ŽILINSKÁ UNIVERZITA V ŽILINE

FAKULTA RIADENIA A INFORMATIKY

BAKALÁRSKA PRÁCA

ŠTUDIJNÝ ODBOR: INFORMATIKA

DOROTA KARDOŠOVÁ

Štatistické modelovanie dát v oblasti vzdelávania

Statistical modeling in education

Žilinská univerzita v Žiline
Fakulta riadenia a informatiky

Vedúci práce: Ing. Lucia Pančíková, PhD.
Registračné číslo: 1111/2019
Ministerské číslo práce: 28360520201111
Žilina, 2020

ŽILINSKÁ UNIVERZITA V ŽILINE

FAKULTA RIADENIA A INFORMATIKY

BAKALÁRSKA PRÁCA

ŠTUDIJNÝ ODBOR: INFORMATIKA

DOROTA KARDOŠOVÁ

Štatistické modelovanie dát v oblasti vzdelávania

Statistical modeling in education

Žilinská univerzita v Žiline
Fakulta riadenia a informatiky

Vedúci práce: Ing. Lucia Pančíková, PhD.
Dátum zadania práce: 6.12.2019
Dátum odovzdania práce: 6.5.2020
Žilina, 2020

ŽILINSKÁ UNIVERZITA V ŽILINE, FAKULTA RIADENIA A INFORMATIKY.

ZADANIE TÉMY BAKALÁRSKEJ PRÁCE.

Študijný program: Informatika

Meno a priezvisko
Dorota Kardošová

Osobné číslo
556516

Názov práce v slovenskom aj anglickom jazyku

Štatistické modelovanie dát v oblasti vzdelávania

Statistical modeling in education

Zadanie úlohy, ciele, pokyny pre vypracovanie (Ak je málo miesta, použite opačnú stranu)

Cieľ bakalárskej práce:
Na základe analýzy reálnych dát a ich väzieb implementovať vhodné štatistické prostriedky na skúmanie a vyhodnocovanie súvislostí, faktorov vplyvu v oblasti vzdelávania

Obsah:
1. Teoretické východiská.
2. Súčasné riešenia s prihliadnutím na vybrané štatistické a predikčné metódy s využitím programovacieho jazyka R.
3. Analýza reálneho súboru dát v oblasti vzdelávania, výber modelovacích a predikčných metód.
4. Návrh, tvorba aplikácie.
5. Implementácia riešenia, overenie a komparácia.
6. Vyhodnotenie a diskusia.

Meno a pracovisko vedúceho BP: Ing. Lucia Pančíková, PhD., KMME, ŽU
Meno a pracovisko tutora BP:

11 FEB. 2020

vedúci katedry
(dátum a podpis)

Zadanie zaregistrované dňa 06.12.2019 pod číslom 1111/2019 podpis

ČESTNÉ VYHLÁSENIE

Čestne vyhlasujem, že som túto bakalársku prácu vypracovala samostatne s využitím dostupných literárnych zdrojov, ktoré boli v práci riadne citované, ako aj teoretických poznatkov a praktických zručností nadobudnutých počas štúdia.

Žilina, 2020

.....

Dorota Kardošová

POĎAKOVANIE

V prvom rade chcem poďakovať vedúcej bakalárskej práce, Ing. Lucii Pančíkovej, PhD., za jej cenné rady, návrhy a odbornú pomoc pri vypracovaní tejto bakalárskej práce. Rovnako by som chcela poďakovať študentom Fakulty riadenia a informatiky za vyplnenie dotazníka a poskytnutie dát. Osobitné ďakujem patrí najmä mojej rodine a priateľom, za ich slová podpory a ich pomoc počas celého štúdia. V neposlednom rade chcem z celého srdiečka poďakovať Sárine, ktorá mi už len svojou prítomnosťou dokázala zlepšiť náladu, 16 rokov mi pomáhala nielen v štúdiu, ale bola tu pre mňa v každej situácii, radostnej či smutnej, a vždy sa tvárila, že dokonale rozumie všetkému, čo jej hovorím.

ABSTRAKT

KARDOŠOVÁ, Dorota: *Štatistické modelovanie dát v oblasti vzdelávania*. [Bakalárska práca]. – Žilinská univerzita v Žiline. Fakulta riadenia a informatiky; Katedra makro a mikroekonomiky. – Vedúci práce: Ing. Lucia Pančíková, PhD. – Stupeň odbornej kvalifikácie: bakalár. – Študijný odbor: Informatika. Žilina, 2020. 87 s.

Táto bakalárska práca bola zameraná na identifikáciu faktorov vplývajúcich na spokojnosť študentov so štúdiom na vysokej škole a na faktory, ktoré môžu ovplyvňovať rozhodnutie študenta pokračovať v inžinierskom stupni štúdia na danej vysokej škole. V rámci tejto práce bola vykonaná analýza reálnych dát zozbieraných prostredníctvom dotazníka od študentov Fakulty riadenia a informatiky, navštevujúcich bakalárske študijné programy ponúkané fakultou. Na základe tejto analýzy dát a teoretických poznatkov boli identifikované faktory ovplyvňujúce rozhodnutie študenta pokračovať v štúdiu a následne boli vytvorené, vyhodnotené a porovnané viaceré regresné modely. Tieto informácie boli využité pri spracovaní aplikácie, prostredníctvom ktorej bolo možné po zadaní základných informácií o študentovi predikovať jeho rozhodnutie pokračovať v štúdiu.

Kľúčové slová: predikcia, logistická regresia, multinomická logistická regresia, klasifikačné rozhodovacie stromy, spokojnosť so štúdiom

ABSTRACT

KARDOŠOVÁ, Dorota: *Statistical modeling in education*. [Bachelor thesis]. – University of Žilina. Faculty of Management Science and Informatics; Department of Macro and Microeconomics. – Supervisor: Ing. Lucia Pančíková, PhD. – Qualification level: bachelor. – Study program: Informatics. Žilina, 2020. 87 p.

This bachelor thesis was focused on identification of factors influencing the satisfaction of students with their university studies and factors, which could influence their decisions to continue with an engineering degree at the given university. Within this thesis an analysis of real data collected from students of Faculty of Management Science and Informatics, attending bachelor study programs offered by the faculty, was conducted. Based on this data analysis and theoretical bases, factors influencing this decision were identified and various regression models were constructed, evaluated, and compared. The information gained was used to create an application that, by filling out basic information about a student, can predict the decision of said student to continue with their studies.

Key words: prediction, logistic regression, multinomial logistic regression, classification decision trees, academic satisfaction

Zoznam obrázkov

Zoznam grafov

Zoznam tabuliek

Zoznam použitých skratiek

AIC	Akaike information criterion
angl.	anglicky
a pod.	a podobne
CDA	Correlation Data Analysis
CDT	Classification Decision Tree
CU	celková úspešnosť
cp	complexity parameter
EDA	Exploratory Data Analysis
FRI	Fakulta riadenia a informatiky
GUI	Graphic User Interface
IDA	Inferential Data Analysis
INF	informatika
JAR	Java Archive
LR	Logistic Regression
MAT	matematika
MLR	Multinomial Logistic Regression
napr.	napríklad
SJL	slovenský jazyk
slov.	slovensky
t. j.	to jest

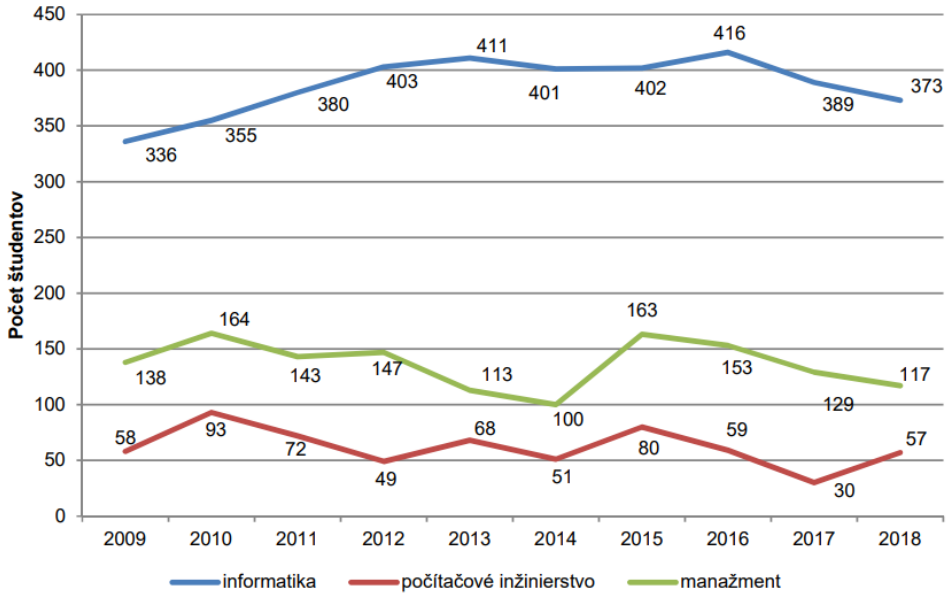
Obsah

Úvod

Napriek tomu, že veľa stredoškolákov si túto skutočnosť neuvedomuje, voľba vysokej školy sa dá považovať za jedno z najdôležitejších rozhodnutí v živote študenta. Niektorí študenti sa už po skončení strednej školy rozhodnú uplatniť na trhu práce, bez ďalšieho získaného vzdelania, avšak čoraz viac študentov si uvedomuje dôležitosť vysokoškolského vzdelania pre ich budúce pôsobenie v spoločnosti a rozhodne sa prihlásiť na vysokú školu. Veľakrát sa však stane, že sa študenti po získaní bakalárskeho titulu rozhodnú nepokračovať v štúdiu na inžinierskom stupni štúdia v danom odbore, na danej škole, alebo sa rozhodnú nepokračovať v inžinierskom štúdiu vôbec, pretože to nepokladajú za dôležité. Prečo je tomu tak?

Táto bakalárska práca je zameraná na identifikáciu faktorov vplývajúcich na spokojnosť študenta so štúdiom na Fakulte riadenia a informatiky a na zistenie informácií o tom, ako tieto faktory vplývajú na rozhodnutie študenta pokračovať v inžinierskom stupni štúdia na fakulte.

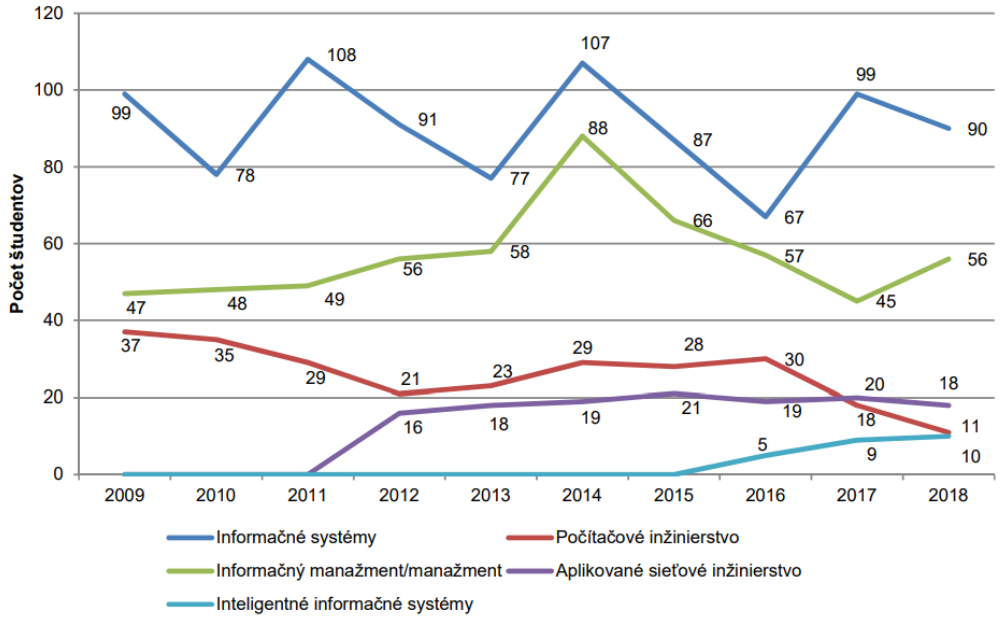
Podľa údajov uvedených vo Výročnej správe fakulty z roku 2018, za posledné dva roky počet prijatých študentov do 1. ročníka bakalárskeho štúdia na Fakultu riadenia a informatiky značne klesol v študijných programoch Informatika aj Manažment. Tento vývoj prijatých študentov za posledných 9 rokov môžeme vidieť na nasledujúcom grafe č.1.



Graf 1 – Vývoj počtu prijatých študentov do 1.ročníka bakalárskeho štúdia v jednotlivých študijných programoch

Zdroj: [50]

Čo sa týka prijatých študentov do 1. ročníka inžinierskeho štúdia na FRI, počet študentov pokračujúcich na študijnom programe Informačné systémy je v priebehu rokov premenlivý. Menší odliv študentov programu Informačné systémy zapríčinilo v roku 2016 otvorenie nového študijného programu Inteligentné informačné systémy. Avšak výrazný pokles v posledných rokoch môžeme vidieť pri študijnom programe Počítačové inžinierstvo. Bližšie informácie o vývoji počtu študentov prvých ročníkov nám môže poskytnúť graf č.2.



Graf 2 – Vývoj počtu prijatých študentov do 1.ročníka inžinierskeho štúdia v jednotlivých študijných programoch

Zdroj: [50]

Podľa informácií získaných zo študijného oddelenia Fakulty riadenia a informatiky sú počty aktuálnych študentov podľa jednotlivých ročníkov a študijných programov nasledovné:

Tabuľka 1 – Zapísaní študenti podľa ročníkov ku 31.10.2019

Študijný odbor	Stupeň štúdia	Spolu	Ročníky		
			1.	2.	3.
počítačové inžinierstvo	1	87	32	27	28
manažment	1	194	54	65	75
informatika	1	621	228	155	238
biomedicínska informatika	2	17	17		
počítačové inžinierstvo	2	27	13	14	
inteligentné informačné systémy	2	14	6	8	
informačný manažment	2	114	63	51	
aplikované sieťové inžinierstvo	2	38	16	22	
informačné systémy	2	121	51	70	
Spolu		**1233**	**480**	**412**	**341**
Opakujúci		**175**	**29**	**54**	**92**

Zdroj: [41]

Fakultu riadenia a informatiky navštevuje v akademickom roku 2019/2020 dokopy 1233 študentov. Počet prvákov na bakalárskom stupni je 314 z celkového počtu 902 študentov bakalárskeho stupňa. Najväčší podiel z celkového počtu majú tohtoroční tretiaci, ktorých je 341. V tejto tabuľke môžeme taktiež vidieť počet opakujúcich študentov, ktorých je 175 a najväčší podiel týchto študentov navštevuje tretí ročník. Celková tabuľka so všetkými podrobnosťami je dostupná v prílohe A.

Dôležitým údajom pre nás je počet študentov na inžinierskom stupni štúdia, ktorých je dokopy 331 a drvivá väčšina týchto študentov navštevuje študijné programy Informačný manažment a Informačné systémy – ktorý je rozdelený na Podnikovú informatiku, Spracovanie dát a Grafické spracovanie dát. V akademickom roku 2019/2020 bol predstavený aj nový študijný program Biomedicínska informatika, o ktorý prejavilo záujem až 17 študentov. Napriek tomu však môžeme vidieť, že kým v akademickom roku 2018/2019 bolo prijatých 185 prvákov, v akademickom roku 2019/2020 je to len 166 nových prvákov. Je dôležité poznamenať, že všetky údaje týkajúce sa akademického roku 2019/2020 sú aktuálne ku 31.10.2019 a môžu sa teda mierne líšiť, a to napríklad v prípade, ak niektorí študenti niekedy v priebehu zimného semestra zanechali štúdium, čo je bežné napríklad v prvom ročníku bakalárskeho štúdia, kedy veľa študentov po prvej neúspešnej skúške štúdium „vzdá“.

Úlohou tejto práce je poukázať na faktory, ktoré by mohli ovplyvňovať rozhodnutie študentov Fakulty riadenia a informatiky pokračovať na druhom stupni vysokoškolského vzdelania. V prípade, že by vďaka tejto práci boli zistené faktory, ktoré pozitívne alebo negatívne výrazne vplývajú na toto rozhodnutie študenta, mohla by pomôcť a podnietiť k uskutočneniu istých zmien, na základe ktorých by sa viac študentov rozhodlo pokračovať v štúdiu na inžinierskom stupni štúdia.

Cieľom tejto práce bola analýza reálnych dát získaných od aktuálnych študentov Fakulty riadenia a informatiky navštevujúcich bakalársky stupeň štúdia, vytvorenie štatistických modelov za použitia relevantných premenných a validácia týchto štatistických modelov, ktorých účelom bola predikcia rozhodnutia študenta pokračovať v štúdiu na druhom stupni vysokoškolského štúdia na Fakulte riadenia a informatiky.

V prvej kapitole je popísaný súčasný stav riešenej problematiky na Slovensku i v zahraničí. Druhá kapitola sa venuje teoretickým východiskám a s ňou súvisí tretia kapitola, ktorá poskytuje informácie o štatistických metódach a modeloch, ktoré boli v tejto práci

použité. Popisu zberu dát, ich úprave a analýze sa venuje kapitola štvrtá. Piata kapitola je venovaná štatistickému modelovaniu na študentských dátach, s využitím rôznych štatistických modelov. V šiestej kapitole je následne poskytnuté vyhodnotenie práce a možné návrhy do budúcna.

1 Súčasný stav riešenej problematiky na Slovensku a vo svete

Miera udržateľnosti (angl. *student retention*) študentov na vysokej škole a miera predčasných odchodov študentov bez získania titulu (angl. *student attrition*) sú problémy, ktorými sa zaoberajú univerzity celosvetovo už niekoľko rokov. V súčasnosti, kedy je trh vyššieho vzdelania globalizovaný a konkurencia univerzít je vysoká, môže vzniknutá povesť nízkej udržateľnosti študentov a vysokej miery odchodov študentov z univerzity bez získania titulu výrazne poškodiť príliv nových študentov na túto univerzitu. [33]

Podľa štúdie z roku 2002 (Cotton, Dollard, Jonge) vykonanej na University of South Australia sú stres a celková spokojnosť výrazne prepojené s mierou udržateľnosti študentov – pokiaľ sú študenti vystavovaní výrazným stresovým faktorom a pociťujú úzkosť, ich akademické výsledky, napredovanie a celková spokojnosť sú negatívne ovplyvnené a je pravdepodobnejšie, že štúdium na škole predčasne ukončia. Je preto potrebné venovať pozornosť vytvoreniu priaznivého pracovného prostredia, ktoré by malo túto úzkosť a nespokojnosť študentov znížiť. [8]

Yu, Digangi a kol. vo svojej štúdii v roku 2007 skúmali potenciálne faktory, ktoré môžu ovplyvňovať rozhodnutie študenta zotrvať na škole s využitím klasifikačných stromov. V rámci tejto štúdie bol využitý klasifikačný rozhodovací strom, pomocou ktorého boli skúmané demografické údaje, akademické výsledky z predošlého vzdelania ako aj aktuálne akademické výsledky. Štúdiou bol dokázaný vzťah medzi zotrvaním na vysokej škole a bydliskom študenta – to, či študent býva na internáte/dochádza, alebo je jeho trvalé bydlisko v blízkosti školy aj z hľadiska financií a z toho vyplývajúcich vyšších nákladov na štúdium vplýva na jeho rozhodnutie zotrvať na škole. [52]

V roku 2015 sa Siming, Niamatullah, Gao, Xu, Shaf snažili identifikovať vzťahy medzi spokojnosťou študentov so štúdiom a odbornosťou, prípravou a prístupom vyučujúcich, vybavením školy, ako aj s príležitosťami a skúsenosťami, ktoré inštitút študentom poskytuje. Pomocou deskriptívnej analýzy, regresie a korelačnej analýzy bolo štúdiou preukázané, že všetky tieto faktory majú vplyv na spokojnosť študentov vysokých škôl a preto sa odporúča na tieto faktory sústrediť svoju pozornosť. [42]

Acheampong, Boyetey, Osei Gyimah a Okyere v roku 2013 pomocou logistickej regresie ukázali, že na spokojnosť študentov na škole majú vplyv aj faktory ako napríklad výučba, odbornosť pedagógov, vybavenie školy, študijný odbor, štúdium v zahraničí alebo vyučované predmety. [1]

Pokiaľ ide o štúdie na Slovensku, v roku 2015 Kravčáková, Kozelová a Župová využili štatistický softvér SPSS na zistenie súvislostí medzi vzťahom ku práci (štúdiu) a sociálno-demografickými charakteristikami. Pri uvažovaní faktorov vplývajúcich na vzťah študenta ku štúdiu aplikovali Herzbergovu teóriu pracovnej motivácie a identifikované faktory rozdelili na motivátory (vnútorné faktory) a dissatisfaktory (vonkajšie faktory). Ako motivátory boli definované napr. atraktívnosť študijných programov, náročnosť predmetov, prax, prístup učiteľov, spravodlivé hodnotenie a medzi dissatisfaktory patrili napr. ponuka študijných programov, uplatniteľnosť na trhu práce, rozvrh, mimoškolské aktivity alebo vybavenie fakulty. [27]

Vychádzajúc z projektu s názvom To dá Rozum, realizovaného organizáciou MESA10 – centrom pre ekonomické a sociálne analýzy, v prípade porovnania študentov študujúcich na Slovensku a študentov zo Slovenska študujúcich v zahraničí, je na študentov na Slovensku oproti štandardom v iných krajinách vyvíjaný oveľa väčší tlak z hľadiska počtu predmetov za semester, kedy skoro 40% študentov uviedlo, že za semester má priemerne 8 a viac predmetov. Na druhej strane u študentov v zahraničí sa toto číslo pohybovalo len niečo málo cez 13% študentov, ktorí majú 8 a viac predmetov za semester. [34]

Analýza školstva na Slovensku týkajúca sa predčasných odchodov z VŠ, vykonaná v rámci projektu To dá Rozum ukázala, že väčšina študentov odchádza už počas bakalárskeho štúdia. Nakoľko väčšina škôl upustila od organizovania prijímacích skúšok a študenti sa prijímajú náborovo, znamená to, že najväčšia selekcia študentov nastáva už v prvom ročníku, kedy štúdium ukončí 30-40% študentov. Študenti v odbore informatiky predčasne zo štúdia odchádzajú aj z dôvodu nedostatku it-čkárov na pracovnom trhu, z čoho vyplýva, že sú na trhu práce veľmi žiadaní bez ohľadu na ich nadobudnuté vzdelanie. Z tohto dôvodu väčšina študentov informatiky zanechá štúdium na inžinierskom stupni a nedokončí ho, alebo sa uspokojí s bakalárskym titulom a na druhý stupeň štúdia nenastúpi vôbec, pretože nepovažujú ďalšie štúdium za potrebné. [34]

Na to, či sa študenti rozhodnú po zisku bakalárskeho titulu pokračovať na druhom stupni vysokoškolského štúdia majú vplyv viaceré faktory. Okrem uplatniteľnosti na trhu práce, ktoré študenti považujú za veľmi dôležité, majú na toto rozhodnutie vplyv aj iné faktory. Výsledky z dotazníka projektu To dá rozum ukázali ako jeden z najčastejšie uvádzaných dôvodov štúdia na vysokej škole očakávania zo strany rodiny. Tento dôvod

uvádzali zo všetkých študijných programov najčastejšie študenti IT a matematiky, či už z respondentov na bakalárskom stupni, ale aj respondenti na inžinierskom stupni štúdia. [11]

Veľké množstvo teoretickej prípravy na bakalárskom stupni štúdia s veľmi obmedzenými možnosťami využiť tieto teoretické vedomosti v praxi môže viesť ku strate motivácie a chuti učiť sa. Tlak zo strany rodiny môže pôsobiť ako silný faktor v prípade, ak študentom chýba motivácia iných foriem, predovšetkým tá vnútorná motivácia študovať. Napriek tomu, že absolventi bakalárskeho štúdia v odbore IT nemajú problém nájsť si zamestnanie už počas bakalárskeho štúdia, väčšinou pociťujú tlak zo strany spoločnosti, ktorá považuje až získanie magisterského alebo inžinierskeho titulu ako úspešné ukončenie vysokoškolského vzdelania. [11]

2 Teoretické východiská

V tejto kapitole sú popísané základné štatistické pojmy, pričom teoretické znalosti popísané v tejto kapitole boli ďalej využité pri vypracovaní praktickej časti práce, v ktorej boli vytvorené predikčné modely.

2.1 Základné pojmy

Základný štatistický súbor je množinou všetkých štatistických jednotiek, inak povedané množinou všetkých objektov, ktoré pozorujeme. Tieto štatistické jednotky majú istú spoločnú charakteristiku. Túto spoločnú vlastnosť všetkých štatistických jednotiek nazývame štatistický znak (napr. pohlavie, vek, vzdelanie, a pod.). Štatistické znaky možno deliť na kvalitatívne (kategoriálne) a kvantitatívne (číselné) znaky. Kvalitatívne znaky vyjadrujú vlastnosť štatistickej jednotky slovne. Delíme ich na ordinálne a nominálne. Nominálne (názvové) znaky vyjadrujú istú nemerateľnú vlastnosť, napr. pohlavie. Ordinálne (poradové) znaky je možné usporiadať do poradia, pričom ale nevieme určiť, o koľko je daná hodnota väčšia (napr. hodnotenie v škole). Kvantitatívne znaky nadobúdajú číselné hodnoty, čo znamená, že ich vieme usporiadať do poradia a vieme aj povedať, o koľko je jedna hodnota väčšia ako druhá. Tieto znaky delíme na diskrétne a spojité. Diskrétne znaky môžu nadobúdať spočítateľne veľa hodnôt z konečnej množiny hodnôt (počet súrodencov) a spojité znaky svoje hodnoty nadobúdajú z definovaného intervalu (príjem). [45]

Nakoľko je zrejmé, že či už z časových, alebo iných dôvodov nie je možné skúmať celý štatistický súbor, je potrebné z tohto súboru vytvoriť výber podľa vopred stanovených pravidiel, alebo tento výber vykonať náhodne. Tento výber nazývame výberový súbor. Ide o tzv. vzorku, ktorá je podmnožinou základného súboru a pomocou ktorej má byť tento základný súbor charakterizovaný. Dôležitá je reprezentatívnosť tohto výberu, a teda výberový súbor musí správne reprezentovať základný súbor. [45][49]

2.2 Analýza dát (Data Analysis)

Dátová analytika je proces spracovania, modelovania a vyhodnocovania dát, prostredníctvom logických, matematických a štatistických metód a úvah s cieľom nájsť v týchto dátach dôležité informácie, ktoré môžu viesť k podpore rozhodovania. Ide o určenie relevantných vzťahov, vzorcov a závislostí v týchto dátach, pričom vopred nevieme aké vzťahy medzi premennými očakávať. Analýza dát sa často rozdeľuje do dvoch fáz – exploratórnej a konfirmačnej. [13][51]

2.2.1 Exploratórna analýza dát (Exploratory Data Analysis)

Exploratórna analýza je prvým krokom pri analýze dát. Ako už z názvu odvodeného z angl. slova *explore* (slov. skúmať, objavovať) vyplýva, ide o skúmanie dát a hľadanie tzv. *clues,* teda záchytných bodov. V tejto fáze je dôležité oboznámiť sa s dátami a zistiť, s akým datasetom vlastne pracujeme. Dôležité je stanoviť si cieľ a určiť otázky, na ktoré chceme zistiť odpovede. V tejto fáze je potrebné prísť na najlepší možný spôsob, ako tieto dáta vhodne využiť, resp. ako z týchto dát získať maximum. EDA zahŕňa proces zisťovania základných charakteristík, zisťovanie štruktúry dát, identifikovanie chýb a chýbajúcich dát, zisťovanie anomálií v dátach, stanovenie hypotéz, odhad parametrov a stanovenie intervalov spoľahlivosti. [14]

Vizualizácia je grafickou reprezentáciou dát. Využitím grafov, tabuliek alebo máp vieme jednoducho „*od oka*“ odhaliť vzťahy, povahu dát a súvislostí medzi týmito dátami. Vizualizácia dát je taktiež dobrým spôsobom, ako komunikovať dáta pre širokú verejnosť. Najčastejším a často aj najzrozumiteľnejším spôsobom vizualizácie je prostredníctvom grafov, pričom je veľmi dôležité zvoliť správny typ grafu, aby bol zrozumiteľný, čitateľný a prehľadný. Na výber máme z viacerých typov grafov, ako napr. koláčový graf, stĺpcový graf alebo histogram, či boxplot. Dáta je taktiež možné zobraziť v tabuľkách, čo nám zjednodušuje určovanie deskriptívnych štatistík, ktoré nám poskytujú dôležité informácie o našom dátovom súbore. Tieto deskriptívne štatistiky rozdeľujeme na charakteristiky polohy, charakteristiky variability a doplňujúce charakteristiky. [9][15][43]

Číselné charakteristiky polohy vyjadrujú polohu znaku, okolo ktorého sú hodnoty rozptýlené. Táto poloha sa meria pomocou stredných hodnôt, ktorých je viac druhov, pričom v istých situáciách sú niektoré stredné hodnoty vhodnejšie ako iné. Tieto stredné hodnoty sú napríklad[25][43]:

- aritmetický priemer, ktorý je vôbec najznámejším priemerom a je definovaný vzťahom

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad (2.1)$$

  kde n je rozsah náhodného výberu a x_i je hodnota i-teho pozorovaného štatistického znaku,
- modus M_o je hodnota štatistického znaku s najväčšou početnosťou, inak povedané, je to najčastejšie sa vyskytujúca hodnota štatistického znaku,

– medián M_e je hodnota, ktorá nám rozdeľuje štatistický súbor na dve približne rovnako početné skupiny (polovice).

Charakteristiky variability nám hovoria o tom, ako veľmi sú hodnoty rozptýlené okolo strednej hodnoty. Napriek tomu, že dáta môžu mať rovnakú strednú hodnotu, môžu mať inú rozptýlenosť. Čím je hodnota variability vyššia, tým menej sa dá stredná hodnota považovať za reprezentatívnu. Túto premenlivosť dát môžeme vyjadrovať rôznymi charakteristikami. Medzi najčastejšie používané charakteristiky variability patria [12][43]:

– rozptyl je najpoužívanejšou charakteristikou variability a udáva nám informáciu o tom, ako veľmi sa údaje odchyľujú od strednej hodnoty, pričom je definovaný vzťahom

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (2.2)$$

– smerodajná odchýlka je druhou odmocninou rozptylu

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (2.3)$$

Všetky tieto charakteristiky nám poskytujú základné informácie o tom, ako náš dátový súbor vyzerá a aké dáta obsahuje. [15]

2.2.2 Inferenčná analýza dát (Inferential Data Analysis)

Inferenčná (konfirmačná) analýza dát (tzv. *testovanie hypotéz*) je druhou fázou analýzy dát a nasleduje po exploratívnej analýze dát. V tejto fáze sú dáta podrobované testom. Pri vykonávaní štatistických testov sa najprv stanoví hypotéza, pričom táto hypotéza je na základe výsledkov týchto testov – ktoré sú vykonané na výberovom súbore – buď prijatá alebo zamietnutá. Okrem testovania hypotéz IDA zahŕňa aj vytváranie odhadov, regresnú analýzu (odhady vzťahov medzi premennými) a analýzu rozptylu (hodnotenie rozdielov medzi plánovaným a skutočným výsledkom). [13]

Testovaním hypotéz sa snažíme generalizovať tvrdenie tým, že overíme, či je daná hypotéza pravdivá, a teda či sa prijíma, alebo nie. Testovanie vykonávame na výberovom súbore, ktorý je istou reprezentáciou základného súboru, inak povedané, zisťujeme ako dobre náš výberový súbor modeluje celkovú populáciu (v našom prípade študentov FRI). Pri tomto testovaní oproti sebe kladieme dve hypotézy, ktoré si navzájom odporujú – nulovú a alternatívnu. Nulová alebo testovaná hypotéza H_0 je hypotéza, ktorej platnosť overujeme.

Proti nej staviame alternatívnu hypotézu H_a, ktorá nulovej hypotéze odporuje a popiera ju. [31][44]

Po vyslovení H_0 a H_a je potrebné stanoviť hladinu významnosti $\alpha \in (0; 1)$, ktorá je definovaná ako tzv. chyba prvého druhu, kedy nulovú hypotézu zamietneme napriek tomu, že je pravdivá. Ďalšími krokmi sú voľba vhodného štatistického testu, výpočet hodnoty testovacieho kritéria, výpočet príslušnej p–hodnoty a vyhodnotenie testu. Ak $p-hodnota \leq \alpha$, nulovú hypotézu H_0 na hladine významnosti α zamietame v prospech alternatívnej hypotézy H_a. Ak $p-hodnota > \alpha$, potom nulovú hypotézu H_0 na hladine významnosti α nezamietame. [43]

2.2.2.1 Pearsonov (chí-kvadrát) test

Chí-kvadrát test nezávislosti je najznámejší a najpoužívanejší test nezávislosti štatistických znakov usporiadaných v kontingenčnej tabuľke. Chí-kvadrát test nezávislosti testuje nulovú hypotézu H_0, naproti ktorej stavia alternatívnu hypotézu H_a. Hypotézy H_0 a H_a sú nasledovné [43]:

H_0: náhodné velíčiny X a Y sú nezávislé

H_a: náhodné velíčiny X a Y sú závislé

Tvrdenie, že náhodné veličiny (premenné) X a Y sú nezávislé znamená, že pravdepodobnosť nastania určitej varianty veličiny X neovplyvňuje nastanie určitej varianty veličiny Y. [43]

Tento test je založený na porovnaní pozorovaných (empirických) početností n_{ij} a očakávaných (teoretických) početností e_{ij} náhodných veličín X a Y. Tieto náhodné veličiny predstavujú znak štatistického súboru (napr. pohlavie, vek...). Náhodná premenná X nadobúda hodnoty $1,2 \dots r$ a náhodná premenná Y nadobúda hodnoty $1,2 \dots s$, kde r je počet riadkov a s je počet stĺpcov kontingenčnej tabuľky. [43]

Empirická početnosť n_{ij} je počet prípadov, kedy sa vyskytla vo výbere dvojica (i, j), a teda sa u prvkov výberu zistil i-ty stupeň znaku X a j-ty stupeň znaku Y. Empirické početnosti sú definované vzťahmi [43]:

$$n_{i\bullet} = \sum_{j=1}^{s} n_{ij} \tag{2.4}$$

$$n_{\bullet j} = \sum_{i=1}^{r} n_{ij}. \tag{2.5}$$

Teoretická početnosť e_{ij} je definovaná nasledovne [43]:

$$e_{ij} = \frac{n_{i\bullet}\, n_{\bullet j}}{n} \tag{2.6}$$

Testovacie kritérium s $(r-1)*(s-1)$ stupňami voľnosti je rozdielom empirickej a teoretickej početnosti. Vzorec je nasledovný [43]:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} \tag{2.7}$$

Nulovú hypotézu H_0 o nezávislosti X a Y zamietame na hladine významnosti α, ak hodnota testovacieho kritéria presiahne príslušný $100(1-\alpha)\%$ kvantil rozdelenia, a teda $\chi^2 \geq \chi^2_{(r-1)*(s-1)}(1-\alpha)$. [43][46]

2.2.2.2 Cramérov kontingenčný koeficient (Cramérovo V)

Cramérov kontigenčný koeficient, skrátene nazývaný aj Cramérovo V, je koeficient určujúci silu vzťahu medzi premennou riadkovou a premennou stĺpcovou v kontingenčnej tabuľke. Tento koeficient je najvhodnejšou mierou asociácie medzi premennými, ktoré sú kategorické a môže nadobúdať hodnoty $< 0; 1 >$, pričom 0 vyjadruje žiadny vzťah medzi premennými a 1 vyjadruje dokonalú asociáciu (vzťah). Hranice asociácie môžu byť nasledovné [37]:

Hodnota 0 ⇒ žiadny vzťah
Hodnoty okolo 0,2 ⇒ slabý vzťah
Hodnoty okolo 0,5 ⇒ stredný vzťah
Hodnoty okolo 0,7 a vyššie ⇒ silný vzťah
Hodnoty 1 ⇒ dokonalý vzťah

Tieto hranice sú len orientačné a vždy závisia od povahy dát. V tejto práci boli hranice stanovené podľa vyššie uvedených hodnôt. [3][37]

Hodnota Cramérovho V sa vypočíta

$$V = \sqrt{\frac{\chi^2/n}{\min(s-1,\ r-1)}}, \tag{2.8}$$

kde χ^2 je hodnota testovacieho kritéria chí-kvadrát testu, n je počet pozorovaní, s je počet stĺpcov a r je počet riadkov kontingenčnej tabuľky. [37]

3 Metodológia

V nasledujúcej kapitole sa pozornosť venuje základným štatistickým a analytickým metódam, ktoré boli využité pri vypracovaní praktickej časti práce na predikciu rozhodnutia študenta pokračovať v bakalárskom stupni štúdia na Fakulte riadenia a informatiky.

3.1 Strojové učenie (Machine Learning)

Strojové učenie je jednou z podoblastí umelej inteligencie, ktorá sa zaoberá algoritmami a metódami, pomocou ktorých sa program učí a následne je schopný na základe informácií, ktoré sa naučil, vhodne reagovať na rôzne vstupné hodnoty bez toho, aby bol na to priamo naprogramovaný. Celý proces učenia začína skúmaním dát, v ktorých program nájde vzory a na základe toho sa rozhoduje. Strojové učenie prebieha v dvoch fázach, vo fáze trénovania a fáze testovania. Fáza trénovania, nazývaná aj fáza učenia, je prvou fázou strojového učenia, počas ktorej je vytvorený matematický model, ktorý je natrénovaný na trénovacej množine. Takto natrénovaný model následne vo fáze testovania aplikujeme na testovacej množine. [2][10]

Podľa spôsobu, akým proces učenia modelov prebieha, rozdeľujeme algoritmy strojového učenia do dvoch skupín [2]:

- učenie s učiteľom (supervised learning)
- učenie bez učiteľa (unsupervised learning)

Do skupiny učenia s učiteľom patrí väčšina algoritmov strojového učenia, napr. lineárna regresia, logistická regresia alebo rozhodovacie stromy. Pri učení s učiteľom sa program učí z tzv. datasetu, ktorý je rozdelený na trénovaciu množinu cvičných dát, na ktorej sa model natrénuje, a testovaciu množinu, na ktorej svoje výsledky otestuje a určí sa jeho úspešnosť. Pre učenie bez učiteľa je špecifické to, že model sa učí len na základe vstupných premenných. Medzi metódy učenia bez učiteľa patrí napr. zhlukovanie a asociácie, ktoré napríklad využívajú spoločnosti ako Amazon a Netflix na odporúčanie svojich produktov zákazníkom, na základe ich predošlej aktivity. [2][47][48]

3.2 Regresná analýza

Regresná analýza je analýza, ktorej cieľom je nájsť, preskúmať a charakterizovať vzájomné vzťahy medzi jednotlivými premennými. Úlohou regresnej analýzy je určiť takú regresnú funkciu, alebo tzv. regresný model, ktorým sme schopní čo najlepšie popísať závislosti medzi premennými. Regresná analýza môže byť jednoduchá alebo viacnásobná.

Pri jednoduchej regresnej analýze sa zaoberáme iba jednou nezávislou premennou, pričom závislá premenná Y bude závisieť od nezávislej premennej X. Pri viacnásobnej regresii uvažujeme väčší počet nezávislých premenných X_1, X_2, ..., X_i. [26]

3.2.1 Lineárna regresia (Linear Regression)

Lineárna regresia je najjednoduchším modelom regresnej analýzy a jej výsledná hodnota je lineárnou kombináciou hodnôt nezávislých premenných X_1, X_2, ..., X_i. Lineárna regresia je štatistická metóda, ktorá sa využíva na vyjadrenie vzťahu medzi závislou premennou Y, ktorá je spojitá, a nezávislými premennými X_1, X_2, ..., X_i. Cieľom lineárnej regresie je modelovať spojitú závislú premennú Y ako matematickú funkciu jednej alebo viacerých premenných X tak, aby tento regresný model mohol byť použitý na predpovedanie hodnoty Y vtedy, keď poznáme iba hodnoty X. Tento základný regresný model lineárnej regresie s jednou nezávislou premennou má tvar matematickej rovnice [7][15]:

$$Y_i = \beta_0 + \beta_1 * X_1 + \varepsilon, \qquad (3.1)$$

kde Y_i vyjadruje hodnotu závislej premennej na základe hodnoty nezávislej premennej X_1, β_0 a β_1 predstavujú parametre modelu a ε je chyba pozorovania. [15]

3.2.2 Logistická regresia (Logistic Regression)

Logistická regresia je regresný model, ktorý sa používa na popis vzťahov medzi premennými v prípade, ak závislá premenná Y nie je spojitá, ale je binárna/kategorická. Binárna závislá premenná Y môže nadobúdať dve hodnoty (zvyčajne 1 a 0). Logistická regresia je model pravdepodobnostný, určuje pravdepodobnosť nastania udalosti ($Y = 1$) na základe jednej alebo viacerých nezávislých premenných. Nakoľko ide o pravdepodobnosť, logistická regresia môže nadobúdať hodnoty iba z intervalu $< 0; 1 >$. Vzťah medzi závislou premennou Y a nezávislou premennou X je teda daný určitou pravdepodobnosťou. Táto pravdepodobnosť je definovaná nasledujúcim vzťahom [6][7]:

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 * X_1)}} \qquad (3.2)$$

V prípade, ak by sme uvažovali väčší počet nezávislých premenných, vzťah medzi závislou premennou Y a týmito nezávislými premennými X_1, X_2, ..., X_i, a pravdepodobnosť nastania udalosti by bola definovaná nasledovne [15]:

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 * X_1 + b_2 * X_2 + \cdots + b_i * X_i)}} \qquad (3.3)$$

Threshold value t (slov. prahová hodnota) pri logistickej regresií vytvára akúsi hranicu medzi pravdepodobnosťami. Model logistickej regresie nám poskytuje pravdepodobnosť, s akou udalosť nastane. V našom prípade je to pravdepodobnosť toho, že študent bude pokračovať v štúdiu. Prahová hodnota nám určuje hranicu medzi pravdepodobnosťami, pričom platí [15]:

$$P(Y = 1) > t \Rightarrow \text{predpovedaj } 1$$

$$P(Y = 0) < t \Rightarrow \text{predpovedaj } 0$$

V našom prípade to znamená, že ak je pravdepodobnosť väčšia ako prahová hodnota t, model predpovedá 1, a teda, že študent bude pokračovať v štúdiu na FRI. V opačnom prípade je predikovaná 0, t. j. študent nebude pokračovať v štúdiu na FRI.

3.2.3 Multinomická logistická regresia (Multinomial Logistic Regression)

Multinomická logistická regresia sa používa na predikciu závislej nominálnej náhodnej premennej pomocou jednej a viacerých nezávislých premenných. Ide o pokročilejšiu techniku, ktorá sa dá považovať za rozšírenie binomickej logistickej regresie, nakoľko závislá premenná X môže mať viac ako dve kategórie/hodnoty, ktoré môže nadobúdať. [30]

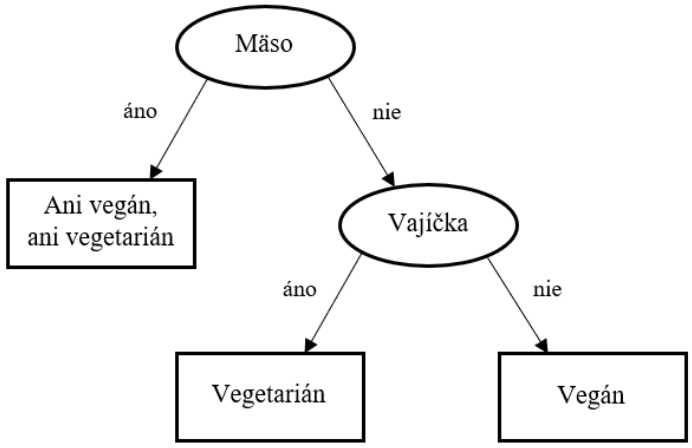
3.2.4 Rozhodovací strom (Decision Tree)

Rozhodovací strom (angl. Decision tree) je jedným z nástrojov na podporu rozhodovania. Rozhodovacie stromy sú jednou z najznámejších foriem klasifikácie, ale okrem klasifikačných stromov existujú aj stromy regresné. V prípade regresných rozhodovacích stromov je výstupná premenná spojitá (numerická hodnota). Pri klasifikačných rozhodovacích stromoch je výstupná premenná kategorická (môže byť binárna). [15]

Rozhodovací strom sa konštruuje na trénovacej množine. Rozhodovací strom je tvorený z uzlov, vetiev a listov. Na vrchole stromu sa nachádza hlavný uzol nazývaný koreň, ktorý obsahuje všetky údaje. Koreň sa ďalej vetví do ďalších uzlov na základe určených pravidiel rozhodovania. Každý z vnútorných uzlov teda predstavuje rozhodnutie na základe vlastnosti objektu a je podmnožinou koreňového uzla. Strom sa vetví až kým objekt nie je zaradený do listového uzla, ktorý ho klasifikuje. V prípade nových objektov, ktoré ešte nie sú klasifikované, tieto objekty postupne prechádzajú od koreňového uzla, cez jednotlivé

vetvy a vnútorné uzly, až kým neskončia v jednom z listov, ktorým môžu byť reprezentované a klasifikované. [32][38]

Ako už z názvu vyplýva, ide o modely, ktoré majú stromovú štruktúru. Vďaka tejto štruktúre sa dá strom jednoducho graficky znázorniť a je ľahko interpretovateľný, čo je jednou z výhod rozhodovacích stromov. Grafickú reprezentáciu a príklad jednoduchého klasifikačného stromu môžeme vidieť na nasledujúcom obrázku. [15]



Obrázok 1 – Príklad klasifikačného rozhodovacieho stromu

Zdroj: Vlastné spracovanie, 2020 podľa [32]

Na obrázku č.1 môžeme vidieť jednoduchú predikciu spôsobu stravovania na základe toho, či respondent konzumuje mäso alebo vajcia. Vďaka grafickému znázorneniu z obr.č.1 sú výsledky predikcie jednoducho interpretovateľné. Pokiaľ osoba nekonzumuje mäso, ale konzumuje vajíčka je to vegetarián. Ak nekonzumuje mäso ani vajíčka, stravuje sa ako vegán. V prípade, ak osoba konzumuje mäso, nie je ani vegán, ani vegetarián.

Okrem grafickej reprezentácie a jednoduchej interpretovateľnosti je výhodou rozhodovacích stromov aj fakt, že nie je potrebná lineárnosť vzťahov. Rozhodovacie stromy majú však aj isté nevýhody. Medzi tieto nevýhody patrí napr. náchylnosť na pretrénovanie modelu (model sa dáta „naučí naspamäť“ a na nových dátach je nepresný). Ďalšou z nevýhod je nestabilnosť modelov rozhodovacích stromov, čo znamená, že aj malá zmena v dátach môže zmeniť model a narušiť štruktúru stromu. [15][21]

Pri vytváraní modelov rozhodovacích stromov je potrebné správne zvoliť parametre minsplit, minbucket a cp. Minsplit a minbucket sú parametre, ktoré nastavujú pravidlá vetvenia a určujú hranice ukončenia vetvenia. Minsplit udáva minimálny počet pozorovaní v uzle na to, aby sa tento uzol mohol ďalej vetviť. Minbucket udáva minimálny počet pozorovaní v listovom uzle. Cp je parameter zložitosti rozhodovacieho stromu. Jeho

hlavnou úlohou je ušetrenie výpočtového času tým, že „odreže“ vetvy, ktoré sú zbytočné – vetvením ktorých by sa nezvýšila celková hodnota koeficientu determinácie R^2 (presnosť modelu) o hodnotu cp. R^2 slúži na popis predikčnej schopnosti vytvoreného modelu. Vyjadruje percento variability dát, ktoré vytvorený model vysvetľuje, a teda o koľko percent je vytvorený model lepší ako tzv. baseline (základný) model. [15][39]

3.2.5 Náhodný les (Random Tree)

Jedným zo spôsobov riešenia problémov spomenutých pri rozhodovacích stromoch je využitie modelu náhodného lesa. Náhodný les (angl. Random Forest) je nástroj strojového učenia, ktorý je zložený z viacerých rozhodovacích stromov. Model náhodného lesa je síce výpočtovo náročný a ťažko interpretovateľný, ale nakoľko nie je náchylný na pretrénovanie, zlepšuje presnosť rozhodovacích stromov. Výstupom náhodného lesa je kombinácia výsledkov jednotlivých rozhodovacích stromov. Pri regresnom náhodnom lese je výstupom priemer hodnôt jednotlivých rozhodovacích stromov. Výstupom klasifikačného náhodného lesa je najčastejšie sa vyskytujúca kategória. [15]

3.3 Validácia modelov

Pokiaľ majú byť modely využité v praxi, je dôležité tieto modely validovať. Validácia je proces, kedy vytvorený model aplikujeme na nové (testovacie) dáta, na ktorých tento model nebol natrénovaný, a sledujeme úspešnosť modelu na týchto testovacích dátach. Pomocou otestovania modelu na testovacích dátach zisťujeme nakoľko je model presný na neznámych dátach, a či nedošlo ku pretrénovaniu modelu. [24]

Na hodnotenie klasifikačných modelov slúži klasifikačná matica, inak nazývaná matica chybovosti, ktorá vyzerá nasledovne [15]:

Tabuľka 2 – Klasifikačná matica

		Predikovaná hodnota	
		Predikovaná 0	Predikovaná 1
Skutočná hodnota	Skutočná 0	TN	FP
	Skutočná 1	FN	TP

Vlastné spracovanie podľa [15]

Pokiaľ by sme vzali ako príklad rozhodnutie študenta pokračovať v štúdiu, pričom toto rozhodnutie by mohlo nadobúdať len dve hodnoty (1 - áno/0 - nie), táto tabuľka by zobrazovala výsledky predikcií tohto rozhodnutia (v stĺpcoch) a skutočné rozhodnutie pokračovať v štúdiu alebo odísť (v riadkoch). Hodnota TN (true negative) udáva počet výsledkov, ktoré boli skutočne negatívne, a teda koľko študentom, ktorí odišli bolo

predpovedané, že odídu. Hodnota FP (false positive) udáva počet výsledkov, ktoré boli nesprávne pozitívne, čo znamená, že bolo predpovedané, že študent bude pokračovať v štúdiu, ale študent zo školy odišiel. Hodnota FN (false negative) udáva počet nesprávne negatívnych výsledkov, kedy bolo predikované, že študent zo školy odíde, ale študent sa rozhodol pokračovať v štúdiu. Hodnota TP (true positive) udáva počet skutočne pozitívnych výsledkov, kedy študentovi bolo predpovedané, že bude v štúdiu pokračovať a tento študent sa tak aj v skutočnosti rozhodol. [22]

Na základe hodnôt z tejto matice vieme vypočítať základné metriky určujúce úspešnosť nášho modelu. Medzi tieto metriky patria senzitivita, špecificita a celková úspešnosť. [15]

Špecificita modelu v našom prípade udáva schopnosť modelu rozpoznať študentov, ktorí sa rozhodli odísť a nepokračovať v štúdiu, a teda pravdepodobnosť toho, že predikovaná hodnota bude negatívna (0 = študent nepokračuje), keď študent neplánuje v štúdiu pokračovať. Špecificitu modelu vypočítame pomocou nasledujúceho vzťahu, vyplývajúceho z matice chybovosti z tabuľky č.2 [22]:

$$TN/(TN + FP) \tag{3.4}$$

Senzitivita modelu udáva schopnosť modelu rozpoznať študentov, ktorí sa rozhodli v štúdiu pokračovať, a teda pravdepodobnosť toho, že predikovaná hodnota bude pozitívna (1 = študent pokračuje) vtedy, keď študent plánuje v štúdiu pokračovať ďalej. Senzitivitu modelu vypočítame na základe tabuľky č.2 zobrazujúcej maticu chybovosti nasledujúcim vzťahom [22]:

$$TP/(FN + TP) \tag{3.5}$$

Aby sa model dal považovať za dobrý je potrebné, aby medzi hodnotami špecificity a senzitivity boli čo najmenšie rozdiely a tieto hodnoty si boli čo najbližšie. [15]

Celkovú úspešnosť modelu môžeme na základe matice chybovosti z tabuľky č.2 vypočítať nasledovným vzťahom [22]:

$$(TN + TP)/(TN + FP + FN + TP) \tag{3.6}$$

Všetky uvedené metriky sú opísané na klasifikačnom probléme, ktorý má len dve kategórie – áno (1) alebo nie (0). Klasifikačný problém riešený v tejto práci má kategórie až

štyri, tým pádom aj klasifikačná matica vyzerá inak. Matica chybovosti v prípade, že závislá premenná má viac ako dve kategórie, má rozmer $i * i$, pričom i je počet kategórií závislej premennej. V tomto prípade sa senzitivita a špecificita počítajú pre každú kategóriu závislej premennej zvlášť. V štatistickom programovacom jazyku R je možné na vytvorenie matice chybovosti použiť funkciu *confusionMatrix()* z balíčka *caret*. Pomocou tejto funkcie nielenže vieme vypísať maticu chybovosti, ale automaticky sú z tejto matice vďaka tejto funkcií vypočítané aj jednotlivé štatistiky. [5]

4 Analýza a úprava dát

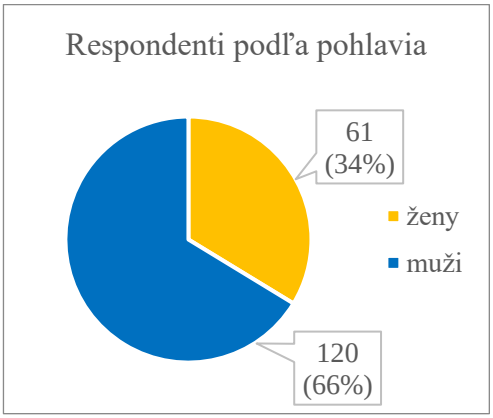
Táto kapitola je venovaná popisu spôsobu získania dát využitých v tejto práci, popisu dát samotných, ich štruktúre a následnej úprave týchto dát. Na analýzu dát bol použitý tabuľkový procesor Excel pri vytváraní grafov a programovací jazyk R, ktorý je určený na štatistickú analýzu dát a ich grafické zobrazenie. Jeho funkcie môžu byť rozšírené nainštalovaním balíčkov a knižníc. Jazyk R bol v tejto kapitole použitý pri zisťovaní miery asociácie premenných výpočtom Cramérovho V a pomocou chí-kvadrát testu. [15]

4.1 Analýza dát z dotazníka

Dáta k vypracovaniu tejto práce boli získané prostredníctvom anonymného dotazníka, zberom odpovedí od súčasných študentov navštevujúcich študijné odbory Informatika, Manažment a Počítačové inžinierstvo, bakalárskeho stupňa štúdia na Fakulte riadenia a informatiky. Dotazník vyplnilo 181 respondentov a zber dát prebehol v mesiacoch február – marec 2020, kedy bol dotazník šírený prostredníctvom sociálnych sietí. Náš základný súbor teda obsahuje 181 pozorovaní, ktorými sú študenti bakalárskych študijných programov Informatika, Manažment a Počítačové inžinierstvo na FRI.

4.1.1 Exploratórna analýza dát z dotazníka

Ako sa dalo očakávať a ako z nasledujúceho grafu č.3 vyplýva, nakoľko Fakultu riadenia a informatiky navštevuje viac chlapcov ako dievčat, tak viac ako polovicu respondentov (66%) tvorili muži.

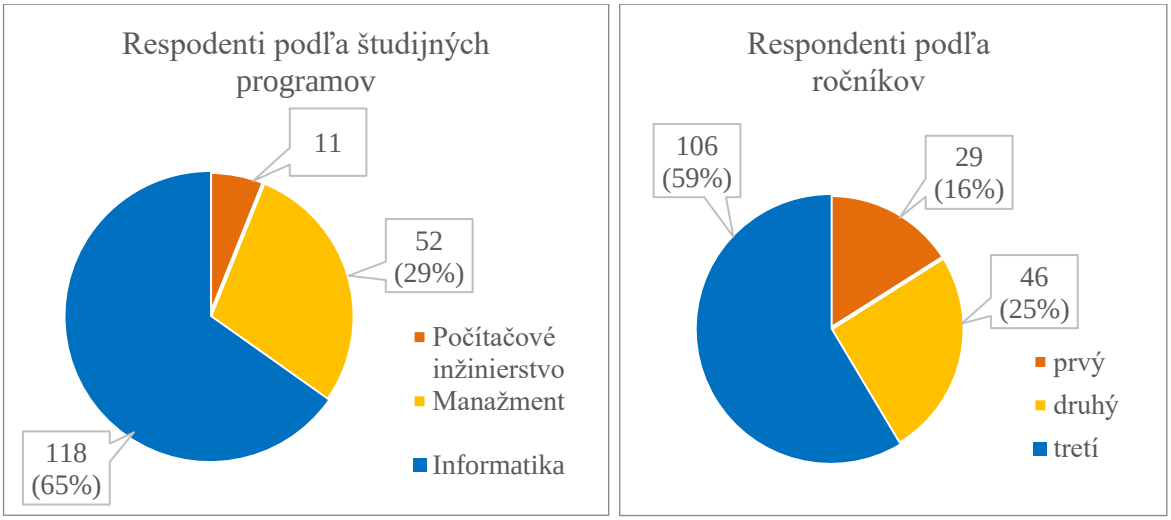


Graf 3 – Respondenti podľa pohlavia

Zdroj: Vlastné spracovanie, 2020

Čo sa týka študijných programov, tak najväčšie zastúpenie respondentov mal študijný program Informatika (65%) a najmenej odpovedí bolo zaznamenaných od študentov

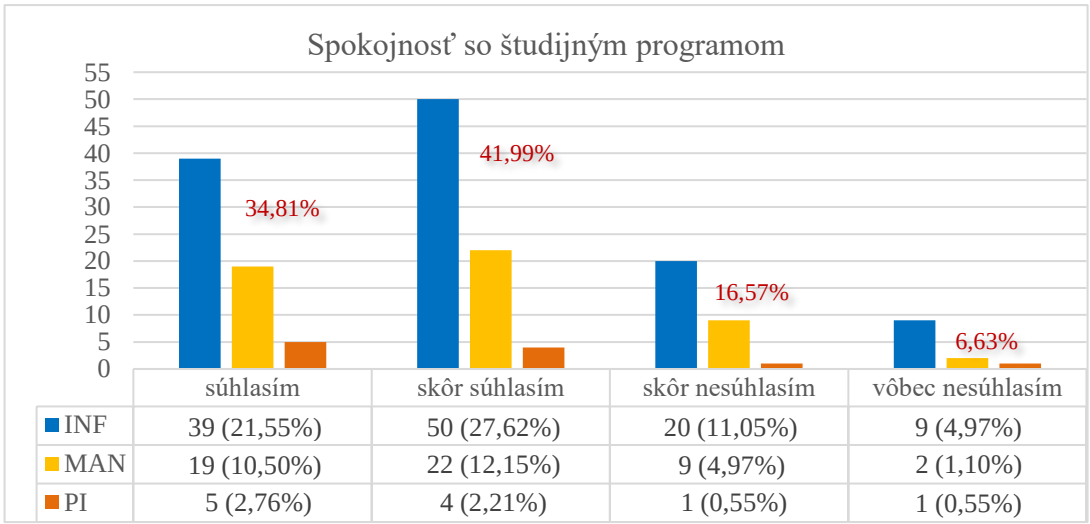
navštevujúcich odbor Počítačové inžinierstvo (len 6%). Viac ako polovica respondentov navštevuje tretí ročník (59%), 25% respondentov chodí do druhého ročníka a prvý ročník navštevuje 16% respondentov. Toto môžeme vidieť na grafe č.4.



Graf 4 – Respondenti podľa študijných programov a ročníkov

Zdroj: Vlastné spracovanie, 2020

Na grafe č.5 môžeme vidieť, ako sú respondenti spokojní so študijným programom, ktorý navštevujú. Celkovo 34,81% respondentov je so svojím študijným programom veľmi spokojných a 41,99% je s programom skôr spokojných ako nespokojných. Počet študentov, ktorí sú so svojím študijným programom nespokojní, tvorí len niečo málo cez 6% z celkového počtu respondentov. Na nasledujúcom grafe č.5 môžeme vidieť spokojnosť respondentov podľa jednotlivých študijných programov.



	súhlasím	skôr súhlasím	skôr nesúhlasím	vôbec nesúhlasím
INF	39 (21,55%)	50 (27,62%)	20 (11,05%)	9 (4,97%)
MAN	19 (10,50%)	22 (12,15%)	9 (4,97%)	2 (1,10%)
PI	5 (2,76%)	4 (2,21%)	1 (0,55%)	1 (0,55%)

Graf 5 – Spokojnosť respondentov so študijným programom

Zdroj: Vlastné spracovanie, 2020

Ide o jednu z premenných, ktorá môže ovplyvňovať rozhodnutie študenta pokračovať v budúcnosti v štúdiu na inžinierskom stupni štúdia na Fakulte riadenia a informatiky. Dá sa predpokladať, že pokiaľ je študent so svojím študijným programom spokojný, bude chcieť pokračovať v štúdiu na inžinierskom stupni štúdia a naopak.

4.1.2 Úprava dát z dotazníka

Aby mohli byť správne použité, bolo potrebné získané dáta pred ich použitím najskôr upraviť. Nakoľko väčšina odpovedí v dotazníku bola povinných, problém s chýbajúcimi dátami bol minimálny. Musím taktiež poznamenať, že respondenti vypĺňali dotazník naozaj dôsledne a svedomite a väčšina respondentov uviedla odpovede aj na otázky, ktoré boli označené za nepovinné. Ak sa vyskytli chýbajúce odpovede, tento problém bol pri jednotlivých premenných riešený nasledovne (otázky ku jednotlivým premenným sú dostupné v prílohe B a im priradené hodnoty v prílohe C). Premennej pVyucbaPrax v otázke týkajúcej sa toho, či je výučba dostatočne orientovaná na prax boli chýbajúce odpovede nahradené modusom. Rovnaký postup bol zvolený aj v prípade premennej pVyuzitieUciva, ktorá vyjadruje nakoľko respondenti využívajú nadobudnuté vedomosti a zručnosti v praxi. Posledná otázka, pri ktorej boli zaznamenané chýbajúce odpovede sa týkala toho, či sú študenti hrdí na to, že študujú na FRI. Nakoľko v tejto otázke bolo možné zvoliť si z dvojice odpovedí áno(1)/nie(0), v premennej fHrdy boli chýbajúce údaje nahradené hodnotou 0,5. Pri premennej vLudiZapis, kde mal v otázke respondent zadať počet ľudí, ktorých poznal pri zápise, bola tejto premennej priradená hodnota 0 v prípade, že respondent vo svojej odpovedi uviedol číslo menšie ako 0.

Respondenti udávali v dotazníku aj svoje doterajšie študijné výsledky zo strednej školy, ale aj výsledky z aktuálneho štúdia na vysokej škole na Fakulte riadenia a informatiky. Pri maturitách v prípade, že respondent z daného predmetu nematuroval bolo toto vyjadrené hodnotou 0. Obdobne tomu bolo aj pri aktuálnych akademických výsledkoch. Nakoľko respondenti navštevujú rozličné študijné programy, nie všetky predmety sa nachádzali študijných osnovách programu, ktorý navštevujú. Z tohto dôvodu v prípade, že študent neabsolvoval predmet, alebo ho nemal v študijnej osnove bolo toto vyjadrené hodnotou 0.

4.1.3 Inferenčná analýza dát z dotazníka

V teoretickej časti v kapitole 2.2.2 sme spomínali, že štatistické znaky, alebo teda premenné, sa delia na kvalitatívne a kvantitatívne. Je to potrebné odlíšiť aj z toho dôvodu, že kvalitatívne premenné sú faktorové a spracovávajú sa odlišne. Kvalitatívne premenné sa

delia na nominálne (názvové) a ordinálne (poradové). Hodnoty ordinálnych kvalitatívnych premenných boli nahradené číselnými hodnotami 0-3 a to nasledovne:

Tabuľka 3 – Hodnoty ordinálnych kvalitatívnych premenných

Slovné vyjadrenie			Hodnota
nie	negatívne	nepotreboval/a som to	0
na prednášky nechodím vôbec	vôbec nesúhlasím	nesúhlasím	
vôbec	nikdy	nie, pretože nemám záujem	
nie, ale premýšľal/a som o tom	neutrálne	čiastočne	1
občas sa prednášky zúčastním	skôr negatívne	skôr nesúhlasím	
plánuje/m ho prerušiť	nie, pretože nemám čas	občas	
áno	skôr pozitívne	občas prídem na akciu	2
áno, ale niektoré prednášky vynechám	skôr súhlasím		
áno, chodím na všetky prednášky	pozitívne	áno, chodím na akcie	3
súhlasím			

Zdroj: Vlastné spracovanie, 2020

V tejto práci skúmame vzťahy medzi rozhodnutím pokračovať v štúdiu na inžinierskom stupni štúdia a jednotlivými premennými. Silu týchto závislostí medzi rozhodnutím pokračovať v štúdiu a jednotlivými premennými bolo potrebné otestovať a bol na to využitý celý dataset obsahujúci 181 pozorovaní.

4.2 Miery asociácie medzi nominálnymi premennými a rozhodnutím pokračovať v štúdiu

Jedným z často využívaných spôsobov vyjadrenia miery asociácie medzi nominálnymi premennými je pomocou Cramér's V, ktoré je bližšie definované v kapitole 2.2.2.2. Hodnoty Cramérovho V pre jednotlivé premenné z nášho datasetu sú zobrazené v nasledujúcej tabuľke č. 4. Vyberáme premenné, ktoré majú na rozhodnutie pokračovať v štúdiu istý vplyv, aj keď len minimálny. Z tohto dôvodu berieme do úvahy všetky premenné, ktoré nadobúdajú hodnotu Cramérovho V väčšiu ako 0,2, ktorá vyjadruje slabú silu asociácie medzi premennými. Tieto premenné sú v tabuľke č.4 vyznačené zelenou farbou. [3][29]

Tabuľka 4 – Asociácie medzi nominálnymi premennými a rozhodnutím pokračovať

x	pohlavie	odbor	kraj	bydlisko	typSS
v	0.2105838	0.2618212	0.2154427	0.1517381	0.09627018

Zdroj: Vlastné spracovanie, 2020

Ako môžeme z tabuľky vidieť, ide o premenné pohlavie, študijný odbor a kraj, z ktorého študent pochádza. Hodnotu Cramérovho kontigenčného koeficientu (vzorec 2.8) sme vypočítali v jazyku R pomocou knižnice *lsr* a v nej obsiahnutej funkcie *cramersV()*. Všetky otázky ku ktorým sú jednotlivé premenné priradené sú dostupné v prílohe B a hodnoty týchto premenných sú dostupné v prílohe C.

4.3 Miery asociácie medzi ordinálnymi premennými a rozhodnutím pokračovať v štúdiu

Na vyjadrenie miery asociácie medzi rozhodnutím pokračovať v štúdiu a jednotlivými ordinálnymi premennými bol použitý chí-kvadrát test. Chí-kvadrát test sa používa na overenie hypotézy, ktorá tvrdí, že medzi dvoma kategorickými premennými neexistuje žiaden vzťah. Tento test porovnáva empirické (pozorované) početnosti z dát s očakávanými početnosťami – očakávané za predpokladu, že medzi premennými neexistuje vzťah. Nulová a alternatívna hypotéza sú nasledovné [4][17][18]:

H_0: Medzi danou ordinálnou premennou a premennou fIng neexistuje závislosť

H_a: Medzi danou ordinálnou premennou a premennou fIng existuje závislosť

Pomocou jazyka R a funkcie *chisq.test()* z knižnice *MASS* boli asociácie medzi jednotlivými ordinálnymi premennými a rozhodnutím pokračovať v štúdiu na FRI otestované chí-kvadrát testom pri hladine významnosti $\alpha = 0.05$. V nasledujúcej tabuľke č.5 sú zelenou farbou vyznačené premenné, ktoré majú vplyv na rozhodnutie študenta pokračovať v štúdiu fIng, a teda ich výsledná $p\text{–}hodnota \leq \alpha$, čo znamená, že nulovú hypotézu H_0 zamietame na hladine významnosti α v prospech alternatívnej hypotézy H_a, ktorá tvrdí, že medzi premennou fIng a danou premennou existuje štatisticky významná závislosť.

Tabuľka 5 – Asociácie medzi ordinálnymi premennými a rozhodnutím pokračovať

x	rocnik	matSJL	matMAT	matINF	matFYZ
p	0.1474	0.7628	0.8522	0.005855	0.4297
x	matOBN	matODB	progrSS	zhoduje	vIneSkoly
p	0.4414	0.6474	0.3223	0.01514	0.1705
x	vPrvaVolba	vSam	vRodicia	sBavi	sZnova
p	0.07929	0.07929	0.09174	0.000000443	0.000006344
x	sINF	sMAN	sPI	sPAS	sAUS
p	0.2857	0.2081	0.2367	0.01794	0.0424
x	sPrednasky	sRodiciaPodpora	sPrenasam	sOpakujem	sPredcasne
p	0.02772	0.2075	0.0006044	0.0001001	0.00000001629

x	sPredcasneKam	sPomoc	sVztahy	sSocZivot	sNezvladam
p	0.02703	0.1642	0.0102	0.0009991	0.0002968
x	sPsycholog	sNarocne	sNaroky	spProgramy	spPrilezitosti
p	0.0001891	0.008752	0.0007134	0.0000001454	0.1515
x	spSemester	spOdradza	spZbytocne	spZastarale	spStudProgram
p	0.000007488	0.02688	0.001166	0.0007382	0.000000002978
x	spVybavenie	spOdbornost	spPristup	spHodnotenie	spTerminy
p	0.01023	0.0124	0.000001858	0.00003689	0.0000277
x	pPraca	pVOdbore	pVyucbaPrax	pVyuzitieUciva	pTitul
p	0.5155	0.8551	0.0003029	0.0002688	0.0004418
x	pFinNarocne	pFinRodicia	fAkcie	fPovest	fOdporucam
p	0.7126	0.2058	0.02082	0.00005135	0.0000000000000613
x	fHrdy	fHodnotenie	fErasmus	fDobraVolba	internat
p	0.00003596	0.0000006474	0.3672	0.0000000005	0.2915

Zdroj: Vlastné spracovanie, 2020

Z tabuľky môžeme vidieť, že všetky premenné, ktoré vyjadrujú spokojnosť, či už so študijným programom, hodnotením, ale aj vybavením školy alebo vyučovanými predmetmi, sú považované za dôležité. Rovnako dôležité sú aj premenné týkajúce sa štúdia a samotnej fakulty, napr. či študent „prenášal“ predmet, náročnosť štúdia, vzťahy nadobudnuté počas štúdia a s tým súvisiace zúčastňovanie sa na akciách fakulty, povesť fakulty a fakt, či je na ňu študent hrdý, hlási sa ku nej a odporúčal by ju aj svojím známym. Celkové znenie otázok ku jednotlivým premenným nájdete v prílohe B a hodnoty premenných sú v prílohe C.

5 Štatistické modelovanie

V nasledujúcej kapitole sú popísané jednotlivé štatistické modely, ktoré boli vytvorené na základe určitých kritérií vychádzajúcich zo štúdií predstavených v kapitole 1. Modely boli vytvárané s využitím voľne dostupného programovacieho jazyka R určeného na štatistickú analýzu dát a ich grafickú vizualizáciu. Základné funkcie jazyka R je možné rozšíriť nainštalovaním rozličných knižníc, prostredníctvom ktorých používateľ získa prístup ku rôznym funkciám a štatistickým metódam, ktoré môže využívať. [15]

Na vytvorenie modelov bol použitý celý dataset obsahujúci 181 záznamov od respondentov študujúcich na bakalárskom stupni Fakulty riadenia a informatiky. Tento dataset bol v jazyku R pomocou funkcie *sample.split()* z knižnice *caTools* a funkcie *subset()* rozdelený na trénovaciu a testovaciu množinu. Trénovacia množina slúžila na natrénovanie predikčného modelu a tento model bol následne otestovaný na množine testovacej. Aby mohol byť model čo najpresnejší, je potrebné, aby sa natrénoval ideálne na čo najväčšej vzorke dát. Model ale musí byť na dostatočne veľkej vzorke dát aj otestovaný, takže ani testovacia množina nemôže byť malá. Po zvážení týchto faktov bol parameter splitRatio vo funkcii *sample.split()* nastavený na hodnotu *0,7*. Do trénovacej množiny bolo potom funkciou *subset()* vložených 70% dát a testovacia množina obsahovala zvyšných 30% dát.

Pred vytvorením modelov bolo potrebné skontrolovať závislosti medzi jednotlivými nezávislými premennými – koreláciu. Korelácia vyjadruje silu lineárnej závislosti (silu vzťahu) medzi premennými. Korelácia je žiadúca medzi závislou premennou a nezávislými premennými, avšak medzi jednotlivými nezávislými premennými je korelácia nežiadúca. Korelačný koeficient môže nadobúdať hodnoty z intervalu $< -1; +1 >$, pričom hodnota $+1$ vyjadruje perfektnú pozitívnu koreláciu, hodnota -1 perfektnú negatívnu koreláciu a hodnota 0 udáva, že medzi premennými neexistuje žiaden vzťah. Čím je hodnota korelačného koeficientu bližšia k 0, tým je vzťah medzi premennými slabší, resp. až neexistujúci. Koreláciu medzi jednotlivými nezávislými premennými sme v jazyku R zisťovali zostrojením korelačnej matice, a to zavolaním funkcie *cor()*. Korelačnú maticu sme zostrojili pred samotným vytvorením modelov aby sme zistili, či použité nezávislé premenné nemajú medzi sebou veľmi silný vzťah a výrazne sa neovplyvňujú. Hodnoty korelačného koeficientu sme interpretovali nasledovne [15]:

$< -1; -0{,}6 > \ \vee \ < 0{,}6; 1 > \ \Rightarrow$ silný vzťah
$(-0{,}6; -0{,}3 > \ \vee \ < 0{,}3; 0{,}6) \ \Rightarrow$ stredne silný vzťah
$(-0{,}3; 0 > \ \vee \ < 0; 0{,}3) \ \Rightarrow$ slabý až neexistujúci vzťah

Ako prípustnú hranicu sme zvolili interval $< -0{,}6; 0{,}6 >$. V prípade, ak by bola absolútna hodnota korelačného koeficientu väčšia ako $0{,}6$, zaradenie týchto premenných do modelu budeme musieť zvážiť, nakoľko zaradenie oboch premenných do modelu môže spôsobiť to, že sa tieto premenné budú navzájom ovplyvňovať v dôsledku čoho budú obe štatisticky nevýznamné.

Cieľom vytvorených modelov bolo na základe informácií získaných od študentov predikovať rozhodnutie študenta pokračovať v štúdiu na druhom stupni štúdia na Fakulte riadenia a informatiky. Toto rozhodnutie študenta bolo našou závislou premennou. Nakoľko ide o problém klasifikačný, bolo potrebné využiť metódy na riešenie klasifikačných problémov. Pri vytváraní modelov sa vychádzalo zo štúdií spomínaných v kapitole 1, z ktorých boli do jednotlivých modelov zvolené adekvátne premenné. Všetky informácie ku jednotlivým premenným využitým v modeloch a im prislúchajúcim otázkam sú dostupné v prílohe B a prílohe C.

Pri skúmaní faktorov vplývajúcich na celkovú spokojnosť študenta so školou a jeho zotrvanie na škole boli v minulosti využité rozličné postupy a metódy, a to napríklad aj klasifikačné rozhodovacie stromy alebo logistická regresia (pozri kapitola 1). Nakoľko v prípade nami riešeného problému má naša závislá premenná fIng, vyjadrujúca rozhodnutie študenta pokračovať v štúdiu na FRI štyri kategórie, pri riešení nášho problému bola využitá multinomická logistická regresia. Binárnu logistickú regresiu sme využili pri zjednodušení nášho problému, kedy sme závislú premennú so štyrmi kategóriami nahradili premennou binárnou. Nakoniec sme skúsili náš problém riešiť pomocou klasifikačných rozhodovacích stromov.

5.1 Modely multinomickej logistickej regresie (MLR)

Jednou z možností predikcie rozhodnutia študenta pokračovať v štúdiu je vytvorením modelov multinomickej logistickej regresie. Multinomická logistická regresia je rozšírením logistickej regresie, pričom narozdiel od logistickej regresie má závislá premenná pri multinomickej regresii viac ako dve kategórie. Závislá premenná v nami vytvorených modeloch multinomickej logisticke regresie je premenná fIng a má štyri kategórie (pozri príloha B a príloha C). Na vytvorenie modelu multinomickej logistickej regresie existuje v jazyku R funkcia *multinom()* z balíčka *nnet*, ktorý je potrebné nainštalovať pomocou príkazu *install.packages("nnet")* a pred použitím načítať príkazom *library(nnet)*.

5.1.1 Model č.1 zohľadňujúci stresové faktory (MLR)

Pri tvorbe modelu č.1 sme vychádzali zo štúdie z roku 2002, v ktorej Cotton, Dollard a Jonge preukázali, že na rozhodnutie študenta pokračovať v štúdiu a neukončiť ho predčasne vplývajú stresové faktory. V našom prípade teda boli do modelu zvolené premenné, ktoré odrážajú psychický stav študenta a teda či má pocit, že sú naňho kladené privysoké nároky, alebo či musel niekedy vyhľadať odbornú pomoc psychológa. Jedná sa o premenné sSocZivot, sPrenasam, sOpakujem, sNezvladam, sPsycholog, sPredcasne, sPredcasneKam, sNarocne, sNaroky, sPomoc a sVztahy.

Pred samotným vytvorením modelu bolo potrebné skontrolovať mieru korelácie medzi nezávislými premennými. Pre tento účel sme v jazyku R zostrojili korelačnú maticu pomocou príkazu $cor(model1)$ a sledovali sme, či sa hodnota korelačného koeficientu nachádza v nami stanovenom intervale $< -0{,}6; 0{,}6 >$. Korelačnú maticu všetkých premenných zvolených do modelu č.1 môžeme vidieť na nasledujúcom obrázku č.2.

	fIng	sPrenasam	sOpakujem	sNezvladam	sSocZivot	sPredcasne	sPredcasneKam
fIng	1.0000000	-0.29769086	-0.26904636	-0.3164618	-0.1986071	-0.3431581	-0.12717871
sPrenasam	-0.2976909	1.00000000	0.52891974	0.3504183	0.1351956	0.2719838	0.05338523
sOpakujem	-0.2690464	0.52891974	1.00000000	0.3625330	0.1152107	0.3328572	0.15737739
sNezvladam	-0.3164618	0.35041832	0.36253297	1.0000000	0.3730675	0.3869653	0.17720725
sSocZivot	-0.1986071	0.13519564	0.11521073	0.3730675	1.0000000	0.2282882	0.15550927
sPredcasne	-0.3431581	0.27198378	0.33285724	0.3869653	0.2282882	1.0000000	0.22789657
sPredcasneKam	-0.1271787	0.05338523	0.15737739	0.1772072	0.1555093	0.2278966	1.00000000
sPomoc	0.1884283	-0.03549628	0.01315547	-0.1224927	-0.1180941	-0.0925067	-0.06861073
sVztahy	0.1505434	0.06033308	0.05210193	-0.1162201	-0.1936987	-0.1395791	-0.12376852
sPsycholog	-0.3158087	0.08045765	0.13828157	0.2628472	0.2440920	0.3353274	0.15042728
sNarocne	-0.1995968	0.19367045	0.19310911	0.4994518	0.3481516	0.2489416	0.03114528
sNaroky	-0.3193848	0.13930945	0.13933997	0.3886740	0.2740236	0.2295076	0.15242408

	sPomoc	sVztahy	sPsycholog	sNarocne	sNaroky
fIng	0.18842830	0.15054338	-0.31580874	-0.19959676	-0.31938480
sPrenasam	-0.03549628	0.06033308	0.08045765	0.19367045	0.13930945
sOpakujem	0.01315547	0.05210193	0.13828157	0.19310911	0.13933997
sNezvladam	-0.12249269	-0.11622009	0.26284720	0.49945183	0.38867401
sSocZivot	-0.11809408	-0.19369869	0.24409196	0.34815157	0.27402357
sPredcasne	-0.09250670	-0.13957910	0.33532745	0.24894161	0.22950758
sPredcasneKam	-0.06861073	-0.12376852	0.15042728	0.03114528	0.15242408
sPomoc	1.00000000	0.36763464	-0.07395376	0.11116924	-0.12039764
sVztahy	0.36763464	1.00000000	-0.17331703	0.07848369	-0.06384924
sPsycholog	-0.07395376	-0.17331703	1.00000000	0.22795699	0.26938146
sNarocne	0.11116924	0.07848369	0.22795699	1.00000000	0.47608873
sNaroky	-0.12039764	-0.06384924	0.26938146	0.47608873	1.00000000

Obrázok 2 – Korelačná matica premenných z modelu č.1

Zdroj: Vlastné spracovanie, 2020

Ako z korelačnej matice vyplýva, žiadna dvojica premenných nemá hodnotu korelačného koeficientu vyššiu ako 0,6, a teda nie je medzi nimi vysoká miera korelácie. Všetky premenné môžu byť do modelu zaradené. Model č.1 multinomickej logistickej regresie bol vytvorený pomocou príkazu *multinom()* z balíčka *nnet*. Základné informácie o modeli a niektoré jeho hodnotiace štatistiky získame zavolaním funkcie *summary()*. Výstup tejto funkcie môžeme vidieť na obrázku č.3 na nasledujúcej strane.

```
> #MODEL C.1
> summary(multinomial)
Call:
multinom(formula = fIng ~ sPrenasam + sOpakujem + sNezvladam +
    sSocZivot + sPredcasne + sPredcasneKam + sPomoc + sVztahy +
    sPsycholog + sNarocne + sNaroky, data = train)

Coefficients:
  (Intercept) sPrenasam  sOpakujem sNezvladam sSocZivot sPredcasne sPredcasneKam     sPomoc
1  -31.401585 -1.539023  0.2022546  -1.077130  2.327841 -0.9226472    16.4915821 -0.2422802
2  -29.619925 -3.923625  2.6561928  -2.070993  3.427979 -1.0857738    14.0619326 -0.2170400
3    4.031535 -2.420146 -0.1435015  -1.083700  1.290040 -1.0181790     0.1372382  0.1583119
    sVztahy sPsycholog   sNarocne  sNaroky
1 1.9051841  -1.686326 -2.1778615 1.712457
2 0.0975072  -2.837626 -0.5448863 2.266140
3 1.5703984  -2.507702 -1.4098977 1.071684

Std. Errors:
  (Intercept) sPrenasam sOpakujem sNezvladam sSocZivot sPredcasne sPredcasneKam    sPomoc
1   0.4383290  1.218003 0.5411423  1.0184799 0.7393526  0.5591766     0.8766592 0.8788532
2   0.6987782  1.860885 1.1199117  1.3784154 1.2627809  0.7740074     1.3975572 1.2016809
3   2.4031283  1.189705 0.5257848  0.9671726 0.6943808  0.5229930     0.7791806 0.8383953
    sVztahy sPsycholog sNarocne  sNaroky
1 0.9361244  0.8062536 1.182069 1.097992
2 1.2135838  1.1524553 1.629840 1.373685
3 0.9543487  0.8286198 1.096654 1.063510

Residual Deviance: 171.8073
AIC: 243.8073
```

Obrázok 3 – Výpis funkcie *summary()* pôvodného modelu č.1 (MLR)

Zdroj: Vlastné spracovanie, 2020

Jedným z hodnotiacich kritérií modelov je Akaikeho informačné kritérium (AIC), ktoré slúži na porovnávanie modelov. Toto kritérium je obsiahnuté priamo vo výpise funkcie summary(). Ak porovnávame dva modely, tak čím je hodnota tohto kritéria daného modelu menšia, tým sa dá tento model považovať za lepší. V prípade modelu č.1 multinomickej logistickej regresie je hodnota $AIC = 243.8073$.

Ďalším hodnotiacim kritériom regresných modelov je koeficient determinácie R^2 využívaný pri lineárnej regresií. Tento koeficient udáva koľko percent variability dát vytvorený model vysvetľuje – inak povedané, udáva o koľko je daný model lepší ako základný (baseline) model. V prípade logistickej regresie je možné koeficient determinácie R^2 čiastočne nahradiť McFaddenovým pseudo-R^2. Nakoľko sa ale nejedná o úplný ekvivalent koeficientu determinácie známeho z lineárnej regresie a tieto dve štatistiky neznamenajú úplne to isté, pri jeho interpretácií sa odporúča opatrnosť. Tento pseudo koeficient determinácie môže nadobúdať hodnoty z intervalu $< 0; 1 >$, pričom už pri hodnotách okolo 0,2 môžeme hovoriť o uspokojivom modeli. Nakoľko nám výpis *summary()* toto hodnotiace kritérium neposkytuje, je potrebné ho vypočítať. Pre výpočet rozličných variant pseudo-R^2 hodnotiacich štatistík existuje v knižnici *DescTools* funkcia *PseudoR2()*, ktorej výstup môžeme vidieť na obrázku č.4. [15][20][35]

```
> PseudoR2(multinomial, which = "all")
      McFadden    McFaddenAdj       CoxSnell      Nagelkerke   AldrichNelson
     0.3471505      0.0735582      0.5213367      0.5922637             NA
VeallZimmermann          Efron McKelveyZavoina           Tjur            AIC
            NA             NA             NA             NA     243.8073233
           BIC         logLik        logLik0             G2
   345.3374596    -85.9036616   -131.5826442     91.3579651
```

Obrázok 4 – Pseudo-R^2 pôvodného modelu č.1 (MLR)

Zdroj: Vlastné spracovanie, 2020

Funkcia *PseudoR2()* vypočíta pri nastavení parametra *which = "all"* aj iné hodnotiace kritériá a pseudo-R^2 štatistiky. McFaddenovo pseudo-R^2 je najpoužívanejším indexom determinácie a preto sme ho pri našej práci zvolili aj my. Hodnota McFaddenovho pseudo-R^2 nášho modelu je 0.3471. Hodnota pseudo-R^2 je vyššia ako 0.2 čo značí, že náš model je viac než uspokojivý. [35]

Koeficient determinácie R^2 slúži na celkové hodnotenie kvality modelu (ako dobre model „sedí" na dáta), avšak hodnota tohto koeficientu pridaním každej novej nezávislej premennej rastie a to aj v prípade, že by nezávislé premenné modelu boli korelované. Je teda dôležité sledovať hodnotu korigovaného koeficientu determinácie, ktorý slúži na porovnávanie modelov na základe premenných v nich obsiahnutých. Na základe hodnoty tohto koeficientu vieme povedať, ktorý model je v skutočnosti lepší. Hodnota korigovaného koeficientu determinácie klesá, ak je do modelu pridaná premenná, ktorá presnosť modelu nezlepšuje. To platí aj v prípade McFaddenovho korigovaného pseudo-R^2 a preto je dôležité túto hodnotu sledovať. Jeho hodnota je v prípade nášho modelu 0.0736. [16][23]

Zvýšiť presnosť modelu je možné odstránením tých premenných z modelu, ktorých parametre nie sú štatisticky významné. Výpis funkcie *summary()* pri modeli multinomickej logistickej regresie nám však opäť – narozdiel od výpisu pri logistickej regresií – neposkytuje p-hodnoty parametrov pri jednotlivých premenných, na základe ktorých vieme určiť ich štatistickú významnosť. Výpočet p-hodnôt je však pomerne jednoduchý. Najskôr je potrebné vypočítať hodnoty z-value. Tie vypočítame ako pomer koeficientov modelu a ich štandardných chýb. Z hodnôt z-value vypočítame p-hodnotu pomocou štandardného normálneho rozdelenia príkazom *p = pnorm((abs(z)), lower.tail = FALSE)*2*. Funkcia *pnorm()* slúži na výpočet integrálu od hodnoty nášho testovacieho kritéria $z - value$ po ∞ (pri nastavení parametra *lower.tail = FALSE)*. Tento integrál predstavuje našu p-hodnotu, ktorú (ako bolo definované v kapitole 2.2.2) použijeme na testovanie hypotéz a určenie štatistickej významnosti parametra. P-hodnoty jednotlivých parametrov pre každú kategóriu závislej premennej fIng môžeme vidieť na obrázku č.5. [19][28]

```
> z = summary(multinomial)$coefficients/summary(multinomial)$standard.errors
> pnorm((abs(z)), lower.tail = FALSE) * 2
  (Intercept)  sPrenasam  sOpakujem sNezvladam   sSocZivot sPredcasne sPredcasneKam
1  0.00000000 0.20638704 0.70858671  0.2902445 0.001641190 0.09894080  6.039309e-79
2  0.00000000 0.03499013 0.01770226  0.1329822 0.006635039 0.16067795  8.149938e-24
3  0.09342136 0.04192737 0.78490846  0.2625080 0.063193456 0.05155503  8.601907e-01
     sPomoc    sVztahy  sPsycholog  sNarocne    sNaroky
1 0.7827957 0.04183249 0.036478131 0.0654144 0.11884821
2 0.8566708 0.93596168 0.013807087 0.7381389 0.09900845
3 0.8502282 0.09986294 0.002475167 0.1985703 0.31360536
```

Obrázok 5 – P-hodnoty parametrov modelu č.1 (MLR)

Zdroj: Vlastné spracovanie, 2020

P-hodnoty parametrov z obrázku č.5 porovnávame s hladinou významnosti α, na ktorej má byť daný parameter štatisticky významný. Ak je $p - hodnota < \alpha$, parameter je na danej hladine významnosti α štatisticky významný. Okrem parametrov pri premenných sNezvladam a sPomoc, sú všetky parametre aspoň pri jednej kategórii závislej premennej fIng štatisticky významné, a to najmenej na hladine významnosti $\alpha = 0.1$.

Premenné sme na základe štatistického významu ich parametrov z modelu postupne odstraňovali, až kým všetky parametre modelu neboli štatisticky významné. Taktiež pri každom odstránení takejto premennej bola skontrolovaná hodnota AIC modelu. V modeli boli ponechané premenné, ktorých parametre boli štatisticky významné aspoň pri jednej kategórii závislej premennej fIng a ich prípadné odstránenie malo negatívny vplyv na hodnotu AIC. Premenné, ktoré boli v modeli ponechané a p-hodnoty ich parametrov sú nasledovné:

```
> z = summary(multinomial)$coefficients/summary(multinomial)$standard.errors
> pnorm((abs(z)), lower.tail = FALSE) * 2
  (Intercept)  sPrenasam  sOpakujem   sSoczivot sPredcasne sPredcasneKam    sVztahy
1   0.0000000 0.12087197 0.87500686 0.005218136 0.09392522  3.349671e-81 0.04689469
2   0.0000000 0.04424096 0.04427812 0.009458302 0.11902064  5.937208e-39 0.90594735
3   0.1693169 0.02665747 0.43845693 0.138533862 0.04867615  9.911109e-01 0.05025974
   sPsycholog
1 0.025110326
2 0.028975388
3 0.001152796
```

Obrázok 6 – P-hodnoty ponechaných parametrov modelu č.1 (MLR)

Zdroj: Vlastné spracovanie, 2020

Z modelu boli odstránené premenné sNezvladam, sPomoc, sNarocne a sNaroky. Hodnota Akaikeho informačného kritéria tohto upraveného modelu č.1 multinomickej logistickej regresie je $AIC = 231.0318$. Hodnota McFaddenovho pseudo-R^2 je menšia ako pri pôvodnom modeli a je rovná 0.3044. Korigovaný koeficient McFadden pseudo-R^2, ktorého hodnota je v prípade upraveného modelu 0.1221 vzrástol takmer dvojnásobne. Na základe nárastu hodnoty tohto koeficientu v upravenom modeli môžeme povedať, že niektoré premenné aj napriek tomu, že zvyšovali celkovú hodnotu pseudo-R^2 v pôvodnom modeli nezlepšovali jeho celkovú presnosť.

Vytvorené modely bolo potrebné otestovať a aplikovať aj na testovacej množine a určiť nakoľko je vytvorený predikčný model úspešný pri predikovaní rozhodnutia študenta pokračovať v štúdiu na Fakulte riadenia a informatiky. V tabuľke č.6 na nasledujúcej strane sú uvedené úspešnosti pôvodného aj upraveného modelu č.1 multinomickej logistickej regresie na trénovacej aj testovacej množine.

Tabuľka 6 – Úspešnosť pôvodného a upraveného modelu č.1 (MLR)

Model	Úspešnosť na trénovacej množine	Úspešnosť na testovacej množine
pôvodný model č.1	70.16%	52.63%
upravený model č.1	70.97%	52.63%

Zdroj: Vlastné spracovanie, 2020

Ako z tabuľky č.6 vyplýva, pri oboch modeloch je úspešnosť na testovacej množine výrazne nižšia ako na množine trénovacej. Oba modely na testovacej množine vykazujú rovnakú úspešnosť. Modely však na testovacej množine nadobúdajú úspešnosť len cca 53%, čo znamená, že modely nie sú na predikciu rozhodnutia študenta pokračovať v štúdiu vhodné.

5.1.2 Model č.2 zohľadňujúci bydlisko študenta (MLR)

Druhý model bol inšpirovaný štúdiou z roku 2007 (Ho Yu, Yu, Digangi a kol.), v ktorej bol dokázaný vzťah medzi zotrvaním na škole a bydliskom študenta, ktorý môže vplývať na financie. Zvolené premenné boli kraj, bydlisko, internat a premenná pFinNarocne.

Rovnako ako pri vytváraní modelu č.1 sme aj pred vytvorením tohto modelu najskôr zostrojili korelačnú maticu (výstup na obr.7) pomocou príkazu $cor(model2)$, aby sme zistili, či medzi zvolenými nezávislými premennými neexistuje silná väzba.

	fIng	pFinNarocne	bydlisko	kraj	internat
fIng	1.000000000	-0.1059339	0.13905963	-0.001019817	0.09048282
pFinNarocne	-0.105933861	1.0000000	-0.10128815	-0.192694945	0.19305220
bydlisko	0.139059627	-0.1012881	1.00000000	0.027459570	-0.11215450
kraj	-0.001019817	-0.1926949	0.02745957	1.000000000	-0.48310958
internat	0.090482819	0.1930522	-0.11215450	-0.483109579	1.00000000

Obrázok 7 – Korelačná matica premenných z modelu č.2

Zdroj: Vlastné spracovanie, 2020

Nakoľko sa nepreukázala silná korelácia medzi jednotlivými nezávislými premennými, všetky zvolené premenné boli do modelu zaradené. Model č.2 multinomickej logistickej regresie bol rovnako ako prvý model vytvorený v jazyku R príkazom *multinom()*. Na obrázku č.8 na nasledujúcej strane môžeme vidieť výstup funkcie *summary()* poskytujúci informácie o vytvorenom modeli.

```
> summary(multinomial)
Call:
multinom(formula = fIng ~ pFinNarocne + bydlisko + kraj + internat,
    data = train)

Coefficients:
  (Intercept) pFinNarocne   bydlisko1     kraj2     kraj3    kraj4     kraj5
1   -19.36320  -0.7576729  0.00690568  94.80628  22.36792 21.50403  90.91689
2   -12.11410  -0.2668193 -0.32699034  85.87131 -60.76984 12.83786 -10.38439
3   -37.49528  -0.7070329  0.64679445 112.64088  40.07884 39.19999 -44.77957
      kraj6     kraj7    kraj8   internat
1 -22.25662 -44.64679 21.59223 -0.6349585
2 -12.53251  67.77532 12.51395 -1.0812616
3  83.95385  92.62087 39.77845  0.3798100

Std. Errors:
  (Intercept) pFinNarocne bydlisko1     kraj2     kraj3     kraj4        kraj5
1   0.8710461   0.4177530 0.7345926 0.4882293 0.9883648 0.6863508 4.360191e-16
2   1.0685101   0.5569614 0.9904824 0.7165963       NaN 0.8455839 6.549443e-17
3   0.8039069   0.3810224 0.6711048 0.4411297 0.9441815 0.6299969 2.165916e-67
         kraj6     kraj7     kraj8  internat
1          NaN       NaN 0.5313403 0.8071277
2          NaN 0.5973685 0.6782572 1.1493204
3 3.580253e-20 0.5973685 0.4901172 0.7310219

Residual Deviance: 228.1191
AIC: 294.1191
```

Obrázok 8 – Výpis funkcie *summary()* pôvodného modelu č.2 (MLR)

Zdroj: Vlastné spracovanie, 2020

Ako môžeme z výpisu *summary()* vidieť, hodnota Akaikeho informačného kritéria modelu č.2 multinomickej regresie je $AIC = 294.1191$. Hodnota McFaddenovho pseudo-R^2 modelu č.2 je rovná 0.1331. Ako z obrázku č.9 vyplýva, aj hodnota korigovaného pseudo-R^2 je veľmi nízka a model vôbec nie je dobrý.

```
> PseudoR2(multinomial, which = "all")
       McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
      0.1331716      -0.1176213       0.2462015       0.2796968              NA
 VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
              NA              NA              NA              NA     294.1191428
             BIC          logLik         logLik0              G2
     387.1884345    -114.0595714    -131.5826442      35.0461456
```

Obrázok 9 – Pseudo-R^2 pôvodného modelu č.2 (MLR)

Zdroj: Vlastné spracovanie, 2020

Pri výpočte p-hodnôt a určovaní štatistickej významnosti parametrov modelu sme postupovali rovnako ako pri prvom modeli multinomickej logistickej regresie v predchádzajúcej kapitole. P-hodnoty parametrov pôvodného modelu č.2 sú nasledovné:

```
> pnorm((abs(z)), lower.tail = FALSE) * 2
    (Intercept) pFinNarocne bydlisko1 kraj2         kraj3         kraj4 kraj5 kraj6
1 1.768404e-109  0.06972600 0.9924994     0 2.135363e-113 1.767975e-215     0   NaN
2  8.567384e-30  0.63189417 0.7412999     0           NaN  4.636374e-52     0   NaN
3  0.000000e+00  0.06350769 0.3351584     0  0.000000e+00  0.000000e+00     0     0
  kraj7        kraj8  internat
1   NaN 0.000000e+00 0.4314639
2     0 5.199379e-76 0.3468159
3     0 0.000000e+00 0.6033701
```

Obrázok 10 – P-hodnoty parametrov modelu č.2 (MLR)

Zdroj: Vlastné spracovanie, 2020

Ako štatisticky nevýznamné sa preukázali parametre pri premenných bydlisko a internat. Ostatné parametre boli štatisticky významné minimálne na hladine významnosti $\alpha = 0.1$, a to aspoň pri jednej z kategórií závislej premennej. Pri faktorovej premennej kraj

sme zaznamenali chýbajúce údaje z dôvodu nedostatku dát v datasete pre jednotlivé kraje (väčšina respondentov bola zo Žilinského kraja). Pri postupnom odstraňovaní premenných, na základe štatistickej významnosti ich parametrov, boli z modelu odstránené premenné bydlisko a internat a v modeli boli ponechané premenné pFinNarocne a kraj. Nakoľko odstránené premenné bydlisko a internat – vychádzajúc zo štúdie predstavenej v úvode kapitoly – môžu mať vplyv na zotrvanie študenta na škole, bol vytvorený aj samostatný model s využitím týchto premenných. Parametre pri premenných sa však ani v tomto modeli nepreukázali ako štatisticky významné a preto sme s týmto modelom ďalej nepracovali.

Pri vytvorení upraveného modelu č.2 multinomickej regresie boli použité premenné kraj a pFinNarocne. Odstránenie premenných bydlisko a internat znížilo hodnotu AIC na $AIC = 288.8358$. Hodnota McFaddenovho pseudo-R^2 je v prípade tohto modelu 0.1076 a korigovaný pseudo-R^2 je rovný -0.09754. Napriek tomu, že po odstránení premenných bydlisko a internat sa hodnota korigovaného pseudo-R^2 zvýšila, stále sa jedná o zápornú hodnotu a koeficient je veľmi nízky. Pôvodný aj upravený model sme aplikovali na testovaciu množinu dát. Úspešnosti týchto modelov na trénovacej a testovacej množine môžeme vidieť v tabuľke č.7.

Tabuľka 7 – Úspešnosť pôvodného a upraveného modelu č.2 (MLR)

Model	Úspešnosť na trénovacej množine	Úspešnosť na testovacej množine
pôvodný model č.2	62.90%	52.63%
upravený model č.2	61.29%	49.12%

Zdroj: Vlastné spracovanie, 2020

Z tabuľky č. 7 vyplýva, že pôvodný model č.2 multinomickej logistickej regresie vykazuje vyššiu úspešnosť na trénovacej aj testovacej množine. Úspešnosti modelov sú však v porovnaní s modelmi č.1 zohľadňujúcimi stresové faktory nižšie, čo znamená, že ani modely vytvorené v rámci tejto kapitoly nie sú na predikciu rozhodnutia študenta pokračovať v štúdiu vhodné. Toto vyplýva aj zo zvyšných hodnotiacich štatistík modelu na základe ktorých sa model už od začiatku javil ako nie veľmi dobrý.

5.1.3 Model č.3 zohľadňujúci spokojnosť študenta (MLR)

Model č.3 vychádza zo štúdie od Siminga a kol. z roku 2015 predstavenej v kapitole 1 v ktorej ukázali, že na celkovú spokojnosť študenta majú vplyv rozličné faktory. Toho istého roku Kravčáková, Minárová a Župová aplikovali Herzbergovu teóriu pracovnej

motivácie pri uvažovaní faktorov vplývajúcich na vzťah študenta ku štúdiu (práci). Tieto faktory na základe tejto teórie rozdelili na motivátory (vnútorné faktory) a dissatisfaktory (vonkajšie faktory). Do modelu boli zvolené premenné týkajúce sa spokojnosti študenta, medzi ktoré patria: spVybavenie, spOdbornost, spPristup, spOdradza, spZbytocne, spZastarale, spStudProgram, spHodnotenie, spTerminy, spPrilezitosti, spSemester a odbor. Aj pri vytváraní tohto modelu bolo potrebné skontrolovať koreláciu medzi nezávislými premennými. Korelačnú maticu môžeme vidieť na obrázku č.11.

	fIng	odbor	spOdradza	spZbytocne	spZastarale	spStudProgram	spVybavenie
fIng	1.0000000	-0.15650599	-0.24562009	-0.21091867	-0.31693535	0.51127392	0.24944660
odbor	-0.1565060	1.00000000	-0.02454686	0.02312498	0.07112197	-0.06574973	0.05088977
spOdradza	-0.2456201	-0.02454686	1.00000000	0.49039445	0.36993992	-0.36842871	-0.11514998
spZbytocne	-0.2109187	0.02312498	0.49039445	1.00000000	0.36824998	-0.29603305	-0.23352874
spZastarale	-0.3169353	0.07112197	0.36993992	0.36824998	1.00000000	-0.36172845	-0.21163617
spStudProgram	0.5112739	-0.06574973	-0.36842871	-0.29603305	-0.36172845	1.00000000	0.33834863
spVybavenie	0.2494466	0.05088977	-0.11514998	-0.23352874	-0.21163617	0.33834863	1.00000000
spOdbornost	0.2325923	-0.04507743	-0.18841922	-0.25926157	-0.31028563	0.30226205	0.39488959
spPristup	0.3562059	-0.20949473	-0.21775419	-0.30694262	-0.34230978	0.44128531	0.37162471
spHodnotenie	0.3586212	-0.05971735	-0.26017598	-0.25012134	-0.27481282	0.35955258	0.12242935
spTerminy	0.3111330	-0.21813055	-0.27660200	-0.22236955	-0.24144142	0.37568773	0.24612711
spPrilezitosti	0.1907395	0.08321015	-0.15170036	-0.25938478	-0.21109517	0.32916309	0.32303928
spSemester	0.4404819	-0.34879171	-0.34477584	-0.30794222	-0.36212345	0.49040950	0.18115298

	spOdbornost	spPristup	spHodnotenie	spTerminy	spPrilezitosti	spSemester
fIng	0.23259226	0.3562059	0.35862120	0.3111330	0.19073950	0.4404819
odbor	-0.04507743	-0.2094947	-0.05971735	-0.2181305	0.08321015	-0.3487917
spOdradza	-0.18841922	-0.2177542	-0.26017598	-0.2766020	-0.15170036	-0.3447758
spZbytocne	-0.25926157	-0.3069426	-0.25012134	-0.2223696	-0.25938478	-0.3079422
spZastarale	-0.31028563	-0.3423098	-0.27481282	-0.2414414	-0.21109517	-0.3621234
spStudProgram	0.30226205	0.4412853	0.35955258	0.3756877	0.32916309	0.4904095
spVybavenie	0.39488959	0.3716247	0.12242935	0.2461271	0.32303928	0.1811530
spOdbornost	1.00000000	0.5305514	0.37086373	0.2500669	0.25936397	0.2328787
spPristup	0.53055139	1.0000000	0.49522195	0.4148392	0.36990061	0.4239883
spHodnotenie	0.37086373	0.4952219	1.00000000	0.4210879	0.28382843	0.3422039
spTerminy	0.25006688	0.4148392	0.42108786	1.0000000	0.24799648	0.3222618
spPrilezitosti	0.25936397	0.3699006	0.28382843	0.2479965	1.00000000	0.1524724
spSemester	0.23287872	0.4239883	0.34220388	0.3222618	0.15247244	1.0000000

Obrázok 11 – Korelačná matica premenných z modelu č.3

Zdroj: Vlastné spracovanie, 2020

Nakoľko sa medzi jednotlivými nezávislými premennými nepreukázala silná korelácia, všetky pôvodne zvolené premenné boli použité pri vytvorení modelu multinomickej logistickej regresie. Informácie o vytvorenom modeli, ktoré nám poskytuje výpis funkcie *summary()* sú na obrázku č.12.

```
> #MODEL C.3
> summary(multinomial)
Call:
multinom(formula = fIng ~ odbor + spPrilezitosti + spSemester +
    spOdradza + spZbytocne + spZastarale + spStudProgram + spVybavenie +
    spOdbornost + spPristup + spHodnotenie + spTerminy, data = train)

Coefficients:
  (Intercept)     odbor1     odbor2 spPrilezitosti spSemester spOdradza spZbytocne spZastarale
1  -10.369115  14.932090 12.4820848     0.49478560  0.2307427 -2.274850   2.551518  0.34242920
2   -8.422251 -31.511489 -6.0932279     1.36487894 -1.7826644 -1.278256   3.336585 -1.27831810
3   -2.270452   1.876192 -0.3972716     0.04843732  0.3082864 -1.477229   1.526868  0.08989989
  spStudProgram spVybavenie spOdbornost  spPristup spHodnotenie  spTerminy
1      1.382534    0.875335  -1.0286817 -0.9405611   -1.0684838 -1.1022306
2     -1.590100    6.542215  -4.0697916 -0.7835980    3.1492469  0.7740699
3      2.257906    1.588633  -0.8292353 -0.6287602    0.2339284 -0.5470675

Std. Errors:
  (Intercept)       odbor1   odbor2 spPrilezitosti spSemester spOdradza spZbytocne spZastarale
1    2.741083 1.648595e+00 1.465193      0.7162612  0.7807341  1.107988  0.8859403   0.6448231
2    7.624097 6.786254e-09 3.295583      1.2633314  1.2739141  1.797797  1.4070936   1.2072095
3    4.253561 2.266855e+00 2.055611      0.6867870  0.6785207  1.041051  0.7369231   0.5575394
  spStudProgram spVybavenie spOdbornost spPristup spHodnotenie spTerminy
1     0.7231233   0.7184671   0.7427459 0.7950217    0.8164920 0.6316494
2     1.8192418   2.5092479   1.7726210 1.1676718    1.5449892 1.2058830
3     0.7064922   0.7366963   0.7620127 0.7661868    0.7485808 0.5766113

Residual Deviance: 142.948
AIC: 226.948
```

Obrázok 12 – Výpis funkcie *summary()* pôvodného modelu č.3 (MLR)

Zdroj: Vlastné spracovanie, 2020

Akaikeho informačné kritérium má pri modeli č.3 hodnotu $AIC = 226.948$. Náš index determinácie McFaddenovo pseudo-R^2 pri tomto modeli nadobúda hodnotu 0.4568. Hodnota tohto koeficientu je veľmi vysoká a dá sa predpokladať, že tento model bude pri predikcií rozhodnutia študenta vhodný. Všetky ostatné charakteristiky poskytuje výstup funkcie *PseudoR2()* na obrázku č.13. Je tu opäť predpoklad, že hodnota korigovaného pseudo-R^2 sa prípadným odstránením premenných, pri ktorých sú parametre štatisticky nevýznamné, zlepší.

```
> PseudoR2(multinomial, which = "all")
       McFadden     McFaddenAdj        CoxSnell       Nagelkerke   AldrichNelson
      0.4568129       0.1376219       0.6207252       0.7051739              NA
VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
             NA              NA              NA              NA     226.9479941
            BIC          logLik         logLik0              G2
    345.3998198     -71.4739970    -131.5826442     120.2172943
```

Obrázok 13 – Pseudo-R^2 pôvodného modelu č.3 (MLR)

Zdroj: Vlastné spracovanie, 2020

Po vytvorení modelu bolo teda opäť potrebné zistiť štatistickú významnosť parametrov premenných modelu, na základe ich p-hodnôt. Tieto p-hodnoty parametrov môžeme vidieť na nasledujúcom obrázku č.14 a interpretujeme ich rovnako ako pri predošlých modeloch.

```
> z = summary(multinomial)$coefficients/summary(multinomial)$standard.errors
> pnorm((abs(z)), lower.tail = FALSE) * 2
  (Intercept)       odbor1       odbor2 spPrilezitosti spSemester  spOdradza  spZbytocne
1 0.0001550411 1.335181e-19 1.608400e-17      0.4896979  0.7675769 0.04005937 0.003976604
2 0.2692946249 0.000000e+00 6.447118e-02      0.2799727  0.1617051 0.47707646 0.017727563
3 0.5934960269 4.078614e-01 8.467537e-01      0.9437739  0.6495764 0.15590516 0.038270034
  spZastarale spStudProgram spVybavenie spOdbornost spPristup spHodnotenie  spTerminy
1   0.5953885   0.055890040 0.223095954   0.1660613 0.2367840   0.19066058 0.08098418
2   0.2896438   0.382093442 0.009127495   0.0216806 0.5021713   0.04151377 0.52093078
3   0.8719012   0.001393766 0.031050198   0.2764992 0.4118538   0.75466370 0.34274114
```

Obrázok 14 – P-hodnoty parametrov modelu č.3 (MLR)

Zdroj: Vlastné spracovanie, 2020

Premenné spPrilezitosti, spSemester, spZastarale a spPristup boli postupne z modelu odstránené, nakoľko sa ich parametre preukázali ako štatisticky nevýznamné. Všetky ostatné parametre boli štatisticky významné minimálne na hladine významnosti $\alpha = 0.1$, a to aspoň pri jednej z kategórií závislej premennej fIng. V modeli boli ponechané premenné odbor, spOdradza, spZbytocne, spStudProgram, spVybavenie, spHodnotenie a spTerminy.

Upravený model má hodnotu Akaikeho informačného kritéria $AIC = 212.9902$. McFaddenovo pseudo-R^2 nadobúda hodnotu 0.4186. Hodnota korigovaného pseudo-R^2 sa pri upravenom modeli zvýšila na 0.1906. Oba vytvorené modely č.3, pôvodný aj upravený, boli aplikované na trénovaciu aj na testovaciu množinu. Úspešnosti oboch týchto modelov sú uvedená v tabuľke č.8 na nasledujúcej strane.

Tabuľka 8 – Úspešnosť pôvodného a upraveného modelu č.3 (MLR)

Model	Úspešnosť na trénovacej množine	Úspešnosť na testovacej množine
pôvodný model č.3	74.19%	66.67%
upravený model č.3	72.58%	73.68%

Zdroj: Vlastné spracovanie, 2020

Úspešnosť týchto modelov multinomickej logistickej regresie je vyššia ako pri predchádzajúcich modeloch. Ako lepší model multinomickej logistickej regresie sa javí upravený model č.3, v ktorom boli zahrnuté len tie premenné týkajúce sa spokojnosti študenta, ktorých parametre boli štatisticky významné. Úspešnosť tohto modelu na testovacej množine je vysoká. Model nadobúda na testovacej množine úspešnosť až 73,68% a môžeme ho považovať za vhodný model na predikciu rozhodnutia študenta pokračovať v štúdiu.

5.1.4 Model č.4 využiteľný v aplikácií (MLR)

Pokiaľ by sme vychádzali zo všetkých štúdií predstavených v kapitole 1, v ktorých boli preukázané faktory, ktoré môžu ovplyvňovať zotrvanie študenta na škole, tak medzi tieto faktory patrili nielen študijné výsledky zo strednej školy, ale aj aktuálne študijné výsledky. Pri vytváraní modelu č.4 sme vzali do úvahy študijné výsledky študenta zo strednej školy, konkrétne známky z maturít, typ strednej školy, ktorú študent navštevoval, ale aj známky získané z predmetov jeho vysokoškolského štúdia na FRI, odbor, ktorý študuje, prípadne to, či študent predmet opakoval.

Tento model je využiteľný v aplikácií určenej na predikciu rozhodnutia študenta pokračovať v inžinierskom štúdiu na FRI, vytvorenej v rámci praktickej časti tejto práce. Táto aplikácia je určená pre študijné oddelenie FRI, ktoré má prístup ku informáciám potrebným pre vytvorenie tohto predikčného modelu. Tie sa týkajú študenta, priebehu jeho stredoškolského aj aktuálneho štúdia a jeho študijných výsledkov. Ide o premenné vek, pohlavie, typSS, odbor, matSJL, matMAT, matINF, sINF, sMAN, sPI, sAUS, sOpakujem, sPrenasam a internat. Aby sme zistili, či medzi premennými existuje korelácia, bola pre tieto premenné zostrojená korelačná matica (obrázok č.15 na nasledujúcej strane).

```
                fIng    pohlavie         vek       typSS       odbor       matSJL       matMAT       matINF
fIng       1.000000000 -0.12920075 -0.23146750  0.09151675 -0.15650599 -0.123957595 -0.06476222 -0.15360936
pohlavie  -0.129200749  1.00000000  0.14818415 -0.20398072  0.23389703  0.138030220  0.13454215  0.12313963
vek       -0.231467501  0.14818415  1.00000000 -0.02741021  0.01780338  0.121982380  0.02798414  0.11866166
typSS      0.091516746 -0.20398072 -0.02741021  1.00000000  0.01723988 -0.083726054  0.42785479  0.44949347
odbor     -0.156505990  0.23389703  0.01780338  0.01723988  1.00000000  0.038409851  0.24767212  0.18794939
matSJL    -0.123957595  0.13803022  0.12198238 -0.08372605  0.03840985  1.000000000  0.13079457  0.06456346
matMAT    -0.064762215  0.13454215  0.02798414  0.42785479  0.24767212  0.130794575  1.00000000  0.38276014
matINF    -0.153609359  0.12313963  0.11866166  0.44949347  0.18794939  0.064563456  0.38276014  1.00000000
sINF      -0.002895633 -0.12092955  0.07895290 -0.14467191 -0.11161847  0.160270131 -0.15383655 -0.14591838
sPI       -0.044487271  0.09333646  0.23992799 -0.04123748 -0.33433693  0.042103050 -0.05404973 -0.01901478
sMAN       0.109824257 -0.23794036  0.04201066 -0.02746397 -0.55339017  0.078290083 -0.15663520 -0.12084452
sAUS      -0.342511693  0.20189252  0.32717026 -0.01421985  0.45487786  0.004023768  0.11040210  0.12573244
sOpakujem -0.269046355  0.22288432  0.44890107 -0.04072563  0.16616181  0.068110313  0.12678913  0.12366996
sPrenasam -0.297690861  0.14081902  0.40893744 -0.06972619  0.18062564  0.113453865  0.07743992  0.07629256
internat   0.090482819 -0.10146163  0.17134338 -0.07934728  0.00627152 -0.079394108 -0.03802641 -0.10904781
                sINF         sPI        sMAN        sAUS   sOpakujem   sPrenasam     internat
fIng      -0.002895633 -0.04448727  0.10982426 -0.342511693 -0.26904636 -0.29769086  0.090482819
pohlavie  -0.120929550  0.09333646 -0.23794036  0.201892523  0.22288432  0.14081902 -0.101461635
vek        0.078952905  0.23992799  0.04201066  0.327170257  0.44890107  0.40893744  0.171343375
typSS     -0.144671910 -0.04123748 -0.02746397 -0.014219853 -0.04072563 -0.06972619 -0.079347284
odbor     -0.111618465 -0.33433693 -0.55339017  0.454877864  0.16616181  0.18062564  0.006271520
matSJL     0.160270131  0.04210305  0.07829008  0.004023768  0.06811031  0.11345386 -0.079394108
matMAT    -0.153836548 -0.05404973 -0.15663520  0.110402101  0.12678913  0.07743992 -0.038026406
matINF    -0.145918377 -0.01901478 -0.12084452  0.125732443  0.12366996  0.07629256 -0.109047808
sINF       1.000000000 -0.01433523  0.12537408  0.137292926  0.19564916  0.25505947  0.059312589
sPI       -0.014335227  1.00000000  0.04192810 -0.020099668  0.10628517  0.16086141  0.038943923
sMAN       0.125374079  0.04192810  1.00000000 -0.347285335 -0.22837835 -0.12739499 -0.025202680
sAUS       0.137292926 -0.02009967 -0.34728534  1.000000000  0.46858363  0.53700259  0.006044844
sOpakujem  0.195649157  0.10628517 -0.22837835  0.468583631  1.00000000  0.52891974 -0.022955849
sPrenasam  0.255059474  0.16086141 -0.12739499  0.537002588  0.52891974  1.00000000  0.098024018
internat   0.059312589  0.03894392 -0.02520268  0.006044844 -0.02295585  0.09802402  1.000000000
```

Obrázok 15 – Korelačná matica premenných z modelu č.4

Zdroj: Vlastné spracovanie, 2020

Všetky zvolené premenné boli použité pri vytvorení modelu č.4, nakoľko medzi jednotlivými nezávislými premennými nebola zistená silná korelácia a nemajú na seba tým pádom tieto premenné výrazný vplyv. Informácie o modeli vytvorenom s využitím týchto premenných nám poskytuje obrázok č.16 výstupu funkcie *summary()*.

```
> summary(multinomial)
Call:
multinom(formula = fIng ~ odbor + pohlavie + vek + typSS + matSJL +
    matMAT + matINF + sINF + sPI + sAUS + sMAN + sOpakujem +
    sPrenasam + internat, data = train)

Coefficients:
  (Intercept)      odbor1    odbor2 pohlavie1        vek     typSS1     matSJL     matMAT
1  -92.297061  98.9077015 100.89369  1.532275 -0.4043720 -0.9893421  0.1654593 0.22202574
2  -36.586296 -24.8396021  27.86765 20.230363 -0.5052144 -1.8156545 -0.1392837 0.03971652
3    7.743783   0.4430392   2.06820  1.971210 -0.4149306  0.3554366 -0.4233151 0.18071243
    matINF        sINF       sPI       sAUS      sMAN  sOpakujem sPrenasam    internat
1 1.565700 -0.01324477 40.74224  0.2287425 1.7390755  0.2659246 -3.252394  0.58147790
2 1.392484  0.49999846 29.49542 -0.4532236 0.9121593  3.1697289 -5.642653 -0.06298725
3 1.014197  0.37747127 20.24951 -0.1284672 1.3481286 -0.1196652 -3.363137  1.47844115

Std. Errors:
  (Intercept)       odbor1   odbor2 pohlavie1       vek   typSS1    matSJL    matMAT
1    4.856714 2.786518e+00 2.616539  1.266098 0.3442420 1.136100 0.4739742 0.5161433
2    4.004734 1.445813e-13 3.769226  4.004770 0.5743682 1.773652 0.7336778 0.6673091
3    6.620445 2.799118e+00 1.958836  1.166929 0.3177757 1.076587 0.4380303 0.4990944
     matINF      sINF          sPI      sAUS      sMAN sOpakujem sPrenasam  internat
1 0.8727617 0.3590902 4.441128e-05 0.2133168 0.9302102 0.5988141  1.576630 0.9557652
2 1.2321417 0.4896819 1.013349e+00 0.4413182 2.3919144 1.7138862  2.627939 1.3057043
3 0.8426583 0.3238236 1.013320e+00 0.1876789 0.8758066 0.5713401  1.443040 0.9005181

Residual Deviance: 169.664
AIC: 265.664
```

Obrázok 16 – Výpis funkcie *summary()* pôvodného modelu č.4 (MLR)

Zdroj: Vlastné spracovanie, 2020

Hodnota AIC tohto modelu je $AIC = 265.664$. McFaddenovo pseudo-R^2 modelu č.4 multinomickej logistickej regresie nadobúda hodnotu 0.3552 a korigovaný McFaddenov koeficient nadobúda zápornú hodnotu, a to -0.00949. Tieto hodnotiace kritéria modelu a iné ďalšie nám poskytuje výstup na obrázku č.17 na nasledujúcej strane.

```
> round(PseudoR2(multinomial, which = "all"),6)
       McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
       0.355295       -0.009495        0.529539        0.601582              NA
VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
             NA              NA              NA              NA      265.663984
            BIC          logLik         logLik0              G2
     401.037499      -84.831992     -131.582644       93.501304
```

Obrázok 17 – Pseudo-R^2 pôvodného modelu č.4 (MLR)

Zdroj: Vlastné spracovanie, 2020

Štatistickú významnosť parametrov rovnako ako pri predošlých modeloch určíme vypočítaním p-hodnoty a jej porovnaním s hladinou významnosti α. P-hodnoty jednotlivých parametrov môžeme vidieť na nižšie uvedenom obrázku č.18.

```
> pnorm((abs(z)), lower.tail = FALSE) * 2
  (Intercept)        odbor1       odbor2    pohlavie1       vek    typSS1    matSJL    matMAT
1 1.579878e-80 5.851980e-276 0.000000e+00 2.261890e-01 0.2401254 0.3838506 0.7270224 0.6670771
2 6.494751e-20  0.000000e+00 1.430483e-13 4.382009e-07 0.3790759 0.3059858 0.8494320 0.9525400
3 2.421309e-01  8.742376e-01 2.910456e-01 9.117561e-02 0.1916429 0.7412857 0.3338411 0.7172917
      matINF      sINF           sPI      sAUS       sMAN  sOpakujem  sPrenasam  internat
1 0.07281943 0.9705773  0.000000e+00 0.2835792 0.06154619 0.65698086 0.03912433 0.5429289
2 0.25842021 0.3072222 2.938942e-186 0.3044313 0.70294272 0.06439437 0.03177914 0.9615249
3 0.22875620 0.2437481  7.690626e-89 0.4936561 0.12373123 0.83409961 0.01977488 0.1006382
```

Obrázok 18 – P-hodnoty parametrov modelu č.4 (MLR)

Zdroj: Vlastné spracovanie, 2020

Na základe p-hodnôt parametrov boli postupne z modelu odstraňované tie premenné, ktorých parametre boli štatisticky nevýznamné, a to až kým neboli všetky parametre modelu štatisticky významné. Následne boli zvyšné premenné použité pri vytvorení upraveného modelu č.4 multinomickej logistickej regresie. V tomto upravenom modeli boli ponechané premenné odbor, pohlavie, matINF, sPI, sMAN, sAUS, sOpakujem a sPrenasam. Hodnota Akaikeho informačného kritéria upraveného modelu je $AIC = 248.7806$. McFaddenov pseudo koeficient determinácie nadobúda hodnotu 0.2826 a korigovaný koeficient sa zvýšil na hodnotu 0.0546. Oba modely, upravený aj pôvodný, boli aplikované nielen na trénovaciu, ale aj na testovaciu množinu a bola zisťovaná ich úspešnosť (pozri tabuľka č.9).

Tabuľka 9 – Úspešnosť pôvodného a upraveného modelu č.4 (MLR)

Model	Úspešnosť na trénovacej množine	Úspešnosť na testovacej množine
pôvodný model č.4	70.97%	49.12%
upravený model č.4	68.55%	57.90%

Zdroj: Vlastné spracovanie, 2020

Z tabuľky č.9 vyplýva, že vyššiu úspešnosť na testovacej množine nadobúda upravený model č.4 multinomickej logistickej regresie. Pri oboch modeloch je úspešnosť na trénovacej množine vyššia ako na množine testovacej. Upravený model č.4 nadobúda na testovacej množine úspešnosť skoro 58%.

5.1.5 Vyhodnotenie úspešnosti modelov multinomickej logistickej regresie

Všetky vytvorené modely multinomickej logistickej regresie boli po ich „natrénovaní“ aplikované na testovacej množine dát. V tabuľke č.10 môžeme vidieť ich úspešnosti.

Tabuľka 10 – Úspešnosť modelov multinomickej logistickej regresie

Model	Úspešnosť na trénovacej množine	Úspešnosť na testovacej množine
pôvodný model č.1	70.16%	52.63%
upravený model č.1	70.97%	52.63%
pôvodný model č.2	62.90%	52.63%
upravený model č.2	61.29%	49.12%
pôvodný model č.3	74.19%	66.67%
upravený model č.3	**72.58%**	**73.68%**
pôvodný model č.4	70.97%	49.12%
upravený model č.4	68.55%	57.90%

Zdroj: Vlastné spracovanie, 2020

Ako sa dalo predpokladať a ako z tabuľky č.10 vyplýva, najväčšiu úspešnosť zo všetkých vytvorených modelov multinomickej logistickej regresie nadobúda na testovacej množine dát upravený model č.3, ktorý zohľadňuje celkovú spokojnosť študenta so školou. Tento model nadobúda na testovacej množine úspešnosť 73,68% a dá sa považovať za model vhodný na predikciu rozhodnutia študenta pokračovať v štúdiu.

V prípade modelu č.4 multinomickej logistickej regresie určeného pre našu aplikáciu, model nadobúda úspešnosť na testovacej množine 57,90%. Z dôvodu pomerne nízkej úspešnosti tohto modelu sme skúsili nami riešený problém zjednodušiť a v nasledujúcej kapitole praktickej časti tejto práce boli tieto modely vytvorené pomocou binárnej logistickej regresie. Naším cieľom bolo zistiť, či sa binárna logistická regresia nepreukáže ako vhodnejší spôsob riešenia nášho problému.

5.2 Modely logistickej regresie (LR)

V tejto kapitole sme náš problém skúsili riešiť pomocou binárnej logistickej regresie. Pri binárnej logistickej regresií môže závislá premenná – v našom prípade rozhodnutie študenta pokračovať v štúdiu na FRI – nadobúdať dve hodnoty. Náš dataset bolo potrebné upraviť a aktuálne štyri kategórie nahradiť kategóriami dvomi (pozri príloha C).

Pri modeloch binárnej logistickej regresie budeme podobne ako pri modeloch multinomickej logistickej regresie vychádzať zo štúdií predstavených v kapitole 1. Do

modelov binárnej logistickej regresie boli zvolené rovnaké nezávislé premenné, ako pri modeloch multinomickej logistickej regresie.

5.2.1 Model č.1 zohľadňujúci stresové faktory (LR)

Pri modeli č.1 vytvoreného pomocou logistickej regresie sme vychádzali zo štúdie z roku 2002, v ktorej Cotton, Dollard a Jonge preukázali, že na rozhodnutie študenta pokračovať v štúdiu a predčasné ukončenie štúdia vplývajú stresové faktory. Do modelu boli zvolené premenné sSocZivot, sPrenasam, sOpakujem, sNezvladam, sPsycholog, sPredcasne, sPredcasneKam, sNarocne, sNaroky, sPomoc a sVztahy. Pomocou týchto premenných a príkazu *glm()* sme v jazyku R vytvorili model logistickej regresie na trénovacej množine. Informácie o modeli môžeme vidieť na nasledujúcom obrázku č.19.

```
> #MODEL C.1
> model = glm(fIng ~ sPrenasam + sOpakujem + sNezvladam + sSocZivot + sPredcasne + sPredcasneKam +
+               sPomoc + sVztahy + sPsycholog + sNarocne + sNaroky, data = train, family = binomial)
> summary(model)

Call:
glm(formula = fIng ~ sPrenasam + sOpakujem + sNezvladam + sSocZivot +
    sPredcasne + sPredcasneKam + sPomoc + sVztahy + sPsycholog +
    sNarocne + sNaroky, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5391  -0.8377   0.4450   0.7109   2.3682

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.4558     1.2866   1.909  0.05629 .
sPrenasam     -0.8947     0.5181  -1.727  0.08417 .
sOpakujem      0.3237     0.3045   1.063  0.28773
sNezvladam     0.2158     0.4536   0.476  0.63428
sSocZivot     -0.8108     0.3531  -2.296  0.02167 *
sPredcasne    -0.4973     0.2509  -1.982  0.04743 *
sPredcasneKam -0.2753     0.4281  -0.643  0.52023
sPomoc         0.2922     0.4055   0.721  0.47112
sVztahy       -0.2747     0.5527  -0.497  0.61918
sPsycholog    -1.2269     0.4315  -2.843  0.00447 **
sNarocne       1.1187     0.4792   2.334  0.01957 *
sNaroky       -0.8777     0.3925  -2.236  0.02533 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 165.52  on 124  degrees of freedom
Residual deviance: 126.11  on 113  degrees of freedom
AIC: 150.11
```

Obrázok 19 – Výpis funkcie *summary()* pôvodného modelu č.1 (LR)

Zdroj: Vlastné spracovanie, 2020

Výpis nám v prípade logistickej regresie poskytuje nielen hodnotu AIC, ale aj informáciu o štatistickej významnosti parametrov modelu. Ako môžeme z tohto výpisu vidieť, nie všetky parametre modelu sú štatistický významné. Štatistickú významnosť parametrov nám vo výpise *summary()* vyjadrujú znaky uvedené pri p-hodnotách týchto parametrov. Význam týchto znakov je nasledovný:

*** ⇒ parameter je štatisticky významný na hladine významnosti 0.001
** ⇒ parameter je štatisticky významný na hladine významnosti 0.01
* ⇒ parameter je štatisticky významný na hladine významnosti 0.05
. ⇒ parameter je štatisticky významný na hladine významnosti 0.1

V stĺpci $P(>|t|)$ na obrázku č.19 sa nachádza p-hodnota definovaná v kapitole 2.2.2, vďaka ktorej vieme aj bez značiek určiť štatistickú významnosť parametrov. V prípade, ak je $p-hodnota$ vyššia ako hladina významnosti α, daný parameter nie je na tejto hladine štatisticky významný.

Hodnotu McFaddenovho pseudo koeficientu determinácie sme opäť vypočítali pomocou funkcie *PseudoR2()*. V prípade tohto modelu má McFaddenov pseudo koeficient determinácie hodnotu 0.23807. Všetky ostatné štatistiky nám poskytuje obrázok 20.

```
> PseudoR2(model, which = "all")
       McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
     0.23807652      0.09307717      0.27039177      0.36839652      0.23968686
VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
     0.42069956      0.30945181      0.37684834      0.29980973    150.11203800
            BIC           logLik          logLik0              G2
  184.05180285    -63.05601900    -82.75899140     39.40594480
```

Obrázok 20 – Pseudo-R^2 pôvodného modelu č.1 (LR)

Zdroj: Vlastné spracovanie, 2020

Z modelu sme postupne odstránili premenné, ktorých parametre sa preukázali ako štatisticky nevýznamné, a to až kým všetky parametre modelu neboli štatisticky významné. Výstup tohto upraveného modelu obsahujúceho iba premenné, ktorých parametre sú štatisticky významné sa nachádza na obrázku č.21.

```
> #UPRAVENY MODEL C.1
> model = glm(fIng ~ sSocZivot + sPredcasne + sPsycholog +
+               sNarocne + sNaroky, data = train, family = binomial)
> summary(model)

Call:
glm(formula = fIng ~ sSocZivot + sPredcasne + sPsycholog + sNarocne +
    sNaroky, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4637  -0.8959   0.5156   0.7102   2.2635

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7486     0.5713   3.060  0.00221 **
sSocZivot    -0.7021     0.3155  -2.225  0.02605 *
sPredcasne   -0.5128     0.2280  -2.249  0.02452 *
sPsycholog   -1.1025     0.4140  -2.663  0.00774 **
sNarocne      1.1312     0.4285   2.640  0.00829 **
sNaroky      -0.9290     0.3831  -2.425  0.01532 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 165.52  on 124  degrees of freedom
Residual deviance: 130.19  on 119  degrees of freedom
AIC: 142.19
```

Obrázok 21 – Výpis funkcie *summary()* upraveného modelu č.1 (LR)

Zdroj: Vlastné spracovanie, 2020

Hodnota Akaikeho informačného kritéria upraveného modelu klesla na 142.19. McFaddenov pseudo koeficient determinácie nadobúda v upravenom modeli hodnotu 0.2134. Hodnota korigovaného McFaddenovho koeficientu upraveného modelu vzrástla na hodnotu 0.1409, čo značí jeho zlepšenie.

Ďalším krokom je vykonanie samotnej predikcie zavolaním funkcie *predict()*, a to v tvare *predictTrain = predict(model, type = "response", newdata = train)*. Pri výsledkoch predikcie je potrebné určiť hranicu a rozhodnúť, aká výstupná hodnota (pravdepodobnosť) bude znamenať to, že študent v štúdiu pokračuje, a ktorá hodnota už nie. Táto hranica sa nazýva threshold value (prahová hodnota). Threshold value je bližšie charakterizovaná v kapitole 3.2.2. Ak je výstupom logistickej regresie hodnota menšia ako threshold value, model bude predpovedať 0 (študent nepokračuje). V prípade, že je táto hodnota väčšia nanajvýš rovná, bude predpovedať 1 (študent pokračuje).

Štandardne je prahová hodnota 0,5. Pri tejto prahovej hodnote zavoláme v R funkciu *table(train$fIng, predictTrain >= 0.5)* a zostrojíme klasifikačnú maticu, kde *train$fIng* sú pôvodné dáta z testovacej množiny a *predictTrain* sú predikcie. Výsledná klasifikačná matica vyzerá nasledovne:

Tabuľka 11 – Klasifikačná matica modelu č.1 pri t = 0,5

	FALSE	TRUE
0	30	17
1	12	66

Zdroj: Vlastné spracovanie, 2020

Klasifikačná matica bola definovaná v kapitole 3.3. Stĺpce TRUE a FALSE predstavujú predikovanú hodnotu premennej fIng, a teda, či študent bude pokračovať v štúdiu na FRI, alebo nie. Hodnoty predikcie sa nastavia na hodnotu FALSE v prípade, ak je pravdepodobnosť toho, že študent bude pokračovať v štúdiu menšia ako 0,5. Hodnota TRUE je v tabuľke nastavená v prípade, ak je pravdepodobnosť toho, že študent bude pokračovať v štúdiu väčšia nanajvýš rovná ako 0,5.

Z tabuľky teda pre lepšie pochopenie vyplýva, že 30 študentov v štúdiu nepokračovalo a model správne predikoval, že pokračovať ani nebudú. Nepokračovalo 17 študentov, ktorým model nesprávne predikoval, že zostanú študovať na FRI. V štúdiu pokračovalo 12 študentov, o ktorých model nesprávne tvrdil, že pokračovať nebudú. Pri 66 študentoch model správne predpovedal, že budú pokračovať v štúdiu na FRI.

Ako bolo spomínané v kapitole 3.3, klasifikačná matica slúži na validáciu klasifikačných modelov, a teda aby sme vedeli povedať, nakoľko je náš vytvorený model úspešný. Z matice vieme vypočítať metriky ako sú senzitivita, špecificita alebo celková úspešnosť modelu. Celkovú úspešnosť (CU) modelu vypočítame tak, že súčet hodnôt na

hlavnej diagonále vydelíme súčtom hodnôt v tabuľke. V jazyku R to vieme spraviť jednoducho pomocou príkazu *sum(diag(tableTrain)) / sum(tableTrain) * 100.* Celková úspešnosť modelu č. 1 na trénovacej množine je 76,8%. Nielen celková úspešnosť je však dôležitá. Je potrebné brať ohľad aj na hodnoty senzitivity a špecificity. Špecificita vyjadruje schopnosť modelu správne rozpoznať študentov, ktorí sa rozhodli odísť a nepokračovať v štúdiu. Senzitivita udáva schopnosť modelu správne zachytiť prípady, kedy sa študenti rozhodli pokračovať v štúdiu. Špecificita modelu je 63,83% a v jazyku R sme ju vypočítali príkazom *tableTrain[1,1] / (tableTrain[1,1] + tableTrain[1,2]) * 100*. Pre výpočet senzitivity sme použili príkaz *(tableTrain[2,2]) / (tableTrain[2,1] + tableTrain[2,2]) * 100*. Senzitivita je 84,62%.

Rozdiely medzi hodnotami špecificity a senzitivity by nemali byť príliš veľké. Vysoká senzitivita znamená, že model produkuje falošné pozitíva, čo v praxi znamená, že model predikuje pokračovanie študenta v štúdiu na FRI, aj keď tento študent v štúdiu pokračovať nebude. V tomto prípade je rozdiel medzi senzitivitou a špecificitou pomerne vysoký, preto je vhodné nájsť lepšiu hodnotu threshold value. V nasledujúcej tabuľke č.12 sú uvedené rôzne hodnoty threshold value a celková úspešnosť, senzitivita a špecificita, ktoré model nadobúda pri týchto hodnotách na trénovacej množine.

Tabuľka 12 – Špecificita, senzitivita a CU modelu č.1 na trénovacej množine (LR)

Threshold value	Celková úspešnosť	Špecificita	Senzitivita
t = 0.5	76.8%	63.83%	84.62%
t = 0.8	64.8%	89.36%	50%
t = 0.7	74.4%	85.11%	67.95%
t = 0.6	**76.8%**	**76.60%**	**76.92%**
t = 0.55	77.6%	72.34%	80.77%
t = 0.57	76.8%	74.47%	78.21%

Zdroj: Vlastné spracovanie, 2020

Z tabuľky vyplýva, že pri zvolenej hodnote threshold value = 0,6 sú rozdiely medzi senzitivitou a špecificitou minimálne, pričom celková úspešnosť modelu 76,8% na trénovacej množine je pomerne vysoká. Túto hodnotu thresholdu je potrebné otestovať na testovacej množine. Predikciu na testovacej množine vykonáme pomocou príkazu *predictTest = predict(model, type = "response", newdata = test)*. Následne zostrojíme klasifikačnú maticu a vypočítame celkovú úspešnosť modelu, senzitivitu a špecificitu modelu na testovacej množine. Na testovacej množine pri prahovej hodnote 0,6 má celková

úspešnosť modelu hodnotu 60,71%, špecificita je 47,37% a senzitivita je 67,57%. Model teda na testovacej množine nadobúda priemernú úspešnosť.

5.2.2 Model č.2 zohľadňujúci bydlisko študenta (LR)

Do modelu č.2 zohľadňujúcom bydlisko študenta, inšpirovanom štúdiou z roku 2007 (Ho Yu, Yu, Digangi a kol.), boli zvolené premenné kraj, bydlisko, internat a premenná pFinNarocne. Pri tvorbe modelu sme postupovali rovnako ako pri vytváraní modelu č.1 v kapitole 5.2.1 a preto podrobnosti postupu tvorby tohto modelu nebudú bližšie popísané. Na obrázku č.22 sa nachádza výpis funkcie *summary()*, na ktorom môžeme vidieť nakoľko sú jednotlivé parametre štatisticky významné na základe ich p-hodnoty a hodnotu Akaikeho informačného kritéria tohto modelu.

```
> #MODEL C.2
> model = glm(fIng ~ pFinNarocne + bydlisko + kraj + internat,
+             data = train, family = binomial)
> summary(model)

Call:
glm(formula = fIng ~ pFinNarocne + bydlisko + kraj + internat,
    family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7330  -1.3142   0.8283   0.9741   1.3837

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.18528    0.93702   1.265   0.2059
pFinNarocne -0.39440    0.22203  -1.776   0.0757 .
bydlisko     0.13712    0.38301   0.358   0.7203
kraj        -0.05937    0.09519  -0.624   0.5328
internat     0.44040    0.46674   0.944   0.3454
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 165.52  on 124  degrees of freedom
Residual deviance: 160.17  on 120  degrees of freedom
AIC: 170.17
```

Obrázok 22 – Výpis funkcie *summary()* pôvodného modelu č.2 (LR)

Zdroj: Vlastné spracovanie, 2020

Z obrázku vyplýva, že jediný štatisticky významný parameter je pri premennej pFinNarocne, a je štatisticky významný na hladine významnosti 0,1. Akaikeho informačné kritérium je $AIC = 170.17$. Hodnota McFaddenovho pseudo-R^2 je veľmi nízka. Korigovaný McFaddenov pseudo-koeficient determinácie nadobúda záporné hodnoty. (pozri obr.23)

```
> PseudoR2(model, which = "all")
      McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
    0.03230475     -0.02811165      0.04187414      0.05705162      0.04102139
VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
    0.07200095      0.04119023      0.05347691      0.04205511    170.17096640
           BIC           logLik          logLik0              G2
  184.31253509    -80.08548320    -82.75899140      5.34701640
```

Obrázok 23 – Pseudo-R^2 pôvodného modelu č.2 (LR)

Zdroj: Vlastné spracovanie, 2020

Pri postupnom odstraňovaní premenných, ktorých parametre boli štatisticky nevýznamné, aj parameter pri premennej pFinNarocne stratil svoju štatistickú významnosť. Model nebolo možné vhodne upraviť ani inými kombináciami zvolených premenných. Taktiež aj hodnoty pseudo-koeficientov determinácie pôvodného modelu sú veľmi nízke. Pri tvorbe modelu zohľadňujúceho bydlisko sa teda ani binárna logistická regresia nepreukázala ako vhodný spôsob riešenia nášho problému a takýto model nie je vhodný pri predikcii rozhodnutia študenta pokračovať v štúdiu.

5.2.3 Model č.3 zohľadňujúci spokojnosť študenta (LR)

V roku 2013 aj Acheampong, Boyetey, Osei Gyimah a Okyere v rámci svojej štúdie pomocou logistickej regresie potvrdili, že na celkovú spokojnosť študenta na škole majú vplyv faktory akými sú napr. spokojnosť s vybavením školy alebo so študijným odborom. Do modelu boli zvolené nasledovné premenné: odbor, spPrilezitosti, spSemester, spVybavenie, spOdbornost, spPristup, spOdradza, spZbytocne, spZastarale, spStudProgram, spHodnotenie, spTerminy. Na nasledujúcom obrázku č.24 môžeme vidieť informácie o vytvorenom modeli z výpisu funkcie *summary()*.

```
> summary(glm(fIng ~ odbor  + spPrilezitosti + spSemester + spOdradza + spZbytocne +
+               spZastarale + spStudProgram + spVybavenie + spOdbornost + spPristup +
+               spHodnotenie + spTerminy, data = train, family = binomial))

Call:
glm(formula = fIng ~ odbor + spPrilezitosti + spSemester + spOdradza +
    spZbytocne + spZastarale + spStudProgram + spVybavenie +
    spOdbornost + spPristup + spHodnotenie + spTerminy, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2951  -0.7535   0.3224   0.6688   2.2612

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -5.38324    2.36272  -2.278   0.0227 *
odbor          -0.06701    0.41387  -0.162   0.8714
spPrilezitosti -0.04035    0.38694  -0.104   0.9170
spSemester      0.59132    0.38684   1.529   0.1264
spOdradza       0.57212    0.39893   1.434   0.1515
spZbytocne     -0.09754    0.40800  -0.239   0.8111
spZastarale    -0.47938    0.31513  -1.521   0.1282
spStudProgram   0.76928    0.36558   2.104   0.0354 *
spVybavenie     0.24875    0.37218   0.668   0.5039
spOdbornost     0.32031    0.40252   0.796   0.4262
spPristup       0.65674    0.42404   1.549   0.1214
spHodnotenie    0.24370    0.34159   0.713   0.4756
spTerminy       0.26297    0.29148   0.902   0.3670
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 165.52  on 124  degrees of freedom
Residual deviance: 111.21  on 112  degrees of freedom
AIC: 137.21
```

Obrázok 24 – Výpis funkcie *summary()* pôvodného modelu č.3 (LR)

Zdroj: Vlastné spracovanie

Z výpisu vyplýva, že jedine parameter premennej spStudProgram je štatisticky významný na hladine významnosti 0,05. Akaikeho informačné kritérium má hodnotu $AIC = 137.21$. Všetky hodnotiace kritéria modelu môžeme vidieť na obrázku č.25. McFaddenovo pseudo-R^2 má hodnotu 0.32808 a hodnota korigovaného koeficientu je taktiež pomerne

vysoká. Môžeme teda predpokladať, že tento model bude mať vysokú presnosť pri predikcii. Pri úprave modelu sme postupovali rovnako ako pri úprave modelu č.1 v kapitole 5.2.1, preto tento postup nebude bližšie popisovaný.

```
> PseudoR2(model, which = "all")
      McFadden    McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
     0.3280835      0.1710009      0.3523661      0.4800828      0.3028588
VeallZimmermann          Efron McKelveyZavoina           Tjur            AIC
     0.5315793      0.3838223      0.5196976      0.3846179    137.2142669
           BIC         logLik        logLik0             G2
   173.9823455    -55.6071334    -82.7589914     54.3037159
```

Obrázok 25 – Pseudo-R^2 pôvodného modelu č.3 (LR)

Zdroj: Vlastné spracovanie, 2020

Po odstránení premenných na základe štatistickej významnosti ich parametrov boli v modeli ponechané premenné spZastarale, spStudProgram a spPristup. Hodnota Akaikeho informačného kritéria upraveného modelu je $AIC = 125.75$. McFaddenovo pseudo-R^2 nadobúda pri upravenom modeli hodnotu 0.2886 a korigované pseudo-R^2 hodnotu 0.2402. Hodnota korigovaného pseudo koeficientu determinácie výrazne narástla.

Ďalším krokom pri predikcii je nájdenie vhodnej prahovej hodnoty, vytvorenie klasifikačnej matice a výpočet celkovej úspešnosti, senzitivity a špecificity modelu na trénovacej množine. V nasledujúcej tabuľke č.13 sa nachádza výber testovaných prahových hodnôt a na základe nich vypočítané im prislúchajúce metriky.

Tabuľka 13 – Špecificita, senzitivita a CU modelu č.3 na trénovacej množine (LR)

Threshold value	Celková úspešnosť	Špecificita	Senzitivita
t = 0.5	80%	70.21%	85.90%
t = 0.8	71.2%	89.36%	60.26%
t = 0.7	78.4%	80.85%	76.92%
t = 0.6	**80%**	**74.47%**	**83.33%**
t = 0.65	78.4%	80.85%	76.92%
t = 0.55	80%	74.47%	83.33%

Zdroj: Vlastné spracovanie, 2020

Rozdiely medzi senzitivitou a špecificitou sú minimálne a celková úspešnosť je vysoká pri prahovej hodnote 0,6. Celková úspešnosť na trénovacej množine je vtedy 80%. Celková úspešnosť modelu na testovacej množine je pri prahovej hodnote 0,6 rovná 66,07%, senzitivita 68,42% a špecificita je 68,86%. Celková úspešnosť je pomerne vysoká aj na testovacej množine. Taktiež medzi hodnotami špecificity a senzitivity je na testovacej množine minimálny rozdiel. Ak však vezmeme do úvahy model č.4 multinomickej logistickej regresie, ktorý na testovacej množine nadobúdal úspešnosť skoro 74%, model č.4 logistickej regresie sa s úspešnosťou 66,07% nepreukázal ako lepší.

5.2.4 Model č.4 využiteľný v aplikácií (LR)

Nasledujúci model je využiteľný v aplikácií vytvorenej v rámci praktickej časti tejto práce. Aplikácia by mala byť využiteľná študijným oddelením a obsahuje premenné zodpovedajúce údajom, ku ktorým majú na tomto oddelení prístup. Ku takýmto premenným patria vek, pohlavie, odbor, typSS, matSJL, matMAT, matINF, sINF, sMAN, sPI, sAUS, internat, sPrenasam, sOpakujem. Výstup funkcie *summary()* vytvoreného modelu logistickej regresie môžeme vidieť na nasledujúcom obrázku č.26.

```
> #MODEL C.4
> summary(model)

Call:
glm(formula = fIng ~ odbor + pohlavie + vek + typSS + matSJL +
    matMAT + matINF + sINF + sPI + sAUS + sMAN + sOpakujem +
    sPrenasam + internat, family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0210  -0.9909   0.6069   0.9023   1.5320

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.69529    3.79326   0.447   0.6549
odbor       -0.61293    0.58952  -1.040   0.2985
pohlavie     0.46497    0.55663   0.835   0.4035
vek         -0.01453    0.17512  -0.083   0.9339
typSS        1.16924    0.55090   2.122   0.0338 *
matSJL      -0.26144    0.26041  -1.004   0.3154
matMAT      -0.03524    0.25634  -0.137   0.8907
matINF      -0.52179    0.34062  -1.532   0.1256
sINF         0.16784    0.16432   1.021   0.3071
sPI         -0.16505    0.31444  -0.525   0.5996
sAUS        -0.22250    0.11380  -1.955   0.0506 .
sMAN        -0.35875    0.23927  -1.499   0.1338
sOpakujem    0.02126    0.30284   0.070   0.9440
sPrenasam   -0.22006    0.57258  -0.384   0.7007
internat     0.56186    0.46112   1.218   0.2230
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 165.52  on 124  degrees of freedom
Residual deviance: 143.10  on 110  degrees of freedom
AIC: 173.1
```

Obrázok 26 – Výpis funkcie *summary()* pôvodného modelu č.4 (LR)

Zdroj: Vlastné spracovanie, 2020

Hodnota AIC modelu č.4 logistickej regresie je $AIC = 173.1$. Štatisticky významné parametre v tomto modeli sú pri premennej typSS na hladine významnosti 0,05 a pri premennej sAUS na hladine významnosti 0,1. Obrázok č.27 nám poskytuje hodnoty pseudo koeficientov determinácie a iné hodnotiace štatistiky modelu. McFaddenovo pseudo-R^2 tohto modelu je 0.13546. Korigovaný McFaddenov pseudo koeficient determinácie nadobúda hodnotu -0.04578.

```
> PseudoR2(model, which = "all")
       McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
     0.13546565      -0.04578354      0.16420842      0.22372653      0.15209399
VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
     0.26695613       0.17183919       0.23110225      0.17127666   173.09598200
            BIC           logLik          logLik0              G2
  215.52068806    -71.54799100    -82.75899140     22.42200080
```

Obrázok 27 – Pseudo-R^2 pôvodného modelu č.4 (LR)

Zdroj: Vlastné spracovanie, 2020

Z modelu boli postupne odstraňované premenné, ktorých parametre boli štatisticky nevýznamné. Medzi ponechanými premennými v modeli boli premenné odbor, typSS, matINF, sMAN, sPI, sAUS a sPrenasam. Hodnota AIC upraveného modelu je $AIC = 163$. McFaddenovo pseudo-R^2 má hodnotu 0.111880. V ďalšom kroku tvorby predikčného modelu boli pri predikcií vyskúšané viaceré hodnoty threshold value a pomocou zostrojenej klasifikačnej matice boli vypočítané metriky slúžiace na validáciu modelu, ktoré môžeme vidieť v tabuľke 14.

Tabuľka 14 – Špecificita, senzitivita a CU modelu č.4 na trénovacej množine (LR)

Threshold value	Celková úspešnosť	Špecificita	Senzitivita
t = 0.5	72%	53.19%	83.33%
t = 0.55	70.4%	53.19%	80.77%
t = 0.6	64.8%	57.45%	69.23%
t = 0.65	63.2%	59.57%	65.38%
t = 0.7	58.4%	78.72%	46.15%
t = 0.66	**61.6%**	**61.70%**	**61.54%**
t = 0.67	60.8%	63.83%	58.97%
t = 0.68	63.2%	70.21%	58.97%

Zdroj: Vlastné spracovanie, 2020

Pri prahovej hodnote 0,66 sú rozdiely medzi špecificitou a senzitivitou minimálne. Celková úspešnosť na trénovacej množine je pri tejto hodnote threshold value rovná 61,6%. Na testovacej množine je celková úspešnosť modelu 66,27%, špecificita je 89,47% a senzitivita je 54,05%.

5.2.5 Vyhodnotenie úspešnosti modelov logistickej regresie

Okrem modelu č.2 boli všetky modely logistickej regresie po ich „natrénovaní“ aplikované na testovaciu množinu dát. Prahová hodnota bola pri každom modeli nastavená na základe celkovej úspešnosti, senzitivity a špecificity, ktoré model nadobúdal na trénovacej množine. V tabuľke č.15 môžeme vidieť úspešnosti modelov.

Tabuľka 15 – Úspešnosť modelov logistickej regresie

Model	Úspešnosť na trénovacej množine	Úspešnosť na testovacej množine
upravený model č.1	76.8%	60.71%
upravený model č.3	80%	66.07%
upravený model č.4	**61.6%**	**66.27%**

Zdroj: Vlastné spracovanie, 2020

Napriek tomu, že sme skúsili riešený problém zjednodušiť a hodnoty kategorickej závislej premennej fIng, sme znížili zo štyroch kategórií na kategórie dve, vytvorené modely logistickej regresie sa pri predikcií rozhodnutia študenta pokračovať v štúdiu nepreukázali ako výrazne efektívnejšie. Model č.3 logistickej regresie dokonca nadobúda nižšiu úspešnosť na testovacej množine (66,07%), než model č.3 multinomickej logistickej regresie, ktorý nadobúdal úspešnosť 73,68%. Najvyššiu celkovú úspešnosť z modelov logistickej regresie nadobúda model č.4 (66,27%).

5.3 Modely klasifikačných rozhodovacích stromov (CDT)

Ďalšou možnosťou predikcie rozhodnutia študenta pokračovať v štúdiu je pomocou rozhodovacích stromov. V tejto kapitole budú vyššie predstavené modely č. 1-4 vytvorené pomocou klasifikačných rozhodovacích stromov. V jazyku R je na vytváranie modelov rozhodovacích stromov dostupný balíček *rpart*, ktorého nainštalovaním získame prístup ku funkcií *rpart()*. Závislou premennou modelov bola – podobne ako pri modeloch multinomickej logistickej regresie – premenná fIng vyjadrujúca rozhodnutie študenta pokračovať v štúdiu, ktorá má štyri kategórie (pozri príloha C). Nezávislé premenné jednotlivých modelov ostali rovnaké.

Pri vytváraní modelov rozhodovacích stromov bolo potrebné vhodne nastaviť parametre minsplit, minbucket a cp. Parameter minsplit určuje minimálny počet pozorovaní v uzle potrebných na to, aby sa uzol mohol ďalej vetviť. Minbucket udáva minimálny počet pozorovaní, ktoré musí obsahovať list. Cp je parameter zložitosti. (pozri kapitola 3.2.4)

Predvolené hodnoty týchto parametrov vo funkcií *rpart()* sú $minsplit = 20$, $minbucket = (\frac{minsplit}{3})$ a $cp = 0.01$. Postupným testovaním rôznych hodnôt parametrov sme sa pri každom modeli snažili nájsť takú kombináciu hodnôt týchto parametrov, pri ktorých náš model vykazoval najvyššiu úspešnosť na testovacej množine. Hodnota parametrov minsplit a minbucket by nemala byť príliš nízka, pretože takéto nastavenie parametra by mohlo viesť ku pretrénovaniu (overfitting) modelu. Pri pretrénovaní model vykazuje vysokú úspešnosť na trénovacej množine dát, ale na množine testovacej je úspešnosť nízka. [15][39]

5.3.1 Model č.1 zohľadňujúci stresové faktory (CDT)

Do modelu č.1 boli zvolené premenné sSocZivot, sPrenasam, sOpakujem, sNezvladam, sPsycholog, sPredcasne, sPredcasneKam, sNarocne, sNaroky, sPomoc

a sVztahy. V prípade modelu č.1 boli v klasifikačnom rozhodovacom strome parametre minsplit, minbucket a cp nastavené nasledovne: $minbucket = 4$, $minsplit = 14$ a $cp = 0$.

5.3.2 Model č.2 zohľadňujúci bydlisko študenta (CDT)

Ku premenným zvoleným do modelu č.2 patria premenné kraj, bydlisko, internat a pFinNarocne. Model č.2 klasifikačného rozhodovacieho stromu sme jeho parametre nastavili nasledovne: $minbucket = 14$, $minsplit = 1$ a $cp = 0$.

5.3.3 Model č.3 zohľadňujúci spokojnosť študenta (CDT)

Do modelu č.3 boli zvolené nasledovné premenné: odbor, spPrilezitosti, spSemester, spVybavenie, spOdbornost, spPristup, spOdradza, spZbytocne, spZastarale, spStudProgram, spHodnotenie, spTerminy. V prípade tohto modelu boli parametre nastavené nasledovne: $minbucket = 1$, $minsplit = 7$ a $cp = 0.02$.

5.3.4 Model č.4 využiteľný v aplikácií (CDT)

Do modelu č.4 boli zvolené premenné vek, pohlavie, odbor, typSS, matSJL, matMAT, matINF, sINF, sMAN, sPI, sAUS, internat, sPrenasam, sOpakujem. Klasifikačnému rozhodovaciemu stromu č.4 využiteľnému v aplikácií sme nastavili nasledovné hodnoty parametrov: $minbucket = 3$, $minsplit = 15$ a $cp = 0.01$.

5.3.5 Vyhodnotenie úspešnosti modelov rozhodovacích stromov

Všetky modely klasifikačných rozhodovacích stromov boli najskôr „natrénované“ na trénovacej množine dát. Každému modelu boli nastavené najoptimálnejšie hodnoty parametrov minsplit, minbucket a cp, pri ktorých daný model vykazoval najvyššiu úspešnosť na testovacej množine. Úspešnosti modelov klasifikačných stromov na trénovacej aj testovacej množine a hodnoty ich parametrov môžeme vidieť v tabuľke č.16.

Tabuľka 16 – Úspešnosť modelov klasifikačných rozhodovacích stromov

Model	Úspešnosť na trénovacej množine	Úspešnosť na testovacej množine
model č.1	66.94%	56.14%
model č.2	60.48%	56.14%
model č.3	**71.77%**	**71.93%**
model č.4	69.35%	61.40%

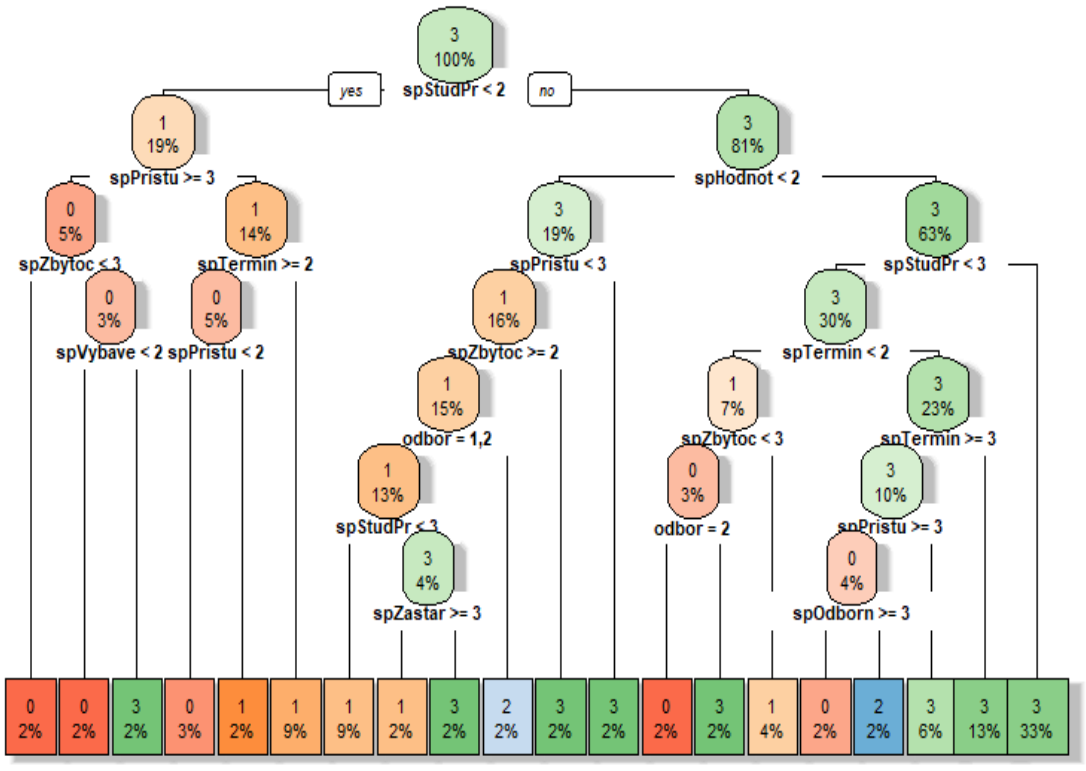
Zdroj: Vlastné spracovanie, 2020

Úspešnosť modelov č.1,2 a 4 klasifikačných rozhodovacích stromov je vyššia ako pri modeloch multinomickej logistickej regresie. Model č.3 nadobúda pri obidvoch spôsoboch

modelovania približne rovnakú úspešnosť. Nakoľko model č.4 klasifikačného rozhodovacieho stromu nadobúda vyššiu úspešnosť ako model č.4 multinomickej logstickej regresie, rozhodli sme sa tento model použiť na demonštráciu výsledkov tejto práce v aplikácií určenej na predikciu rozhodnutia študenta pokračovať v inžinierskom štúdiu na FRI, vytvorenej v rámci praktickej časti tejto práce. Tento model bol na našej testovacej množine dát úspešný na 61,40%.

5.3.6 Model CDT pre poskytnutie návrhov a odporúčaní

Jednou z výhod rozhodovacích stromov je ich jednoduchá grafická reprezentácia. V jazyku R vieme strom graficky zobraziť funkciou *prp()* z balíčka *rpart.plot*. Vypracovali sme regresný strom s použitím premenných z modelu č.3. Parametrom minsplit, minbucket a cp sme nastavili hodnoty: $minsplit = 4$, $cp = 0.015$, a $minbucket = 2$. Grafickú reprezentáciu (výstup funkcie *prp()*) tohto stromu môžeme vidieť na obrázku č.28.



Obrázok 28 – Klasifikačný rozhodovací strom so všetkými premennými

Zdroj: Vlastné spracovanie, 2020

Tento model nadobúda na testovacej množine úspešnosť 68,42%. Na základe tohto modelu boli identifikované faktory spokojnosti najvýznamnejšie vplývajúce na rozhodnutie študenta pokračovať v štúdiu na inžinierskom stupni na FRI. Podľa tohto modelu boli vytvorené návrhy a odporúčania, ktoré by mohli pozitívne vplývať na spokojnosť študenta s jednotlivými aspektami štúdia a ovplyvniť tak jeho rozhodnutie pokračovať v štúdiu.

6 Aplikácia

Návrh a vytvorenie aplikácie boli súčasťou tejto práce, a to ako spôsob demonštrácie výsledkov tejto práce. Táto aplikácia je určená pre študijné oddelenie Fakulty riadenia a informatiky a po zadaní informácií, ktoré sú na študijnom oddelení dostupné predikuje rozhodnutie študenta pokračovať v inžinierskom stupni štúdia na Fakulte riadenia a informatiky. Medzi tieto informácie patrí vek, stredná škola, či študent býva na internáte, známky z maturity zo SJL, MAT a INF, a známky z dôležitých predmetov jednotlivých študijných programov, ktorými sú Informatika 1, Manažment 1 a Počítačové inžinierstvo. Taktiež je možné zadať informáciu o tom, či študent opakoval niekedy predmet alebo ročník. Všetky zdrojové kódy aplikácie, skripty a použité dáta sú súčasťou práce a sú uvedené v prílohe E. Informácie pre správne spustenie aplikácie sú dostupné v podkapitole 6.3.

6.1 Java

Na vytvorenie aplikácie bol zvolený programovací jazyk Java a pri práci bolo použité vývojové prostredie NetBeans IDE. Jedným z cieľov tvorby aplikácie bolo vytvoriť prehľadné grafické užívateľské rozhranie, čo nám NetBeans veľmi jednoducho umožňuje. Taktiež bolo potrebné prepojiť Javu s programovacím jazykom R, aby bolo možné spustiť skripty napísané v jazyku R vo vnútri Javy a vytvoriť tak predikčné modely pre demonštráciu výsledkov práce. Na toto prepojenie je voľne dostupné obrovské množstvo JAR knižníc. V našej práci sme využili voľne dostupnú knižnicu *RCaller*, ktorá je bližšie opísaná v nasledujúcej podkapitole.

6.1.1 Prepojenie R s Javou

Nakoľko štatistické modely a použitý skript boli vytvorené v programovacom jazyku R určenom na štatistickú analýzu dát, ale grafické užívateľské rozhranie a kostra aplikácie boli vytvorené v programovacom jazyku Java, bolo potrebné tieto dva programy istým spôsobom prepojiť. Samotná Java neobsahuje žiadne knižnice na prácu s jazykom R, resp. skriptami vytvorenými v jazyku R, no túto funkcionalitu umožňuje viacero voľne dostupných JAR knižníc. V aplikácií vytvorenej v rámci tejto bakalárskej práce bola na prepojenie medzi jazykom R a Javou použitá voľne dostupná knižnica *RCaller*. *RCaller* je softvérová knižnica vytvorená na zjednodušenie volania R skriptov z prostredia Javy. Práca s touto knižnicou je veľmi jednoduchá a pre jej používanie stačí mať nainštalovanú Javu a R. Práca so softvérovou knižnicou *RCaller* si v R-ku vyžaduje inštaláciu knižnice *RUniversal* príkazom *install.packages("Runiversal")*. [36][40]

6.2 Návrh tried

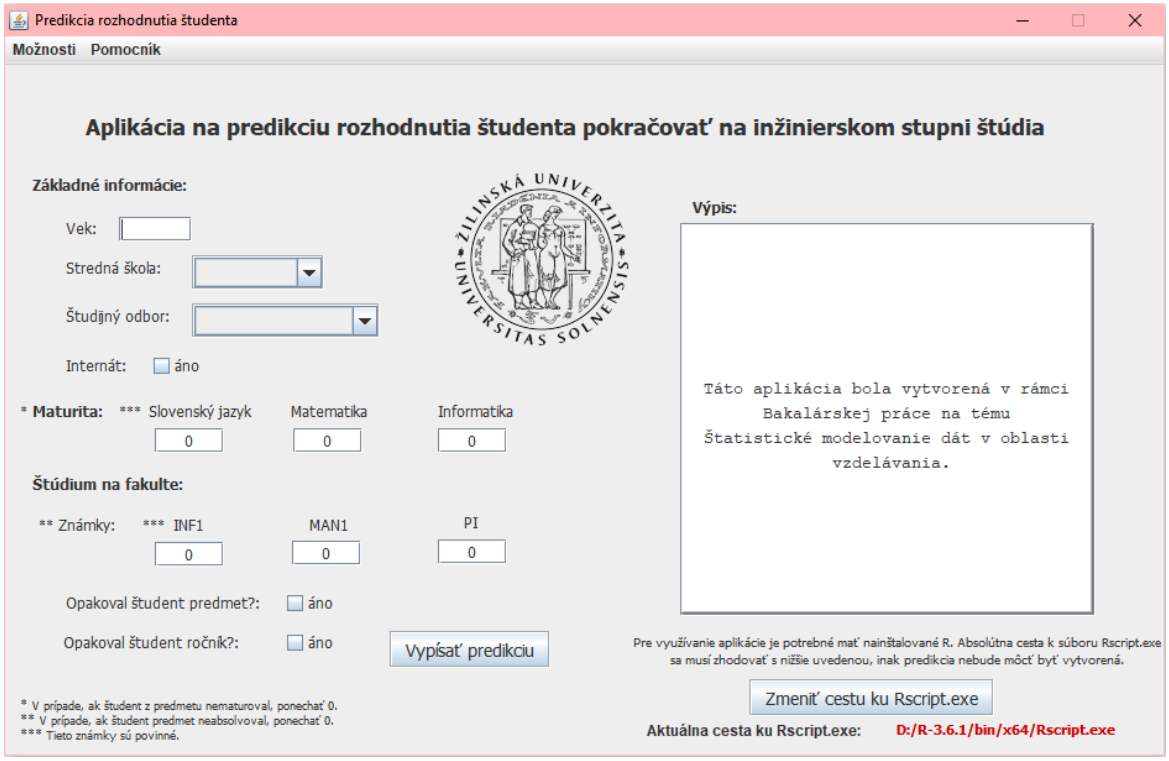
Celá implementácia aplikácie pozostáva zo štyroch Java tried, popisu ktorých sa budeme venovať v nasledujúcich podkapitolách, kde budú postupne opísané funkcionality týchto tried. UML diagram tried aplikácie je dostupný k nahliadnutiu v prílohe D.

6.2.1 Trieda app

Trieda **app** je typu JFrame form a predstavuje grafické používateľské rozhranie našej aplikácie. NetBeans IDE JFrame obsahuje pri vytvorení panel jPanel a poskytuje jednoduchý spôsob tvorby GUI pridávaním grafických prvkov do tohto panela, akými sú napr. textové polia jTextField alebo tlačítka jButton, a to všetko jednoducho, cez záložku „Design“. Všetky tieto prvky je možné editovať a nastavovať im akcie, ktoré sa majú vykonať pri vykonaní udalosti týkajúcej sa daného prvku. Grafické rozhranie je inicializované príkazom *this.initComponents();*.

6.2.1.1 Náhľad GUI

Pri tvorbe aplikácie bolo teda vytvorené jednoduché grafické užívateľské rozhranie, aby bolo ovládanie aplikácie čo najjednoduchšie a údaje sa nemuseli zadávať výpismi cez konzolu. Náhľad tohto grafického užívateľského rozhrania sa nachádza na nasledujúcom obrázku.



Obrázok 29 – GUI aplikácie

Zdroj: Vlastné spracovanie, 2020

Na tomto obrázku môžeme vidieť, že údaje užívateľ zadáva pomocou jednoduchých textových polí, rozbaľovacích polí a check boxov, pričom výsledok predikcie rozhodnutia študenta pokračovať v štúdiu sa vypíše do textovej oblasti v pravej časti aplikácie, a to po stlačení tlačidla „Vypísať predikciu“.

6.2.2 Trieda data

Trieda **data** obsahuje funkcie, ktoré slúžia na prácu s dátami. Dáta sú po zadaní údajov cez GUI postupne ukladané do ArrayListov *informacie* a *ciselneInformacie*. Na prístup ku prvkom týchto ArrayListov sú v triede **data** vytvorené funkcie *getInformacie()* a *getCiselneInformacie()*. Tieto gettre boli využité v triede app pri vkladaní údajov z GUI do ArrayListov pomocou *„názov gettra“.add()*. Dôležitou funkciou tejto triedy je funkcia *ulozCsv()*, ktorá slúži na zápis dát do súboru *student.csv*. Tento súbor je potom v R skripte načítaný a pomocou neho je vykonaná predikcia rozhodnutia študenta pokračovať v štúdiu. Vo funkcií je odchytávaná výnimka *FileNotFoundException* v prípade, ak by daný súbor nebol nájdený. Funkcia zápisu je volaná triedou **app** po úspešnom zadaní informácií o študentovi do GUI. Poslednou významnou funkciou triedy je funkcia *vypisInformacii(),* ktorá slúži na výpis základných informácií o študentovi do textovej oblasti užívateľského rozhrania.

6.2.3 Trieda connect

Trieda **connect** zabezpečuje prepojenie medzi Javou a programom R, na čo využíva softvérovú JAR knižnicu RCaller. Zjednodušene môžeme povedať, že funkcia *connectToR()* triedy **connect** zavolá skript vytvorený v jazyku R, vykoná predikciu a výslednú hodnotu predikcie vloží do premennej *vysledok* v Jave. Na začiatku sa vytvoria inštancie triedy ***RCaller*** a ***RCode***. Nakoľko všetky výpočty zabezpečuje program R, je potrebné inštancií ***RCaller*** nastaviť absolútnu cestu ku spúšťaciemu súboru *RScript.exe*. Absolútna cestu k tomuto súboru sa nastavuje funkciou *setRscriptExecutable()*. Riadky skriptu je možné pridávať po jednom priamo v Jave do inštancie triedy ***RCode*** pomocou príkazu *addRCode("prikaz v R")*, alebo pomocou príkazu *R_source(„absolútna cesta ku skriptu“)* je možné použiť vopred pripravený R skript. Nakoľko vytvorený skript *skript.R* je pomerne dlhý a jednoduchšie editovateľný v samotnom súbore, bola použitá druhá možnosť. Pre spustenie skriptu bolo potrebné inštancií ***RCaller*** „odovzdať“ skript príkazom *setRCode();* a vykonať ho príkazom *runAndReturnResult("premenná skriptu obsahujúca výsledok predikcie");*. Výslednú hodnotu predikcie (jedná sa o hodnotu celočíselnú) do Javy získame

pomocou príkazu *getParser().getAsIntArray("premenná skriptu obsahujúca výsledok predikcie")[0];* a vkladáme ju do premennej *vysledok*. V triede **connect** je vytvorená funkcia *vypisVysledku()*, ktorá na základe hodnoty premennej *vysledok* slovne vypíše výsledok predikcie rozhodnutia študenta pokračovať v inžinierskom stupni štúdia na FRI. Používa sa pre výpis tohto výsledku do textovej oblasti GUI. [40]

6.2.4 Trieda main

Každá aplikácia obsahuje triedu **main**. V triede **main** sa vytvára nová inštancia triedy **app**, teda nášho grafického užívateľského rozhrania. Príkazom *app.setVisible(true);* sa toto rozhranie nastaví na viditeľné a zobrazí sa užívateľovi, ktorý s ním môže ďalej pracovať.

6.3 Informácie pre správne spustenie aplikácie

Aplikácia bude fungovať správne a v zdrojových kódoch nebude potrebné nič meniť ak budú údaje z prílohy E – z DVD priloženého ku práci – v počítači uložené na **DATA D:/BakalarskaPraca**. V opačnom prípade je pre korektné spustenie aplikácie potrebné upraviť absolútne cesty ku jednotlivým súborom a „pracovné“ priečinky nasledovne:

1. V súbore skriptu **skript.R** je potrebné v príkaze *setwd("D:/BakalarskaPraca")* zmeniť absolútnu cestu „pracovného“ priečinka na priečinok, kde je súbor **studentiFri.csv** aktuálne uložený. Súbor **student.csv**, v ktorom sa nachádzajú údaje o študentovi po ich zadaní v GUI sa defaultne ukladá do priečinka *"D:/BakalarskaPraca"*. V prípade, ak by sme menili „pracovný“ priečinok na priečinok, kde sa aktuálne nachádza **skript.R**, je potrebné ho pred načítaním súboru **student.csv** nastaviť na *setwd("D:/BakalarskaPraca"),* v opačnom prípade predikcia nebude vytvorená.
2. Je potrebné mať v počítači nainštalované R. Absolútna cesta ku spúšťaciemu súboru **Rscipt.exe** je defaultne nastavená na *"D:/R-3.6.1/bin/x64/Rscript.exe"*. Táto absolútna cesta sa musí zhodovať s umiestnením súboru **Rscript.exe** v počítači. Je možné ju zmeniť priamo v aplikácií. V prípade, že by bola táto absolútna cesta zle nastavená, predikciu by nebolo možné vykonať. V zdrojovom kóde sa táto absolútna cesta nachádza v premennej *rScript* v triede **connect**.
3. V tej istej triede **connect** je potrebné premennej *sourceCode* zmeniť absolútnu cestu na cestu k súboru **skript.R**. Defaultne je nastavená na *"D:/BakalarskaPraca/skript.R"*.

Program je možné spustiť pomocou spúšťacieho .jar súboru nachádzajúceho sa v priečinku *BakalarskaPraca/BakalarskaPracaApp/dist/BakalarskaPracaApp.jar* (je potrebné mať nainštalovanú Javu).

7 Vyhodnotenie a diskusia

V rámci tejto práce boli vytvorené predikčné modely, pomocou ktorých bola riešená otázka zotrvania študenta na vysokej škole a boli identifikované faktory, ktoré na toto rozhodnutie študenta môžu vplývať. Vytvorené predikčné modely boli okrem identifikácie faktorov využité aj pri vypracovaní aplikácie, ktorá bola súčasťou praktickej časti tejto práce. Táto aplikácia je určená na predikciu rozhodnutia študenta pokračovať v inžinierskom stupni štúdia – konkrétne na Fakulte riadenia a informatiky – a to po zadaní údajov o študentovi dostupných na študijnom oddelení fakulty, týkajúcich sa jeho predchádzajúcich aj aktuálnych študijných výsledkov a priebehu štúdia.

Dáta, ktoré boli použité pri vytvorení modelov v tejto práci boli zozbierané od študentov Fakulty riadenia a informatiky navštevujúcich bakalársky stupeň štúdia. Spôsob modelovania a vytvorené modely boli inšpirované štúdiami uvádzanými v prvej kapitole práce a boli v nich zohľadňované rôzne skupiny faktorov, ktoré na rozhodnutie študenta môžu vplývať. Medzi tieto faktory patria stresové faktory, bydlisko študenta alebo spokojnosť študenta s jednotlivými aspektami ich štúdia. Definované teda boli štyri modely zohľadňujúce isté skupiny premenných, vychádzajúce z jednotlivých štúdií. Tieto modely boli následne vytvorené s využitím multinomickej logistickej regresie, logistickej regresie a klasifikačných rozhodovacích stromov.

7.1 Vyhodnotenie modelov

V kapitole 5 boli postupne vypracované modely s využitím logistickej regresie, multinomickej logistickej regresie a klasifikačných rozhodovacím stromov. Ako najpresnejší model sa preukázal model č.3 klasifikačného rozhodovacieho stromu zohľadňujúci spokojnosť študenta s jednotlivými aspektami jeho štúdia, medzi ktoré patria napríklad vybavenie školy alebo spokojnosť s rozložením semestra a hodnotením. Tento model vykazoval na testovacej množine úspešnosť 71,93%. Model č.2 zohľadňujúci bydlisko študenta sa naopak preukázal ako nie veľmi dobrý pri všetkých troch spôsoboch modelovania. Tento model nadobúdal v prípade klasifikačného rozhodovacieho stromu úspešnosť len niečo málo cez 56%.

Na základe vytvorených modelov ale môžeme povedať, že faktory spokojnosti študenta s rôznymi aspektami štúdia, ktorých vplyv na rozhodnutie študenta pokračovať v štúdiu a jeho celkovú spokojnosť so štúdiom bol potvrdený vo viacerých štúdiách, boli aj v tejto práci preukázané ako vplyvné.

7.2 Odporúčania a návrhy

Pri tvorbe modelov klasifikačných rozhodovacích stromov bol vytvorený aj model určený pre vypracovanie odporúčaní a návrhov, ktoré by mohli pozitívne vplývať na spokojnosť študenta s rôznymi aspektami jeho štúdia a ovplyvniť jeho rozhodnutie pokračovať v štúdiu aj na inžinierskom stupni štúdia. Model bol vytvorený s využitím premenných z modelu č.3 týkajúcich sa spokojnosti študenta, a to napríklad s vybavením školy, alebo so študijným programom, ktorý navštevuje. Grafická reprezentáciu tohto stromu sa nachádza na obrázku č.28 a poskytuje nám pohľad na premenné najvýznamnejšie vplývajúce na rozhodnutie študenta pokračovať v štúdiu. Informácie ku premenným, hodnoty týchto premenných a otázky im prislúchajúce sú dostupné v prílohe B a prílohe C.

Na základe tohto modelu sme zistili, že pokiaľ študent nie je spokojný so svojím študijným programom, ktorý navštevuje (premenná spStudProgram) má to negatívny vplyv na jeho rozhodnutie pokračovať v štúdiu. Takíto študenti nepokračujú v štúdiu vôbec alebo sa rozhodnú zmeniť odbor. Vybavenie (premenná spVybavenie), či už laboratórií alebo iných učební študenti taktiež považujú za dôležité. Toto je pochopiteľné aj z dôvodu, že Fakulta riadenia a informatiky je fakulta informačno-technologická.

Odbornosť (spOdbornost) a prístup pedagógov (spPristup) sú taktiež dôležité faktory. Podobne je tomu tak aj v prípade objektívneho hodnotenia študentov (spHodnotenie). V rámci projektu To dá rozum bolo dokázané, že veľa teoretickej prípravy na bakalárskom stupni štúdia s minimálnymi možnosťami využiť tieto teoretické vedomosti v praxi môže viesť ku strate motivácie k učeniu. Faktory vplývajúce na vzťah študenta k štúdiu môžeme rozdeliť na vnútorné (motivátory) a vonkajšie (dissatisfaktory). Medzi motivátory patria práve aj prístup pedagógov a spravodlivé hodnotenie. Študenti, ktorí sú s týmito faktormi spokojní – vyplývajúc z obrázku č.28 – v štúdiu pokračujú.

Pokiaľ študenti považujú niektoré predmety v osnovách ich študijného programu za zbytočné (premenná spZbytocne) má to taktiež negatívny vplyv na ich spokojnosť a rozhodnutie pokračovať v štúdiu. Z obrázku vyplýva, že študenti odboru Informatika, ktorí považujú niektoré predmety z osnov za zbytočné nepokračujú v štúdiu vôbec. Toto by sa zhodovalo aj s informáciami vyplývajúcimi z projektu To dá rozum (pozri kapitola 1), v ktorom bolo zistené, že študenti informatiky zvyknú nepokračovať v inžinierskom štúdiu aj z toho dôvodu, že to považujú za zbytočné. Takisto oproti štandardom v iných krajinách skoro 40% študentov slovenských vysokých škôl uviedlo, že za semester má priemerne 8

a viac predmetov. Aj z dôvodu vysokého počtu predmetov môžu študenti niektoré považovať za nedôležité.

7.3 Diskusia

Vytvorením štatistických modelov boli identifikované faktory vplývajúce na rozhodnutie študenta pokračovať v štúdiu na inžinierskom stupni. V rámci praktickej časti tejto práce boli vytvorené modely multinomickej logistickej regresie, logistickej regresie a klasifikačných rozhodovacích stromov. Jednotlivé modely a premenné v nich použité boli inšpirované štúdiami predstavenými v prvej kapitole tejto práce. Pochopiteľne najvyššiu úspešnosť a presnosť vykazovali modely zohľadňujúce celkovú spokojnosť študenta, či už s vybavením školy, ale aj so študijným programom alebo napríklad príležitosťami, ktoré škola ponúka. Takéto modely vykazovali najvyššiu celkovú úspešnosť pri všetkých troch spôsoboch modelovania.

Problémom pri vytváraní niektorých modelov bola pomerne malá veľkosť vzorky dát. Ako príklad môžeme uviesť faktorovú premennú kraj, pri ktorej sme zaznamenali pri niektorých krajoch chýbajúce údaje z dôvodu nedostatku dát v datasete pre jednotlivé kraje. Toto bolo spôsobené aj tým, že väčšina respondentov bola zo Žilinského kraja. Modely by bolo možné v budúcnosti zlepšiť získaním väčšieho množstva dát, a to uskutočnením prieskumu v doteraz neskúmaných študijných programoch. Nakoľko fakulta predstavila dva nové bakalárske študijné programy, bolo by vhodné prieskum uskutočniť aj pri študentoch navštevujúcich tieto nové študijné programy.

Vzorka dát bola aj pomerne nerovnomerná, pričom väčšinu respondentov tvorili študenti tretieho ročníka, navštevujúci študijné programy Informatika a Manažment. Tento problém by bolo možné v budúcnosti riešiť lepšou distribúciou dotazníka medzi jednotlivé študijné programy a ročníky. Následne, keby bola vzorka dát väčšia by bolo možné analýzu vykonať aj zvlášť pre každý študijný program, prípadne ročník.

Je pravdepodobné, že na rozhodnutie študenta pokračovať v štúdiu na druhom stupni štúdia vplýva viacero faktorov, ako aj faktory, ktoré v tejto práci neboli prostredníctvom zozbieraných dát odhalené a nemohli byť následne pomocou vytvorených modelov overené. Na prácu by bolo možné v budúcnosti nadviazať rozšírením dotazníka, pridaním ďalších otázok, uskutočnením prieskumu v iných odboroch a ročníkoch, a vytvorením modelov zohľadňujúcich ďalšie faktory, ktoré v rámci tejto práce neboli identifikované a mohli by ovplyvňovať rozhodnutie študenta pokračovať v štúdiu.

Záver

Veľa študentov stredných škôl považuje správnu voľbu vysokej školy za jedno z najdôležitejších rozhodnutí v živote, ktoré v značnej miere ovplyvní jeho budúce fungovanie v spoločnosti. Študenti si vyberajú školy na základe množstva faktorov, pričom na tie pre nich najzaujímavejšie si rozošlú prihlášky a čakajú na pozvánku na prijímacie skúšky, prípadne neskôr na zápis do prvého ročníka. Veľa takýchto prihlášok chodí každý rok aj na študijné oddelenie Fakulty riadenia a informatiky a každý rok v septembri veľké množstvo nových prvákov, budúcich bakalárov začne novú etapu v živote ako vysokoškolskí študenti s vidinou titulu. Čo ich však po získaní bakalárskeho titulu vedie k rozhodnutiu nepokračovať na druhom stupni vysokoškolského štúdia?

Cieľom tejto bakalárskej práce bolo na základe teoretických poznatkov a analýzy reálnych dát získaných prostredníctvom dotazníka od študentov Fakulty riadenia a informatiky identifikovať faktory, ktoré môžu vplývať na spokojnosť študentov na vysokej škole. Pomocou teórie a predchádzajúcich štúdií zaoberajúcich sa touto tematikou, v ktorých boli identifikované faktory vplývajúce na mieru udržateľnosti študentov, mieru zotrvania študentov na škole a celkovú spokojnosť študentov, boli – vychádzajúc z týchto štúdií – vytvorené predikčné modely na predikciu rozhodnutia študenta pokračovať v inžinierskom stupni štúdia, konkrétne na Fakulte riadenia a informatiky. Tieto modely boli vytvorené využitím viacerých spôsobov predikčného modelovania.

Ako najpresnejší model sa preukázal model klasifikačného rozhodovacieho stromu zohľadňujúci celkovú spokojnosť študenta so štúdiom, študijným programom, vybavením školy alebo napríklad príležitosťami, ktoré škola ponúka. Toto bolo potvrdené vo viacerých zahraničných štúdiách a dalo sa predpokladať, že tieto faktory sa preukážu ako dôležité aj v našom prípade. Na základe tohto zistenia bol zostrojený klasifikačný rozhodovací strom, prostredníctvom ktorého boli ponúknuté návrhy a odporúčania, ktoré by mohli pozitívne vplývať na spokojnosť študenta a tým pádom aj na jeho rozhodnutie pokračovať v štúdiu.

Súčasťou tejto práce bolo aj vytvorenie aplikácie na predikciu rozhodnutia študenta pokračovať v inžinierskom stupni štúdia na Fakulte riadenia a informatiky. Aplikácia slúži na demonštráciu výsledkov práce a je schopná predikovať toto rozhodnutie študenta na základe informácií dostupných na študijnom oddelení Fakulty riadenia a informatiky týkajúcich sa aktuálneho priebehu štúdia, aktuálnych študijných výsledkov študenta, ale aj študijných výsledkov zo strednej školy.

Referencie

[1] ACHEAMPONG, Edward, Dominic BUER BOYETEY, Frank OSEI GYIMAH a Eric OKYERE. Assessing student satisfaction: An application of Logistic Regression Analysis to methodist University College Ghana (MUCG) Data [online]. 2013 [cit. 2020-03-16]. Dostupné z: https://www.researchgate.net/publication/276899085

[2] Algoritmy strojového učenia I. - Učenie s učiteľom. *Umelá Inteligencia.sk* [online]. [cit. 2020-02-22]. Dostupné z: https://umelainteligencia.sk/algoritmy-strojoveho-ucenia/

[3] Analýza kontingenčních tabulek. *Jak analyzovat kontignenční tabulky* [online]. [cit. 2020-04-02]. Dostupné z: https://ksoc.ff.cuni.cz/wp-content/uploads/sites/76/2018/09/Jak-analyzovat-kontignen%C4%8Dn%C3%AD-tabulky.pdf

[4] Association Tests for Ordinal Tables. *R Handbook* [online]. [cit.2020-03-17]. Dostupné z: https://rcompanion.org/handbook/H_09.html

[5] BEANE, Robbie. Lesson 4.2 - Multinomial Logistic Regression. *RPubs - 4.2 - Multinomial Logistic Regression* [online]. [cit. 2020-03-29]. Dostupné z: https://rpubs.com/beane/n4_2

[6] Binárna logistická regresia. *Štatistika v PSPP* [online]. 2017 [cit. 2020-02-22]. Dostupné z: https://statistikapspp.sk/binarna-logisticka-regresia-2/

[7] BRIGANT, Vladimír. *Evoluční návrh simulátoru založeného na celulárních automatech* [online]. Brno, 2011 [cit. 2020-03-31]. Dostupné z: https://core.ac.uk/download/pdf/44392782.pdf. Bakalárska práca. Vysoké učení technické v Brně, Fakulta informačních technologií, Ústav počítačové grafiky a multimédií.

[8] COTTON, Sarah J. a kol. Stress and Student Job Design: Satisfaction, Well-Being, and Performance in University Students. *International Journal of Stress Management* [online]. 2002, s. 147-162 [cit. 2020-03-14]. Dostupné z: https://www.researchgate.net/publication/225470183

[9] Data visualization beginner's guide: a definition, examples, and learning resources. *Tableau Software* [online]. [cit. 2020-03-12]. Dostupné z: https://www.tableau.com/learn/articles/data-visualization

[10] Definícia. *Strojové učenie - Definícia* [online]. [cit. 2020-02-22]. Dostupné z: https://smnd.sk/mcibula/zakl_info/definicia.html

[11] Dôvody pokračovania na druhý stupeň VŠ. *To Dá Rozum* [online]. [cit. 2020-03-14]. Dostupné z: https://analyza.todarozum.sk/docs/401066001ui1a/

[12] Exploračná analýza. *Exploračná analýza: Charakteristiky variability* [online]. [cit. 2020-03-12]. Dostupné z: https://amos.ukf.sk/mod/book/view.php?id=8408&chapterid=3165

[13] Exploratory Analysis vs. Confirmatory Analysis. *NDMU Online* [online]. [cit. 2020-02-25]. Dostupné z: https://online.ndm.edu/news/analytics/exploratory-analysis-vs-confirmatory-analysis/

[14] Exploratory and Confirmatory Analysis: What's the Difference? l Sisense. *Sisense* [online]. [cit. 2020-02-28]. Dostupné z: https://www.sisense.com/blog/exploratory-confirmatory-analysis-whats-difference/

[15] FALÁT, Lukáš. 2019. *Dáta, informácie, znalosti*. [edit. 2020-04-06]. Prednáška.

[16] FAQ: What are pseudo R-squareds? *IDRE Stats* [online]. 2011 [cit. 2020-04-23]. Dostupné z: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/

[17] Chi-squared test for association in R. [online]. [cit. 2020-03-17]. Dostupné z: https://www.sheffield.ac.uk/polopoly_fs/1.714597!/file/stcp-karadimitriou-chisqR.pdf

[18] Choosing the Correct Statistical Test in SAS, Stata, SPSS and R. *IDRE Stats* [online]. [cit. 2020-03-17]. Dostupné z: https://stats.idre.ucla.edu/other/mult-pkg/whatstat/

[19] Introduction to dnorm, pnorm, qnorm, and rnorm for new biostatisticians. *Sean Kross* [online]. October 1, 2015 [cit. 2020-04-09]. Dostupné z: http://seankross.com/notes/dpqr/

[20] IZAKOVIČOVÁ, Michaela. *Využitie metód analýzy ordinálnych dát v analýzach systémov celoživotného vzdelávania* [online]. Bratislava, 2015 [cit. 2020-04-21]. Dostupné z: http://webcache.googleusercontent.com/search?q=cache:Bf6PbHHDWWcJ:ftp://193.87.31.84/0206541/Dizertacka_Izakovicova_final.pdf. Dizertačná práca. Ekonomická Univerzita v Bratislave, Fakulta hospodárskej informatiky.

[21] JAMES, Gareth, Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI. *An Introduction to Statistical Learning: with Applications in R* [online]. 7th. New York: Springer Science+Business Media, 2013 [cit. 2020-03-23]. ISBN 978-1-4614-7138-7. Dostupné z: http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf

[22] JANOUŠOVÁ, Eva a Ladislav DUŠEK. *Analýza dat pro Neurovědy* [online]. 2013 [cit. 2020-03-28]. Dostupné z: https://is.muni.cz/el/med/jaro2013/DSAN01/um/40398795/Analyza_dat_pro_Neurovedy_-_blok_6.pdf

[23] JIM, Frost. How to Interpret Adjusted R-Squared and Predicted R-Squared in Regression Analysis. *Statistics By Jim* [online]. 2019 [cit. 2020-04-23]. Dostupné z: https://statisticsbyjim.com/regression/interpret-adjusted-r-squared-predicted-r-squared-regression/

[24] KAPASNÝ, Juraj. *Multinomická logistická regrese, Trojcestné ROC, VUS* [online]. Brno, 2015 [cit. 2020-03-28]. Dostupné z: https://is.muni.cz/th/ei7h1/diplomka_Kapasny.pdf. Diplomová práca. Masarykova univerzita, Přírodovědecká fakulta, Ústav matematiky a statistiky.

[25] KNEŽO, Dušan, Miriam ANDREJIOVÁ a Gabriela IŽARÍKOVÁ. *ZÁKLADNÉ ŠTATISTICKÉ METÓDY* [online]. 2011 [cit. 2020-03-10]. Dostupné z: https://www.sjf.tuke.sk/kamai/literatura/stat_metody.pdf

[26] Korelačná a regresná analýza. *Matematikabezproblemov.webjet.sk* [online]. [cit. 2020-02-22]. Dostupné z: http://matematikabezproblemov.webjet.sk/domov/studijne-materialy/matematika-vs/pravdepodobnost-statistika/korelacna-regresna-analyza/

[27] KRAVČÁKOVÁ, Gabriela, Stanislava KOZELOVÁ a Eliška ŽUPOVÁ. *Vzťah k práci (štúdiu) vysokoškolských študentov* [online]. 2011 [cit. 2020-04-03]. Dostupné z:https://www.researchgate.net/publication/308020463_Vztah_k_praci_studiu_vysokoskolskych_studentov

[28] Logistic Regression in R. *RPubs - Logistic, Ordinal, and Multinomial Regression in R* [online]. 2017 [cit. 2020-04-09]. Dostupné z: https://rpubs.com/rslbliss/r_logistic_ws

[29] Measures of Association for Nominal Variables. *R Handbook* [online]. [cit. 2020-03-16]. Dostupné z: https://rcompanion.org/handbook/H_10.html

[30] Multinomial Logistic Regression using SPSS Statistics. *How to perform a Multinomial Logistic Regression in SPSS Statistics | Laerd Statistics* [online]. [cit. 2020-02-22]. Dostupné z: https://statistics.laerd.com/spss-tutorials/multinomial-logistic-regression-using-spss-statistics.php

[31] ORŠANSKÝ, Pavol. Testovanie štatistických hypotéz a korelačná analýza. *Strojnícka fakulta ŽU* [online]. [cit. 2020-03-12]. Dostupné z: http://fstroj.uniza.sk/kam/orsansky/pdf/4_prednaska.pdf

[32] PARALIČ, Ján. *Objavovanie znalostí v databázach* [online]. Košice: Elfa, 2003 [cit. 2020-03-23]. ISBN 80-89066-60-7. Dostupné z: https://spu.fem.uniag.sk/cvicenia/ksov/zach/Data%20mining/Literatura/ObjavovanieZnalostivDB.pdf

[33] PIDGEON, Aileen M., Nyketa L. DAVIES a Peta STAPLETON. Factors Influencing University Students' Academic Experience: An International Study. *International Journal of Multidisciplinary Perspectives in Higher Education* [online]. 2017, s. 1-8 [cit. 2020-03-14]. Dostupné z: https://www.ojed.org/index.php/jimphe/issue/view/38/Factors%20Influencing%20University%20Students%E2%80%99%20Academic%20Experience%3A%20An%20International%20Study

[34] Predčasné ukončenie VŠ štúdia. *To Dá Rozum* [online]. [cit. 2020-03-15]. Dostupné z: https://analyza.todarozum.sk/docs/432871002mr0a/

[35] PseudoR2. *PseudoR2 function | R Documentation* [online]. [cit. 2020-04-21]. Dostupné z: https://www.rdocumentation.org/packages/DescTools/versions/0.99.34/topics/PseudoR2

[36] Rcaller 2.0 - Calling R from Java. *Practical Code Solutions* [online]. [cit. 2020-04-11]. Dostupné z: http://stdioe.blogspot.com/2011/07/rcaller-20-calling-r-from-java.html

[37] RIMARČÍK Marián. *Štatistika pre prax* [online]. Marián Rimarčík, 2007 [cit. 2020-04-02]. ISBN 80-96981-31-1.

[38] Rozhodovacie stromy. *Strojové učenie - Rozhodovacie stromy* [online]. [cit. 2020-02-23]. Dostupné z: https://smnd.sk/mcibula/alg/DT.html

[39] Rpart. *Function | R Documentation* [online]. [cit. 2020-03-28]. Dostupné z: https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart.control

[40] SATMAN, M. Hakan. RCaller: A Software Library for Calling R from Java. *British Journal of Mathematics & Computer Science* [online]. 2014, 06 June 2014, **4**(15), 2188-2196 [cit. 2020-04-11]. DOI: 10.9734/BJMCS/2014/10902. Dostupné z: http://www.sciencedomain.org/abstract/4838

[41] SIČOVÁ, Mária. 2020. Študijné oddelenie Fakulty riadenia a informatiky. *Zapísaní študenti*. 2020-02-26. [cit. 2020-02-29]. Osobná komunikácia.

[42] SIMING, Luo, NIAMATULLAH, Jianying GAO, Dan XU a Khurrum SHAFI. *Factors Leading to Students' Satisfaction in the Higher Learning Institutions* [online]. 2015 [cit. 2020-03-16]. Dostupné z: https://files.eric.ed.gov/fulltext/EJ1083362.pdf

[43] STANKOVIANSKA, Ida. 2017. *Pravdepodobnosť a štatistika*. [cit. 2020-04-02]. Prednáška.

[44] Survey Data Analysis: Descriptive vs. Inferential Statistics. *Cvent Blog* [online]. [cit. 2020-03-12]. Dostupné z: https://www.cvent.com/en/blog/events/survey-data-analysis-descriptive-vs-inferential-statistics

[45] Štatistické pojmy. *Štatistický úrad SR* [online]. [cit. 2020-02-23]. Dostupné z: https://www7.statistics.sk/wps/portal/ext/services/statsimple/statterms/

[46] Testování nezávislosti (Pearsonův chí-kvadrát test). *Matematická biologie učebnice: Testování nezávislosti (Pearsonův chí-kvadrát test)* [online]. [cit. 2020-04-01]. Dostupné z: https://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--analyza-a-management-dat-pro-zdravotnicke-obory--testovani-hypotez-o-kvalitativnich-promennych--analyza-kontingencnich-tabulek--testovani-nezavislosti-pearsonuv-chi-kvadrat-test

[47] Učenie bez učiteľa. *Strojové učenie - Fungovanie - Učenie bez učiteľa* [online]. [cit. 2020-02-22]. Dostupné z: https://smnd.sk/mcibula/fungovanie/fungovanie-unsl.html

[48] Učenie s učiteľom. *Strojové učenie - Fungovanie - Učenie s učiteľom* [online]. [cit. 2020-02-22]. Dostupné z: https://smnd.sk/mcibula/fungovanie/fungovanie-sl.html

[49] Úvod do predmetu. *Základy štatistiky* [online]. [cit. 2020-02-23]. Dostupné z: https://www.unipo.sk/public/media/30187/t%C3%A9ma%201.pdf

[50] VÝROČNÁ SPRÁVA O ČINNOSTI ZA ROK 2018. *Žilinská univerzita v Žiline: Fakulta riadenia a informatiky* [online]. Žilina: EDIS, 2019 [cit. 2020-03-01]. Dostupné z: https://www.fri.uniza.sk/uploads/files/1557921573-FRI-vyrocna-sprava-za-rok-2018.pdf

[51] What is Data Analysis ? *PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices* [online]. [cit. 2020-03-12]. Dostupné z: https://www.predictiveanalyticstoday.com/data-analysis/

[52] YU, Chong Ho a kol. *A data-mining approach to differentiate predictors of retention* [online]. 2007 [cit. 2020-03-15]. Dostupné z: https://files.eric.ed.gov/fulltext/ED496657.pdf

Prílohy

Príloha A: Tabuľka zapísaných študentov podľa ročníkov

Študijný odbor	Stupeň štúdia	Spolu	Ženy	Ročníky					
				1.		2.		3.	
				Spolu	Ženy	Spolu	Ženy	Spolu	Ženy
počítačové inžinierstvo	1	87	8	32	2	27	2	28	4
manažment	1	194	101	54	23	65	32	75	46
informatika	1	621	84	228	23	155	23	238	38
biomedicínska informatika	2	17	4	17	4	-	-		
počítačové inžinierstvo	2	27	2	13	-	14	2		
inteligentné informačné systémy	2	14	1	6	1	8	-		
informačný manažment	2	114	66	63	34	51	32		
aplikované sieťové inžinierstvo	2	38	6	16	2	22	4		
informačné systémy	2	121	8	51	4	70	4		
Spolu		1233	280	480	93	412	99	341	88
Opakujúci		175	24	29	2	54	8	92	14

Príloha B: Dotazník

	Premenné	Otázky	Odpovede
Základné informácie	ID	-	kedy bola odpoveď vytvorená
	pohlavie	Pohlavie	žena / muž
	vek	Vek	číselný údaj
	rocnik	Ročník	prvý / druhý / tretí
	odbor	Študijný odbor	Informatika / Manažment / Počítačové inžinierstvo
	kraj	Z ktorého kraja pochádzaš?	Žilinský kraj / Trnavský kraj / Košický kraj / Bratislavský kraj / Trenčiansky kraj / Prešovský kraj / Nitriansky kraj / Banskobystrický kraj / iné
	bydlisko	Bývaš..?	v meste / na dedine
	internat	Bývaš na internáte?	áno / nie
	typSS	Aký typ strednej školy si navštevoval/a?	gymnázium / odborná škola
	matSJL	Aká bola tvoja známka z maturity zo slovenského jazyka?	1 / 2 / 3 / 4 / z tohto predmetu som nematuroval/a
	matMAT	Aká bola tvoja známka z maturity z matematiky?	
	matINF	Aká bola tvoja známka z maturity z informatiky?	
	matFYZ	Aká bola tvoja známka z maturity z fyziky?	
	matOBN	Aká bola tvoja známka z maturity z občianskej náuky?	
	matODB	Aká bola tvoja známka z maturity z odborných predmetov?	
	progrSS	Učili ste sa na hodinách informatiky na SŠ základy nejakého programovacieho jazyka?	áno / nie
	zhoduje	Zhoduje sa to, čo si študoval/a na strednej škole, resp. z čoho si maturoval/a s tým, čo študuješ teraz?	áno / čiastočne / nie
	vIneSkoly	Rozmýšľal/a si pred nástupom na Fakultu riadenia a informatiky aj o iných vysokých školách?	áno / nie
	vPrvaVolba	Bola Fakulta riadenia a informatiky tvoja prvá voľba spomedzi vysokých škôl?	
	vSam	Vybral/a si si Fakultu riadenia a informatiky sám/sama?	
	vRodicia	Rozhodol/la si sa ísť študovať kvôli rodičom?	áno / čiastočne / vôbec
	vLudiZapis	Koľko ľudí si poznal/a pri zápise? (ktorí už na škole študovali, alebo ste šli študovať spolu)	číselný údaj
Štúdium	sBavi	Baví ťa to, čo študuješ?	áno / čiastočne / vôbec
	sZnova	Vybral/a by si si tento odbor znova?	áno / nie
	sINF	Aká bola tvoja známka z predmetu Informatika 1?	A / B / C / D / E / predmet opakujem / predmet nemám v študijnej osnove
	sMAN	Aká bola tvoja známka z predmetu Manažment 1?	
	sPI	Aká bola tvoja známka z predmetu Počítačové inžinierstvo?	A / B / C / D / E / predmet opakujem / predmet nemám v študijnej osnove / predmet som ešte nemal/a
	sPAS	Aká bola tvoja známka z predmetu Pravdepodobnosť a štatistika?	
	sAUS	Aká bola tvoja známka z predmetu Algoritmy a údajové štruktúry 1?	
	sPrednasky	Navštevuješ prednášky?	áno, chodím na všetky / áno, ale niektoré vynechám / občas sa nejakej zúčastním / na prednášky nechodím vôbec
	sRodiciaPodpora	Rodičia ma v štúdiu podporujú a záleží im na mojom vzdelaní a na získaní titulu.	súhlasím / skôr súhlasím / skôr nesúhlasím / nesúhlasím
	sPrenasam	"Prenášal/a" si niekedy nejaký predmet?	áno / nie
	sOpakujem	Opakoval/a si niekedy ročník?	áno / nie / pravdepodobne budem
	sPredcasne	Zvažoval/a si niekedy predčasne ukončiť štúdium?	áno / nie / plánujem ho prerušiť
	sPredcasneKam	Ukončil nejaký tvoj kamarát predčasne štúdium?	áno / nie / plánuje ho prerušiť
	sPomoc	Mal si pocit, že ako študenti ťaháte za jeden povraz a pomáhate si pri štúdiu?	áno / čiastočne / vôbec
	sVztahy	Vzťahy, ktoré som si vytvoril/a počas štúdia na fakulte považujem za:	pozitívne / neutrálne / negatívne
	sSocZivot	Mal/a si pocit, že škola výrazne obmedzuje tvoj sociálny život?	áno / čiastočne / nie
	sNezvladam	Mal/a si pocit, že štúdium nezvládaš?	často / občas / nikdy
	sPsycholog	Vyhľadal/a si niekedy počas štúdia odbornú pomoc psychológa?	áno / nie, ale premýšľal/a som o tom / nie, nepotreboval/a som to
	sNarocne	Považuješ štúdium za náročné?	áno / čiastočne / vôbec
	sNaroky	Máš pocit, že sú na študentov kladené príliš vysoké nároky?	

Spokojnosť	spProgramy	Som spokojný/a so študijnými programami, ktoré fakulta ponúka.	súhlasím / skôr súhlasím / skôr nesúhlasím / vôbec nesúhlasím
	spPrilezitosti	Som spokojný/a s príležitosťami, ktoré fakulta ponúka. (či už pracovné alebo nejaké iné)	
	spSemester	Som spokojný/a s rozložením predmetov v semestroch a považujem ich za vyvážené, čo sa náročnosti týka.	
	spOdradza	Niektoré predmety ma pri štúdiu odrádzajú.	
	spZbytocne	Niektoré predmety považujem za zbytočné z hľadiska praxe.	
	spZastarale	Niektoré predmety považujem za zastaralé.	
	spStudProgram	Som spokojný/a so študijným programom, ktorý študujem.	
	spVybavenie	Som spokojný/a s vybavením školy.	
	spOdbornost	Som spokojný/a s odbornosťou pedagógov.	
	spPristup	Som spokojný/a s prístupom a ochotou pedagógov.	
	spHodnotenie	Hodnotenie znalostí študentov považujem za objektívne.	
	spTerminy	Počas skúškového obdobia bol počet skúškových termínov dostačujúci.	
Práca	pPraca	Pracuješ popri škole?	áno / nie
	pVOdbore	Pracuješ v odbore, ktorý študuješ?	áno / nie / čiastočne / popri štúdiu nepracujem
	pVyucbaPrax	Myslíš, že výučba bola dostatočne orientovaná na prax?	súhlasím / skôr súhlasím / skôr nesúhlasím / vôbec nesúhlasím
	pVyuzitieUciva	To čo študujem, reálne využívam aj v praxi.	
	pTitul	Pre výkon mojej práce je dôležité získať inžiniersky titul.	
	pFinNarocne	Štúdium považujem za finančne náročnú záležitosť.	
	pFinRodicia	Rodičia ma pri štúdiu podporujú finančne.	áno / čiastočne / nie
Vzťah ku fakulte	fAkcie	Zúčastňuješ sa na akciách, ktoré fakulta organizuje? (Fričkovica, FRI punč, FRI ples...)	áno občas nie, pretože nemám čas nie, pretože nemám záujem
	fPovest	Fakulta riadenia a informatiky má dobrú povesť.	súhlasím skôr súhlasím skôr nesúhlasím vôbec nesúhlasím
	fOdporucam	Odporučil/a by si štúdium na fakulte svojim známym?	áno / nie
	fHrdy	Si hrdý/a na to, že študuješ na Fakulte riadenia a informatiky?	
	fHodnotenie	Fakultu ako celok hodnotím	pozitívne skôr pozitívne skôr negatívne negatívne
	fInaSkola	Rozmýšľal/a si nad tým, že by si pokračoval v štúdiu na inej škole?	áno, aj plánujem študovať inde áno, ale nie som o tom presvedčený/á nie, chcem pokračovať v štúdiu na FRI nie, vôbec neplánujem ďalej študovať
	fErasmus	Zúčastnil/a si sa počas štúdia programu Erasmus+?	áno a bola to skvelá skúsenosť áno, ale už by som nešiel/nešla nie, nemal/a som o také niečo záujem nie, nemal/a som čas zúčastniť sa
	fVZahranici	Rozmýšľal/a si nad tým, že by si pokračoval/a v štúdiu v zahraničí?	áno, aj tam študovať plánujem / áno, ale nie som o tom presvedčený/á / plánujem študovať na Slovensku / nie, vôbec neplánujem ďalej študovať
	fDobraVolba	Bola z celkového pohľadu Fakulta riadenia a informatiky pre teba dobrá voľba?	áno / nie
	fZnova	Ak by si si mohol/mohla znovu vybrať, šiel/šla by si študovať na Fakultu riadenia a informatiky?	áno, presne na tento istý odbor / áno, ale na iný odbor / nie, zvolil / a by som štúdium na inej univerzite / nešiel / nešla by som študovať vôbec
	fIng	Plánuješ pokračovať v inžinierskom štúdiu na Fakulte riadenia a informatiky?	áno / áno, ale menim odbor / nie, plánujem študovať na inej škole / neplánujem pokračovať v štúdiu vôbec

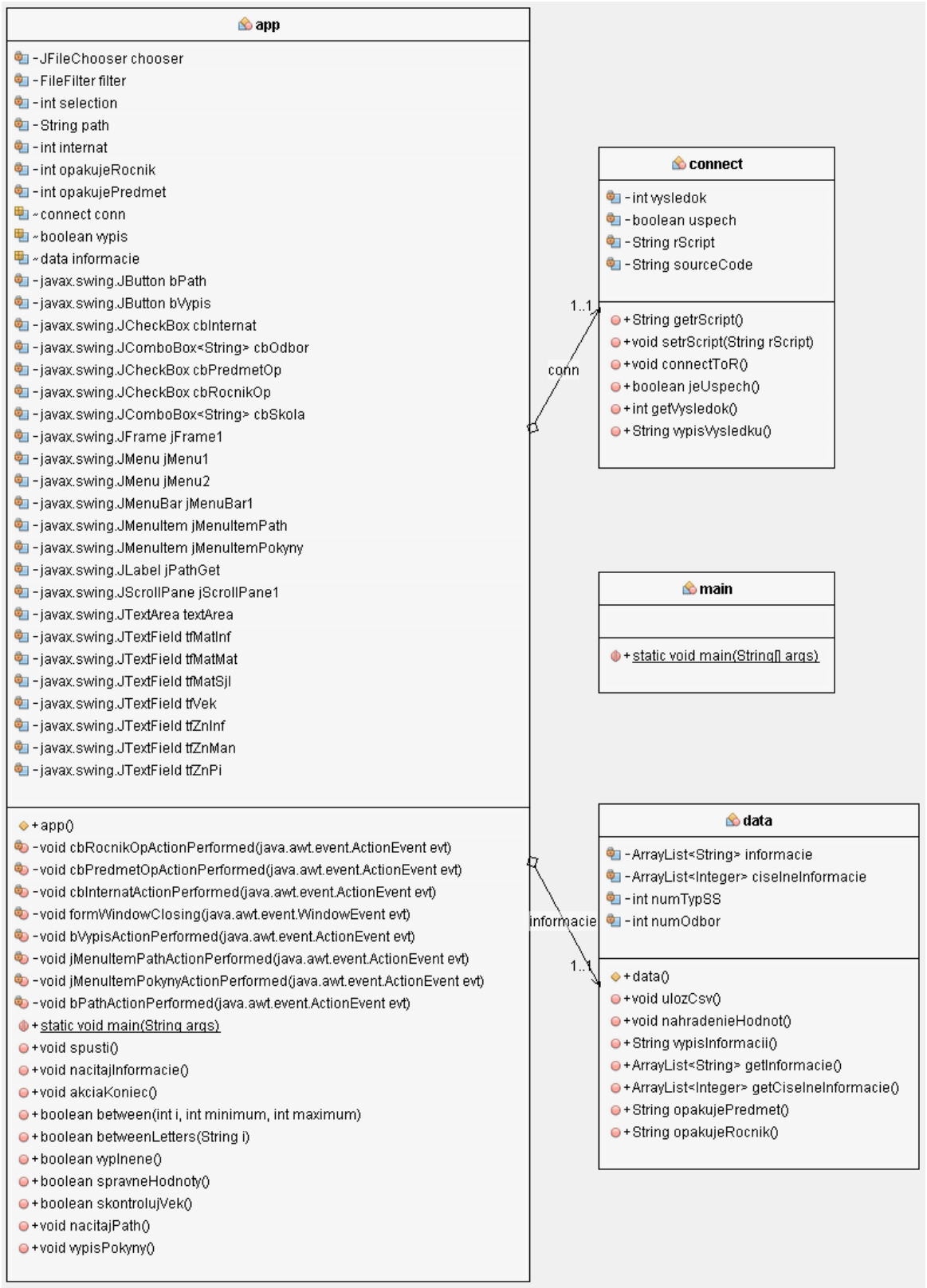
Príloha C: Premenné a im priradené hodnoty

	Premenné	Odpovede	Vyjadrenie premenných
Základné informácie	ID	kedy bola odpoveď vytvorená	-
	pohlavie	žena / muž	0 / 1
	vek	číselný údaj	-
	rocnik	prvý / druhý / tretí	1 / 2 / 3
	odbor	Informatika / Manažment / Počítačové inžinierstvo	2 / 1 / 0
	kraj	Žilinský kraj / Trnavský kraj / Košický kraj / Bratislavský kraj / Trenčiansky kraj / Prešovský kraj / Nitriansky kraj / Banskobystrický kraj / iné	8 / 7 / 6 / 5 / 4 / 3 / 2 / 1 / 0
	bydlisko	v meste / na dedine	1 / 0
	internat	áno / nie	1 / 0
	typSS	gymnázium / odborná škola	1 / 0
	matSJL	1 / 2 / 3 / 4 / z tohto predmetu som nematuroval/a	1 / 2 / 3 / 4 / 0
	matMAT		
	matINF		
	matFYZ		
	matOBN		
	matODB		
	progrSS	áno / nie	1 / 0
	zhoduje	áno / čiastočne / nie	2 / 1 / 0
	vIneSkoly	áno / nie	1 / 0
	vPrvaVolba		
	vSam		
	vRodicia	áno / čiastočne / vôbec	2 / 1 / 0
	vLudiZapis	číselný údaj	-
Práca	pPraca	áno / nie	1 / 0
	pVOdbore	áno / nie / čiastočne / popri štúdiu nepracujem	2 / 1 / 0 / -
	pVyucbaPrax	súhlasím / skôr súhlasím / skôr nesúhlasím / vôbec nesúhlasím	3 / 2 / 1 / 0
	pVyuzitieUciva		
	pTitul		
	pFinNarocne		
	pFinRodicia	áno / čiastočne / nie	2 / 1 / 0

	Premenné	Odpovede	Vyjadrenie premenných
Štúdium	sBavi	áno / čiastočne / vôbec	2 / 1 / 0
	sZnova	áno / nie	1 / 0
	sINF	A / B / C / D / E / predmet opakujem / predmet nemám v študijnej osnove	1 / 2 / 3 / 4 / 5 / 6 / 0
	sMAN		
	sPI	A / B / C / D / E / predmet opakujem / predmet nemám v študijnej osnove / predmet som ešte nemal/a	1 / 2 / 3 / 4 / 5 / 6 / 0 / 0
	sPAS		
	sAUS		
	sPrednasky	áno, chodím na všetky / áno, ale niektoré vynechám / občas sa nejakej zúčastním / na prednášky nechodím vôbec	3 / 2 / 1 / 0
	sRodiciaPodpora	súhlasím / skôr súhlasím / skôr nesúhlasím / nesúhlasím	3 / 2 / 1 / 0
	sPrenasam	áno / nie	1 / 0
	sOpakujem	áno / nie / pravdepodobne budem	2 / 0 / 1
	sPredcasne	áno / nie / plánujem ho prerušiť	2 / 0 / 1
	sPredcasneKam	áno / nie / plánuje ho prerušiť	2 / 0 / 1
	sPomoc	áno / čiastočne / vôbec	2 / 1 / 0
	sVztahy	pozitívne / neutrálne / negatívne	2 / 1 / 0
	sSocZivot	áno / čiastočne / nie	2 / 1 / 0
	sNezvladam	často / občas / nikdy	2 / 1 / 0
	sPsycholog	áno / nie, ale premýšľal/a som o tom / nie, nepotreboval/a som to	2 / 1 / 0
	sNarocne	áno / čiastočne / vôbec	2 / 1 / 0
	sNaroky		
Spokojnosť	Premenné	Odpovede	Vyjadrenie premenných
	spProgramy	súhlasím / skôr súhlasím / skôr nesúhlasím / vôbec nesúhlasím	3 / 2 / 1 / 0
	spPrilezitosti		
	spSemester		
	spOdradza		
	spZbytocne		
	spZastarale		
	spStudProgram		
	spVybavenie		
	spOdbornost		
	spPristup		
	spHodnotenie		
	spTerminy		

	Premenné	Odpovede	Vyjadrenie premenných
Vzťah ku fakulte	fAkcie	áno občas nie, pretože nemám čas nie, pretože nemám záujem	3 / 2 / 1 / 0
	fPovest	súhlasím skôr súhlasím skôr nesúhlasím vôbec nesúhlasím	3 / 2 / 1 / 0
	fOdporucam	áno / nie	1 / 0
	fHrdy		
	fHodnotenie	pozitívne skôr pozitívne skôr negatívne negatívne	3 / 2 / 1 / 0
	flnaSkola	áno, aj plánujem študovať inde áno, ale nie som o tom presvedčený/á nie, chcem pokračovať v štúdiu na FRI nie, vôbec neplánujem ďalej študovať	3 / 2 / 1 / 0
	fErasmus	áno a bola to skvelá skúsenosť áno, ale už by som nešiel/nešla nie, nemal/a som o také niečo záujem nie, nemal/a som čas zúčastniť sa	3 / 2 / 1 / 0
	fVZahranici	áno, aj tam študovať plánujem / áno, ale nie som o tom presvedčený/á / plánujem študovať na Slovensku / nie, vôbec neplánujem ďalej študovať	3 / 2 / 1 / 0
	fDobraVolba	áno / nie	1 / 0
	fZnova	áno, presne na tento istý odbor / áno, ale na iný odbor / nie, zvolil / a by som štúdium na inej univerzite / nešiel / nešla by som študovať vôbec	3 / 2 / 1 / 0
Závislá premenná	flng	áno / áno, ale menim odbor / nie, plánujem študovať na inej škole / neplánujem pokračovať v štúdiu vôbec	3 / 2 / 1 / 0
	flng	áno / nie	* Logistická regresia 1 / 0

Príloha D: UML diagram



Príloha E: Obsah DVD

Priložené DVD obsahuje:

Práca v elektronickej podobe (Formát PDF)

R skripty

Zdrojové kódy aplikácie

Dáta z dotazníka