci 4.5 a zároveň v ní existuje záznam ke každému studentovi. Dále pak existují atributy jako *predchozi_studium*, které jsou ukládány jako texty, ke kterým neexistuje žádný předem daný formát ani povinný obsah textu. Texty píší samotní uchazeči, jejich automatická zpracovatelnost by tedy byla velice obtížná.

Otázkou tedy zůstává, zda informace získané z přihlášek, můžeme pro účely predikce využít, tak abychom zvýšili přesnost predikcí. S touto problematikou se důkladně seznámíme v kapitole 5.

4.5 Informace o studentech a jejich studiu

Celkem bylo k dispozici 16 605 záznamů o jednotlivých studentech na fakultě a 16 621 záznamů různých započatých studií. Je evidováno 9 401 započatých studií bakalářského programu Informatika, prezenční formy v českém jazyce (zkráceně BICS) a 2 703 započatých studií magisterského programu Informatika, prezenční formy v českém jazyce (zkráceně MICS). Informace týkající se studentů a jejich studiu poskytují tabulky *studium_dim* 4.1 a *student_dim* 4.2. Záznamy z obou tabulek jsou propojeny přes společný identifikátor *studium_id*.

Díky identifikátoru *peridno* z tabulky *student_dim*, který identifikuje unikátní studenty, můžeme zjistit, že na fakultě studovalo 8 217 unikátních studentů prezenční formy v češtině – 8 150 na BICS a 2 453 na MICS. Někteří z nich však fakultu nedokončili a přihlásili se znovu, takových studentů je 1 280 – 1 064 na BICS a 216 na MICS. Rekordmanem je student, který na fakultě studoval celkem šestkrát na BICS. Nejčastěji ale studenti opakují pouze jednou, konkrétně 909 studentů u BICS a 184 u MICS. Existuje několik možností, jak tento problém řešit. Jedna cesta je počítat pouze s unikátními studenty a jejich výsledky z let minulých nebrat v potaz, tak bychom ale přišli o informace o jejich neúspěšných studiích. Další možností je počítat pouze s výsledky z prvního studia „studentů opakovačů“ na fakultě a výsledky z dalších studií nebrat v potaz, neboť již mohou být ovlivněny předchozím studiem. V takovém případě bychom se ale připravili o možnost predikování výsledků při jejich dalších studiích. Cesta, která byla zvolena v této práci je kompromisem předchozích dvou možností. Spočívá v tom, že je každé započaté studium bráno samostatně, jakoby se jednalo o jedinečného studenta. Na nepovedená studia z předchozích let je tak nahlíženo jako na neúspěšné studenty. Na případné konečné úspěšné studium je nahlíženo jako na unikátního úspěšného studenta. Pokud si daný student nechá uznat známky z předchozího studia, jsou zapsány do semestru *A000*, případně *A00*. Ztrácíme tak sice informaci o tom, v jakých semestrech student předměty absolvoval, případně na jaký pokus, v opačném případě bychom ale museli míchat dohromady více různých studií.

Další důležité rozhodnutí spočívalo ve vybrání skupin studentů, kterými se práce bude zabývat. Po konzultaci s vedoucí bakalářské práce byli pro průzkum dat vybráni pouze studenti BICS a studenti MICS. Po samotném průzkumu dat bylo rovněž rozhodnuto, že prediktivní modelovaní bude zaměřeno pouze na studenty BICS z důvodu malého množství dat již dostudovaných studentů MICS, jejichž rok nástupu spadá do akreditace 2016, popřípadě nové akreditace 2020 (pouze 895 takových studentů). Předchozí akreditace před rokem 2016 byla velmi odlišná, co do skladby povinných předmětů a jejich doporučených průchodů, jak podrobně popisuji v sekci 3.1.3, tak do samotné vyučované látky v nich.

Tabulka 4.1 Atributy tabulky studium_dim

Název atributu	Vyplněnost v procentech
studium_id	100
typ_programu	100
forma_studia	83,70
datum_zahajeni	100
datum_ukonceni	86,40
rocnik	99,92
studijni_stav	100
studuje	100
zamereni_id	0,02
fakulta_id	100
nazev_fakulty	100
jazyk_vyuky	96,70
ukonceni_zpusob	86,40
platby_stav	53,14
pruchod	3,87
rok_maturity	71,41
odkud_skola_kod	65,96

Tabulka 4.2 Atributy tabulky student_dim

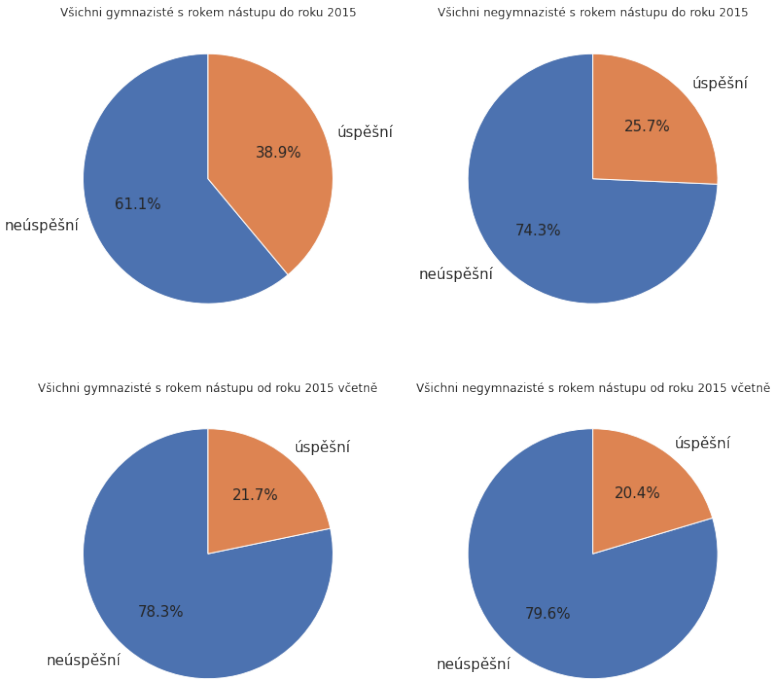
Název atributu	Vyplněnost v procentech
studium_id	100
peridno	100
pohlavi_kod	98,39
datum_narozeni	98,82
misto_narozeni	97,53
statni_prislusnost_nazev	98,75
stat_narozeni_nazev	80,37
typ_adresy	100
kod_obce	74,06
kod_okresu	75,74
kod_zeme	100
psc	76,22

4.5.1 Vliv střední školy

Jak je vidět z grafu 4.2, to jestli student absolvoval gymnázium mělo poměrně velký vliv ve dřívějších letech, zde konkrétně vidíme srovnání studentů s rokem nástupu před rokem 2015 a po roku 2015 včetně, ve kterém proběhla akreditace. Studenti, kteří nastoupili do studia před rokem 2015 a absolvovali gymnázium, byli úspěšní až ve 38,9 % případů. Naopak studenti z ostatních středních škol byli úspěšní pouze ve 25,7 % případů. Můžeme si povšimnout, že úspěšnost studentů s rokem nástupu od roku 2015 včetně je již poměrně vyrovnaná mezi gymnazisty a absolventy ostatních středních škol.

Tyto výsledky ale nemusí být zcela přesné. Informace o tom, na jaké střední škole student studoval, je totiž zjištěna primárně pomocí IZO kódů. V tabulce *prihlaska_dim* sice existují další atributy popisující typ střední školy, ale jak popisuji v sekci 4.4.1, DWH ČVUT zaznamenává velmi malé množství přihlášek a zároveň i ty, které eviduje, nejsou kompletně vyplněny. Ani vyplnění IZO kódů však není kompletní, jak je popsáno v sekci 4.4.1, a dalším problémem je

fakt, že MŠMT neposkytuje strojově zpracovatelná data, ve kterých by šlo podle IZO kódů jednoduše dohledat typ střední školy. Jediná možnost tak byla využít tabulky středních škol[2] a pomocí metod *web scrapingu* získat IZO kódy škol, které v názvu obsahují slovo gymnázium a slova od něj odvozená. To šlo jednoduše udělat pouze u gymnázií, které ve svých názvech slovo gymnázium používají velmi často kvůli prestiži. U ostatních typů středních škol (SOŠ a SOU) tomu tak nebylo, neboť s novým školním zákonem[3] od roku 2004 není povinnost mít v názvu typ školy. Zároveň existují školy, které pod jedním IZO kódem evidují více různých typů středních škol, v takovém případě byla škola brána jako gymnázium, i pokud obsahovala i ostatní typy středních škol.



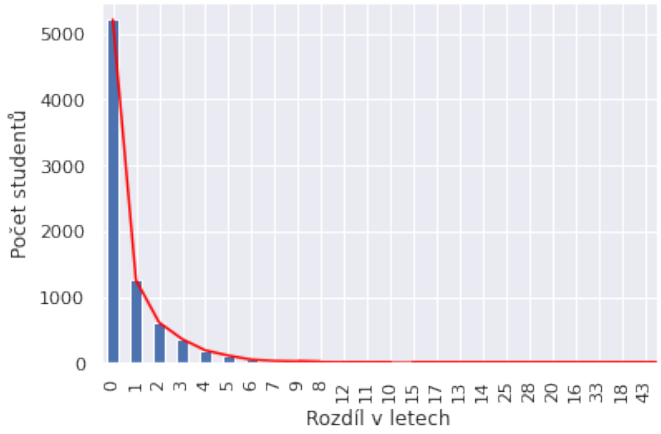
Obrázek 4.2 Vliv absolvování gymnázia na úspěch ve studiu – BICS

4.5.2 Vliv rozdílu mezi rokem nástupu a rokem maturity

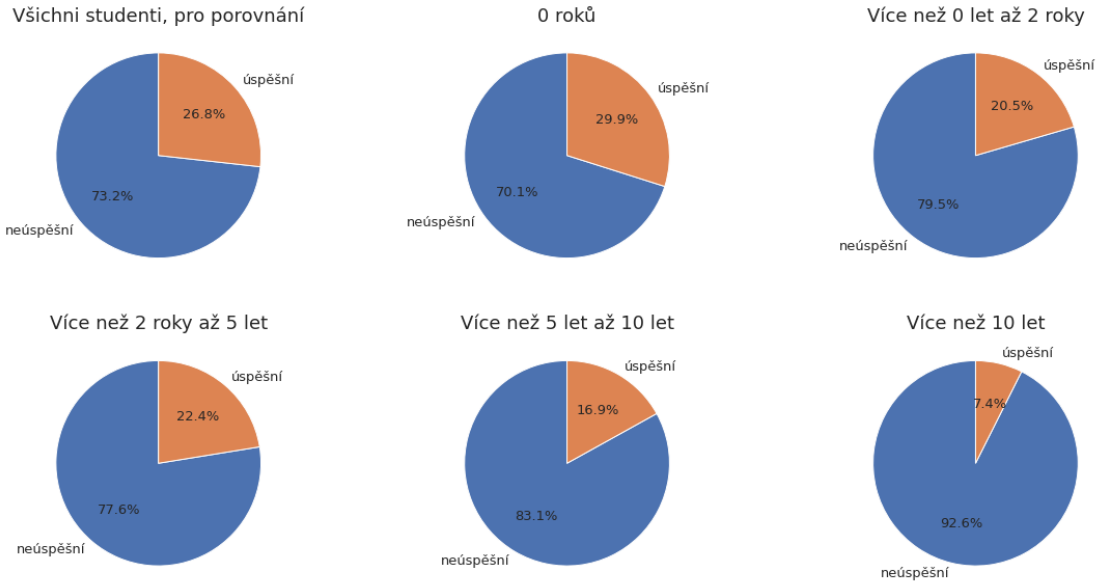
Nejčastěji studenti zahajují své studium na FIT ve stejný rok, ve které složí maturitní zkoušky. Konkrétně se jedná o 66 % všech dostudovaných studentů BICS od založení fakulty, 82 % studentů nastoupí na fakultu do jednoho roku a 90 % studentů do dvou let. Graf 4.3 zobrazuje klesající tendenci počtu studentů v závislosti na rozdílu mezi rokem nástupu a rokem složení maturitních zkoušek.

Graf 4.4 indikuje vliv rozdílu mezi rokem nástupu a rokem maturity u studentů BICS. Nejvyšší úspěšnost mají studenti, kteří nastoupí na fakultu ve stejném roce, ve kterém maturovali – 29,9 %. U studentů, kteří mají rok až dva mezeru, klesá úspěšnost o 9,4 procentních bodů a úspěšně dokončí studium pouze necelých 21 % z nich. Výrazné snížení úspěšnosti můžeme vidět také u studentů s rozdílem pět až deset let, jejich studijní úspěšnost se pohybuje pouze kolem necelých 17 %. Studenti s vyšším rozdílem již dostudují úspěšně pouze v 7,4 % případů.

[2] https://rejstriky.msmt.cz/rejskol/
[3] https://ppropo.mpsv.cz/zakon_561_2004/



Obrázek 4.3 Četnost rozdílů mezi rokem maturity a rokem nástupu v letech – BICS



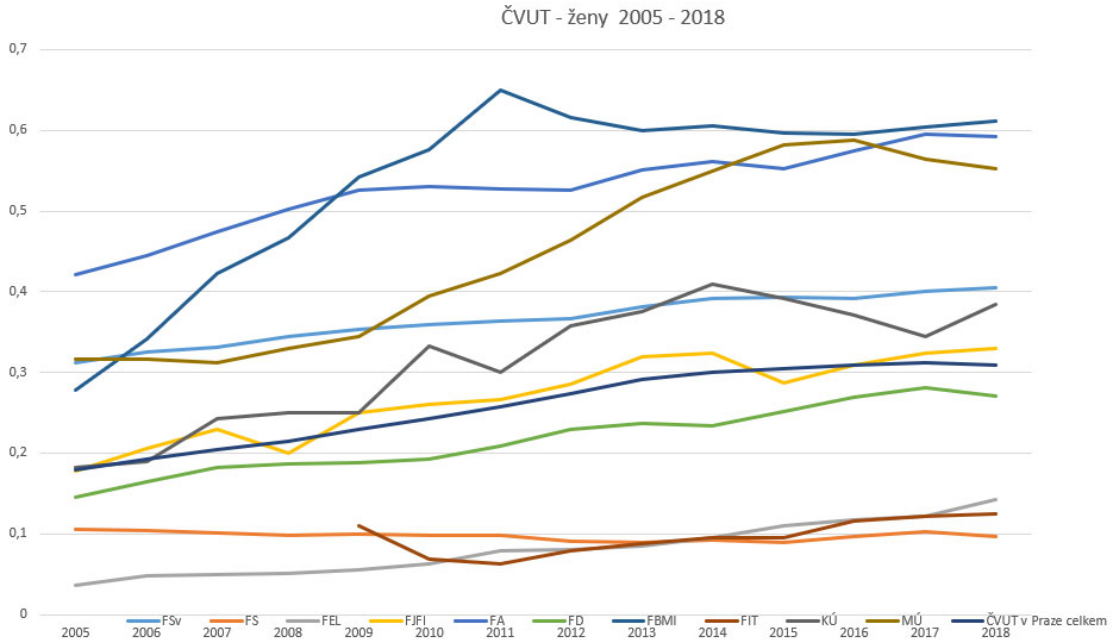
Obrázek 4.4 Úspěšnost studentů v souvislosti s rozdílem mezi rokem maturity a nástupu do studia – BICS

4.5.3 Vliv pohlaví

Dlouhodobým trendem v České republice je vyšší podíl žen na vysokých školách. K roku 2021 na českých vysokých školách studovalo 56 % žen, což je nárůst zhruba o osm procentních bodů oproti roku 2001 [2]. Na technických vysokých školách je tomu ovšem přesně opačně.

Univerzitní časopis Pražská Technika vydal v březnu roku 2019 článek [36] zabývající se poměrem žen a mužů na ČVUT. Článek pozitivně kvituje zvyšování podílu žen na univerzitě. Procentuální zastoupení žen na ČVUT bylo v roce 2007 20 %, k roku 2018 (k němuž se článek vztahoval) se zastoupení žen zvýšilo na 31 %. Nejnovější výroční zpráva ČVUT [37] vztahující se k roku 2020, uvádí zhruba 30,5% zastoupení žen. Nejlépe co do poměru žen je na tom Fakulta

biomedicínského inženýrství, k roku 2018 na ní studovalo 61 % žen. Naopak nejhůře je na tom Fakulta strojní s pouhými 10 % žen. Fakulta informačních technologií obsadila druhé místo od konce s 12% zastoupením žen. Vývoj procentuálního zastoupení žen na fakultách ČVUT od roku 2005 do roku 2018 můžeme vidět na grafu 4.5.



Obrázek 4.5 Procentuální zastoupení žen na fakultách ČVUT v letech

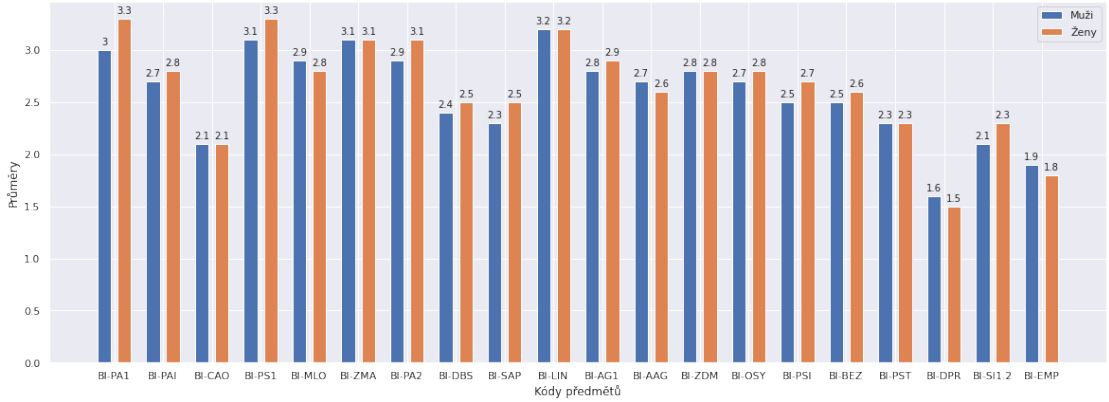
Je tedy vidět, že na naše fakulta na tom není co do podílu žen mezi studenty zrovna nejlépe, nabízí se ale otázka, jakých výkonů dosahují ženy, které se pro studium informatiky přece jen rozhodnou a na fakultu nastoupí. Z dat, které fakulta o svých studentech udržuje, vychází, že pouze zhruba 22 % žen úspěšně dokončí prezenční bakalářský studijní program. Z mužů bakalářské studium úspěšně dokončí zhruba 27 %. Na magisterském programu se výkony mužů a žen více vyrovnávají – z žen dokončí úspěšně studium téměř 57 % a z mužů 60 %. Rozdíl v úspěšnosti tedy není extrémní, přesto je ale znatelný.

Pojďme se nyní podívat na rozdíly v průměru známek mužů a žen ve společných předmětech, ty zachycuje graf 4.6. Na grafu si můžeme povšimnout rozdílu v průměrech zejména u předmětů, které jsou obecně považovány za předměty programovací. Jedná se o předměty BI-PA1 (průměr známek mužů je 3, průměr známek žen je 3,3), dále pak předmět BI-PS1 (muži 3,1 a ženy 3,3) a BI-PA2 (muži 2,9 a ženy 3,1). Mezi další předměty, ve kterých mají muži viditelně lepší průměr, jsou například BI-SAP, BI-PSI, či BI-SI1.2. Ženy jsou naopak o něco lepší v předmětech jako BI-MLO, BI-AAG, BI-DPR nebo BI-EMP. V matematických předmětech jsou průměry vyrovnané.

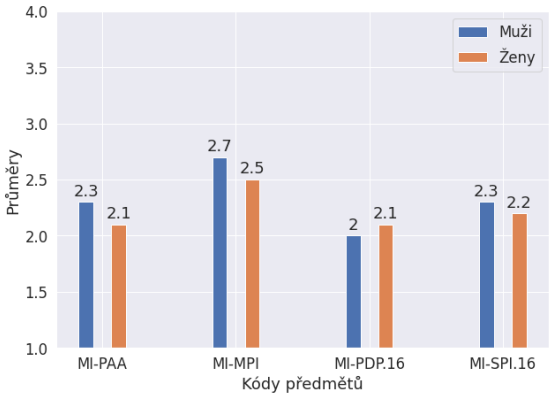
Na grafu 4.7 si můžeme všimnout zajímavého obratu v magisterském programu v prezenční formě. Ve třech ze čtyř povinných předmětů programu mají ženy lepší průměr nežli muži.

4.5.4 Jsou Češi lepší než cizinci?

Studia v českém jazyce na FIT využívá poměrně hodně studentů s trvalým bydlištěm mimo Českou republiku – 19,4 % všech studentů, kteří studují nebo studovali v českém jazyce od založení fakulty. Nejčastěji přijíždějí studovat na fakultu studenti ze Slovenska, kteří tvoří 8,8 % studentů, následují studenti z Ruska (4,6 %) a Ukrajiny (2,4 %). Úspěšnost Čechů v BICS je



Obrázek 4.6 Průměry známek žen a mužů v PP předmětech BICS



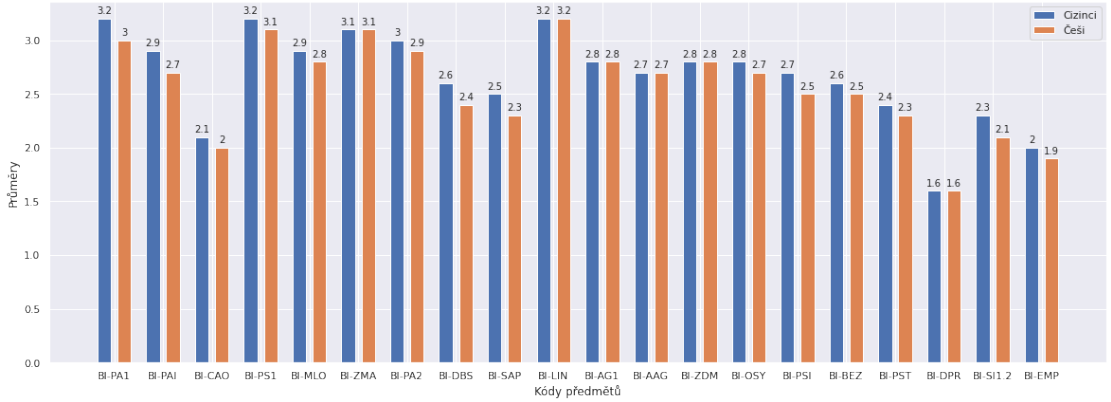
Obrázek 4.7 Průměry známek žen a mužů v PP předmětech MICS

27,2 %, u cizinců to činí 22,1 %. Na magisterském studiu se poté bavíme o 61,7 % u Čechů a 47,9 % u mimočeských studentů. Z výše jmenovaných cizích státních příslušností jsou v bakalářském programu nejúspěšnější studenti ze Slovenska s průchodností 30,1 % a nejméně úspěšní potom studenti s ruskou národností s úspěšností pouhých 15 %. Stejně tak i na magisterském programu jsou nejúspěšnější studenti ze Slovenska – 63,2 % a nejméně úspěšní studenti z Ruska – 27,5 %.

Zaměřme se nyní na rozdíl mezi průměry z povinných společných předmětů. Z grafu 4.8 můžeme vyčíst, že největší rozdíl v průměrech známek je vidět u předmětů BI-PA1, BI-PAI, BI-DBS, BI-SAP, BI-PSI a BI-SI1.2. Zajímavý je také fakt, že čtyři z šesti těchto předmětů mají doporučený rok zapsání v prvním roce studia.

4.6 Klasifikace

Záznamy o výsledcích studentů v jednotlivých předmětech uchovává tabulka *klasifikace_fact*. Seznam jejích atributů, společně s jejich vyplněností udává tabulka 4.3. Celkem datový sklad eviduje 354 514 záznamů klasifikace. 241 711 záznamů patří studentům BICS a 63 646 záznamů studentům MICS.



Obrázek 4.8 Průměry známek Čechů a cizinců v PP předmětech bakalářského studia

Tabulka 4.3 Atributy tabulky klasifikace_fact

Název atributu	Vyplněnost v procentech
semestr_id	100
predmet_id	98,98
studium_id	100
zapocteno	69,51
zakonceno	66,79
znamka	58,46
poradi_zapisu	98,98

4.6.1 Problémy s daty klasifikace

V průběhu vypracovávání mé práce muselo být bráno v potaz několik problémů vztahujících se k tabulce *klasifikace_fact*. Častým případem je, že studentovi není explicitně dopsána známka 4 (F) do klasifikace, pokud nesplní požadavky pro úspěšné dokončení předmětu. V takovém případě je v datech jako známka z předmětu vyplněna hodnota `NaN`. Pro představu se jedná o 27,5 % případů všech záznamů klasifikace bakalářských PP předmětů a 75 % případů všech hodnocení nesplnění bakalářských PP předmětů. Zároveň pokud student nesplní požadavky předmětu pro jeho absolvování a předmět tedy nezakončí, je atribut *zakonceno* vyplněn jako hodnota `NaN`. Hodnoty atributů *zakonceno* a *zapocteno* byly v rámci předzpracování dat přetypovány a doplněny podle podmínek, které popisuji v kapitole 5.

Dalším problémem je, pokud si student nechá uznat úspěšné dokončení předmětu na základě absolvování podobného předmětu na jiné fakultě. V některých případech neexistuje záznam klasifikace daného uznaného předmětu a musíme se spolehnout na ruční prohledávání ostatních předmětů, které student absolvoval. Předměty z jiných fakult ale nejsou často uchovány v tabulce *predmet_dim*, ze které bychom byli schopni zjistit název předmětu (známe pouze číselný identifikátor). Po konzultaci s pracovníky datového skladu byly do tabulky *predmet_dim* přidány ještě dva předměty Fakulty elektrotechnické ČVUT, které jsou uznávány místo předmětů z naší fakulty. Jedná se o předměty: A0M32IBE (Informační bezpečnost) – předmět uznatelný místo předmětu BI-BEZ a A0B01LAG (Lineární Algebra) – předmět uznatelný místo BI-LIN.

Stejný problém byl i s uznáváním předmětů, které existovaly v minulosti, nebyly povinné pro všechny studenty a po jejich absolvování bylo studentům uznáno absolvování povinného předmětu. Takovým příkladem je předmět BI-ZSI, který byl v minulosti uznatelný místo předmětu BI-SI1.2. Informace o tom, jaké předměty byly uznatelné místo povinných předmětů, ale nejsou

dobře dostupné. V průběhu vypracovávání práce bylo nutné precizně projít studenty s chybějícími záznamy v povinných předmětech a prozkoumat ostatní předměty, které absolvovali, zaměřit se na podobné struktury v datech ostatních takových studentů a na základě zjištěných informací ještě ověřit fakta se studenty nebo kantory, kteří si staré běhy předmětů pamatovali.

Další nesnáz spočívala v uznávání předmětů u studentů, kteří se přihlásili znovu do studia. Pokud si takoví studenti chtějí nechat uznat klasifikace předmětů z minulých studií, jsou jim patřičné klasifikace zapsány, ale do semestru *A000*, případně *A00*. Ztrácíme tak informaci o tom, v jakém semestru student předměty absolvoval.

Nepříjemnou skutečností byl také fakt, že jsem většinu doby pracovala s nekompletními daty a až po konzultacích s pracovníky skladu mi byl poskytnut zbytek záznamů klasifikace – zda se jedná o všechna data nejde ale jednoduše ověřit. Snadná ověřitelnost je pouze u studentů, kteří již dostudovali úspěšně a víme tedy, že v průběhu studia museli absolvovat všechny povinné předměty. Celkově se jednalo o zhruba 70 000 záznamů klasifikace, které mi byly poskytnuty až v závěru práce.

Kapitola 5
Předzpracování dat

Pro předzpracování dat byla využita primárně softwarová knihovna *pandas*[1]. Jedná se o knihovnu napsanou pro programovací jazyk *Python*, která slouží k manipulaci s daty a datovou analýzu. Jak již bylo zmíněno výše, tato práce se zaměřuje na tři typy predikce – predikce známek z povinných předmětů, predikce úspěšného dokončení studia a predikce úspěšného dokončení jednotlivých semestrů. Pro každou predikci je nutné vytvořit odpovídající dataset, na kterém budou vytvořeny prediktivní modely. V této kapitole se zaměříme na proces tvorby datasetů.

5.1 Dataset matrix_bak_2015

Modely určené k predikci známek z povinných předmětů a predikci úspěšného dokončení studia jsou postaveny nad datasetem *matrix_bak_2015*. Řádky tabulky jsou tvořeny záznamy o jednotlivých studentech, index tabulky tvoří atribut *studium_id*, sloupce představují jednotlivé příznaky. Pro vytvoření matice byli bráni pouze studenti, kteří již dostudovali, mají rok nástupu vyšší nebo roven roku 2015 a kteří zároveň nejsou z nové akreditace 2021. Důvody, které stály za tímto rozhodnutím vychází ze skutečností, které jsou detailně popsány v sekcích 3.1.3 a 3.3.

Příznaky *matrix_bak_2015* mohou být rozděleny do tří kategorií: výsledná proměnná (tedy proměnná, kterou predikujeme) – *dostudoval_uspesne*, příznaky předmětové (příznaky obsahující známky studentů z jednotlivých předmětů) a příznaky s osobními daty studenta (tedy příznaky obsahující sociodemografické údaje, informace o důležitých datech souvisejících se studiem, či informace o způsobu přijetí). Příznak *dostudoval_uspesne* obsahoval hodnotu 0, pokud student nedostudoval úspěšně, a hodnotu 1, pokud ano. Zda jednotlivý studenti úspěšně dostudovali, bylo zjištěno z tabulky *studium_dim* z atributu *ukonceni_zpusob*. Příznaky z dalších dvou kategorií podrobně rozebereme v následujících sekcích.

5.1.1 Příznaky předmětů

Příznaky předmětů obsahují všechny povinné společné předměty spadající do BICS. Nejprve bylo nutné doplnit známku 4 a případně chybějící hodnoty *zapocteno* a *zakonceno*, pokud nebyly vyplněny kantory při absolvování předmětu a místo nich jsou v záznamu patřičné klasifikace hodnoty `NaN`. Pro lepší orientaci byly hodnoty *zapocteno* změněny z hodnoty Z (zápočet byl udělen) na 1 a z hodnoty N (zápočet nebyl udělen) na 0. Ze stejného důvodu byla hodnota atributu *zakonceno* doplněna na hodnotu 0, pokud byla předtím rovna hodnotě `NaN` a student předmět nedokončil úspěšně. Pokud atribut *znamka* záznamu nebyl roven 4 a zároveň nebyl roven

[1] https://pandas.pydata.org/

hodnotě `NaN`, byl atribut *zakonceno* doplněn na hodnotu 1. *Zapocteno* bylo doplněno na hodnotu 0, pokud byla splněna alespoň jedna z níže popsaných podmínek:

- *znamka* je rovna 4,
- *zakonceno* je rovno 0.

Zapocteno bylo doplněno na hodnotu 1, pokud byla splněna alespoň jedna z níže popsaných podmínek:

- *znamka* nebyla rovna 4 a zároveň nebyla `NaN`,
- *zakonceno* je rovno 1.

Známka 4 byla doplněna, pokud byla splněna alespoň jedna z níže popsaných podmínek:

- nejedná se o aktuální semestr $\wedge$ *zapocteno* není rovno 1 $\wedge$ *zakonceno* není rovno 1 $\wedge$ *znamka* je `NaN`,
- nejedná se o aktuální semestr $\wedge$ *zapocteno* je rovno 1 $\wedge$ *zakonceno* není rovno 1 $\wedge$ *znamka* je `NaN`.

Pro každý z povinných předmětů byl vytvořen příznak, který obsahoval číselné označení známek, které studenti v předmětech obdrželi. Neboť v průběhu let docházelo k různému přejmenovávání předmětů, jak shrnuje sekce 3.1.3, bylo nutné sjednotit záznamy tak, aby tabulka obsahovala všechny známky z povinných předmětů přes všechny verze předmětů. Uznávány byly taktéž známky z odpovídajících předmětů, které byly vyučovány v kombinované formě studia, popř. odpovídající předměty vyučované v anglickém jazyce. Taktéž byly uznávány předměty nové akreditace 2021, jejichž uznatelnost byla oboustranná s předměty staré akreditace, více o propojení nové a staré akreditace popisuji v sekcích 3.3 a 3.1.3. Zároveň příznak obsahuje známky, které student získal v posledním zápisu předmětu (pokud měl předmět zapsaný víckrát a měl tedy více hodnocení k danému předmětu).

5.1.2 Příznaky s osobními daty studenta

Pro účely predikce byly na základě rešerše (popsané v kapitole 2) a zároveň prozkoumání problematiky na fakultě (kapitola 3) a poskytnutých dat (kapitola 4) vytvořeny následující příznaky:

- **pohlavi_kod** – obsahuje hodnotu 0, pokud se jedná o muže, nebo hodnotu 1, pokud se jedná o ženy. Korelace mezi pohlavím studenta a jeho studijními výsledky byla popsána v sekci 4.5.3. Kód pohlaví byl získán z atributu *pohlavi_kod*, který je součástí tabulky *student_dim*. Atribut *pohlavi_kod* sice není plně vyplněn, ale byl vyplněn u všech záznamů studentů z BICS, takže nebylo nutné doplňovat chybějící hodnoty.
- **je_cech** – informace o státní příslušnosti studenta je zjištěna z atributu *statni_prislusnost_nazev* z tabulky *student_dim*, pokud je atribut roven hodnotě *Česko*, je hodnota příznaku *je_cech* rovna 1, pokud není atribut vyplněn je rovna -1, jinak je rovna 0. Rozdíly ve výsledcích Čechů a cizinců jsou popsány v sekci 4.5.4.
- **datum_zahajeni** – obsahuje roky, ve kterých student započal své studium. Rozdíly v průchodnostech v průběhu let se věnuje sekce 3.2.1. Rok zahájení je zjištěn z atributu *datum_zahajeni*, který je součástí tabulky *studium_dim*.
- **rok_maturity** – obsahuje rok, ve kterém student složil maturitní zkoušky. Rok maturity je zjištěn z atributu *rok_maturity*, který je součástí tabulky *studium_dim*.

- **maturita_nastup_rozdil** – obsahuje rozdíl mezi rokem maturity a rokem nástupu. Souvislost mezi rozdílem roku maturity a rokem nástupu je popsána v sekci 4.5.2.
- **gymnazium** – obsahuje 0, pokud student absolvuje jinou střední školu nežli gymnázium, hodnotu 1, pokud se jedná o gymnazistu, nebo hodnotu -1, pokud chybí záznamy o střední škole. To, jakou střední školu student navštěvoval, je zjištěno pomocí IZO kódů, které jsou obsaženy v tabulce *studium_dim*. Zároveň je informace o střední škole doplněna, pokud ke studentovi existuje přihláška, ve které jsou vyplněny atributy související se střední školou. Problematice špatné vyplněnosti informací o předchozím vzdělání se věnují sekce 4.4.1 a 4.5.1, která se rovněž věnuje vlivu typu střední školy na výsledky studenta.
- **praha** – nabývá hodnoty 0, pokud má student trvalé bydliště mimo Prahu, hodnoty 1, pokud je student trvalým bydlištěm z Prahy nebo hodnoty -1, pokud chybí informace o trvalém bydlišti studenta. Informace o trvalém bydlišti je zjištěna z atributů, které jsou součástí tabulky *student_dim*: *kod_obce* (musí obsahovat hodnotu `554782`, aby hodnota příznaku *praha* byla 1), případně z atributu *kod_okresu* (musí obsahovat hodnotu `3100`, aby hodnota příznaku *praha* byla 1), pokud ani jeden z těchto dvou atributů není vyplněn je hodnota *praha* nastavena na -1.
- **scio** – je roven hodnotě 0, pokud u studenta není evidován výsledek scio testů, nebo hodnotě 1, pokud evidován je. Výsledky ze scio testů jsou zjištěny z tabulky *prihlaska_dim* pomocí atributů: *rozhodnuti_kod* (pokud je hodnota atributu 13), *hodnoceni_pz_cast7* (pokud není `NaN`) a *scio_test* (pokud není `NaN`).
- **olym** – je roven hodnotě 0, pokud u studenta není evidován záznam o absolvování olympiád, které jsou uznávány místo přijímacích zkoušek, nebo hodnotě 1, pokud u studenta záznam o těchto olympiádách evidován je. Výsledky z olympiád jsou zjištěny z tabulky *prihlaska_dim* pomocí atributů: *rozhodnuti_kod* (pokud je hodnota atributu 11), *hodnoceni_pz_cast6* (pokud není `NaN`), *olympiady* (pokud není `NaN`).
- **zkouska** – nabývá hodnoty 0, pokud není o studentovi udržována informace o tom, že absolvoval přijímací zkoušku, nebo hodnoty 1, pokud taková informace udržována je. Výsledky ze fakultních přijímacích zkoušek jsou zjištěny z tabulky *prihlaska_dim* pomocí atributů: *rozhodnuti_kod* (pokud je hodnota atributu 10 nebo 12 – přijat na základě odvolání) a *hodnoceni_pz_cast2* (pokud není `NaN`).
- **prominuti** – obsahuje hodnotu 0, pokud ke studentovi existuje informace o prominutí přijímací zkoušky, nebo hodnotu 1, pokud taková informace existuje. Výsledky o prominutí přijímací zkoušky jsou zjištěny z tabulky *prihlaska_dim* pomocí atributu: *hodnoceni_pz_cast1* (pokud není `NaN`).

V následujících sekcích bude popsáno, jak byl dataset *matrix_bak_2015* přizpůsoben potřebám jednotlivých predikcí.

5.1.3 Doplnění chybějících hodnot

Neboť pro účely prediktivního modelování byla vybrána knihovna *scikit-learn*[2], jejíž metody pro vytváření prediktivních modelů dokážou pracovat pouze s hodnotami příznaků, které nejsou `NaN`, bylo nutné najít způsob, kterým chybějící příznaky týkající se známek předmětů ve výsledném datasetu *matrix_bak_2015* doplnit. Vyplněnost předmětových atributů zachycuje tabulka 5.1. Z pochopitelných důvodů jsou nejvíce vyplněny známky z předmětů z prvního roku studia, naopak nejmenší vyplněnost má předmět BI-DPR, který si zapisují studenti až těsně před státnicemi v semestru, ve kterém pracují na své bakalářské práci.

[2] https://scikit-learn.org/stable/

Jako způsob doplnění chybějících známek bylo vybráno doplnění chybějících hodnot pomocí kNN, konkrétně byla využita třída *KNNImputer* z knihovny *scikit-learn*. Po doplnění bylo nutné ještě zaokrouhlit doplněné známky, tak aby se jednalo o diskrétní příznaky.

Tabulka 5.1 Předmětové příznaky tabulky matrix_bak_2015

Název atributu	Vyplněnost v procentech
BI-PA1	98,22
BI-PAI	98,25
BI-CAO	98,30
BI-PS1	98,32
BI-MLO	98,30
BI-ZMA	98,15
BI-PA2	56,53
BI-DBS	63,47
BI-SAP	62,55
BI-LIN	60,65
BI-AG1	40,29
BI-AAG	45,88
BI-ZDM	43,62
BI-OSY	32,34
BI-PSI	37,04
BI-BEZ	34,50
BI-PST	27,01
BI-DPR	23,25
BI-SI1.2	31,81
BI-EMP	36,00

5.2 Predikce úspěšného dokončení studia

Pro predikci úspěšného dokončení studia byl využit dataset *matrix_bak_2015*. Před samotným vytváření modelů bylo nutné ještě přetypovat příznaky na kategorické nominální a kategorické ordinální, k tomu bylo využito třídy *CategoricalDtype*[3], která je součástí knihovny *pandas*. Jako kategorické nominální byly přetypovány všechny nepředmětové příznaky kromě výsledné proměnné *dostudoval_uspesne* a příznaků *rok_maturity*, *maturita_nastup_rozdil* a *datum_zahajeni*. Jako kategorické ordinální byli přetypovány všechny předmětové příznaky.

Navíc byl vytvořen příznak *prumer*, jednalo se o průměr všech předmětů, u kterých má student vyplněnou známku.

Dále bylo důležité rozhodnout, které příznaky budou vybrány pro vytvoření výsledného prediktivního modelu. Z předmětových příznaků, byl vynechán příznak BI-DPR k tomu vedly tyto důvody:

1. jak již bylo zmíněno, jedná se o předmět, který je oficiálně doporučovaný k absolvování v šestém semestru, prakticky je doporučovaný k zapsání v tom semestru, ve kterém student píše svou bakalářskou práci. S tím souvisí velmi nízká vyplněnost známek z předmětu (většina studentů končí své studium již v prvních letech studia, jak rozebírá sekce 3.2.1) – je vyplněno pouze 915 záznamů z celkových 3 936,
2. pouze 1,42 % neúspěšných studentů absolvovalo předmět úspěšně. Je pravděpodobné, že s tím také souvisí skutečnost, že byl příznak při vytváření modelů považován jako důležitý, i když se

[3] https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.api.types.CategoricalDtype.html/

nejedná o předmět, který by nějakým zásadním způsobem ovlivňoval to, že student skutečně nedostuduje. Hrozilo tak přeučení modelů, neboť modely vytváříme na již dostudovaných studentech.

Zároveň se v průběhu vytváření modelů u všech metod ukázala většina nepředmětových příznaků jako příznaky s velmi nízkou důležitostí (tzv. *feature importance*). Nejlépe byly hodnoceny příznaky související s rokem nástupu a rokem absolvování maturity, tedy: *maturita_nastup_rozdil*, *datum_zahajeni* a *rok_maturity*. Z těchto příznaků byl vybrán do výsledného datasetu příznak *maturita_nastup_rozdil*, který dosahoval nejvyšší *feature importance*. Jednotlivé *feature importance* jsou získány z atributů *feature_importances_* příslušných *scikit-learn* tříd dílčích metod. Konkrétně pro metody Rozhodovacího stromu, Náhodného lesa a AdaBoost je použita tzv. *impurity-based importance*[38], ta je založená na *Mean decrease in impurity (MDI)*. Důležitost příznaků metody podpůrných vektorů je zjištěna pomocí funkce *permutation_importance* z knihovny *scikit-learn*[4]. Pro metodu XGBoost je použita metoda *get_score*, ta pro určení důležitosti jednotlivých příznaků využívá parametrů jako *weight* – kolikrát je příznak použit při vytváření modelu, *gain* – průměrný zisk při větvení, *total_gain* – celkový zisk, aj. Kompletní seznam je dostupný v dokumentaci třídy XGBoost[5].

U atributu *datum_zahajeni* hrozilo přeučení modelu, neboť byli bráni v potaz pouze dostudovaní studenti. Následkem toho je fakt, že studenti s rokem nástupu 2018 a výš, dosahovali mnohem vyšší neúspěšnosti, neboť ještě velké množství těchto studentů stále studuje. Stejný problém byl i s příznakem *rok_maturity*. Ze sekce 4.5.2 víme, že 66 % studentů nastoupí na fakultu ve stejném roce, ve kterém složí maturitní zkoušky, 90 % do dvou let od maturity. Rok maturity tak koresponduje s příznakem *datum_zahajeni*.

5.3 Predikce známek z povinných předmětů

Pro predikování známek z povinných předmětů bylo nutné pro predikci každého předmětu vytvořit speciální dataset, ve kterém výslednou proměnnou tvořil predikovaný předmět. Tyto datasety byly vytvořeny třemi způsoby z datasetu *matrix_bak_2015*:

1. Dataset zůstal zachován.
2. Byly ponechány pouze předmětové příznaky.
3. Předměty byly rozděleny do čtyř shluků způsobem, který bude detailně popsán v následující části textu. Výsledné datasety, na kterých byly vytvářeny prediktivní modely pro jednotlivé předměty, byly tvořeny pouze předměty, které patřily do stejných shluků jako predikovaný předmět.

Následně bylo provedeno přetypování příznaků na kategorické – k tomu bylo využito třídy *CategoricalDtype*. Na kategorické nominální byly převedeny příznaky *pohlavi_kod*, *je_cech*, *gymnazium*, *praha*, *scio*, *olym*, *zkouska* a *prominuti*. Na kategorické ordinální byly převedeny příznaky týkající se známek z předmětů.

5.3.1 Rozdělení předmětů do jednotlivých shluků

Cílem rozdělení předmětů bylo zjistit, zda bude predikce známek v předmětech přesnější, pokud budeme vytvářet prediktivní modely pouze nad datasety, které budou tvořeny předměty stejného typu. Tedy zda bude stačit, pokud výslednou známku z vybraného předmětu budeme predikovat na základě známek z daného předmětu studentů, kteří mají nejpodobnější klasifikaci

[4] https://scikit-learn.org/stable/modules/permutation_importance.html/
[5] https://xgboost.readthedocs.io/en/stable/python/python_api.html#xgboost.Booster.get_fscore/

v předmětech typu vybraného předmětu. Typ předmětu byl určen zejména tím, jestli se jednalo o předmět matematický nebo programovací a zároveň podle rozložení jednotlivých známek v předmětu.

Pro shlukování předmětů bylo využito třídy *KMeans*[6] z knihovny *scikit-learn*. Shlukování bylo provedeno nad maticí, která obsahovala rozložení jednotlivých známek v předmětech a navíc dva sloupce – *matematika* a *programovani*, které nabývaly hodnoty 1, pokud byl předmět matematický, resp. programovací, a hodnoty 0, pokud nebyl. Tabulka 5.2 zobrazuje tuto matici. Před samotným shlukováním byla matice ještě znormalizována pomocí třídy *MinMaxScaler*[7] z knihovny *scikit-learn*, která provádí min-max normalizaci a výsledné sloupce matice tak mohly nabývat hodnot mezi 0 a 1.

Tabulka 5.2 Rozložení známek BICS povinných společných předmětů a informace o tom, zda se jedná o předmět matematický, či programovací

	1.0	1.5	2.0	2.5	3.0	4.0	matematika	programováni
BI-PA1	316	266	520	707	638	2676	0	1
BI-PAI	155	460	1014	1197	789	975	0	0
BI-CAO	2004	1138	84	261	209	872	0	0
BI-PS1	322	326	562	801	869	2106	0	1
BI-MLO	219	449	876	969	238	2173	1	0
BI-ZMA	177	422	706	771	173	2966	1	0
BI-PA2	310	269	404	547	169	1742	0	1
BI-DBS	359	632	679	610	265	740	0	0
BI-SAP	904	326	287	528	364	825	0	0
BI-LIN	70	142	403	650	351	2107	1	0
BI-AAG	157	319	482	475	205	1118	0	0
BI-ZDM	91	175	459	705	172	1052	1	0
BI-AG1	243	157	270	499	278	1086	0	1
BI-OSY	250	48	72	544	388	459	0	1
BI-PSI	75	210	425	490	277	436	0	0
BI-BEZ	192	251	347	419	210	367	0	0
BI-PST	90	237	338	406	198	235	1	0
BI-DPR	475	263	106	79	7	139	0	0
BI-SI1.2	188	352	485	355	80	207	0	0
BI-EMP	624	660	247	114	17	259	0	0

Předměty byly rozděleny do čtyř shluků:

1. BI-MLO, BI-ZMA, BI-LIN, BI-ZDM, BI-PST – odpovídá matematickým předmětům,
2. BI-PA1, BI-PS1, BI-PA2, BI-AG1, BI-OSY – odpovídá programovacím předmětům,
3. BI-CAO, BI-DPR, BI-EMP – odpovídá lehčím předmětům,
4. BI-PAI, BI-DBS, BI-SAP, BI-AAG, BI-PSI, BI-BEZ, BI-SI1.2 – zbytek předmětů.

5.3.2 Výsledné datasety

Nejlepších výsledků bylo dosaženo při predikci nad datasetem, který byl tvořen pouze předmětovými příznaky. Tabulka 5.3 ukazuje průměrnou RMSE hodnotu prediktivních modelů nad jednotlivými typy datasetů. Z toho důvodu byl tento způsob vybrán pro vytváření datasetů

[6] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html/
[7] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html/

po jednotlivých předmětech pro výslednou predikci. Celý proces prediktivního modelování nad vybranými datasety, jež byly vytvořeny vybraným způsobem, i výsledná evaluace je detailně popsána v sekci 6.3.

Tabulka 5.3 Průměrná RMSE hodnota prediktivních modelů nad datasety, vytvořených jednotlivými způsoby

	Průměrná hodnota RMSE
1. způsob – zachování původního datasetu	0,458
2. způsob – pouze předmětové příznaky	0,430
3. způsob – rozdělení do shluků	0,540

5.4 Predikce úspěšného dokončení jednotlivých semestrů

Ze stejných důvodů popisovaných v sekci 5.1, která se zabývá tvorbou *matrix_bak_2015*, byly pro tvorbu datasetů pro predikování úspěšného dokončení jednotlivých semestrů vybráni pouze studenti, kteří mají datum nástupu vyšší nebo roven roku 2015 a zároveň nespadají do nové akreditace, tedy rok nástupu je menší než 2021.

Byl vytvořen samostatný dataset pro každý semestr, tedy konkrétně pro první, druhý, třetí, čtvrtý a pátý semestr. Šestý semestr nebyl součástí, neboť by dataset obsahoval prakticky stejné předmětové příznaky jako dataset, kterým predikujeme úspěch celého studia. Každý z vytvořených datasetů obsahuje pouze informace z předchozích dokončených semestrů. Tedy ku příkladu dataset, který je použit pro predikci úspěšného dokončení druhého semestru, obsahoval pouze informace související s klasifikací předmětů, které mají doporučený průchod v prvním semestru. Řádky datasetů jsou tvořeny záznamy o jednotlivých studentech, u kterých již víme zda daný semestr absolvovali úspěšně či neúspěšně. Důležité také je, že nejsou bráni v potaz studenti, kteří úspěšně dokončí studium za méně nežli tři roky – nejedná se o standardní průchod studiem. Index tabulky tvoří atribut *studium_id*, sloupce představují jednotlivé příznaky. Přesněji součástí každého datasetu jsou tyto informace:

- výsledná proměnná (*dostudoval_semestr*) – obsahovala 1, pokud student absolvoval úspěšně daný semestr, hodnotu 0, pokud ne.
- příznaky obsahující známky z povinných předmětů, jejichž doporučený zápis je v předchozích semestrech. Tyto příznaky jsou ve tvaru *kód_předmětu_znamka*. Známka v konkrétních předmětech je vyplněna, pouze pokud student skutečně předmět absolvoval v daném semestru svého studia nebo v semestrech předchozích, pokud si student nechá uznat předměty z předchozího studia, jsou známky z těchto předmětů zapsány v semestrech s doporučeným průchodem studia (nemáme informaci o skutečném semestru absolvování předmětu). Pokud předmět ještě neabsolvoval, je vyplněna hodnota -1.
- příznaky obsahující pořadí zápisu daného předmětu. Opět pouze předměty, jejichž doporučený zápis je v předchozích semestrech a zároveň obsahoval nejaktuálnější zápis předmětu. Pokud si student předmět ještě nezapsal je vyplněna -1.
- příznak obsahující počet kreditů, které student získal v jednotlivých semestrech předcházejících zadanému semestru.
- příznaky obsahující osobní data studenta – tyto příznaky jsou použity pro predikování úspěšného absolvování prvního semestru, u kterého se jedná o jediné informace, které o studentovi máme. Jedná se o příznaky: *pohlavi_kod*, *je_cech*, *datum_zahajeni*, *rok_maturity*, *matu-rita_nastup_rozdil*, *gymnazium*, *praha*, *scio*, *olym*, *zkouska*, *prominuti*. Příznaky spadající do

této kategorie byly získány stejným způsobem, který byl již popsán v sekci 5.1.2, jež se zabývá získáním příznaků s osobními daty studenta při tvorbě *matrix_bak_2015*.

Informace o tom, zda student dokončil úspěšně, tedy obsah příznaku *dostudoval_semestr*, je získána z dat následujícím způsobem. Pro každého studenta, který nedostudoval úspěšně (tj. atribut *ukonceni_zpusob* není roven 1 a zároveň atribut *studuje* je roven hodnotě K – student již nestuduje, oba tyto atributy patří do tabulky *studium_dim*), je vypočítán tzv. neúspěšný semestr. Pomocí atributu *datum_ukonceni* z tabulky *studium_dim* je zjištěn měsíc ukončení studia. Pokud měsíc ukončení spadá do měsíců březen, duben, květen, červen, červenec, srpen, září, je jako semestr ukončení považován letní semestr v ročníku studia, ve kterém student skončil (atribut *rocnik* z tabulky *studium_dim*). Jestliže měsíc ukončení studia spadá do ostatních měsíců, jako semestr ukončení je považován zimní semestr v ročníku, ve kterém student skončil. Tento způsob je využit pouze u semestrů, ve kterých není kontrola počtu splněných kreditů, tedy třetí a pátý. Pro semestry, ve kterých je podmínka postupu počet kreditů (konkrétní počet kreditů shrnuji v sekci 3.2), tedy v letních semestrech a v prvním roce i v zimním, je kontrolováno, jestli v daném semestru student dosáhl požadovaného počtu kreditů.

Ač je toto řešení dostačující v naprosté většině případů, najdou se i výjimky. Ku příkladu někteří studenti, kteří si nechají uznat velké množství absolvovaných předmětů z minulých studií, prochází, zřejmě díky výjimce, do dalších semestrů, i když nemají získáno dostatečné množství kreditů pro postup. Problematické jsou i semestry, které spadaly do tzv. „covidového období“, o kterém se více rozepisuji v sekci 3.4. V těchto semestrech totiž docházelo ke snižování hranice počtu získaných kreditů pro postup studiem. Další problematickou skupinou jsou studenti, kteří si pozastaví své studium a v následujícím semestru nemají zapsaný žádný předmět, tedy nesplní podmínku dostatečného množství kreditů. Otázkou tedy bylo, jak naložit s těmito okrajovými případy.

Jedním z řešení, které se nabízelo, bylo nebrat v potaz podmínku dostatečného množství kreditů a zaměřit se pouze na měsíc ukončení. Datum ukončení studia se ale ukázal jako poměrně nespolehlivý, z průzkumu dat totiž vyšlo najevo, že studijní oddělení v nezanedbatelném množství případů ukončuje studium, až když již probíhá období dalšího semestru. A to i v případě, že student nesplnil podmínky pro postup (například nesplnil ani jeden předmět). Další možností bylo zaměřit se na to, zda mají studenti v následujícím semestru zapsán nějaký předmět. Tato cesta se ale také ukázala jako lichá, neboť častým případem je, že se studentům předměty do dalšího semestru zapisují automaticky při předběžných zápisech. Zároveň ale muselo být bráno v potaz i to, zda vůbec počítat s těmito okrajovými případy, neboť se většinově daly považovat za odlehlé hodnoty, které nesplňují standardní podmínky postupu ve studiu, a které by učení modelů nijak nepřispěly, spíše naopak. Na základě všech těchto skutečností bylo rozhodnuto nechat proces rozhodování úspěšnosti v původním stavu a nezahrnovat do jeho rozhodovacích podmínek další, které by řešily určování těchto okrajových případů.

Následně bylo ještě nutné přetypovat příznaky na příznaky kategorické nominální a kategorické ordinální, k tomu bylo využito třídy *CategoricalDtype*. Na kategorické nominální byly převedeny příznaky *pohlavi_kod*, *je_cech*, *gymnazium*, *praha*, *scio*, *olym*, *zkouska* a *prominuti*. Na kategorické ordinální byly převedeny příznaky týkající se známek z předmětů.

Kapitola 6
Prediktivní modelování

V této kapitole se zaměříme na samotné prediktivní modelování výsledných datasetů. Jednotlivým predikcím jsou vyčleněny příslušné sekce. Součástí je také diskuse zabývající problematikou možného rozšíření práce o studenty nové akreditace.

6.1 Predikce úspěšného dokončení studia

Pro predikování úspěšného dokončení studia byl využit upravený dataset *matrix_bak_2015*. Vybraný dataset obsahoval 3 936 záznamů, následně byl rozdělen na trénovací a testovací sadu dat v poměru 4 : 1 pomocí metody *train_test_split*[1] z knihovny *scikit-learn learn*. Data byla rozdělena *stratify* způsobem, kdy každá skupina obsahuje po rozdělení poměrově stejné zastoupení hodnot výsledné proměnné a parametr *shuffle* byl nastaven na `True`, tedy jednotlivé záznamy byly před rozdělením promíchány. Predikovanou proměnnou tvoří příznak *dostudoval_uspesne*, jenž je binární příznak. Studenti, kteří nedostudovali úspěšně, jsou reprezentovaní hodnotou 0, v opačném případě, kdy dostudovali úspěšně naopak hodnotou 1. Rozložení hodnot v datasetu ukazuje tabulka 6.1.

Tabulka 6.1 Popis trénovacího a testovacího datasetu pro predikci úspěšného dokončení studia

	Trénovací	Testovací
počet záznamů	3 148	788
procento záznamů z původního datasetu	79,98	20,02
procentuální zastoupení *dostudoval_uspesne* = 1	29,03	21,07
procentuální zastoupení *dostudoval_uspesne* = 0	78,97	78,93

6.1.1 Trénování modelů

Prediktivní modely byly vytvořeny a laděny na trénovacím datasetu pomocí 5-násobné křížové validace, jako metrika vyhodnocující kvalitu kombinace hyperparametrů bylo zvoleno F1 skóre. Pro tvorbu a trénování modelů bylo vybráno na základě provedené rešerše sedm metod. Všechny metody, které byly vybrány, jsou vypsány níže:

- **Rozhodovací strom**,

[1] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html/

- **Náhodný les**,
- **Metoda podpůrných vektorů (SVM)**,
- **AdaBoost**,
- **XGBoost**,
- **Vícevrstvý perceptron (MLP)**,
- **kNN**.

6.1.2 Vyhodnocení

Vyhodnocení výsledků predikce bylo měřeno pomocí Klasifikační přesnosti na testovací množině dat. Přesnost byla měřena jak na kompletních datech testovací množiny, tak bylo nutné otestovat přesnost predikce po jednotlivých dokončených semestrech. Právě přesnost po jednotlivých semestrech je zásadní a skutečně vypovídající o kvalitě modelu. Pokud bychom měřili přesnost pouze na kompletních datech, nevěděli bychom, zda model dobře funguje i na datech studentů, kteří ještě neabsolvovali všechny povinné předměty.

Pro měření po jednotlivých semestrech byl navrhnut následující způsob. Nejprve bylo nutné otestovat, zda model dokáže určit úspěšné dokončení studia i u člověka, který ještě vůbec nastoupil na fakultu (jestli lze využít pouze sociodemografických dat). K tomuto účelu bylo potřeba nějakým způsobem v původním datasetu znehodnotit všechny příznaky týkající se známek z předmětů. Předmětové příznaky tak byly nahrazeny průměrnými známkami v daných předmětech a příznak průměr byl přepočítán na přepsaných hodnotách známek.

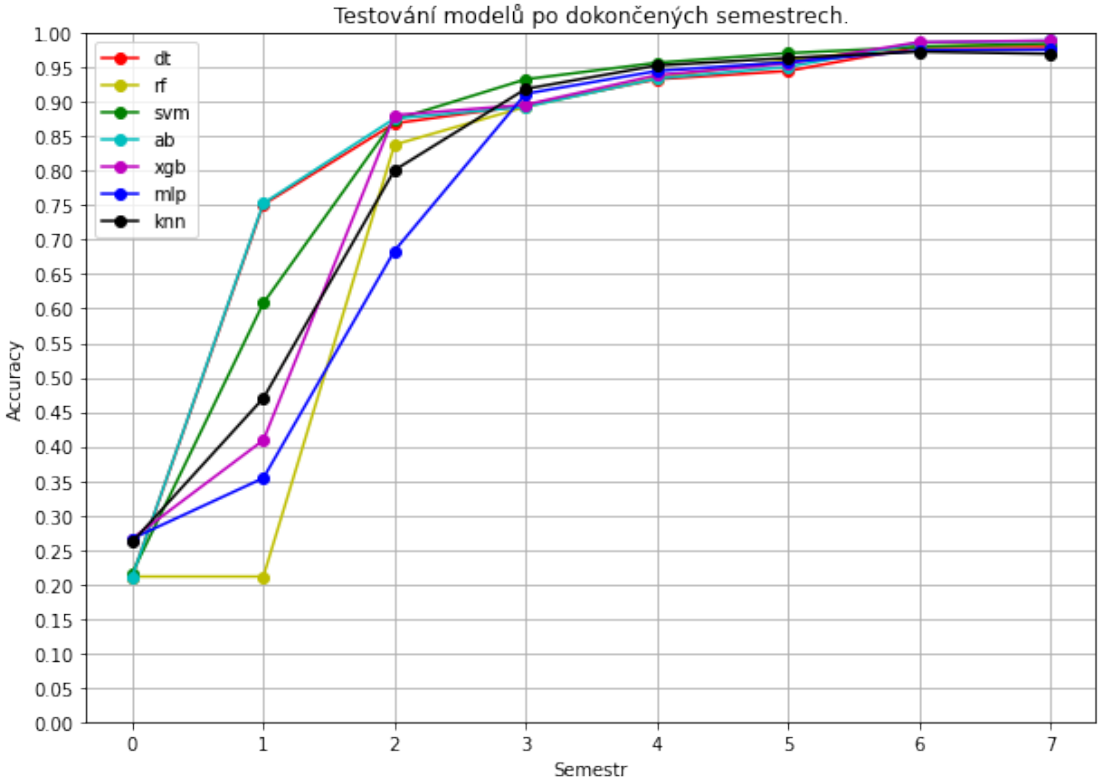
Stejným způsobem se postupovalo i v případech dalších dokončených semestrů. Tedy například v datasetu, který obsahoval pouze známky z dokončených prvních dvou semestrů, byly ponechány známky z těchto dvou semestrů, zbylá hodnocení byla nahrazena průměrnými známkami v daných předmětech.

Následně bylo ovšem nutné zohlednit fakt, že dataset *matrix_bak_2015* obsahuje finální hodnoty známek studentů (např. již po druhém zápisu předmětu). Aby testovací datasety lépe reflektovaly skutečnost, byly využity datasety, které již byly vytvořeny pro predikci průchodu jednotlivými semestry (navíc byl vytvořen i dataset uchovávající informace o výsledcích z pátého semestru). Tyto datasety totiž obsahují informace o známkách studentů po jednotlivých semestrech, které absolvovali. Například pokud student v prvním semestru dostane z předmětu BI-ZMA známku 4, bude tato známka uložena do testovacího datasetu, který reflektuje známky z dokončeného prvního semestru. Pokud si daný student další rok, tedy ve svém třetím semestru, opraví známku na 2, bude tato informace uložena do testovacího datasetu, který obsahuje známky z dokončených prvních třech semestrů. Ve finále byly ještě přepočítány hodnoty příznaku *prumer*.

Celkem se tedy přesnost všech modelů testovala na osmi různých verzích testovacího datasetu. První dataset obsahuje jen sociodemografická data, dalších pět datasetů představují známky studentů po prvních pěti semestrech. Sedmý dataset obsahuje finální známky ze všech předmětů kromě zmíněných BI-EMP a BI-SI1.2 (tyto předměty mají doporučený zápis v Bílé knize v semestru sedmém, jak je popsáno v sekci 3.1.3). Poslední osmý dataset je vlastně sedmý dataset, který navíc obsahuje známky z předmětů BI-EMP a BI-SI1.2. Zde narážíme na úskalí u studentů, kteří studují déle nežli standardní dobu. Ztrácíme u nich totiž informaci o tom, v jakém semestru si známky z předmětů opraví a může se u nich tedy objevit skok mezi datasetem pro dokončený pátý semestr a šestý datasetem, který obsahuje již finální známky. Jelikož se ale zaměřujeme zejména na studenty se standardní délkou studia a standardním průchodem dle Bílé knihy, byly testovací datasety ponechány v tomto stavu.

Teď již k výsledkům jednotlivých modelů. Nejlepších výsledků dosahoval model AdaBoost. Z tabulky 6.2 můžeme vyčíst výsledky jednotlivých modelů při testování po dílčích dokončených semestrech. Obrázek 6.1 zaznamenává výsledky graficky pro lepší představu.

Můžeme si povšimnout, že model AdaBoost dosahoval nadprůměrných výsledků oproti ostatním modelům. Jedinou výjimkou je predikce na datasetu pro predikci průchodu prvním semestrem, který byl založen pouze na sociodemografických datech, zde model AdaBoost zaostává za modely tvořených metodami SVM, XGBoost, MLP a kNN. Ovšem již na datech z prvního dokončeného semestru vidíme prudký nárůst přesnosti modelu AdaBoost, na těchto datech je oproti ostatním modelům nejpřesnější, srovnatelné přesnosti dosahuje pouze model rozhodovacího stromu. Model XGBoost je po druhém dokončeném semestru lehce přesnější než model AdaBoost, oba modely vyrovnávají svou přesnost na kompletním datasetu – oba mají přesnost 98,9 %. Na datech po dokončení dalších semestrů jsou již oba modely poměrně vyrovnané. Model AdaBoost byl zvolen, jelikož se snažíme o co nejlepší včasnou přesnost a zde právě model AdaBoost dosahuje lepších výsledků nežli modely ostatní.



Obrázek 6.1 Testování modelů pro predikování úspěchu studia po dokončených semestrech. Jednotlivé modely jsou zde reprezentovány těmito zkratkami: **dt** – Rozhodovací strom, **rf** – Náhodný les, **svm** – Metoda podpůrných vektorů, **ab** – AdaBoost, **xgb** – XGBoost, **mlp** – Vícevrstvý perceptron, **knn** – kNN.

Zajímavý je také fakt, že ani jeden z modelů nedokáže kvalitně predikovat úspěch studenta pouze na základě sociodemografických dat o studentech. Přijatelných výsledků (přesnost nad 80 %) dosahuje většina modelů (až na model kNN) po absolvování druhého semestru.

Tabulka 6.3 zaznamenává důležitost 10 příznaků s nejvyšší naměřenou *feature importance*. Zajímavý je výskyt předmětu BI-CAO na prvních příčkách, ačkoliv se jedná o předmět, který je všeobecně považován za spíše jednodušší oproti ostatním. Ukazuje se, že to, jakou známku student v tomto předmětu získá, je zásadní pro predikci toho, zda vystuduje úspěšně či nikoliv. Za pozornost také stojí důležitost příznaku *maturita_nastup_rozdil*, tedy rozdílu mezi rokem nástupu a rokem absolvování maturitní zkoušky. Nulovou *feature importance* naopak zazna-

Tabulka 6.2 Výsledky jednotlivých modelů při predikci úspěšného dokončení studia

Dokončených semestrů	Použitá metoda						
	Rozhodovací strom	Náhodný les	SVM	AdaBoost	XGBoost	MLP	kNN
0	0,211	0,211	0,216	0,211	0,266	0,266	0,261
1	0,751	0,211	0,608	0,753	0,409	0,354	0,470
2	0,869	0,838	0,874	0,876	0,881	0,684	0,801
3	0,895	0,893	0,933	0,893	0,896	0,912	0,919
4	0,933	0,935	0,957	0,934	0,939	0,945	0,953
5	0,945	0,961	0,971	0,951	0,956	0,958	0,964
6	0,980	0,981	0,980	0,986	0,987	0,975	0,973
7	0,98	0,984	0,985	0,989	0,989	0,976	0,970

menávají předměty BI-PS1, BI-BEZ a BI-SI1.2.

Tabulka 6.3 Deset příznaků s nejvyšší feature importance u modelu metody AdaBoost u predikce úspěšného dokončení studia

Příznak	Feature importance
maturita_nastup_rozdil	0,13
BI-PST	0,11
prumer	0,11
BI-CAO	0,09
BI-OSY	0,09
BI-PA2	0,07
BI-DBS	0,07
BI-EMP	0,04
BI-PSI	0,04
BI-AAG	0,04

6.1.3 Možné problémy

Jak již bylo zmíněno, model byl naučen na datasetu, který obsahuje již pouze nestudující studenty, a jsme tak schopni rozdělit úspěšné a neúspěšné studenty. Problém, který ale díky tomu může nastat, spočívá v případech, kdy predikujeme úspěšnost u studenta, který má sice z nějakého předmětu známku F, ale v budoucích letech si známku ještě opraví. Model totiž neměl šanci pracovat s podobnými studenty, neboť měl k dispozici pouze finální výsledky. Obecně má model tendenci podceňovat schopnost studentů si špatné známky opravit.

6.2 Predikce dokončení semestrů

Pro predikci úspěchu každého semestru byl vytvořen vlastní prediktivní model na odpovídajícím datasetu (tvorba datasetů je popsána v sekci 5.4). V této části se zaměříme na to, co bylo pro vytváření a evaluaci všech modelů společné, v následujících sekcích se zaměříme na specifika a výsledky predikcí jednotlivých semestrů.

Všechny datasety byly rozděleny na trénovací a testovací data v poměru 4 : 1. K tomu bylo využito metody *train_test_split* z knihovny *scikit-learn*. Data byla rozdělena *stratify* způsobem a parametr *shuffle* byl nastaven na `True`.

Prediktivní modely byly vytvořeny a laděny na trénovacím datasetu pomocí 5-násobné *stratified* křížové validace, jako metrika vyhodnocující kvalitu kombinace hyperparametrů byl zvolen F1 skóre. K tomu bylo využito třídy *GridSearchCV*[2] z knihovny *scikit-learn*.

Pro tvorbu a trénování modelů bylo vybráno na základě provedené rešerše šest metod. Všechny metody, které byly vybrány, jsou vypsány níže:

- **Rozhodovací strom**,
- **Náhodný les**,
- **AdaBoost**,
- **XGBoost**,
- **Vícevrstvý perceptron (MPL)**,
- **kNN**.

Evaluace výsledných datasetů byla provedena pomocí klasifikační přesnosti.

6.2.1 První semestr

Původní dataset obsahoval 4 546 záznamů. Tabulka 6.4 shrnuje vlastnosti vytvořeného trénovacího a testovacího datasetu.

Tabulka 6.4 Popis trénovacího a testovacího datasetu pro první semestr

	Trénovací	Testovací
počet záznamů	3 636	910
procento záznamů z původního datasetu	80	20
procentuální zastoupení *dokoncil_semestr* = 1	53,44	53,41
procentuální zastoupení *dokoncil_semestr* = 0	46,56	46,60

Nejlepších výsledků dosahoval model metody AdaBoost, ten predikoval postup prvním semestrem s přesností na testovací sadě 56,8 %. Tabulka 6.6 ukazuje klasifikační přesnosti ostatních modelů. Predikce úspěšnosti prvního semestru je značně nepřesná, pokud máme k dispozici pouze sociodemografická data o studentech. Zaměřme se ještě na *feature importance* jednotlivých příznaků pro model metody AdaBoost. Jak je vidět z tabulky 6.6, model využívá pouze čtyř příznaků *rok_maturity*, *datum_zahajeni*, *gymnazium* a *maturita_nastup_rozdil*.

Tabulka 6.5 Klasifikační přesnost na testovací sadě jednotlivých modelů u predikce úspěšnosti prvního semestru

Využitá metoda	Klasifikační přesnost
Rozhodovací strom	52,4
Náhodný les	56,2
AdaBoost	56,8
XGBoost	55,2
Vícevrstvý perceptron	46,4
kNN	51,9

[2] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html/

Tabulka 6.6 Feature importance u modelu metody AdaBoost u predikce úspěšnosti prvního semestru

Příznak	Feature importance
rok_maturity	0,33
datum_zahajeni	0,27
gymnazium	0,27
maturita_nastup_rozdil	0,13
pohlavi_kod	0,00
je_cech	0,00
praha	0,00
scio	0,00
olym	0,00
zkouska	0,00
prominuti	0,00

6.2.2 Druhý semestr

Původní dataset obsahoval 2 429 záznamů. Tabulka 6.7 shrnuje vlastnosti vytvořeného trénovacího a testovacího datasetu.

Tabulka 6.7 Popis trénovacího a testovacího datasetu pro druhý semestr

	Trénovací	Testovací
počet záznamů	1 943	486
procento záznamů z původního datasetu	80	20
procentuální zastoupení *dokoncil_semestr* = 1	64,28	64,40
procentuální zastoupení *dokoncil_semestr* = 0	35,72	35,60

Nejlepších výsledků dosahoval model náhodného lesa, ten měl na testovací sadě přesnost 87,7 %. Výsledky všech modelů shrnuje tabulka 6.8. Tabulka 6.9 shrnuje významnost příznaků při vytváření modelu. Je vidět, že nejvyšší významnost je přikládána příznaku, který obsahuje počet kreditů získaných studentem v prvním semestru. Obecně byl tento příznak považován za velmi důležitý i pro ostatní vytvořené modely. Například model AdaBoost využíval pouze dvou příznaků – *pocet_kreditu_semestr_1* s *feature importance* 0,8 a *BI-CAO_znamka* s důležitostí 0,2.

Tabulka 6.8 Klasifikační přesnost na testovací sadě jednotlivých modelů u predikce úspěšnosti druhého semestru

Využitá metoda	Klasifikační přesnost
Rozhodovací strom	87,4
Náhodný les	87,7
AdaBoost	86,6
XGBoost	86,2
Vícevrstvý perceptron	87,2
kNN	85,0

6.2.3 Třetí semestr

Původní dataset obsahoval 2 049 záznamů. Tabulka 6.10 shrnuje vlastnosti vytvořeného trénovacího a testovacího datasetu. Z tabulky si můžeme povšimnout, že dataset obsahuje velmi malé

Tabulka 6.9 Nenulové feature importance u modelu náhodného lesa u predikce úspěšnosti druhého semestru

Příznak	Feature importance
pocet_kreditu_semestr_1	0,94
BI-PA1_znamka	0,02
BI-ZMA_znamka	0,02
BI-PS1_znamka	0,01
BI-CAO_znamka	0,01

procento studentů, kteří třetím semestrem neprojdou. Vzhledem k tomu, že se jedná o zimní semestr, ve kterém se při tvorbě datasetu nekontrolovaly získané kredity, ale pouze datum ukončení (problémy s datem ukončení jsou popsány v sekci 5.4), je možné, že byli někteří studenti identifikováni špatně. Velmi malé procento neúspěšných studentů se promítlo i do vytváření jednotlivých modelů.

Tabulka 6.10 Popis trénovacího a testovacího datasetu pro třetí semestr

	Trénovací	Testovací
počet záznamů	1 639	410
procento záznamů z původního datasetu	80	20
procentuální zastoupení *dokoncil_semestr* = 1	93,47	93,41
procentuální zastoupení *dokoncil_semestr* = 0	6,53	6,59

Tabulka 6.11 Klasifikační přesnost na testovací sadě jednotlivých modelů u predikce úspěšnosti třetího semestru

Využitá metoda	Klasifikační přesnost
Rozhodovací strom	93,4
Náhodný les	93,4
AdaBoost	93,2
XGBoost	93,2
Vícevrstvý perceptron	93,9
kNN	93,4

Výsledky jednotlivých metod shrnuje tabulka 6.11. Z tabulky je vidět, že u predikce úspěchu ve třetím semestru mají hned tři modely stejnou přesnost na testovací sadě dat a to 93,4 %, jak ale víme, toto číslo odpovídá procentuálnímu zastoupení úspěšných studentů v testovací sadě dat. Bylo tedy nutné zjistit, za jakých podmínek modely predikují neúspěch a s jakou přesností. Po průzkumu dat bylo zjištěno, že pouze dva modely nepredikovaly ve všech případech úspěch. Jednalo se o modely AdaBoost a model vícevrstvého perceptonu. Jako nejlepší se tedy ukázal model AdaBoost, který lépe identifikoval neúspěšné studenty. Ač tedy model AdaBoost vykazoval menší přesnost než ostatní modely na testovací sadě dat, byl zvolen jako výsledný.

6.2.4 Čtvrtý semestr

Původní dataset obsahoval 1 616 záznamů. Tabulka 6.12 shrnuje vlastnosti vytvořeného trénovacího a testovacího datasetu.

Výsledky evaluace modelů shrnuje tabulka 6.13. Modely rozhodovacího stromu a náhodného lesa dosahovaly nejlepších výsledků na testovací sadě dat. Jejich klasifikační přesnost byla 88,6 %.

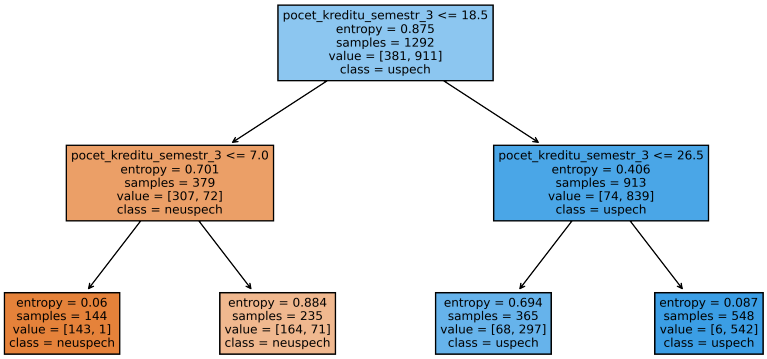
Tabulka 6.12 Popis trénovacího a testovacího datasetu pro čtvrtý semestr

	Trénovací	Testovací
počet záznamů	1 292	324
procento záznamů z původního datasetu	79,95	20,05
procentuální zastoupení *dokoncil_semestr* = 1	70,51	70,68
procentuální zastoupení *dokoncil_semestr* = 0	29,49	29,32

Tabulka 6.13 Klasifikační přesnost na testovací sadě jednotlivých modelů u predikce úspěšnosti čtvrtého semestru

Využitá metoda	Klasifikační přesnost
Rozhodovací strom	88,6
Náhodný les	88,6
AdaBoost	87,7
XGBoost	86,1
Vícevrstvý perceptron	84,9
kNN	85,8

Na obrázků 6.2 můžeme viděl vytvořený rozhodovací strom. Oranžově jsou značeny uzly, ve kterých převažují neúspěšní studenti, modře uzly, ve kterých převažují úspěšní studenti. Z obrázku si můžeme povšimnout, že model využíval pouze jediného příznaku *pocet_kreditu_semestr_3*, tedy počtu kreditů dosažených ve třetím semestru. V kořenu stromu je podmínka, říkající, že pokud student nezískal ve třetím semestru alespoň 18,5 kreditů, je klasifikován jako neúspěšný, v opačném případě jako úspěšný.



Obrázek 6.2 Rozhodovací strom u predikce postupu čtvrtým semestrem

Model náhodného lesa se jeví více stabilněji díky využívání většího množství příznaků nežli pouze počtu kreditů ze třetího semestru, jak lze vyčíst z tabulky 6.14 zobrazující 10 příznaků s nejvyšší důležitostí. Je vidět, že model taktéž považuje počet získaných kreditů ze třetího semestru za nejdůležitější příznak – *feature importance* 0,35. Kromě toho ovšem využívá i řady dalších. Jako druhý nejdůležitější příznak byla zvolena známka z předmětu BI-ZDM s důležitostí

0,13 a jako třetí počet kreditů získaný ve druhém semestru s důležitostí 0,12. Pětici nejdůležitějších příznaků potom uzavírají známky z předmětů BI-AG1 (důležitost 0,09) a BI-LIN (důležitost 0,08).

Tabulka 6.14 Deset příznaků s nejvyšší feature importance u modelu metody Náhodný les u predikce úspěšnosti čtvrtého semestru

Příznak	Feature importance
pocet_kreditu_semestr_3	0,35
BI-ZDM_znamka	0,13
pocet_kreditu_semestr_2	0,12
BI-AG1_znamka	0,09
BI-LIN_znamka	0,08
BI-AAG_znamka	0,06
BI-AG1_poradi_zapisu	0,04
BI-ZDM_poradi_zapisu	0,04
BI-ZMA_znamka	0,03
pocet_kreditu_semestr_1	0,01

6.2.5 Pátý semestr

U predikce pátého semestru byly zjištěny závažné problémy, na základě nichž bylo rozhodnuto vynechat predikci pátého semestru. Nejdříve popišme charakteristiky datasetu. Dataset obsahoval 1 140 záznamů. Tabulka 6.15 shrnuje vlastnosti vytvořeného trénovacího a testovacího datasetu. Jak je z tabulky vidět, dataset pro predikci pátého semestru obsahoval extrémně málo studentů, kteří nedostudovali úspěšně (například v testovací sadě dat se jedná o pouhých 6 studentů). Možných příčin je několik. Za prvé se jedná o semestr v posledním ročníku standardní délky studia. Ve třetím ročníku již obecně není zaznamenáno takové množství studentů, kteří nedostudovali úspěšně. Zároveň se jedná o zimní semestr, na jehož konci není kontrola splněného množství kreditů, tudíž neúspěšnost studentů není tak vysoká jako v semestrech letních. Samozřejmě musíme také brát v potaz fakt, že informace o tom, zda student semestrem projde, je zjištěna pouze na základě data o ukončení studia – problémy s tím související jsou popsány detailně v sekci 5.4, víme tedy, že datum ukončení studia není velmi spolehlivým měřítkem.

Tabulka 6.15 Popis trénovacího a testovacího datasetu pro pátý semestr

	Trénovací	Testovací
počet záznamů	912	228
procento záznamů z původního datasetu	80	20
procentuální zastoupení *dokoncil_semestr* = 1	97,48	97,37
procentuální zastoupení *dokoncil_semestr* = 0	2,52	2,63

Velmi malé množství neúspěšných studentů se projevilo na tvorbě prediktivních modelů, řada z nich totiž stejně jako u predikce průchodu třetím semestrem předpovídala úspěch u všech studentů. Modely, které předpovídali také neúspěch, i když ve velmi malém množství případů, byly postaveny na metodách AdaBoost a kNN. Ukázalo se ovšem, že se jednalo o přeučení na trénovacím datasetu. Oba modely totiž dokázaly dobře identifikovat v testovacích datech pouze jediný záznam neúspěšného studenta. K přeučení zřejmě také napomohlo celkově malé množství dat původního datasetu. Z těchto důvodů bylo rozhodnuto vynechat predikci dokončení pátého semestru.

6.3 Predikce známek

Pro predikování známek v jednotlivých povinných předmětech byl pro každý předmět vytvořen odpovídající dataset, jak je popsáno v sekci 5.3. V každém datasetu je jako predikovaná proměnná vybrán konkrétní předmět, u všech ostatních předmětů jsou vyplněny chybějící předměty pomocí pomocí třídy *KNNImputer* z knihovny *scikit-learn* a jednotlivé příznaky jsou přetypovány na kategorické ordinální prostřednictvím třídy *CategoricalDtype*. Před samotným trénováním modelů jsou datasety rozděleny na trénovací a testovací sadu dat v poměru 4 : 1 pomocí metody *train_test_split* z knihovny *scikit-learn*, parametr *shuffle* byl nastaven na `True`.

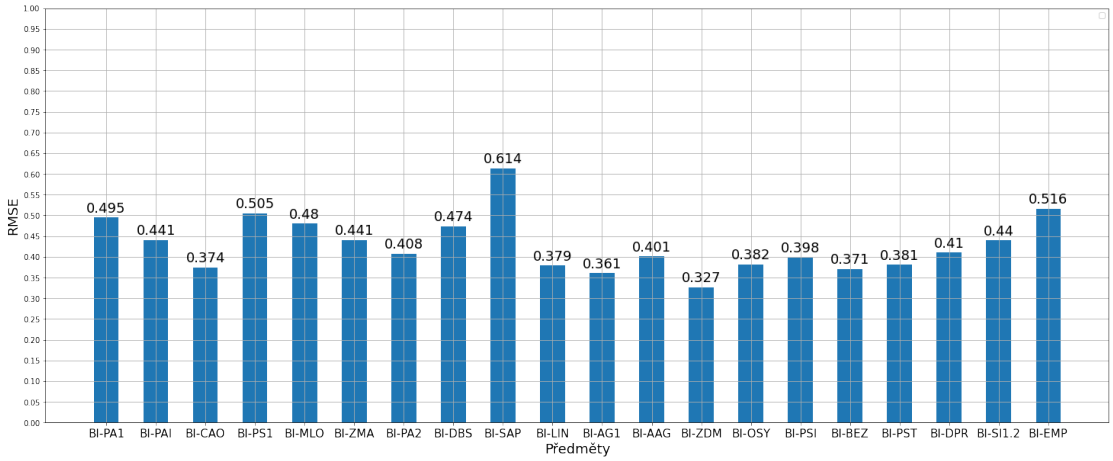
6.3.1 Trénování modelu

Na základě provedené rešerše bylo rozhodnuto, že predikce známek z povinných předmětů bude považována za regresní úlohu i přesto, že výsledné proměnné známek mají pouze omezený počet výsledných hodnot. Musíme ale vzít v potaz fakt, že mezi známkami existuje pořadí a tedy nezáleží pouze na tom, zda budeme predikovat správně výslednou známku, ale také na tom, jak moc se naše predikce od výsledné známky bude odlišovat. Uveďme pro lepší vhled do problematiky jednoduchý příklad. Pokud by skutečná hodnota výsledné proměnné u predikovaného předmětu byla 1 (známka A, tedy nejlepší možná) je velký rozdíl, zda bude naše predikovaná hodnota známky 1.5 (B), či 4 (F – student neuspěl).

Jako metoda pro tvorbu prediktivních modelů bylo vybráno kNN, přesněji implementace třídou *KNeighborsRegressor* z knihovny *scikit-learn*. Prediktivní modely byly vytvořeny a laděny na trénovacím datasetu pomocí 5-násobné křížové validace. K tomu bylo využito třídy *GridSearchCV* z knihovny *scikit-learn*.

6.3.2 Vyhodnocení

Výsledná RMSE hodnota na příslušných testovacích sadách dat je zobrazena na grafu 6.3. Nejlepších výsledků dosahují predikce známek v předmětech BI-ZDM, BI-AG1, BI-BEZ, BI-CAO a BI-LIN. Naopak nejméně přesné jsou predikce v předmětech BI-SAP, BI-EMP a BI-PS1. Průměrná RMSE hodnota je 0,430.



Obrázek 6.3 Výsledné hodnoty RMSE jednotlivých modelů predikce známek

6.3.3 Možné problémy

Vzhledem k poměrně vysoké nevyplněnosti předmětů s průchodem doporučeným v pozdějších semestrech, bylo nutné velké množství záznamů známek z těchto předmětů doplnit. Doplnění známek bylo provedeno pomocí třídy *KNNImputer* z knihovny *scikit-learn*. Výsledné predikce tím tak mohou být ovlivněny. Z těchto důvodů považuji za relevantní hlavně predikce výsledků u předmětů s doporučeným průchodem v prvním roku studia, neboť jejich vyplněnost byla vysoká, jak je popsáno v sekci 5.1.3.

6.4 Diskuse

V budoucí práci se bude nutné zaměřit na novou akreditaci 2021. K tomu bude ovšem potřeba mít dostupná data z minimálně několika let výuky – jak již bylo zmíněno v sekci 3.3, jedná se o akreditaci s nejzásadnějšími změnami od vzniku fakulty. Na základě těchto dat bude muset být provedena důkladná analýza, pomocí níž bude rozhodnuto, jak propojit nově vzniklé předměty s předměty starými. Pro dobrou predikci v prvních letech akreditace, kdy ještě nebude možné postavit modelování pouze na nových datech, kterých bude nedostatečné množství, budou totiž data z momentální akreditace 2015 potřeba. Bude potřeba vytvořit dataset sdružující studenty staré i nové akreditace, příznaky budou tvořeny předměty nové akreditace, bude tedy nutné najít způsob, jak doplnit hodnoty v nových předmětech studentům, kteří je v rámci staré akreditaci neměli šanci absolvovat. U některých předmětů bude propojení snadné, neboť jejich formát se s novou akreditací změnil minimálně, u jiných, a to zejména matematických předmětů, bude propojení poměrně obtížné.

Možné řešení u sloučených předmětů (např. BI-DML.21, které vzniklo sloučením BI-MLO a BI-ZDM) by mohlo spočívat ve hledání vah pro známky z původních předmětů, tak aby vážený průměr známek těchto předmětů mohl být považován za známku v předmětu novém.

U některých předmětů staré akreditace došlo k rozdělení do více předmětů a rozšíření vyučované látky. Příkladem nám mohou být předměty staré akreditace BI-ZMA a BI-LIN. U předmětu BI-ZMA došlo k rozdělení do dvou nových předmětů BI-MA1.21 a BI-MA2.21. Je stanoveno, že pokud student absolvuje předmět BI-ZMA, může mu být uznán předmět BI-MA1.21. Doplnění známek v předmětu BI-MA2.21 by mohlo spočívat například ve hledání souvislostí mezi známkami z ostatních matematických předmětů a známkou z předmětu BI-MA2.21. K tomu ale bude nutné mít k dispozici adekvátní množství záznamů studentů nové akreditace. U předmětu BI-LIN došlo k rozdělení předmětu do BI-LA1.21 a BI-LA2.21. Povinným předmětem pro všechny specializace je však pouze předmět BI-LA1.21. Platí, že po absolvování předmětu BI-LIN může být studentovi předmět BI-LA1.21 uznán.

Zajímavé by také bylo, pokud by se nová práce měla zaměřit i na oborové předměty/specializované předměty. Jelikož došlo k velkým změnám u jednotlivých oborů a vzniklo navíc i velké množství zcela nových, nebude již zřejmě možné u většiny nových specializací použít data studentů staré akreditace. O to větší zásoba záznamů studentů nové akreditace bude ke kvalitní predikci potřeba. Minimálně v prvních letech nové akreditace budou tedy predikce jen těžko proveditelné.

Kapitola 7
Automatizace

Před samotnou automatizací bylo nutné rozhodnout, které modely budou použity. Na tomto místě je nutné si nejprve uvědomit, pro jaké studenty budeme predikce vytvářet. Víme, že nás zajímají studenti spadající do akreditace 2015, tedy takoví studenti, jejichž datum nástupu patří do akademických let 2015/2016 až 2020/2021. Od roku 2021/2022 všichni nově přihlášení studenti automaticky spadají do nové akreditace 2021. Veškeré informace o akreditacích na Fakultě informačních technologií jsou popsány v sekci 3.3. V okamžiku tvorby práce je v běhu letní semestr B212. Z toho vyplývá, že studenti, u kterých můžeme predikovat jejich studijní výkony se momentálně nachází minimálně ve druhém ročníku studia, konkrétně právě dokončují čtvrtý semestr studia. Predikce průchodu prvním, druhým i třetím semestrem tedy nepřipadá v úvahu, neboť nejsou studenti, pro které bychom predikce vytvářeli. Ze sekce 3.2.1 víme, že největší množství studentů nedokončí své studium právě v prvních letech studia. Z těchto důvodů byly pro automatizaci vybrány predikce úspěšného dokončení studia a predikce dokončení čtvrtého semestru studia.

Automatizace je psána (stejně jako celá předchozí práce) v jazyce *Python* a skládá se zejména z následujících souborů:

config.py V souboru jsou uloženy konfigurační proměnné, například jméno a heslo k autentizaci u endpointů.

download_endpoints.py Skript je zodpovědný za stažení čerstvých dat z potřebných endpointů a uložení na serveru.

preprocessing.py Skript z dat filtruje aktuální studenty, předzpracovává data způsobem uvedeným v předchozích kapitolách a ukládá finální datasety k predikci.

model.py Skript provede predikci úspěšnosti dokončení studia na modelu metody AdaBoost a pro aktuální studenty druhého ročníku predikci úspěšnosti dokončení čtvrtého semestru na modelu náhodného lesa. Výsledek uloží do souboru.

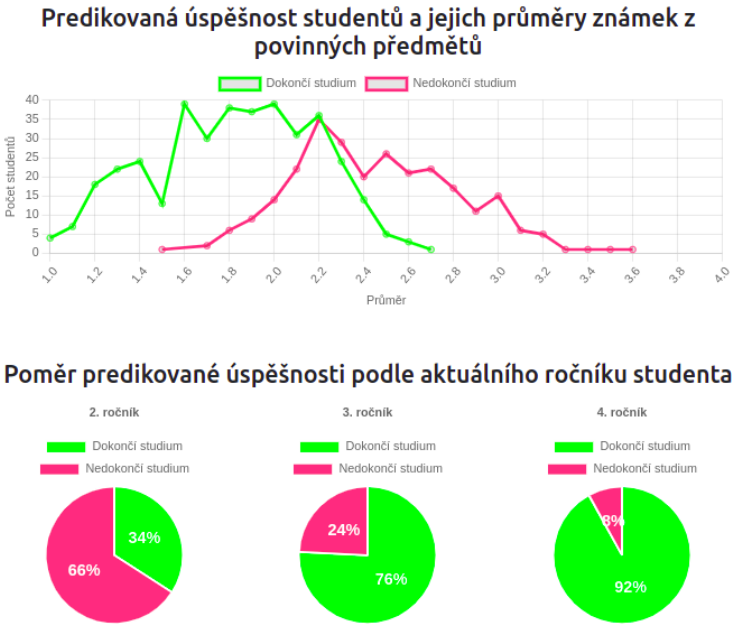
generate_html.py Skript zpracuje výsledky predikce a vygeneruje finální statické HTML stránky z šablon pomocí šablonovacího nástroje *Jinja*[1].

main.py Skript volá postupně funkce z výše uvedených jednotlivých skriptů, po jeho spuštění dojde k aktualizaci stránek novými daty.

Výsledné skripty a HTML webová prezentace byla nasazena na CloudFIT[2], momentálně k vytvořené stránce mají přístup vedoucí a oponent práce s možností rozšíření přístupu i dalším

[1] https://jinja.palletsprojects.com/en/3.1.x/
[2] https://cloud.fit.cvut.cz/

kantorům či studentům z fakulty. Obsah stránky je automaticky aktualizován každý týden – k tomu slouží utilita *cron*, která pravidelně spouští skript `main.py`. Predikce jsou vytvářeny pro studenty, u kterých ještě nevíme jejich studijní výsledky, tedy studenti, kteří nebyli součástí trénovacích a testovacích datasetů při tvorbě patřičných modelů. Součástí stránky jsou predikované výsledky studentů, jak můžeme vidět na obrázcích 7.2 a 7.4, a také grafy shrnující výkony a predikovanou průchodnost čtvrtým semestrem, jak si můžeme prohlédnout na obrázcích 7.1 a 7.3.

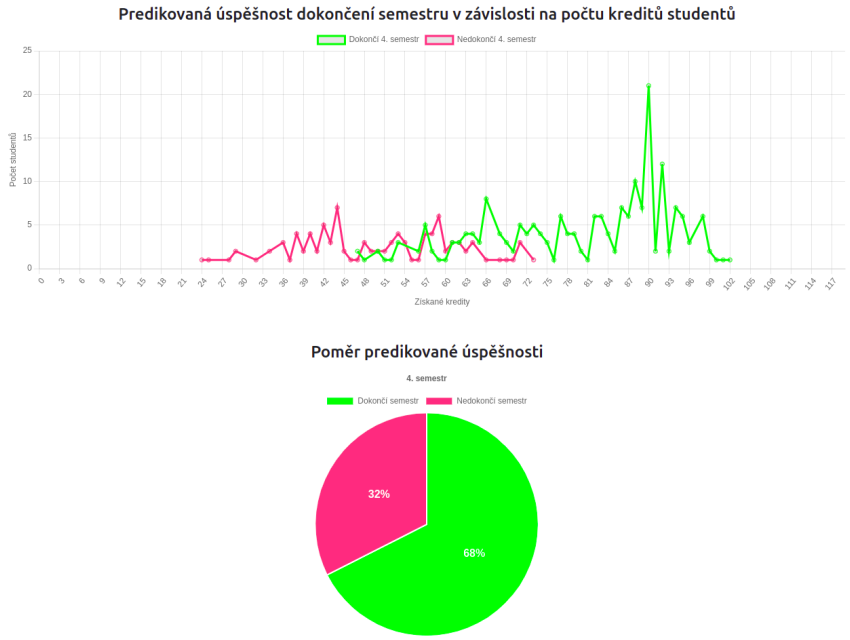


Obrázek 7.1 Ukázka grafů shrnující výsledky studentů při průchodu studiem

Kompletní data predikce úspěšnosti studentů

Studium ID	Ročník	Průměr	BI-PA1	BI-PAI	BI-CAO	BI-PS1	BI-MLO	BI-ZMA	BI-PA2	BI-DBS	BI-SAP	BI-LIN	BI-AG1	BI-AAG	BI-ZDM	BI-OSY	BI-PSI	BI-BEZ	BI-PST	BI-SI1.2	BI-EMP	Dostuduje?
[illegible]506	2.0	2.46	2.5	1.0	1.0	2.0	2.5	2.0	4.0	1.5	2.5	4.0	-	4.0	2.5	-	-	-	-	-	-	×
[illegible]606	3.0	2.28	3.0	2.5	1.0	2.0	1.5	2.5	2.0	2.5	2.0	3.0	3.0	2.5	2.5	-	3.0	1.5	3.0	2.0	1.5	✓
[illegible]506	2.0	2.36	3.0	2.5	2.0	1.5	1.5	2.5	2.5	2.5	1.5	2.0	3.0	2.5	3.0	-	2.5	2.5	2.5	3.0	2.0	✓
[illegible]306	5.0	2.32	3.0	2.5	1.5	3.0	3.0	1.5	2.5	3.0	1.0	2.0	3.0	2.0	2.5	2.5	3.0	1.5	2.5	2.5	1.5	✓
[illegible]306	2.0	2.95	4.0	1.0	3.0	1.5	2.0	4.0	-	3.0	2.0	4.0	-	4.0	4.0	-	-	-	-	-	-	×
[illegible]206	2.0	2.16	1.0	1.5	1.5	2.0	1.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.0	3.0	2.0	1.5	✓
[illegible]606	2.0	2.19	1.5	2.0	1.0	1.5	2.5	2.0	4.0	1.5	1.5	3.0	2.5	3.0	2.5	-	-	-	-	-	-	×
[illegible]406	2.0	1.23	2.0	1.5	1.0	1.0	1.0	1.5	1.5	1.0	1.0	1.0	1.0	1.5	1.0	-	-	-	-	-	-	✓
[illegible]506	3.0	2.28	3.0	2.0	1.0	2.5	1.5	2.5	2.0	2.5	2.5	3.0	2.5	2.5	2.5	2.5	2.0	2.0	3.0	1.5	-	✓
[illegible]406	2.0	2.69	4.0	2.0	1.0	2.5	2.5	2.0	4.0	2.0	3.0	3.0	4.0	2.5	2.5	-	-	-	-	-	-	×
[illegible]706	2.0	2.00	2.0	2.5	1.0	2.0	1.5	1.5	2.5	1.5	3.0	2.0	2.5	2.0	2.0	-	-	-	-	-	-	✓
[illegible]106	3.0	2.00	2.5	1.5	1.0	2.0	1.5	1.5	4.0	1.5	2.5	2.0	3.0	2.0	2.0	-	-	1.5	2.5	2.0	1.0	×
[illegible]806	2.0	2.32	1.0	1.5	1.0	1.0	1.5	2.0	4.0	1.5	1.5	4.0	4.0	4.0	4.0	-	-	-	-	-	1.5	×
[illegible]106	2.0	2.27	4.0	2.0	1.0	2.0	1.0	1.5	4.0	1.5	1.5	2.0	4.0	3.0	2.0	-	-	-	-	-	-	×

Obrázek 7.2 Ukázka zobrazení výsledků predikce úspěšného dokončení studia



Obrázek 7.3 Ukázka grafů shrnující výsledky studentů při průchodu čtvrtým semestrem

Kompletní data predikce úspěšnosti studentů

Studium ID	BI-PA1	BI-PAI	BI-CAO	BI-PS1	BI-MLO	BI-ZMA	BI-PA2	BI-DBS	BI-SAP	BI-LIN	BI-AG1	BI-AAG	BI-ZDM	Kredity 1. semestr	Kredity 2. semestr	Kredity 3. semestr	Dokončí semestr?
[redacted]506	2.5	1.0	1.0	2.0	2.5	2.0	4.0	1.5	2.5	4.0	-	4.0	2.5	24	12	25	✓
[redacted]506	3.0	2.5	2.0	1.5	1.5	2.5	2.5	2.5	1.5	2.0	3.0	2.5	3.0	37	20	0	×
[redacted]306	3.0	1.0	3.0	1.5	2.0	4.0	-	3.0	2.0	4.0	-	4.0	4.0	13	16	11	×
[redacted]206	1.0	1.5	1.5	2.0	1.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.5	19	31	8	×
[redacted]606	1.5	2.0	1.0	1.5	2.5	2.0	4.0	1.5	1.5	3.0	2.5	3.0	2.5	24	19	23	✓
[redacted]406	2.0	1.5	1.0	1.0	1.0	1.5	1.5	1.0	1.0	1.0	1.0	1.5	1.0	30	32	30	✓
[redacted]406	2.5	2.0	1.0	2.5	2.5	2.0	4.0	2.0	3.0	3.0	4.0	2.5	2.5	18	23	23	✓
[redacted]706	2.0	2.5	1.0	2.0	1.5	1.5	2.5	1.5	3.0	2.0	2.5	2.0	2.0	24	28	36	✓
[redacted]806	1.0	1.5	1.0	1.0	1.5	2.0	4.0	1.5	1.5	4.0	4.0	4.0	4.0	30	17	14	×
[redacted]106	2.5	2.0	1.0	2.0	1.0	1.5	4.0	1.5	1.5	2.0	4.0	3.0	2.0	24	23	27	✓
[redacted]506	1.5	2.0	1.0	1.5	1.5	2.0	2.5	1.5	1.5	2.5	2.5	2.0	2.5	30	32	26	✓
[redacted]006	1.0	3.0	1.0	1.0	2.5	2.5	2.0	1.5	1.0	4.0	4.0	3.0	4.0	19	23	17	×
[redacted]106	1.5	1.5	1.5	2.5	1.5	2.5	-	2.5	3.0	3.0	-	4.0	4.0	19	23	19	✓
[redacted]406	2.0	1.5	1.0	1.0	1.5	2.0	1.5	1.5	1.5	4.0	4.0	4.0	4.0	24	25	20	✓
[redacted]006	2.0	2.0	1.0	2.0	1.5	2.5	2.0	1.5	1.0	4.0	-	-	-	19	19	19	✓
[redacted]706	2.0	2.0	1.0	2.0	1.5	1.5	4.0	2.0	1.5	2.0	4.0	3.0	2.5	21	25	20	✓
[redacted]606	2.5	2.0	1.0	2.5	2.5	2.5	1.0	1.0	2.0	2.0	2.5	2.5	2.5	18	20	0	×

Obrázek 7.4 Ukázka zobrazení výsledků predikce průchodu čtvrtým semestrem

Kapitola 8

Závěr

Cílem práce bylo prozkoumání možností predikování výsledků studentů na vybraných datech získaných z datového skladu ČVUT na základě rešerše problematiky *educational data miningu*. Primárně byla k dispozici data ze systému KOS, který udržuje studijní prospěch studenta. To se povedlo a byly vytvořeny prediktivní modely, které predikují úspěšné dokončení studia, známky v povinných předmětech a průchodnost studentů v prvních čtyřech semestrech studia.

Na začátku práce bylo velmi důležité dokázat proniknout do fungování fakulty v průběhu let. Velká část informací však nebyla dobře dostupná a s přibývajícími lety fakulty ubývalo i starších studentů, u kterých bylo možno získat vhled do minulosti fakulty. Pochopení této problematiky bylo ale zásadní pro následnou práci s daty i vytváření prediktivních modelů a jejich interpretaci. Velmi důležité bylo pochopit, jak na sebe navazovaly jednotlivé verze povinných předmětů, co přinesly jednotlivé akreditace, či jaké důsledky na studijní výsledky měl rok a půl výuky v distančním režimu způsobený epidemií COVIDu. Ukázalo se například, že vliv distanční výuky se negativně projevil zejména v akademickém roce 2020/2021, naopak v prvním půlroce distanční výuky se u řady předmětů výkon studentů spíše zlepšil.

Dále bylo potřeba prozkoumat možnosti pro výběr sociodemografických příznaků, které by mohly mít vliv na výsledky studentů. Jako velmi důležitý se ukázal rozdíl mezi rokem nástupu studenta a rokem absolvování maturitní zkoušky, ten byl ostatně použit i při predikci úspěšného dokončení studia. Z průzkumu vyplynulo, že čím větší je časové okno mezi maturitou a nástupem, tím méně úspěšní studenti jsou. Největší šanci na úspěch mají studenti, kteří nastoupí ve stejném roce, ve kterém také odmaturují. Zaznamenány byly i rozdíly ve studijním prospěchu žen a mužů – ukázalo se ženy mají nižší průchodnost studiem nežli muži. Na bakalářském studiu je signifikantní horší průměr známek žen ve většině povinných předmětů, zajímavý je obrat na studiu magisterském, kde mají naopak ve většině povinných předmětů lepší průměr ženy. Taktéž byl zkoumán rozdíl prospěchu mezi českými a zahraničními studenty, kteří studují v českém jazyce. Jako nejúspěšnější se ukázali být studenti ze Slovenska a naopak jako výrazně nejhorší studenti z Ruska. Zajímavý byl také rozdíl mezi studenty z gymnázií a studenty z ostatních středních škol. Ukázalo se, že gymnazisté s rokem nástupu před rokem 2015 byli zřetelně úspěšnější oproti ostatním studentům, tento rozdíl je ale u studentů s rokem nástupu po roce 2015 takřka minimální.

Pro předzpracování dat bylo nutné se vypořádat s nedobrou kvalitou poskytnutých dat, jejich absencí v průběhu práce a obtížnou interpretací některých atributů tabulek. Předzpracování zabralo nejvíce času vypracování práce, zároveň s měnícími se poskytnutými daty v průběhu práce bylo nutné se k předzpracování vracet a provádět jej znovu. Pro každou jednotlivou predikci byl vytvořený odpovídající dataset, který reflektoval poznatky získané z rešeršní práce.

Při vytváření prediktivních modelů se muselo přistupovat adekvátním způsobem ke každé predikci zvlášť a hledat různé způsoby, jak dosáhnout co nejlepších výsledků. Stejně tak muselo

být bráno v potaz i dobré vyhodnocení výsledků, tak aby nedošlo k přeučení modelů. Například u predikce úspěchu ve studiu bylo nutné najít cestu, jak netestovat výsledky pouze na kompletních datech, ale také na datech studentů bez známek ze všech dokončených semestrů. Jedině tak se daly získat výsledky odpovídající skutečným datům. Nakonec se podařilo dopracovat k velmi dobrým výsledkům, které by mohly přispět k lepšímu pochopení problematiky neúspěšnosti studentů.

Práce se zaměřovala zejména na studijní výsledky studentů bakalářského programu Informatika a ubírala se do třech směrů. Úspěšné dokončení studia se povedlo predikovat s přesností 75,3 % již po prvním dokončeném semestru studia. Po roce studia již bylo možné provádět predikci s přesností 87,7 %. Po druhém ukončeném roku má nejlepší metoda přesnost 93,4 %. Jako nejlepší se ukázal model metody AdaBoost. U predikce úspěšných dokončení jednotlivých semestrů se podařilo na základě výsledků z předchozích semestrů určit s přesností 87,7 % úspěch dokončení druhého semestru, 93,2 % u třetího semestru a 88,6 % u čtvrtého. Predikce průchodu prvním semestrem byla postavena na datasetu, který se skládal z pouze sociodemografických dat, které víme o studentovi ještě před jeho nástupem na fakultu. Tato predikce dosahovala klasifikační přesnosti 56,8 % na modelu metody AdaBoost, ukázalo se tak, že pouze sociodemografická data o studentech nejsou dostatečným měřítkem studentova budoucího akademického výkonu. To se ostatně projevilo i u predikce úspěšného dokončení celého studia, kdy naprostá většina dat této skupiny nedosahovala vysoké *feature importance*. U predikování známek z povinných předmětů se pomocí regresních metod povedlo predikovat výsledky předmětů s průměrnou RMSE chybou 0,43.

Dalším úkolem práce bylo vytvořit automatizaci procesu stažení aktualizovaných dat, jejich předzpracování a aplikaci vybraných metod predikování. Toho bylo dosaženo a práce je nyní dostupná pro vedoucí a oponenta práce. Automatizovány jsou momentálně pro ukázku predikce úspěšného dokončení studia a predikce úspěšného dokončení čtvrtého semestru. Predikce jsou prováděny pro odpovídající skupiny aktuálních studentů, u kterých ještě nevíme jejich úspěšnost. Výsledky, které jsou momentálně přístupné přes vytvořenou webovou stránku, tak mohou být využity pro pomoc příslušným studentům.

Do budoucna bude důležité zaměřit se na novou bakalářskou akreditaci a její vliv na výsledky studentů. Jelikož letos došlo ke spuštění nové akreditace, nejsou ještě k dispozici patřičná data, na kterých by bylo možné práci postavit. Kvůli absenci dat o oborech studentů se práce zaměřovala pouze na povinné předměty, tudíž by mohlo být rovněž zajímavé se v budoucí práci zaměřit i na obory a specializace.

Příloha A

Tabulka A.1 Ukázka atributů a jejich vyplněnosti tabulky prihlaska_dim

Název atributu	Vyplněnost v %
cislo_prihlasky	100,0
odkud_uchazec_prihlasen_kod	46,389
prihlaska_kod	100,0
studium_id	100,0
rozhodnuti_kod	99,902
obor_ss_kod	25,625
ss_izo	35,938
ss_typ	35,682
ss_predmet1	1,279
ss_predmet2	1,201
ss_predmet3	0,02
ss_predmet4	0,02
ss_predmet5	0,0
ss_predmet1_znamky	0,63
ss_predmet2_znamky	0,551
ss_predmet3_znamky	0,02
ss_predmet4_znamky	0,02
ss_predmet5_znamky	0,0
ss_predmet1_prumer	0,12
ss_predmet2_prumer	0,12
ss_predmet3_prumer	0,06
ss_predmet4_prumer	0,06
ss_predmet5_prumer	0,0
ss_prumer_prumeru	0,06
ss_prumer_znamek	0,63
bakalar_prumer	0,12
predchozi_studium	21,787
doplnujici_udaj1	0,0
doplnujici_udaj2	0,0
doplnujici_udaj3	0,0
doplnujici_udaj4	0,512

Tabulka A.1 Ukázka atributů a jejich vyplněnosti tabulky prihlaska_dim

Název atributu	Vyplněnost v %
typ_prihlasky	100,0
maturita_znamky	0,0
maturita_prumer	0,0
maturita_predmety	0,039
maturita_rok	53,257
cislo_mistnosti_pz	0,0
datum_pz	0,0
prominuti_pz	0,354
hodnoceni_pz	70,518
hodnoceni_pz_cast1	1,397
hodnoceni_pz_cast2	9,801
hodnoceni_pz_cast3	41,822
hodnoceni_pz_cast4	15,981
hodnoceni_pz_cast5	41,822
hodnoceni_pz_cast6	2,913
hodnoceni_pz_cast7	16,985
hodnoceni_pz_cast8	5,708
hodnoceni_pz_cast9	3,739
hodnoceni_pz_cast10	0,0
hodnoceni_pz_cast11	0,0
hodnoceni_pz_cast12	0,0
zapsal_ke_studiu	42,255
status	0,0
stipendium	0,039
stipendium_cizi	0,0
kolej_zadost	0,512
termin	0,0
prechodne	0,315
studijni_skupina	18,54
nove_prijaty	73,332
navazujici_stud_program	16,414
trvaly_pobyt_v_cr	90,888
financovani	100,0
stupen_predchoziho_vzdelani	97,343
pz1_datum	0,0
pz1_cas	0,0
pz1_mistnost	0,0
pz2_datum	0,0
pz2_cas	0,0
pz2_mistnost	0,0
pz3_datum	0,0
pz3_cas	0,0
pz3_mistnost	0,0
pisemny_test	0,0
anketa_odkud_kod	2,007
anketa_odkud_poznamka	0,177
cislo_uchazece	97,54

Tabulka A.1 Ukázka atributů a jejich vyplněnosti tabulky prihlaska_dim

Název atributu	Vyplněnost v %
scio_test	20,213
olympiady	3,267
postizeni_kod	0,63
postizeni_poznamka	0,079
placeno_r	0,0
placeno_poznamka	0,0
placeno_mail	0,0
staz_stat	1,043
staz_poznamka	0,0
maturita_uroven	0,0
zajem	0,0
zprava_pro_uchazece	0,0
priorita	0,0
datum_registrace	96,772
datum_rozhodnuti	97,54
misto_kod	65,794
okres_kod	55,422
zeme_kod	65,105
psc	57,843

Bibliografie

1. MŠMT. *Studijní úspěšnost na českých vysokých školách v roce 2018* [online]. [B.r.] [cit. 2022-04-10]. Dostupné z: `https://www.msmt.cz/uploads/odbor_30/TF/Analyticke_materialy/Studijni_uspesnost_na_ceskych_vysokych_skolach_v_roce_2018.pdf`.
2. MŠMT. *Odbor statistiky, analýz a rozvoje eEducation* [online]. [B.r.] [cit. 2022-04-10]. Dostupné z: `https://dsia.msmt.cz//vystupy/vu_vs_f1.html`.
3. HAN, Jiawei; KAMBER, Micheline. *Data mining: concepts and techniques*. 2nd. San Francisco: Morgan Kaufmann, 2012. ISBN 9780123814791.
4. BAKER, RSJD et al. Data mining for education. *International encyclopedia of education* [online]. 2010, roč. 7, č. 3, s. 112–118 [cit. 2022-04-11]. Dostupné z: `https://www.upenn.edu/learninganalytics/ryanbaker/Encyclopedia%5C%20Chapter%5C%20Draft%5C%20v10%5C%20-fw.pdf`.
5. DUTT, Ashish; ISMAIL, Maizatul Akmar; HERAWAN, Tutut. A Systematic Review on Educational Data Mining. *IEEE Access*. 2017, roč. 5, s. 15991–16005. Dostupné z DOI: `10.1109/ACCESS.2017.2654247`.
6. PARK, Yeonjeong; YU, Ji Hyun; JO, Il-Hyun. Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute. *The Internet and Higher Education*. 2016, roč. 29, s. 1–11. ISSN 1096-7516. Dostupné z DOI: `https://doi.org/10.1016/j.iheduc.2015.11.001`.
7. PARACK, Suhem; ZAHID, Zain; MERCHANT, Fatima. Application of data mining in educational databases for predicting academic trends and patterns. In: *2012 IEEE International Conference on Technology Enhanced Education (ICTEE)* [online]. 2012, s. 1–4 [cit. 2022-04-11]. Dostupné z DOI: `10.1109/ICTEE.2012.6208617`.
8. HRUBÁ, Eliška. *Analýza výsledků absolventů středních škol na VŠ. Diplomová práce, České vysoké učení technické v Praze, Fakulta informačních technologií*. 2014.
9. BEN-GAL, Irad. Outlier detection. In: *Data mining and knowledge discovery handbook*. Springer, 2005, s. 131–146.
10. ASIF, Raheela; MERCERON, Agathe; ALI, Syed Abbas; HAIDER, Najmi Ghani. Analyzing undergraduate students' performance using educational data mining. *Computers & Education*. 2017, roč. 113, s. 178. ISSN 0360-1315. Dostupné z DOI: `https://doi.org/10.1016/j.compedu.2017.05.007`.
11. DATA SCIENCE PROCESS ALLIANCE. *What is CRISP DM?* [Online]. [B.r.] [cit. 2022-04-11]. Dostupné z: `https://www.datascience-pm.com/crisp-dm-2/`.

12. HUBER, Steffen; WIEMER, Hajo; SCHNEIDER, Dorothea; IHLENFELDT, Steffen. DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. *Procedia CIRP*. 2019, roč. 79, s. 403–408. ISSN 2212-8271. Dostupné z DOI: https://doi.org/10.1016/j.procir.2019.02.106. 12th CIRP Conference on Intelligent Computation in Manufacturing Engineering, 18-20 July 2018, Gulf of Naples, Italy.
13. WOODIE, Alex. *Data Prep Still Dominates Data Scientists' Time, Survey Finds* [online]. [B.r.] [cit. 2022-04-26]. Dostupné z: https://www.datanami.com/2020/07/06/data-prep-still-dominates-data-scientists-time-survey-finds/.
14. ANUNAYA, Sadhvi. *Data Preprocessing in Data Mining -A Hands On Guide* [online]. 2021 [cit. 2022-04-11]. Dostupné z: https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/.
15. VERLEYSEN, Michel; FRANÇOIS, Damien. The curse of dimensionality in data mining and time series prediction. In: *International work-conference on artificial neural networks*. 2005, s. 758–770.
16. SUNDARAM, Ramya Bhaskar. *An End-to-End Guide to Understand the Math behind XGBoost* [online]. [B.r.] [cit. 2022-05-03]. Dostupné z: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/.
17. MURALIDHAR, KSV. *What is Stratified Cross-Validation in Machine Learning?* [Online]. [B.r.] [cit. 2022-05-03]. Dostupné z: https://towardsdatascience.com/what-is-stratified-cross-validation-in-machine-learning-8844f3e7ae8e.
18. SAMMUT, Claude; WEBB, Geoffrey I. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.
19. JOSHI, Renuka. *Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures* [online]. [B.r.] [cit. 2022-04-25]. Dostupné z: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.
20. BROWNLEE, Jason. *Regression Metrics for Machine Learning* [online]. [B.r.] [cit. 2022-04-25]. Dostupné z: https://machinelearningmastery.com/regression-metrics-for-machine-learning/.
21. HELAL, Sumyea; LI, Jiuyong; LIU, Lin; EBRAHIMIE, Esmaeil; DAWSON, Shane; MURRAY, Duncan J. Identifying key factors of student academic performance by subgroup discovery. *International Journal of Data Science and Analytics*. 2019, roč. 7, č. 3, s. 227–245.
22. QUADRI, Mr MN; KALYANKAR, NV. Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*. 2010.
23. FELISONI, Daniel Darghan; GODOI, Alexandra Strommer. Cell phone usage and academic performance: An experiment. *Computers & Education*. 2018, roč. 117, s. 175–187. ISSN 0360-1315. Dostupné z DOI: https://doi.org/10.1016/j.compedu.2017.10.006.
24. COSTA, Evandro B.; FONSECA, Baldoino; SANTANA, Marcelo Almeida; FERREIRA, Fabrísia; REGO, Joilson. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*. 2017, roč. 73, s. 247–256. ISSN 0747-5632. Dostupné z DOI: https://doi.org/10.1016/j.chb.2017.01.047.
25. PRIYA, S.; ANKIT, T.; DIVYANSH, D. Student Performance Prediction Using Machine Learning. In: *Advances in Parallel Computing Technologies and Applications*. IOS Press, 2021, s. 167–174.
26. ASIF, Raheela; MERCERON, Agathe; ALI, Syed Abbas; HAIDER, Najmi Ghani. Analyzing undergraduate students' performance using educational data mining. *Computers & Education*. 2017, roč. 113, s. 177–194. ISSN 0360-1315. Dostupné z DOI: https://doi.org/10.1016/j.compedu.2017.05.007.

27. BYDŽOVSKÁ, Hana. A Comparative Analysis of Techniques for Predicting Student Performance. *International Educational Data Mining Society*. 2016. Dostupné také z: https://eric.ed.gov/?id=ED592637.

28. STATISTICS.COM. *Similarity Matrix* [online]. [B.r.] [cit. 2022-04-14]. Dostupné z: https://www.statistics.com/glossary/similarity-matrix/.

29. FIT ČVUT. *O fakultě FIT v čase* [online]. [B.r.] [cit. 2022-04-15]. Dostupné z: https://fit.cvut.cz/cs/fakulta/o-fakulte/fit-v-case.

30. JIŘINA, Marcel. *Směrnice děkana FIT ČVUT č. 30/2018 pro realizaci bakalářského a magisterského studijního programu Informatika na Fakultě informačních technologií ČVUT v Praze* [online]. [B.r.] [cit. 2022-04-15]. Dostupné z: https://fit.cvut.cz/studenti/bakalar-magistr/predpisy/smernice_bsp_msp.pdf.

31. PETRÁČEK, Vojtěch. *IV. změny Studijního a zkušebního řádu pro studenty ČVUT* [online]. [B.r.] [cit. 2022-04-15]. Dostupné z: https://www.cvut.cz/sites/default/files/content/74c76d2e-7f4d-4cb1-ac28-b0765c7f88f2/cs/20210910-studijni-a-zkusebni-rad-pro-studenty-cvut-v-praze-iv-zmeny-ucinnost-od-20-9-2021.pdf.

32. MŠMT. *AKREDITACE VZDĚLÁVACÍ INSTITUCE* [online]. [B.r.] [cit. 2022-04-24]. Dostupné z: https://www.msmt.cz/vzdelavani/dalsi-vzdelavani/akreditace-vzdelavaci-instituce.

33. FIT ČVUT. *Nová akreditace bakalářského studijního programu a změny ve výuce matematiky* [online]. [B.r.] [cit. 2022-04-24]. Dostupné z: https://fit.cvut.cz/cs/studium/pruvodce-studiem/bakalarske-a-magisterske-studium/nova-akreditace-bsp.

34. FRIEDJUNGOVÁ, Magda. *Predikce studijních výsledků studentů bakalářského programu Informatika FIT ČVUT*. 2016. Dostupné také z: https://dspace.cvut.cz/bitstream/handle/10467/65122/F8-DP-2016-Friedjungova-Magda-thesis.pdf?sequence=1%5C&isAllowed=y.

35. TAYLOR, David. *Difference Between Fact Table and Dimension Table* [online]. [B.r.] [cit. 2022-04-17]. Dostupné z: https://www.guru99.com/fact-table-vs-dimension-table.html.

36. ŠIRANCOVÁ, Katarína. *Ta Technika* [online]. [B.r.] [cit. 2022-04-19]. Dostupné z: https://www.cips.cvut.cz/2019/08/ta-technika/.

37. ODBOR ROZVOJE REKTORÁTU ČVUT A ČESKÁ TECHNIKA. *výroční zpráva o činnosti 2020* [online]. [B.r.] [cit. 2022-04-25]. Dostupné z: https://media.cvut.cz/sites/media/files/content/publications/e94be8c6-9c64-4c01-893c-c2da45fb97ec/946bee17-e7e5-432a-852a-5a4d39f99ea5.pdf.

38. SCIKIT-LEARN DEVELOPERS. *Permutation feature importance* [online]. [B.r.] [cit. 2022-05-03]. Dostupné z: https://scikit-learn.org/stable/modules/permutation_importance.html.

Obsah přiloženého média

- `readme.txt` ..... stručný popis obsahu média a spuštění souborů
- `src`
  - `notebooks` ..... Jupyter notebooky s předzpracováním a modelováním
  - `automatization` ..... skripty k automatickému generování html stránek
  - `requirements.txt` ..... potřebné knihovny ke spuštění souborů
  - `thesis` ..... zdrojová forma práce ve formátu LaTeX
- `text` ..... text práce
  - `thesis.pdf` ..... text práce ve formátu PDF