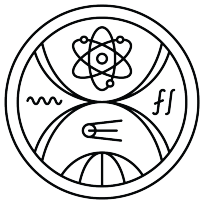
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY



ANALÝZA DÁT SKÚMAJÚCICH PORUCHY OSOBNOSTI KLINICKÝCH A NEKLINICKÝCH PACIENTOV

DIPLOMOVÁ PRÁCA

2026 Bc. Ema SVITEKOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ANALÝZA DÁT SKÚMAJÚCICH PORUCHY OSOBNOSTI KLINICKÝCH A NEKLINICKÝCH PACIENTOV

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	Matematika
Katedra:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Lívia Rosová, PhD.

Bratislava 2026 **Bc. Ema SVITEKOVÁ**





Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Ema Sviteková
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Analýza dát skúmajúcich poruchy osobnosti klinických a neklinických pacientov
Analysis of personality disorder data in clinical and non-clinical patients

Anotácia: Pri diagnostike porúch osobnosti sa využíva na psychiatrické poruchy pohľad aj cez dimenzie osobnosti, ktorý na rozdiel od kategorického prístupu umožňuje náhľad aj na celkové fungovanie osobnosti. K tomuto sa využívajú rôzne typy dotazníkov, ktoré vypĺňa či už pacient sám alebo s pomocou odborníka. Viacrozmerné štatistické metódy poskytujú nástroje na skúmanie výsledkov týchto meraní, hľadaní súvislostí a rozlišovaní medzi zdravou a klinickou populáciou.

Vedúci: Mgr. Lívia Rosová, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: doc. Mgr. Igor Melicherčík, PhD.

Dátum zadania: 09.01.2025

Dátum schválenia: 15.01.2025

prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Čestné vyhlásenie Čestne vyhlasujem, že celú diplomovú prácu na tému „Analýza dát skúmajúcich poruchy osobnosti klinických a neklinických pacientov“, vrátane všetkých jej príloh a obrázkov, som vypracovala samostatne, a to s použitím literatúry uvedenej v priloženom zozname a nástrojov umelej inteligencie. Vyhlasujem, že nástroje umelej inteligencie som použila v súlade s príslušnými právnymi predpismi, akademickými právami a slobodami, etickými a morálnymi zásadami za súčasného dodržania akademickej integrity a že ich použitie je v práci vhodným spôsobom označené.

Miesto a dátum: ..... Titul, meno a priezvisko: .....

Poďakovanie Chcela by som poďakovať svojej vedúcej diplomovej práce Mgr. Lívii Rosovej, PhD. za jej ochotu, pomoc, venovaný čas, cenné rady a podnetné pripomienky, ktoré mi pomohli pri písaní tejto práce. Ďakujem aj môjmu manželovi a rodine za ich každodennú podporu.

Abstrakt

SVITEKOVÁ, Ema: Analýza dát skúmajúcich poruchy osobnosti klinických a neklinických pacientov [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Lívia Rosová, PhD., Bratislava, 2026, 118 s.

Diplomová práca sa zameriava na analýzu výsledkov sebahodnotiacich dotazníkov prostredníctvom viacrozmerných štatistických metód s cieľom rozlíšiť klinickú a neklinickú populáciu v kontexte porúch osobnosti. Analýza je založená na dotazníkoch *Osobnostný dotazník pre DSM-5 (PID-5)*, zameraného na maladaptívne osobnostné črty, a *Sebahodnotiaca škála úrovne fungovania osobnosti (LPFS-SR)*, ktorý hodnotí globálnu mieru narušenia fungovania osobnosti. Dáta pochádzajú zo vzorky 421 klinických pacientov a 676 respondentov bez diagnostikovanej poruchy osobnosti, pozostávajúcej z dobrovoľníkov a študentov psychológie. Teoretická časť práce predstavuje vybrané štatistické metódy vhodné na analýzu viacrozmerných osobnostných dát, so zameraním na klasifikačné prístupy. Praktická časť zahŕňa porovnanie klinickej a neklinickej populácie, ako aj overenie validity a reliability použitých dotazníkov. V závere sú aplikované viaceré klasifikačné modely, ktorých cieľom je čo najlepšie rozlíšenie oboch populácií. Výsledky práce poukazujú na možnosti využitia viacrozmerných štatistických metód pri analýze sebahodnotiacich dotazníkov v klinickej praxi.

Kľúčové slová: poruchy osobnosti, PID-5, LPFS-SR, reliabilita, validita, viacrozmerné štatistické metódy

Abstract

SVITEKOVÁ, Ema: Analysis of personality disorder data in clinical and non-clinical patients [Diploma Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Lívia Rosová, PhD., Bratislava, 2026, 118 p.

This diploma thesis focuses on the analysis of self-report questionnaire data using multivariate statistical methods with the aim of distinguishing between clinical and non-clinical populations in the context of personality disorders. The analysis is based on the *Personality Inventory for DSM-5 (PID-5)*, which assesses maladaptive personality traits, and the *Level of Personality Functioning Scale – Self-Report (LPFS-SR)*, which evaluates the global level of impairment in personality functioning. The data were collected from a sample of 421 clinical patients and 676 respondents without a diagnosed personality disorder, consisting of volunteers and psychology students. The theoretical part of the thesis presents selected statistical methods suitable for the analysis of multidimensional personality data, with a particular focus on classification approaches. The empirical part includes a comparison of the clinical and non-clinical populations, as well as an examination of the validity and reliability of the applied questionnaires. In the final section, several classification models are applied with the aim of achieving the best possible discrimination between the two populations. The results highlight the potential usefulness of multivariate statistical methods for the analysis of self-report questionnaires in clinical practice.

Keywords: personality disorders, PID-5, LPFS-SR, reliability, validity, multivariate statistical methods

Obsah

Úvod **11**

1 Predstavenie dotazníkov **13**

1.1 PID-5 13

1.2 LPFS-SR 17

2 Metódy analýzy viacrozmerných dát **19**

2.1 Analýza dát 19

2.1.1 Shapirov-Wilkov test 19

2.1.2 Dvojvýberový t-test 20

2.1.3 Wilcoxonov test 20

2.1.4 Reliabilita dát 21

2.1.5 Validita dát 22

2.2 Štatistické metódy 24

2.2.1 Metóda hlavných komponentov 24

2.2.2 Metódy selekcie premenných 26

2.2.3 Klasifikačné metódy 29

2.2.4 Logistická regresia 31

2.2.5 Lineárna diskriminačná analýza 33

2.2.6 Klasifikačný strom 35

3 Analýza dát **38**

3.1 Analýza rozdielov medzi skupinami 38

3.1.1 Porovnanie podľa spôsobu zberu dát 38

3.1.2 Porovnanie podľa pohlavia 41

3.2 Deskriptívna analýza dát 45

3.2.1 PID-5_domény 46

3.2.2 PID-5_aspekty 47

3.2.3 LPFS 47

3.3 Reliabilita dát 49

3.3.1 PID-5_aspekty 49

3.3.2 PID-5_domény . . . 52
3.3.3 LPFS . . . 52
3.4 Validita dát . . . 54
3.4.1 PID-5_aspekty . . . 54
3.4.2 PID-5_domény . . . 57
3.4.3 LPFS . . . 60
3.4.4 Vzťahy medzi dotazníkmi . . . 63

4 Štatistické modely **66**
4.1 Metóda hlavných komponentov . . . 66
4.2 Metódy selekcie premenných . . . 69
4.2.1 Krokové metódy . . . 70
4.2.2 LASSO . . . 70
4.2.3 Finálny výber aspektov . . . 71
4.3 Logistická regresia . . . 72
4.4 Lineárna diskriminačná analýza . . . 79
4.5 Klasifikačný strom . . . 82
4.6 Porovnanie modelov . . . 87

Záver **90**

Príloha A **97**

Príloha B **100**

Príloha C **103**

Príloha D **105**

Príloha E **107**

Príloha F **109**

Príloha G **111**

Príloha H **112**

Príloha I **113**

Príloha J **114**

Príloha K **115**

Príloha L **117**

Úvod

Poruchy osobnosti predstavujú komplexné narušenie spôsobu, akým človek myslí, cíti a koná. Môžu sa rozvíjať postupne u každého z nás a nenápadne ovplyvňovať naše každodenné fungovanie. Práve preto je včasná a presná diagnostika mimoriadne dôležitá. V súčasnosti sa pri diagnostike porúch osobnosti využíva aj pohľad cez dimenziu osobnosti, ktorý poskytuje komplexnejší pohľad na fungovanie osobnosti a umožňuje identifikovať riziko vzniku poruchy ešte skôr, než sa plne prejaví. Takýto prístup si však vyžaduje použitie vhodne konštruovaných diagnostických nástrojov a spoľahlivo interpretovateľných výsledkov. Získanie takýchto výsledkov je možné aplikovaním viacrozmerných štatistických metód, ktoré majú významné uplatnenie aj v psychológii. Tieto metódy sa bežne využívajú pri tvorbe sebahodnotiacich nástrojov, pričom sa v tomto kontexte často uplatňuje napríklad exploračná faktorová analýza (EFA) [14].

Náhľad na celkové fungovanie osobnosti ešte pred klinickým vyšetrením poskytujú rôzne typy sebahodnotiacich dotazníkov. V práci analyzujeme dva typy takýchto dotazníkov. Prvým z nich je *Osobnostný dotazník pre DSM-5* (*Personality Inventory for DSM-5*, ďalej len *PID-5*), ktorý bol navrhnutý s cieľom zachytiť maladaptívne osobnostné črty [14]. Český preklad tohto dotazníka bol overený štúdiou [26], ktorá potvrdila jeho dobrú vnútornú konzistenciu a schopnosť rozlišovať medzi klinickou a neklinickou populáciou.

Druhým dotazníkom je *Sebahodnotiaca škála úrovne fungovania osobnosti* (*Levels of Personality Functioning Scale – Self-Report*, ďalej len *LPFS-SR*), ktorý predstavuje sebahodnotiaci nástroj zameraný na hodnotenie globálnej miery narušenia fungovania osobnosti [8]. Výskum [31] zameraný na český preklad dotazníka *LPFS-SR* ukázal, že aj tento dotazník dosahuje vysokú vnútornú konzistenciu a vykazuje dobrú diskriminačnú schopnosť medzi populáciami.

Práca je rozdelená do štyroch kapitol. V prvej kapitole predstavíme vybrané dotazníky používané pri diagnostike porúch osobnosti. Zároveň opíšeme spôsob výpočtu výsledných skóre, ktoré sú následne využívané v ďalších analýzach. Druhá kapitola je venovaná teoretickému pozadiu štatistických metód, ktoré budú v praktickej časti aplikované na získané dáta.

V tretej kapitole sa zameriame na analýzu dát. Prostredníctvom rôznych vizualizácií poskytneme hlbší náhľad do dát, pričom primárna pozornosť bude venovaná porovnaniu

rozdielov medzi klinickou a neklinickou populáciou. Okrem vizualizácie jednotlivých domén sa kapitola venuje aj analýze rozdielov medzi vybranými skupinami. Cieľom je zistiť, či sa údaje získané prostredníctvom online formulára štatisticky významne líšia od údajov získaných papierovou formou, a zároveň preskúmať rozdiely v skóre jednotlivých domén medzi pohlaviami. Taktiež overíme, či české verzie použitých dotazníkov predstavujú na základe našich dát spoľahlivý a validný nástroj na hodnotenie fungovania osobnosti, pričom preskúmame aj vzťahy medzi jednotlivými dotazníkmi.

V poslednej, štvrtej kapitole aplikujeme na získané dáta zvolené klasifikačné metódy, ktoré následne medzi sebou porovnáme. Vzhľadom na znalosť skutočnej klasifikácie respondentov sú v práci použité klasifikačné modely využívajúce tzv. učenie s učiteľom [7]. Ich aplikácia umožňuje skúmať vzťahy medzi osobnostnými črtami a zároveň predikovať, či respondent patrí do klinickej alebo neklinickej populácie. Okrem samotnej klasifikácie sa kapitola zameriava aj na identifikáciu aspektov a domén, ktoré najvýraznejšie prispievajú k správnej klasifikácii. Získané výsledky sú následne interpretované v kontexte klinickej praxe.

1 Predstavenie dotazníkov

Úvodná kapitola sa zameriava na predstavenie dvoch sebahodnotiacich dotazníkov používaných pri diagnostike psychických porúch: *PID-5* [15] a *LPFS-SR* [18]. Okrem samotnej štruktúry dotazníkov sa venuje aj spôsobu, akým sa z odpovedí respondentov získavajú jednotlivé skóre a ako sa na ich základe určujú osobnostné aspekty a domény.

1.1 PID-5

Osobnostný inventár *PID-5* je sebahodnotiaci dotazník určený na zachytenie maladaptívnych osobnostných charakteristík, ktoré sú považované za klinicky významné [14]. Dotazník meria 25 osobnostných aspektov, ktoré sa ďalej zoskupujú do piatich širších oblastí (tzv. domén). Obsahuje 220 položiek, ktoré sa hodnotia na štvorbodovej škále (0 = „nepravdivé“, 1 = „čiastočne nepravdivé“, 2 = „čiastočne pravdivé“, 3 = „pravdivé“). Respondent vypĺňa dotazník samostatne pri úvodnom rozhovore a pri každej položke volí odpoveď, ktorá ho najvýstižnejšie charakterizuje. Niektoré položky je potrebné kódovať opačne pred výpočtom skóre (t. j. $0 \to 3$, $1 \to 2$, $2 \to 1$, $3 \to 0$).

Každý z 25 aspektov pozostáva zo 4 až 14 položiek. Prvým krokom výpočtu skóre daného aspektu pre každého respondenta je reverzia príslušných položiek. Následne sa vypočíta **hrubé skóre** ako súčet všetkých položiek patriacich do daného aspektu. Ak respondent nezodpovedal viac ako 25 % položiek daného aspektu, skóre sa považuje za neplatné. Pri chýbajúcich odpovediach do 25 % sa použije tzv. **distribučné skórovanie**:

$$\text{Distribučné skóre} = \frac{(\text{čiastočné hrubé skóre}) \times (\text{počet položiek v aspekte})}{\text{počet zodpovedaných položiek}},$$

pričom **čiastočné hrubé skóre** predstavuje súčet odpovedí na zodpovedané položky.

Na záver sa vypočíta **priemerné skóre** ako podiel **hrubého skóre** a počtu položiek daného aspektu. **Priemerné skóre** sa pohybuje v intervale 0 až 3 a práve toto skóre je použité ako vstup do ďalších analýz.

Pri výpočte skóre piatich domén sa výsledné hodnoty získavajú spriemerovaním troch aspektov, ktoré danú oblasť tvoria. Každé priemerné skóre aspektov a domén možno interpretovať tak, že dlhodobo zvýšené skóre môže poukazovať na potenciálne problematické oblasti fungovania respondenta, ktoré si môžu vyžadovať ďalšie klinické posúdenie psychológa, terapiu a následnú starostlivosť. Na sledovanie zmien v závažnosti osobnostnej

dysfunkcie v čase sa odporúča realizovať opakované merania v pravidelných intervaloch. Podrobnejšie informácie o dotazníku *PID-5* sú uvedené v [15].

Nasledujúci zoznam obsahuje všetkých 25 osobnostných aspektov (ďalej v texte označovaných ako *PID-5_ aspekty*) a 5 domén (ďalej označovaných ako *PID-5_ domény*). Pri každom z nich spomenieme pôvodný anglický názov, preložený názov do slovenčiny a skratku, ktorá bude v rámci celej diplomovej práce primárne používaná. Zároveň sú pri jednotlivých aspektoch a doménach uvedené ich stručné charakteristiky podľa [1].

PID-5_aspekty

Anhedonia (*Anhedónia*), ďalej len **anhe.**, predstavuje zníženú schopnosť prežívať radosť a nedostatok záujmu alebo energie pre bežné aktivity.

Anxiousness (*Úzkosť*), ďalej len **anxi.**, je spojená s prežívaním nadmerného strachu, napätia a nervozity.

Attention Seeking (*Vyžadovanie pozornosti*), ďalej len **atte.**, je charakterizované silnou potrebou byť neustále v centre pozornosti a záujmu druhých.

Callousness (*Bezcitnosť*), ďalej len **call.**, sa prejavuje ignorovaním dopadu vlastného správania na druhých a nedostatkom záujmu o pocity či problémy iných.

Deceitfulness (*Nečestnosť*), ďalej len **dece.**, označuje tendenciu klamať, podvádzať, zveličovať alebo si vymýšľať udalosti v rámci komunikácie.

Depressivity (*Depresívnosť*), ďalej len **depr.**, je spojená s prežívaním dlhodobého smútku, pocitu beznádeje a pesimizmu ohľadom budúcnosti.

Distractibility (*Nepozornosť*), ďalej len **dist.**, sa vyznačuje ťažkosťami pri udržiavaní pozornosti a zvýšenou náchylnosťou na rozptyľovanie vonkajšími podnetmi.

Eccentricity (*Výstrednosť*), ďalej len **ecce.**, sa prejavuje nezvyčajným, zvláštnym alebo bizarným správaním, vzhľadom a rečou.

Emotional Lability (*Emočná labilita*), ďalej len **emot.**, sa vyznačuje výraznou labilitou emočného prežívania a častými zmenami nálad.

Grandiosity (*Povýšenosť*), ďalej len **gran.**, súvisí s prehnaným presvedčením o vlastnej nadradenosti voči iným.

Hostility (*Nepriateľstvo*), ďalej len **host.**, označuje tendenciu prežívať a prejavovať hnev a podráždenosť, sprevádzanú krutým až pomstychtivým správaním voči druhým.

Impulsivity (*Impulzivita*), ďalej len **impu.**, predstavuje tendenciu konať bez dôkladného zváženia dôsledkov vlastného správania.

Intimacy Avoidance (*Vyhýbanie sa intimite*), ďalej len **inti.**, je spojené s neochotou vytvárať blízke a emocionálne vzťahy.

Irresponsibility (*Nezodpovednosť*), ďalej len **irre.**, sa prejavuje ignorovaním záväzkov, sľubov a dôsledkov vlastného správania.

Manipulativeness (*Manipulatívnosť*), ďalej len **mani.**, označuje snahu dosahovať vlastné ciele prostredníctvom zvádzania, šarmu alebo podplácania.

Perceptual Dysregulation (*Dysregulácia myslenia a vnímania*), ďalej len **perc.**, je charakterizovaná nezvyčajnými alebo narušenými procesmi myslenia a vnímania.

Perseveration (*Vytrvalosť*), ďalej len **pers.**, predstavuje tendenciu zotrvať pri určitom spôsobe konania napriek opakovaným neúspechom.

Restricted Affectivity (*Obmedzená afektívnosť*), ďalej len **rest.**, sa vyznačuje oslabeným vyjadrovaním emócií v emocionálne intenzívnych situáciách.

Rigid Perfectionism (*Perfekcionizmus*), ďalej len **rigi.**, súvisí s nadmernou kontrolou vlastného výkonu a prehnanou potrebou dokonalosti.

Risk Taking (*Vyhľadávanie rizika*), ďalej len **risk.**, označuje záľubu v nebezpečných aktivitách bez primeraného zváženia ich dôsledkov.

Separation Insecurity (*Strach z opustenia*), ďalej len **sepa.**, sa prejavuje strachom z opustenia a nadmernou závislosťou na druhých osobách.

Submissiveness (*Podriadenosť*), ďalej len **subm.**, sa vyznačuje nadmernou ústupčivosťou a prispôsobovaním sa želaniam iných na úkor vlastných potrieb.

Suspiciousness (*Podozrievavosť*), ďalej len **susp.**, vyjadruje tendenciu k nedôvere voči ostatným osobám.

Unusual Beliefs And Experiences (*Nezvyčajné presvedčenia a zážitky*), ďalej len **unus.**, sa vyznačujú presvedčením o vlastných netradičných schopnostiach a neobvyklým vnímaním reality.

Social Withdrawal (*Sociálna izolácia*), ďalej len **soci.**, predstavuje tendenciu vyhýbať sa sociálnym interakciám a preferovať samotu.

PID-5_domény

Antagonism (*Antagonizmus*), ďalej len **Anta.**, súvisí s egocentrickým a bezohľadným prístupom k druhým, sprevádzaným očakávaním špeciálneho zaobchádzania, manipuláciou a nízkou mierou empatie. Zahŕňa aspekty **dece.**, **gran.** a **mani.**

Detachment (*Odstup*), ďalej len **Deta.**, odráža sociálnu izoláciu, emocionálnu uzavretosť a zníženú schopnosť prežívať radosť alebo vytvárať blízke medziľudské vzťahy. Medzi kľúčové aspekty patria **anhe.**, **inti.** a **soci.**

Disinhibition (*Dezinhibícia*), ďalej len **Disi.**, sa prejavuje impulzívnym správaním, ktoré je riadené aktuálnymi myšlienkami, pocitmi a vonkajšími podnetmi. Tvorí ju kombinácia aspektov **dist.**, **impu.** a **irre.**

Negative Affectivity (*Negatívna afektivita*), ďalej len **Nega.**, sa vyznačuje častým a intenzívnym prežívaním širokého spektra negatívnych emócií. Zahŕňa aspekty **anxi.**, **emot.** a **sepa.**

Psychoticism (*Psychoticizmus*), ďalej len **Psyc.**, odráža neštandardné, zvláštne alebo excentrické správanie a spôsoby vnímania reality. Tvoria ho aspekty **ecce.**, **perc.** a **unus.**

Ostatné premenné

Okrem samotných osobnostných aspektov a domén boli merané a zaznamenané aj údaje o tom, či bol dotazník získaný online alebo papierovou formou (označené ako **Zber**). Ďalej bolo pri každom respondentovi zaznamenané, či je klinickým pacientom (**Pati.**), jeho pohlavie (**gend.**) a vek (**age.**). Tieto údaje budú spolu so získanými skóre použité v ďalších častiach pri analýze dát, ako aj pri štatistických metódach.

V získaných dátach sa okrem týchto údajov a odpovedí na jednotlivé položky oboch dotazníkov nachádzajú aj dodatočné informácie o respondentoch, ako napríklad vzdelanie, rodinný stav, zamestnanie, zdravotný stav a ďalšie. Tieto údaje, z dôvodu malého počtu dát, do ďalších analýz nezahrnieme.

1.2 LPFS-SR

Dotazník *LPFS-SR* predstavuje sebaposudzovaciu metódu, ktorá hodnotí štyri navzájom súvisiace funkcie osobnosti: *Identita*, *Sebariadenie*, *Empatia* a *Intimita*. Pozostáva z 80 položiek a je vyplnený respondentom pred samotnou konzultáciou s klinickým pracovníkom.

Každá z uvedených štyroch domén pozostáva zo 16 až 23 položiek, ktoré sú hodnotené na štvorbodovej škále (1 = „nepravdivé“, 2 = „čiastočne pravdivé“, 3 = „prevažne pravdivé“, 4 = „pravdivé“). Položky nie sú reverzne kódované, avšak pri skórovaní sa používajú *váhy*, ktoré môžu nadobúdať kladné aj záporné hodnoty. Záporné váhy umožňujú zníženie výsledného skóre danej položky.

Po priradení hrubého skóre každej položke (1 až 4 podľa odpovede respondenta) sa vypočíta **vážené skóre** položky na základe príslušnej váhy. Položky sú následne rozdelené do štyroch domén, pričom pre každú z nich sa vypočíta súčet vážených skóre prislúchajúcich položiek. Celkové skóre sa určí ako súčet všetkých štyroch domén:

$$S_{\text{celkom}} = S_{\text{Identita}} + S_{\text{Sebariadenie}} + S_{\text{Empatia}} + S_{\text{Intimita}}.$$

Výsledné skóre umožňuje klinickému pracovníkovi posúdiť závažnosť dysfunkcie osobnosti pacienta. V prípade dotazníka *LPFS-SR* sa tieto hodnoty porovnávajú s normami stanovenými na neklinickej populácii tvorenej 306 respondentmi. Vyššie hodnoty poukazujú na väčšiu mieru narušenia osobnostného fungovania. Uvádza sa, že hodnoty približne **1 štandardná odchýlka nad priemerom** môžu naznačovať miernejšie ťažkosti, zatiaľ čo hodnoty **1,5 až 2 štandardné odchýlky nad priemerom** poukazujú na klinicky významné poruchy osobnostného fungovania, ktoré si vyžadujú ďalšie odborné posúdenie a prípadnú liečbu. Podrobnejšie informácie sú uvedené v [18].

Aj v prípade dotazníka *LPFS-SR* uvádzame zoznam domén (ďalej označovaných ako dataset *LPFS*) spolu s pôvodným anglickým názvom, príslušnou skratkou a stručnou charakteristikou podľa [1].

Empathy (*Empatia*), ďalej len **Empa.**, súvisí so schopnosťou vnímať emócie a potreby druhých, tolerovať odlišné názory a chápať vplyv vlastného správania na druhých.

Identity (*Identita*), ďalej len **Iden.**, sa vyznačuje schopnosťou vnímať vlastnú osobnosť, udržiavať realistické sebahodnotenie a prežívať rôzne emócie.

Intimacy (*Intimita*), ďalej len **Inti.**, opisuje schopnosť vytvárať a udržiavať blízke a stabilné medziľudské vzťahy, vrátane vzájomného rešpektu.

Self-direction (*Sebariadenie*), ďalej len **Self.**, reprezentuje schopnosť stanovovať si realistické ciele, udržiavať motiváciu a učiť sa z vlastných chýb.

Total (*Celkové skóre*), ďalej len **Tota.**, predstavuje index celkovej závažnosti poruchy fungovania osobnosti.

2 Metódy analýzy viacrozmerných dát

Kapitola *Metódy analýzy viacrozmerných dát* predstavuje teoretické pozadie štatistických metód vhodných na analýzu viacrozmerných dát získaných prostredníctvom sebahodnotiacich dotazníkov. V prvej časti kapitoly sú predstavené rôzne štatistické testy a postupy na overovanie reliability a validity. Následne sa druhá časť kapitoly venuje konkrétnym štatistickým metódam zameraným na redukciu dimenzionality dát a klasifikáciu respondentov.

2.1 Analýza dát

V tejto časti kapitoly sa zameriame na vysvetlenie testu normality (*Shapirov-Wilkov test*), testov slúžiacich na overenie rovnosti stredných hodnôt (*dvojvýberový t-test* a *Wilcoxonov test*) a metód overovania reliability dát (*Cronbachova alfa*) a validity dát (*konfirmačná faktorová analýza*).

2.1.1 Shapirov-Wilkov test

Shapirov-Wilkov test normality testuje nulovú hypotézu, že pozorované dáta pochádzajú z normálneho rozdelenia, proti alternatívnej hypotéze, že dáta z normálneho rozdelenia nepochádzajú.

Nech $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ označuje vektor usporiadaných náhodných pozorovaní z ľubovoľného spojitého rozdelenia. Ak sú tieto pozorovania z normálneho rozdelenia s neznámou strednou hodnotou μ a disperziou σ^2, potom Y_i možno vyjadriť ako

$$Y_i = \mu + \sigma X_i, \quad \text{pre } i = 1, 2, \ldots, n. \tag{1}$$

Nech $X_1 \leq X_2 \leq \cdots \leq X_n$ predstavuje usporiadaný náhodný výber zo štandardizovaného normálneho rozdelenia. Vektor $\mathbf{m} = (m_1, m_2, \ldots, m_n)^\top$ je tvorený strednými hodnotami usporiadaných štatistík a matica $\mathbf{V} = (v_{ij})$ typu $n \times n$ predstavuje ich kovariančnú maticu.

Testovú štatistiku Shapirovho-Wilkovho testu definujeme nasledovne

$$W = \frac{(\sum_{i=1}^{n} a_i Y_i)^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}, \tag{2}$$

kde vektor váh $\mathbf{a}$ je daný vzťahom

$$\mathbf{a} = \frac{\mathbf{m}^\top \mathbf{V}^{-1}}{\sqrt{\mathbf{m}^\top \mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m}}}. \tag{3}$$

Hodnota testovej štatistiky W nadobúda hodnoty z intervalu $[0, 1]$. Malé hodnoty testovej štatistiky vedú k zamietnutiu nulovej hypotézy o normalite rozdelenia, zatiaľ čo hodnota blízka 1 naznačuje dobrý súlad dát s normálnym rozdelením. Výhodou Shapirovho-Wilkovho testu normality je jeho aplikovateľnosť aj na malé súbory dát [25], [39].

2.1.2 Dvojvýberový t-test

Dvojvýberový t-test je založený na porovnávaní dvoch navzájom nezávislých výberov pochádzajúcich z normálneho rozdelenia s rovnakou disperziou, t. j. $X_1, \ldots, X_n \sim N(\mu_1, \sigma^2)$ a $Y_1, \ldots, Y_m \sim N(\mu_2, \sigma^2)$. Testujeme hypotézu

$$H_0 : \mu_1 - \mu_2 = \Delta \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 \neq \Delta. \tag{4}$$

Nech $n \geq 2$, $m \geq 2$ a $\sigma^2 > 0$. Označme $\bar{X}, S_X^2$ a $\bar{Y}, S_Y^2$ výberové charakteristiky daných výberov. Testová štatistika

$$T = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} \tag{5}$$

má za platnosti nulovej hypotézy Studentovo t-rozdelenie s $n + m - 2$ stupňami voľnosti.

Hypotézu H_0 zamietame na hladine významnosti α, ak platí $|T| \geq t_{n+m-2}(\frac{\alpha}{2})$. V prípade, že disperzie dvoch navzájom nezávislých výberov nie sú rovnaké, používa sa asymptotický t-test [2, s. 74-75].

2.1.3 Wilcoxonov test

Dvojvýberový Wilcoxonov test, známy aj ako *Mann-Whitneyho test*, patrí medzi neparametrické testy, ktoré nevyžadujú predpoklad o konkrétnom type rozdelenia dát (napr. normalite) a pracujú s poradím pozorovaní namiesto ich pôvodných hodnôt.

Uvažujme dva navzájom nezávislé náhodné výbery $X_1, X_2, \ldots, X_m$ a $Y_1, Y_2, \ldots, Y_n$ zo spojitých rozdelení s distribučnými funkciami F a G. Testujeme nulovú hypotézu, že oba výbery pochádzajú z rovnakého rozdelenia, t. j.

$$H_0 : F = G \quad \text{vs.} \quad H_1 : F \neq G. \tag{6}$$

Všetky pozorovania z oboch výberov sa najprv spoja. Následne sa všetkých $m + n$ pozorovaní $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ usporiada vzostupne a každému pozorovaniu sa priradí poradie *(angl. ranks)* od 1 do $m+n$. Sleduje sa rozdelenie týchto poradí medzi oba výbery. Ak platí nulová hypotéza o zhode rozdelení, očakávame, že poradia budú medzi výbermi rozdelené približne rovnomerne.

Nech R_i označuje poradie pozorovania X_i v spoločnom usporiadaní. Testová štatistika je definovaná ako

$$S = \sum_{i=1}^{m} R_i, \tag{7}$$

kde S predstavuje súčet poradí pre prvý výber $X_1, X_2, \ldots, X_m$. Veľmi malé alebo veľmi veľké hodnoty tejto štatistiky poukazujú na rozdiel medzi rozdeleniami porovnávaných výberov. Podrobnejšie informácie o dvojvýberovom Wilcoxonovom teste je možné nájsť v [2, s. 237].

2.1.4 Reliabilita dát

Reliabilita dát v psychologickom výskume vyjadruje mieru presnosti a konzistentnosti meracieho nástroja. Ide o rozsah, v akom nástroj poskytuje stabilné výsledky pri opakovaných meraniach za rovnakých podmienok [12].

Vnútorná konzistencia opisuje mieru, do akej jednotlivé položky merajú ten istý konštrukt, resp. dimenziu. Najčastejšie používaným ukazovateľom na hodnotenie vnútornej konzistencie a spoľahlivosti meracích nástrojov je Cronbachova alfa, ktorá vyjadruje mieru vzájomnej súvislosti položiek v rámci meracieho nástroja. V psychológii sa využíva najmä na overovanie, či testy a dotazníky konzistentne merajú psychologické dimenzie, ako sú napríklad osobnostné črty.

Cronbachova alfa sa vypočíta podľa vzťahu

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_x^2}\right), \tag{8}$$

kde k predstavuje počet položiek v škále, $\sum_{i=1}^{k} \sigma_i^2$ označuje súčet rozptylov jednotlivých položiek a σ_x^2 celkový rozptyl škály [4].

Cronbachova alfa môže nadobúdať hodnoty približne od 0 do 1, v niektorých prípadoch aj záporné hodnoty. Čím bližšie je jej hodnota k 1, tým vyššia je vnútorná konzistencia,

čo naznačuje, že položky merajú rovnaký konštrukt. Naopak, nízka hodnota alfy naznačuje, že niektoré položky nemusia merať rovnaký konštrukt. Takáto hodnota môže byť spôsobená malým počtom položiek alebo slabou vzájomnou koreláciou medzi nimi. Za prijateľné hodnoty Cronbachovej alfy sa považujú hodnoty od 0,70 a vyššie. Viac informácií o Cronbachovej alfe je uvedených v [35].

2.1.5 Validita dát

Validita dát predstavuje mieru, do akej merací nástroj skutočne meria to, čo má merať. Inými slovami, posudzuje, či nástroj reálne zachytáva konštrukt, ktorý máme v úmysle pomocou neho merať. Zároveň vyjadruje mieru, do akej sú interpretácie a použitie výsledkov získaných týmto nástrojom oprávnené [13].

Jedným z typov validity je tzv. konštruktová validita *(angl. construct validity)*. Často využívaným nástrojom na hodnotenie konštruktovej validity je práve konfirmačná faktorová analýza (CFA).

CFA vychádza z vopred špecifikovaného modelu, ktorý určuje štruktúru jednotlivých faktorov. Vďaka tomu je možné explicitne testovať hypotézy týkajúce sa faktorovej štruktúry meracieho nástroja. Úlohou CFA je po špecifikovaní faktorov zosúladiť pozorovanú a teoretickú faktorovú štruktúru dát s cieľom určiť mieru zhody medzi predpokladaným faktorovým modelom a empirickými údajmi.

Presná špecifikácia modelu spočíva v určení počtu faktorov a v priradení jednotlivých položiek ku konkrétnym faktorom. Každá položka by mala byť spravidla priradená k jednému faktoru, pričom každý faktor má byť reprezentovaný aspoň jednou položkou. Model zároveň vopred definuje mieru vzťahov medzi premennými, ako aj korelácie medzi jednotlivými faktormi.

Na posúdenie toho, ako dobre model zodpovedá dátam, sa využívajú štatistiky označované ako indexy zhody *(angl. fit indices)*.

χ^2 test

χ^2 test testuje nulovú hypotézu, podľa ktorej sa očakávaná kovariančná matica štatisticky významne nelíši od pozorovanej kovariančnej matice. Nižšie hodnoty testovej štatistiky poukazujú na lepšiu zhodu modelu s dátami, zatiaľ čo vyššie hodnoty v kombinácii s nízkou p-hodnotou naznačujú, že model neposkytuje adekvátnu zhodu. Tento index je

veľmi citlivý na veľkosť výberu. Pri veľkých vzorkách často vedie k zamietnutiu nulovej hypotézy, pričom nie je možné jednoznačne určiť, či je toto zamietnutie spôsobené zlým prispôsobením modelu alebo veľkosťou výberu.

RMSEA (Root Mean Square Error of Approximation)

RMSEA predstavuje strednú kvadratickú chybu aproximácie. Hodnotí, ako dobre model aproximuje populačnú kovariančnú maticu. Patrí medzi najinformatívnejšie ukazovatele zhody, najmä vďaka tomu, že penalizuje modely s väčším počtom parametrov. Zároveň sa snaží korigovať tendenciu χ^2 testu zamietať modely pri veľkých vzorkách.

Hodnota indexu sa považuje za veľmi dobrú, ak je menšia alebo rovná 0,05, za prijateľnú, ak je menšia ako 0,08, a za nevyhovujúcu pri vyšších hodnotách. Spôsob výpočtu *RMSEA* je uvedený v [30].

SRMR (Standardized Root Mean Square Residual)

Index *SRMR* je definovaný ako druhá odmocnina z priemeru štvorcov štandardizovaných rozdielov medzi pozorovanými a modelom implikovanými koreláciami. Hodnota indexu do 0,08 sa považuje za akceptovateľnú, pričom hodnoty pod 0,05 predstavujú veľmi dobrú zhodu modelu.

CFI (Comparative Fit Index)

Index *CFI* patrí medzi tzv. komparatívne indexy zhody, ktoré porovnávajú zhodu navrhovaného modelu so zhodou tzv. nulového modelu, teda modelu, ktorý predpokladá nekorelovanosť všetkých meraných premenných. Patrí medzi najčastejšie reportované indexy zhody, keďže je jedným z ukazovateľov najmenej ovplyvnených veľkosťou výberu.

Hodnoty *CFI* sa pohybujú v intervale od 0 do 1, pričom hodnoty vyššie ako 0,95 poukazujú na veľmi dobrú zhodu modelu s dátami, resp. na výrazné zlepšenie oproti nulovému modelu. Hodnoty nižšie ako 0,80 sa považujú za nevyhovujúce. Spôsob výpočtu tohto indexu je uvedený v [30].

PNFI (Parsimonious Normed Fit Index)

PNFI patrí medzi indexy zhody, ktoré hodnotia kvalitu modelu z hľadiska jeho jednoduchosti a zároveň penalizujú jeho nadmernú zložitosť. Hoci pre tento index neboli odporúčané žiadne prahové hodnoty, v praxi sa jeho hodnoty často pohybujú približne okolo 0,50. Vo všeobecnosti platí, že vyššie hodnoty indikujú lepšiu zhodu modelu pri zohľadnení jeho jednoduchosti.

Ako uvádza [9], nie je nevyhnutné reportovať všetky dostupné indexy zhody. Odporúča sa uvádzať kombináciu vybraných indexov (napr. χ^2 test, *RMSEA*, *SRMR*, *CFI* a *PNFI*), ktoré zachytávajú rôzne aspekty zhody modelu a sú relatívne menej citlivé na veľkosť výberu a odhady parametrov, pričom χ^2 test je potrebné interpretovať s ohľadom na jeho citlivosť na veľkosť výberu. Ak CFA nepreukáže dostatočne dobrú zhodu medzi teoretickou a pozorovanou faktorovou štruktúrou, odporúča sa preskúmať možnosti zlepšenia prispôsobenia modelu.

Jedným z možných prístupov je individuálne posúdenie faktorov a ich položiek s cieľom identifikovať položky s nízkou mierou vysvetlenej variability. Koeficient determinácie (R^2) vyjadruje mieru vzťahu medzi konkrétnou položkou a faktorom, pričom zároveň určuje, aká časť variability pozorovanej položky je vysvetlená príslušným faktorom. Vyššie hodnoty R^2 naznačujú silnejší vzťah medzi faktorom a pozorovanou položkou, ako aj lepšiu zhodu modelu s empirickými dátami. Naopak, položky s hodnotou R^2 nižšou ako 0,20 sa často považujú za potenciálne problematické, keďže poukazujú na vysoký podiel nevysvetlenej variability. Podrobnejšie informácie, ako aj informácie o iných indexoch zhody, sa nachádzajú v [9], [30] a [34].

2.2 Štatistické metódy

Časť *Štatistické metódy* opisuje metódu extrakcie premenných (*PCA*), metódy selekcie premenných (*spätná selekcia* a *LASSO*) a vybrané klasifikačné metódy, konkrétne *logistickú regresiu*, *lineárnu diskriminačnú analýzu* a *klasifikačný strom*.

2.2.1 Metóda hlavných komponentov

Metóda hlavných komponentov *(angl. Principal Component Analysis, PCA)* je metóda extrakcie premenných, ktorá umožňuje reprezentovať rozsiahly súbor navzájom korelovaných premenných pomocou menšieho počtu reprezentatívnych, navzájom nekorelovaných premenných. Tieto nové premenné, nazývané hlavné komponenty, spoločne vysvetľujú väčšinu variancie pôvodného dátového súboru. Hlavným cieľom metódy PCA je nájsť také lineárne kombinácie pôvodných premenných, ktoré zachytia čo najviac informácie obsiahnutej v pôvodných dátach. Tento cieľ zodpovedá redukcii dimenzionality dát pri zachovaní maximálnej možnej variability [10, s. 499], [29].

Predpokladajme, že máme n pozorovaní vektora náhodných premenných $\boldsymbol{X}$ pozostávajúceho z p premenných. Cieľom je znížiť dimenzionalitu dát z p na q, kde $q < p$. Metóda PCA dosahuje tento cieľ nájdením lineárnych kombinácií $a_1^T\boldsymbol{X}, a_2^T\boldsymbol{X}, \ldots, a_q^T\boldsymbol{X}$, ktoré sa nazývajú hlavné komponenty [11].

Hlavné komponenty sa získavajú riešením optimalizačného problému

$$\begin{aligned} \max_{a} \quad & a^T S a \\ & \|a\| = 1, \end{aligned}$$

kde S je výberová kovariančná matica vektora $\boldsymbol{X}$. Riešením tohto optimalizačného problému sú vlastné vektory kovariančnej matice S, označené $a_1, \ldots, a_q$, ktoré zodpovedajú q najväčším vlastným číslam tejto matice [28].

Prvý hlavný komponent $a_1^T\boldsymbol{X}$, ktorý zodpovedá najväčšiemu vlastnému číslu kovariančnej matice, vysvetľuje najväčší možný rozptyl dát. Druhý hlavný komponent $a_2^T\boldsymbol{X}$ zodpovedá druhému najväčšiemu rozptylu a je definovaný ako komponent, ktorý maximalizuje rozptyl dát za podmienky, že je nekorelovaný s prvým hlavným komponentom. Analogickým spôsobom sa konštruujú aj ďalšie hlavné komponenty. Takto získané hlavné komponenty sú navzájom ortogonálne a zoradené podľa klesajúcej hodnoty vlastných čísel, čo zodpovedá klesajúcej vysvetlenej variancii. Nová q-rozmerná reprezentácia dát je teda tvorená prvými q hlavnými komponentmi, ktoré zachytávajú najvýznamnejšie smery variability. Z geometrického hľadiska predstavujú hlavné komponenty novú súradnicovú sústavu získanú rotáciou pôvodného systému, pričom nové osi sú orientované v smeroch maximálnej variability dát [10, s. 501], [29].

Jednou zo zásadných otázok metódy PCA je určenie počtu hlavných komponentov, ktoré sa ponechajú na ďalšiu analýzu. Cieľom výberu počtu komponentov je nájsť vhodný kompromis medzi jednoduchosťou modelu a jeho schopnosťou dostatočne reprezentovať pôvodné dáta.

Najjednoduchším prístupom je stanovenie prahovej hodnoty vysvetlenej variancie (napr. 80 %) a výber najmenšieho počtu komponentov, ktoré túto hodnotu dosiahnu. Vysvetlená časť variancie prvými q komponentmi je definovaná ako pomer súčtu prvých q vlastných čísel k súčtu všetkých vlastných čísel. Ďalším často používaným kritériom je tzv. *Kaiserovo pravidlo*, podľa ktorého sa ponechávajú tie hlavné komponenty, ktorých vysvetlená

časť variancie je väčšia ako priemerná vysvetlená časť variancie všetkých komponentov [28].

Základným využitím PCA je redukcia dimenzionality dát, vďaka čomu sa hlavné komponenty často používajú ako medzikrok v rozsiahlejších analýzach, napríklad v regresnej analýze ako náhrada pôvodných premenných s cieľom zmierniť problémy multikolinearity. Metóda PCA zároveň umožňuje vizualizáciu viacrozmerných dát v dvojrozmernom priestore pomocou biplotov, identifikáciu odľahlých pozorovaní a zhlukov a zjednodušenie diskriminačnej analýzy pri vysokom počte korelovaných premenných [11].

2.2.2 Metódy selekcie premenných

Existuje viacero metód výberu premenných. V tejto práci sa budeme zaoberať dvoma odlišnými typmi metód: spätnou selekciou, ktorá patrí do triedy tzv. krokových metód, a metódou LASSO, ktorá patrí do triedy tzv. penalizovaných regresií.

Metódy selekcie premenných predstavujú postupy, pri ktorých sa z množiny vysvetľujúcich premenných vyberá taká ich podmnožina, ktorá má byť zahrnutá do regresného modelu. Pri hľadaní optimálnej podmnožiny vysvetľujúcich premenných je potrebné nájsť rovnováhu medzi vychýlenosťou a rozptýlenosťou modelu, tzv. *bias-variance trade-off*. Na jednej strane je cieľom zabezpečiť, aby regresný model čo najlepšie zachytával skutočné vzťahy v dátach, čo si zvyčajne vyžaduje väčší počet regresorov. V takomto prípade so zvyšujúcim sa počtom premenných zahrnutých do modelu vychýlenosť *(angl. bias)* klesá [28].

Na druhej strane, ako uvádza [20], je žiaduce zahrnúť do modelu čo najmenší počet premenných, keďže každá irelevantná premenná znižuje presnosť odhadov regresných koeficientov aj predikovaných hodnôt. S rastúcim počtom vysvetľujúcich premenných v modeli tak narastá jeho rozptýlenosť *(angl. variance)*. Zároveň je potrebné dbať na vyváženosť modelu z hľadiska počtu dát, keďže ak je počet pozorovaní výrazne menší než počet vysvetľujúcich premenných, môže dôjsť k pretrénovaniu modelu.

Cieľom metód výberu premenných je teda dosiahnuť rovnováhu medzi jednoduchosťou modelu a kvalitou jeho prispôsobenia dátam.

Spätná selekcia

Spätná selekcia *(angl. backward stepwise selection)* patrí medzi krokové metódy výberu

premenných. Podľa [10, s. 229] krokové metódy predstavujú kompromis medzi úplným prehľadávaním priestoru všetkých možných modelov a výpočtovou efektivitou, keďže v každom kroku zvažujú iba obmedzený počet kandidátnych modelov.

Metóda spätnej selekcie je založená na tom, že proces výberu optimálnej podmnožiny premenných začína plným modelom, ktorý obsahuje všetky dostupné vysvetľujúce premenné. V každom ďalšom kroku sa z modelu odstráni najmenej prínosná premenná. V danom kroku sa porovnávajú všetky modely vzniknuté odstránením jednej premennej, pričom sa vyberie model s najlepšou hodnotou zvoleného kritéria. Algoritmus sa opakuje dovtedy, kým už nie je možné odstrániť žiadnu premennú tak, aby došlo k zlepšeniu hodnoty zvoleného kritéria.

Pri aplikácii spätnej selekcie vzniká množina modelov, z ktorých každý obsahuje inú podmnožinu vysvetľujúcich premenných. Preto je potrebné tieto modely medzi sebou porovnávať. Reziduálny súčet štvorcov, ktorý sa so zvyšujúcim počtom premenných monotónne zlepšuje, nie je vhodným kritériom na výber optimálneho modelu. Z tohto dôvodu sa pri selekcii premenných používajú informačné kritériá, pričom za optimálny model sa považuje ten, ktorý dosahuje minimálnu hodnotu zvoleného kritéria.

Výhodou spätnej selekcie je jej výpočtová efektivita, vďaka ktorej je použiteľná aj pri veľkom počte premenných. Naopak, nevýhodou je požiadavka, aby počet pozorovaní bol väčší než počet premenných, keďže je potrebné odhadnúť plný počiatočný model [10, s. 231], [28].

Ako je uvedené v [20], informačné kritériá penalizujú počet vysvetľujúcich premenných s cieľom dosiahnuť čo najlepšiu kvalitu prispôsobenia modelu pri čo najmenšej zložitosti. Najčastejšie používanými kritériami sú *AIC* a *BIC*, ktoré sa líšia najmä mierou penalizácie zložitosti modelu.

Akaikeho informačné kritérium (AIC) predstavuje odhad testovacej chyby modelu s fixnou penalizáciou rovnou 2. Je definované vzťahom

$$\mathrm{AIC} = -2\log(L) + 2q, \tag{9}$$

kde L je maximálna hodnota funkcie vierohodnosti a q označuje počet parametrov modelu.

Bayesovské informačné kritérium (BIC) má podobnú štruktúru, avšak penalizácia závisí od logaritmu veľkosti vzorky $\ln(n)$. V dôsledku prísnejšej penalizácie zložitosti

modelu vedie v praxi k výberu jednoduchších modelov. Je definované vzťahom

$$\mathrm{BIC} = -2\log(L) + q\ln(n). \tag{10}$$

LASSO

Ďalšou metódou selekcie premenných je metóda LASSO *(angl. Least Absolute Shrinkage and Selection Operator)*, ktorá patrí medzi penalizované regresné metódy. Na rozdiel od krokových metód, ktoré analyzujú iba vybrané podmnožiny premenných, penalizované regresie poskytujú odhad modelu zahŕňajúci všetky premenné prostredníctvom zavedenia obmedzenia, resp. regularizácie odhadov regresných koeficientov, čo je ekvivalentné ich zmenšovaniu smerom k nule.

Metóda LASSO využíva na penalizáciu veľkosti koeficientov l_1 normu, definovanú ako súčet absolútnych hodnôt regresných koeficientov. Jej hlavnou výhodou je, že pri dostatočne veľkej hodnote penalizačného parametra λ sú niektoré koeficienty presne rovné nule, čo vedie k automatickej selekcii premenných.

Odhady regresných koeficientov sa získavajú minimalizáciou nasledujúcej funkcie

$$\min_{\beta} \quad \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|, \tag{11}$$

ktorá pozostáva zo súčtu reziduálneho súčtu štvorcov (RSS) a penalizačného člena. Parameter $\lambda \in (0, \infty)$ určuje mieru regularizácie [10, s. 241], [24].

Jednou z kľúčových úloh metódy LASSO je voľba optimálnej hodnoty parametra λ. Na jeho určenie je možné použiť informačné kritériá alebo k-násobnú krížovú validáciu. V praxi sa častejšie využíva krížová validácia, keďže výpočet stupňov voľnosti pre penalizované modely je zložitejší [20].

K-násobná krížová validácia

Pri k-násobnej krížovej validácii sa dáta rozdelia na k približne rovnako veľkých častí. V každej iterácii sa $k-1$ častí použije na tréning modelu a zvyšná časť na jeho validáciu. Pre validačnú vzorku sa vypočíta chyba odhadu

$$\mathrm{MSE}_j = \frac{1}{n_j}\left|\left|Y_j - X_j^T\hat{\beta}\right|\right|_2^2, \tag{12}$$

pričom optimálna hodnota λ je tá, pre ktorú je hodnota MSE minimálna ($\lambda_{\min}$). Alternatívne sa používa hodnota λ_{1se}, ktorá predstavuje najväčšiu hodnotu λ, pri ktorej je chyba modelu v intervale jednej smerodajnej odchýlky od minimálnej hodnoty MSE [20], [24].

2.2.3 Klasifikačné metódy

Ak vysvetľovaná premenná y nie je kvantitatívna, ale kvalitatívna, proces predikcie kvalitatívnej odozvy pre jednotlivé pozorovania sa nazýva klasifikácia. V prípade, že rozlišujeme iba dve kategórie, hovoríme o *binárnej klasifikácii*, zatiaľ čo klasifikácia s viac ako dvoma kategóriami sa označuje ako *multikategorická klasifikácia*. Základným prvkom klasifikácie je konštrukcia klasifikátora, ktorého úlohou je predikovať triedu objektu na základe jeho vektora príznakov (vysvetľujúcich premenných) [10, s. 129].

Klasifikačné metódy sa snažia napodobniť kategorizáciu poskytnutú „učiteľom“, aby bolo možné automaticky klasifikovať nové objekty aj bez jeho ďalšej pomoci. Z tohto dôvodu sa tieto metódy zaraďujú medzi učenie s učiteľom *(angl. supervised learning)*. Hlavným cieľom klasifikácie je teda zostrojiť taký klasifikátor, ktorý dokáže spoľahlivo priraďovať triedy aj novým, doposiaľ nepozorovaným objektom, a to výlučne na základe ich známych vektorov príznakov [7].

Konštrukcia klasifikátora $\mathcal{C}$ je založená na trénovacej množine dát pozostávajúcej z n objektov opísaných vektormi príznakov $x_1, \ldots, x_n \in \mathbb{R}^p$ spolu s ich známymi správnymi klasifikáciami $c_1, \ldots, c_n \in \{1, \ldots, m\}$. Ako uvádza [10, s. 129], klasifikačné metódy často najskôr odhadujú pravdepodobnosti príslušnosti objektu k jednotlivým triedam, ktoré následne slúžia ako základ pre samotné rozhodnutie o klasifikácii.

Okrem samotnej konštrukcie klasifikátora je nevyhnutnou súčasťou klasifikácie aj odhad jeho spoľahlivosti. Tá sa často vyjadruje pomocou konfúznej matice *(angl. confusion matrix)*, ktorá sumarizuje správne a chybné klasifikácie. Prvok matice C_{ij} predstavuje pravdepodobnosť, že objekt patriaci do triedy i je klasifikovaný do triedy j. Platí

$$C_{ij} = q_i P(j \mid i) \quad \text{pre všetky } i, j \in \{1, \ldots, m\}, \tag{13}$$

kde $P(j \mid i)$ je podmienená pravdepodobnosť, že objekt z triedy i bude klasifikovaný do triedy j a q_i predstavuje apriórnu pravdepodobnosť, že náhodne zvolené pozorovanie pochádza z triedy i [7].

V prípade binárnej klasifikácie má konfúzna matica tvar

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix},$$

kde

- C_{11} predstavuje správne pozitívne výsledky *(angl. True Positives)*,
- C_{12} predstavuje falošne negatívne výsledky *(angl. False Negatives)*,
- C_{21} predstavuje falošne pozitívne výsledky *(angl. False Positives)*,
- C_{22} predstavuje správne negatívne výsledky *(angl. True Negatives)*.

Jednou zo základných metrík hodnotenia spoľahlivosti klasifikátora je presnosť *(angl. accuracy)*, definovaná ako

$$\frac{C_{11} + C_{22}}{C_{11} + C_{12} + C_{21} + C_{22}}.$$

Ďalšími často používanými metrikami sú *senzitivita* a *špecificita*. Senzitivita (schematicky $P(+|+)$) vyjadruje schopnosť klasifikátora správne identifikovať pozitívne prípady:

$$\frac{C_{11}}{C_{11} + C_{12}}.$$

Špecificita (schematicky $P(-|-)$) vyjadruje schopnosť správne identifikovať negatívne prípady:

$$\frac{C_{22}}{C_{21} + C_{22}}.$$

Vykreslením bodov $(1 - \text{špecificita}; \text{senzitivita})$ získame ROC krivku. Plocha pod touto krivkou, označovaná ako AUC, vyjadruje schopnosť klasifikátora rozlišovať medzi jednotlivými triedami. Čím je hodnota AUC bližšie k 1, tým lepší je klasifikátor. Ako uvádza [16], hodnoty AUC nižšie ako 0,70 sa považujú za slabú zhodu modelu s pozorovanými dátami, hodnoty v intervale 0,70–0,79 za prijateľnú zhodu a hodnoty nad 0,80 za výbornú zhodu.

Cieľom konštrukcie klasifikátora je dosiahnuť, aby spoľahlivo klasifikoval nielen dáta použité na jeho učenie, ale aj nové, doposiaľ nepozorované dáta. Z tohto dôvodu sa dostupné dáta často rozdeľujú na trénovaciu a testovaciu množinu, typicky v pomere 80:20 alebo 70:30. Model sa natrénuje na trénovacej množine a jeho kvalita sa následne vyhodnotí na testovacej množine [28].

Jednou z najčastejšie používaných metód odhadu spoľahlivosti klasifikátora je krížová validácia *(angl. cross-validation)*. Najrozšírenejšou variantou je k-násobná krížová validácia, pri ktorej sa trénovacia množina náhodne rozdelí na k podmnožín. V každej iterácii sa jedna z nich použije ako validačná množina, zatiaľ čo zvyšné podmnožiny slúžia na

trénovanie klasifikátora. Na validačných dátach sa následne vyhodnotia metriky spoľahlivosti klasifikácie. Konečný odhad jednotlivých metrík sa získa spriemerovaním výsledkov zo všetkých iterácií. Ako uvádza [7], táto metóda predstavuje vhodný kompromis medzi presnosťou odhadu a výpočtovou náročnosťou.

2.2.4 Logistická regresia

V aplikačných oblastiach sa často stretávame s premennými typu výskyt javu (áno / nie), úspech (úspech / neúspech) alebo zdravotný stav (chorý / zdravý). Jednou z klasifikačných metód, ktorá umožňuje modelovať vzťah medzi binárnou závislou premennou a jednou alebo viacerými vysvetľujúcimi premennými, je logistická regresia [19].

Základným rozdielom medzi logistickou a lineárnou regresiou je charakter závislej premennej. V logistickej regresii má závislá premenná binomické rozdelenie, pričom platí $Y \sim \text{Bin}(1, p_x)$, kde p_x predstavuje pravdepodobnosť výskytu sledovaného javu. Logistická regresia modeluje logaritmus šancí výskytu javu, ktorý je definovaný vzťahom

$$\log\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \tag{14}$$

kde x_i sú vysvetľujúce premenné a β_i regresné koeficienty. Odhady parametrov logistickej regresie sa získavajú na základe dostupných dát, najčastejšie pomocou metódy maximálnej vierohodnosti [10, s. 137].

Štatistická významnosť jednotlivých regresných koeficientov sa často testuje pomocou Waldovho testu, ktorý overuje nulovú hypotézu $H_0 : \beta_i = 0$. Testová štatistika má tvar

$$Z = \frac{\widehat{\beta}_i}{se(\widehat{\beta}_i)}, \tag{15}$$

pričom za platnosti nulovej hypotézy má asymptoticky normálne rozdelenie $N(0, 1)$ [17].

Keďže pravdepodobnosť je ohraničená intervalom $(0, 1)$, logistická regresia využíva spojovaciu funkciu, ktorá tento interval mapuje na celý reálny obor. Touto spojovacou funkciou je *logit* funkcia definovaná ako

$$\text{Logit}(p_x) = \log\left(\frac{p_x}{1-p_x}\right). \tag{16}$$

Použitie logitovej transformácie umožňuje modelovať vzťah medzi závislou premennou a vysvetľujúcimi premennými pomocou lineárnej kombinácie parametrov [38].

Samotnú pravdepodobnosť výskytu javu možno ekvivalentne vyjadriť v tvare

$$p_x = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}. \tag{17}$$

Táto transformácia zabezpečuje, že modelované pravdepodobnosti pre všetky hodnoty vysvetľujúcich premenných vždy ležia v intervale $(0, 1)$ [10, s. 137].

Šanca *(angl. odds)* predstavuje pomer pravdepodobnosti výskytu javu k pravdepodobnosti jeho nevýskytu, teda výraz $\frac{p_x}{1-p_x}$ v modeli (14). Na rozdiel od pravdepodobnosti, ktorá je ohraničená intervalom $(0, 1)$, šanca nadobúda hodnoty z intervalu $(0, \infty)$ [38].

Dôležitým aspektom logistickej regresie je interpretácia regresných koeficientov. Intercept β_0 reprezentuje logaritmus šancí výskytu udalosti v referenčnej skupine, teda u jedincov, ktorí majú referenčné hodnoty všetkých vysvetľujúcich premenných. Regresné koeficienty β_i vyjadrujú zmenu logaritmu šancí udalosti pri jednotkovej zmene príslušnej vysvetľujúcej premennej [16].

Po exponenciálnej transformácii koeficientov získavame pomer šancí *(angl. odds ratio)*. Hodnota e^{β_0} predstavuje šancu výskytu udalosti v referenčnej skupine, zatiaľ čo e^{β_i} udáva, koľkonásobne sa šanca výskytu udalosti zmení pri jednotkovom náraste danej premennej v porovnaní s referenčnou skupinou. Hodnota väčšia ako 1 znamená nárast šance udalosti, zatiaľ čo hodnota menšia ako 1 indikuje jej pokles. Ako však upozorňuje [19], pomer šancí nevyjadruje priamo zmenu pravdepodobnosti, a preto vysoký pomer šancí nemusí nevyhnutne znamenať vysokú pravdepodobnosť výskytu udalosti.

Hlavnou výhodou logistickej regresie je jej dobrá interpretovateľnosť prostredníctvom pomerov šancí a schopnosť súčasne kontrolovať vplyv viacerých vysvetľujúcich premenných. Vďaka týmto vlastnostiam patrí medzi najčastejšie používané viacrozmerné modely na analýzu binárnych výsledkov v medicínskom výskume [38].

Jedným z najvýznamnejších problémov logistickej regresie je výber vhodných vysvetľujúcich premenných do modelu. Zaradenie príliš veľkého počtu premenných môže viesť k falošne významným výsledkom a k zníženiu štatistickej sily modelu. V literatúre sa často odporúča mať aspoň 10 pozorovaní na jednu premennú, z čoho vyplýva, že pri menších datasetoch je vhodné vykonať selekciu premenných ešte pred samotnou aplikáciou modelu [33].

Následne je možné porovnávať pôvodný model so zjednodušeným submodelom pomo-

cou testu o submodeli, ktorý testuje hypotézy

$$H_0: \text{ submodel postačuje} \quad \text{vs.} \quad H_1: \text{ submodel nepostačuje.} \tag{18}$$

Testová štatistika je založená na rozdiele medzi mierou nekvality submodelu a pôvodného modelu:

$$-2\ln\left(\frac{L_{\text{submodel}}}{L_{\text{plný}}}\right) = \text{DEVIANCE}_{\text{submodel}} - \text{DEVIANCE}_{\text{plný}} \sim \chi^2_{k-s}, \tag{19}$$

kde L predstavuje maximum funkcie vierohodnosti, k je počet parametrov v plnom modeli a s počet parametrov v submodeli.

Ďalším kľúčovým problémom je pretrénovanie modelu *(angl. overfitting)*, teda situácia, keď sa model príliš prispôsobí trénovacím dátam a následne zlyháva pri predikcii na nových dátach. Z tohto dôvodu sa odporúča model validovať [33].

Bootstrap

Metóda bootstrap predstavuje silný a univerzálny štatistický nástroj, ktorý sa využíva na posúdenie neistoty spojenej s odhadmi modelu a na hodnotenie jeho predikčnej schopnosti. Základnou myšlienkou bootstrapu je opakované vzorkovanie z pôvodného dátového súboru s opakovaním, namiesto získavania nových nezávislých výberov z populácie.

V každom bootstrapovom kroku sa náhodným výberom s opakovaním zostaví vzorka rozsahu n z pôvodného dátového súboru. V dôsledku výberu s opakovaním sa jednotlivé pozorovania môžu vo vzniknutej vzorke vyskytovať viackrát, pričom niektoré pozorovania nemusia byť zahrnuté vôbec. Takto vytvorená vzorka slúži ako trénovacia množina, zatiaľ čo pozorovania, ktoré neboli do výberu zahrnuté, tvoria testovaciu množinu (tzv. *out-of-bag* pozorovania). Na základe každej takejto trénovacej množiny sa následne odhadujú parametre alebo vybrané charakteristiky modelu. Celý postup sa opakuje vopred stanovený počet iterácií, čím vzniká množina bootstrapových odhadov umožňujúca posúdenie stability a spoľahlivosti modelu [10, s. 209].

2.2.5 Lineárna diskriminačná analýza

Lineárna diskriminačná analýza *(angl. Linear Discriminant Analysis, LDA)* patrí medzi štatistické metódy, ktoré sa okrem klasifikácie využívajú aj na redukciu dimenzie dát. Na rozdiel od mnohých iných metód redukcie dimenzie zohľadňuje informáciu o triedach, do ktorých sú jednotlivé objekty zaradené [36].

Cieľom metódy LDA je nájsť takú lineárnu transformáciu pôvodných príznakov do priestoru nižšej dimenzie, aby boli jednotlivé triedy čo najlepšie oddelené a zároveň sa zachovalo čo najviac informácie potrebnej na ich rozlíšenie [3].

Pred samotnou aplikáciou metódy LDA uvedieme teoretické pozadie, z ktorého táto metóda vychádza. Predpokladajme, že poznáme hustoty tried $p_1, \ldots, p_m$, ako aj apriórne pravdepodobnosti $q_1, \ldots, q_m$. Pravdepodobnosť nesprávnej klasifikácie, teda že objekt z triedy j bude klasifikovaný do triedy i, je definovaná vzťahom

$$P(i \mid j) = \int_{R_i} p_j(x)\, dx. \tag{20}$$

Stredná strata klasifikátora $R_1, \ldots, R_m$ je potom daná vzťahom

$$E_{R_1,\ldots,R_m} = \sum_{i=1}^{m} \sum_{j=1}^{m} L(i \mid j)\, q_j \int_{R_i} p_j(x)\, dx, \tag{21}$$

kde $L(i \mid j)$, tzv. stratová funkcia, predstavuje stratu klasifikácie v prípade, že objekt patriaci do triedy j je klasifikovaný do triedy i. Cieľom je minimalizovať túto stratu vzhľadom na klasifikátor $R_1, \ldots, R_m$, teda nájsť taký rozklad priestoru, ktorý minimalizuje hodnotu (21). Týmto spôsobom získame teoreticky optimálne klasifikačné pravidlo, tzv. všeobecný Bayesov klasifikátor.

Na minimalizáciu výrazu (21) postačuje zvoliť ľubovoľný merateľný rozklad $R_1^*, \ldots, R_m^*$ spĺňajúci

$$R_i^* \subset \left\{ x \in \mathbb{R}^p : h_i(x) = \min_k h_k(x) \right\}, \qquad i = 1, \ldots, m, \tag{22}$$

kde

$$h_i(x) = \sum_{j=1}^{m} L(i \mid j)\, q_j\, p_j(x), \qquad i = 1, \ldots, m. \tag{23}$$

Bayesov klasifikátor priraďuje pozorovanie do tej triedy, pre ktorú je hodnota $h_i(x)$ minimálna. V praxi však hustoty $p_1, \ldots, p_m$ spravidla nepoznáme, a preto ich musíme odhadnúť z dát, čím získame klasifikátor aproximujúci Bayesovo pravidlo [7].

Pri aplikácii metódy LDA sa predpokladá, že hustoty $p_1, \ldots, p_m$ pochádzajú z viacrozmerného normálneho rozdelenia s rôznymi strednými hodnotami, avšak so spoločnou nesingulárnou kovariančnou maticou Σ, t. j. $p_i(x) \sim \mathcal{N}_p(\mu_i, \Sigma)$. Ako uvádza [10, s. 147], práve rovnosť kovariančných matíc pre všetky triedy vedie k tomu, že rozhodovacie hranice medzi triedami majú lineárny charakter, čo je kľúčová vlastnosť metódy LDA.

Uvažujme prípad dvoch tried ($m = 2$). Potom možno výraz (22) zapísať v tvare

$$R_1^* = \left\{ x \in \mathbb{R}^p : \frac{p_1(x)}{p_2(x)} \geq \frac{q_2 L(1 \mid 2)}{q_1 L(2 \mid 1)} \right\}, \tag{24}$$

pričom $R_2^* = \mathbb{R}^p \setminus R_1^*$. Logaritmovaním podielu hustôt dostávame

$$\ln\left(\frac{p_1(x)}{p_2(x)}\right) = (\mu_1 - \mu_2)^T \Sigma^{-1} x + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 - \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1.$$

Po zavedení označení

$$a := \Sigma^{-1}(\mu_1 - \mu_2), \qquad b := \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 - \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1, \qquad k := \frac{q_2 L(1|2)}{q_1 L(2|1)}, \tag{25}$$

získavame lineárne klasifikačné pravidlo $R_1^* = \left\{ x \in \mathbb{R}^p : a^T x + b - \ln(k) \geq 0 \right\}$ a $R_2^* = \mathbb{R}^p \setminus R_1^*$, kde rozhodovacia hranica je hyperrovina v $\mathbb{R}^p$ [7].

Parametre LDA sa v praxi odhadujú z pozorovaných dát. Stredné hodnoty tried sa odhadujú pomocou výberových priemerov jednotlivých tried a spoločná kovariančná matica ako vážený priemer výberových kovariančných matíc. Apriórne pravdepodobnosti tried sa zvyčajne odhadujú ako podiel počtu pozorovaní v danej triede k celkovému počtu pozorovaní [10, s. 144].

Medzi hlavné výhody metódy LDA patrí jej jednoduchosť, nízka výpočtová náročnosť a dobrá interpretovateľnosť výsledkov. Metóda je vhodná najmä v prípadoch, keď sú triedy približne lineárne separovateľné a je k dispozícii dostatočný počet pozorovaní na spoľahlivý odhad kovariančnej matice. Aj pri porušení predpokladu normality môže LDA poskytovať použiteľný klasifikátor. Naopak, výrazne odlišná kovariančná štruktúra tried alebo nelineárna separácia môžu viesť k zníženiu jej výkonnosti [10, s. 153].

2.2.6 Klasifikačný strom

Spôsob, pri ktorom je možné množinu rozhodovacích pravidiel usporiadať do stromovej štruktúry, sa nazýva klasifikačný strom. Klasifikačný strom predstavuje klasifikátor vo forme hierarchickej štruktúry pozostávajúcej z koreňového (počiatočného) uzla, neterminálnych uzlov a terminálnych uzlov.

Každý neterminálny uzol obsahuje rozhodovaciu podmienku založenú na jednej premennej, pričom hrany stromu zodpovedajú možným výsledkom tohto rozhodovania. Jednej z dvoch hrán je pritom priradené označenie *True* a druhej *False*. Terminálny uzol

predstavuje konečnú predikciu, teda priradenie triedy. Proces klasifikácie prebieha postupným prechodom od koreňového uzla smerom nadol po ceste určenej splnením alebo nesplnením jednotlivých podmienok v neterminálnych uzloch až po dosiahnutie terminálneho uzla, ktorý objektu priradí príslušnú triedu [27].

Ako je uvedené v [7], konštrukcia klasifikačného stromu je založená na objektoch v trénovacej množine, t. j. na objektoch so známou správnou klasifikáciou. Samotný postup vytvárania stromu sa označuje ako rast stromu alebo rekurzívne rozdeľovanie.

Jedným z kľúčových krokov pri tvorbe klasifikačného stromu je výber vhodnej podmienky pre neterminálny uzol. Základný princíp voľby podmienky spočíva podľa [7] v tom, že sa pre daný uzol vytvorí konečný počet prípustných podmienok, ktoré by mu mohli byť priradené. Pre každú z týchto podmienok sa následne definuje miera kvality rozdelenia a vyberie sa tá, ktorá dosahuje najlepšiu hodnotu. Ako uvádza [27], zvolí sa rozhodnutie založené na takej premennej, ktorá najlepšie rozdelí aktuálnu množinu dát vzhľadom na cieľovú triedu. Zásadnou otázkou pritom zostáva, ako vhodne definovať mieru diskriminačnej kvality jednotlivých podmienok v uzle.

Ukazuje sa, že najčastejšie používané kritériá vychádzajú z konceptu nečistoty tried v uzle. Index nečistoty vyjadruje mieru nejednoznačnosti klasifikácie v prípade, že by bol daný uzol považovaný za terminálny. Cieľom je maximalizovať zlepšenie čistoty uzlov, resp. minimalizovať ich nečistotu. Z toho vyplýva, že optimálna podmienka je taká, ktorá rozdelí trénovacie objekty do podmnožín tak, aby bola klasifikácia v týchto podmnožinách čo najjednoznačnejšia [27].

Jednou z najpoužívanejších mier nečistoty je *Giniho index*, definovaný vzťahom

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{26}$$

kde $\hat{p}_{mk}$ označuje podiel trénovacích pozorovaní v m-tej oblasti patriacich do k-tej triedy. Ak sú hodnoty $\hat{p}_{mk}$ blízke nule alebo jednej, Giniho index nadobúda nízke hodnoty, čo znamená, že uzol obsahuje prevažne pozorovania jednej triedy.

Alternatívnou mierou nečistoty je *entropia*, ktorá je definovaná ako

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}. \tag{27}$$

Keďže platí $0 \leq \hat{p}_{mk} \leq 1$, výraz $-\hat{p}_{mk} \log \hat{p}_{mk}$ je nezáporný. Entropia nadobúda hodnotu

blízku nule v prípade, že sú všetky $\hat{p}_{mk}$ blízke nule alebo jednej, teda ak je uzol takmer čistý [10, s. 336].

Ak by klasifikačný strom rástol bez obmedzení, výsledkom by bol príliš zložitý model vedúci k pretrénovaniu. Z tohto dôvodu rast stromu prebieha rekurzívne len do splnenia niektorého zo zastavovacích kritérií. Medzi tieto kritériá patrí napríklad rovnaká trieda všetkých objektov v uzle, minimálny počet pozorovaní v uzle alebo nedostatočné zlepšenie kvality rozdelenia pri ďalšom delení. Okrem toho sa často aplikuje aj tzv. prerezávanie stromu *(angl. pruning)*, ktoré sa vykonáva po dokončení rastu stromu. Prerezávanie zmenšuje výsledný strom odstránením uzlov, ktoré neprispievajú k zlepšeniu kvality modelu. Jeho cieľom je dosiahnuť kompromis medzi presnosťou a jednoduchosťou modelu.

Medzi hlavné výhody klasifikačných stromov patrí skutočnosť, že ide o neparametrické metódy, ktoré dokážu efektívne pracovať s veľkými a zložitými dátovými súbormi bez potreby predpokladať ich konkrétnu parametrickú štruktúru. Ďalšou významnou výhodou je ich jednoduchá a intuitívna interpretácia, a to najmä vďaka grafickému znázorneniu stromovej štruktúry. Na druhej strane medzi ich hlavné nevýhody patrí nižšia predikčná presnosť v porovnaní s niektorými inými metódami. Okrem toho sú citlivé na šum v dátach a aj malé zmeny v trénovacej množine môžu viesť k výrazne odlišnej štruktúre výsledného stromu [27], [32].

3 Analýza dát

Táto kapitola sa venuje analýze dát získaných prostredníctvom jednotlivých dotazníkov na vzorke 421 klinických pacientov vo veku 17 až 73 rokov (priemerný vek 39,32 roka) a 676 respondentov bez diagnostikovanej poruchy osobnosti vo veku 16 až 75 rokov (priemerný vek 32,93 roka), ktorú tvorili dobrovoľníci a študenti psychológie. Najskôr sme porovnali rozdiely medzi skupinami respondentov, konkrétne podľa pohlavia a spôsobu zberu dát. Následne sme pomocou deskriptívnych štatistík preskúmali základné charakteristiky jednotlivých dátových súborov. V závere kapitoly sme overili reliabilitu dát použitím analýzy vnútornej konzistentnosti a validitu dát pomocou korelačnej analýzy a konfirmačnej faktorovej analýzy. Všetky analýzy a pokročilé štatistické metódy boli vykonané v softvéri R [23].

Pred samotnou analýzou sme najskôr upravili dostupné dáta. Vzhľadom na nízky počet chýbajúcich hodnôt v odpovediach respondentov aj v skóre jednotlivých domén boli pozorovania s chýbajúcimi hodnotami z ďalšej analýzy vylúčené. Z dát sme následne odstránili aj respondenta s uvedeným vekom *2 roky*.

3.1 Analýza rozdielov medzi skupinami

3.1.1 Porovnanie podľa spôsobu zberu dát

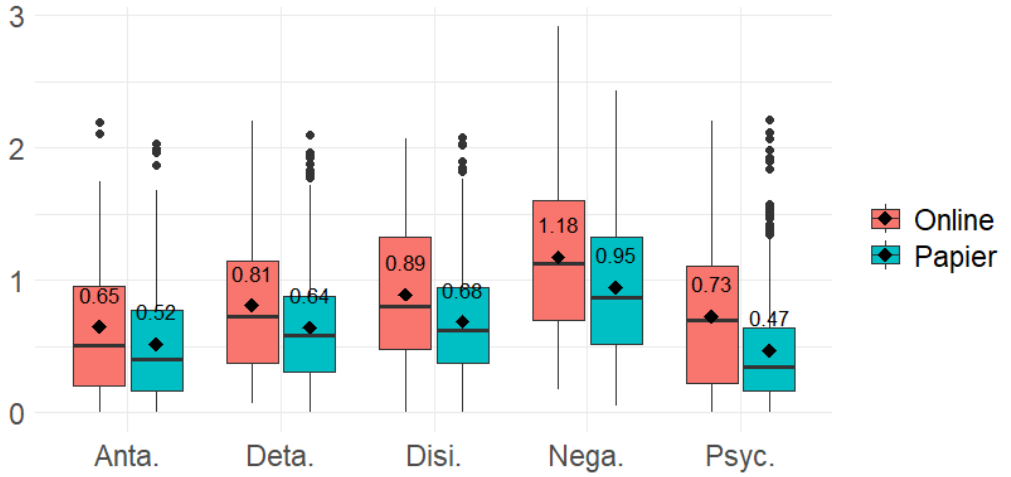
V tejto časti sme sa zamerali na rozdiely medzi údajmi získanými online formou a údajmi zozbieranými prostredníctvom papierových dotazníkov. Rozdiely sme analyzovali pre datasety *PID-5_ domény*[1] a *LPFS*. Keďže dáta získané online formou sú tvorené výlučne skupinou nepacientov, porovnanie sme realizovali iba na tejto vzorke.

PID-5_domény

Na porovnanie spôsobu zberu dát sme použili dataset, ktorý bol po úprave tvorený 665 pozorovaniami, z ktorých 109 bolo zozbieraných online a 556 papierovou formou. Na základe boxplotov zobrazených na Obr. 1 môžeme pozorovať, že pri všetkých doménach dosahujú v priemere vyššie hodnoty údaje získané online.

Ako prvé sme otestovali normalitu dát získaných online a papierovou formou pomocou

[1] *Poznámka:* Dataset *PID-5_ aspekty* sme do analýzy nezaradili, keďže jeho aspekty slúžia ako vstup do odvodených domén v datasete *PID-5_ domény*.



Obr. 1: Boxploty domén Papier vs. Online pre dataset *PID-5_ domény* (zdroj: vlastné spracovanie v softvéri R).

Shapirovho-Wilkovho testu. Na základe výsledkov testu normality (tabuľka 1) sa ukázalo, že pri všetkých doménach bola p-hodnota menšia ako 0,05, čo znamená, že dáta nevykazujú normálne rozdelenie. Z tohto dôvodu sme na otestovanie štatistickej významnosti rozdielov medzi spôsobmi zberu dát použili *Wilcoxonov test*.

Tabuľka 1 s výsledkami testu rozdielov medzi spôsobom zberu dát ukazuje, že pri všetkých doménach existuje štatisticky významný rozdiel. Z toho vyplýva, že naša prvotná intuícia založená na boxplotoch sa potvrdila. Dáta získané online sa štatisticky významne líšia od dát získaných papierovou formou. Keďže sa ukázalo, že pri online zbere dát sa navyšovalo skóre jednotlivých domén, dáta získané online sme z ďalších analýz vynechali.

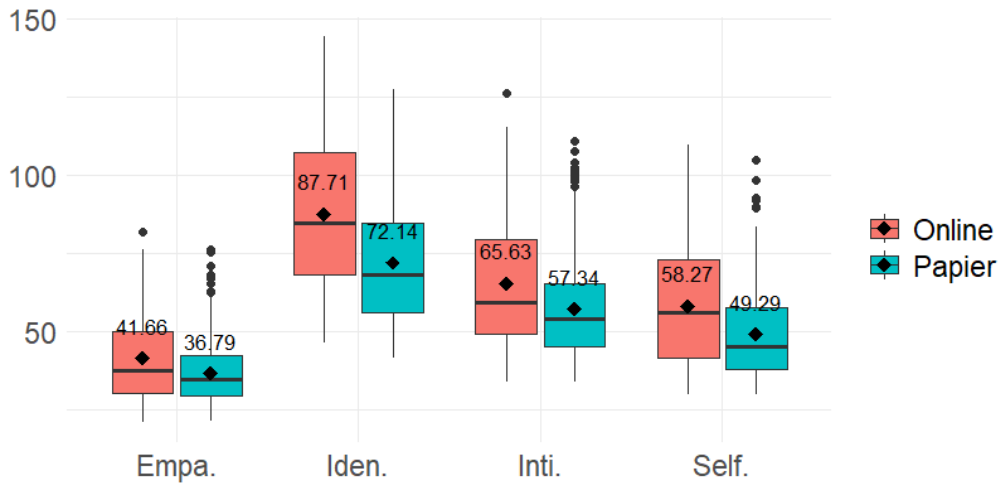
Tabuľka 1: P-hodnoty *Shapirovho-Wilkovho testu (SW)* a *Wilcoxonovho testu (W)*.

	Anta.	**Disi.**	**Deta.**	**Nega.**	**Psyc.**
SW Online	$5,38 \times 10^{-7}$	0,001	0,0002	0,006	$2,23 \times 10^{-5}$
SW Papier	$5,38 \times 10^{-17}$	$6,24 \times 10^{-11}$	$3,33 \times 10^{-14}$	$2,69 \times 10^{-11}$	$2,65 \times 10^{-20}$
W	0,048	0,0001	0,004	0,001	$2,28 \times 10^{-5}$

LPFS

V prípade datasetu *LPFS* sme po úprave získali dataset s 605 pozorovaniami (109 online

a 496 papier). Na základe boxplotov na Obr. 2 pozorujeme, že aj v tomto prípade sa dáta získané online odlišujú od dát získaných papierovou formou.



Obr. 2: Boxploty domén Papier vs. Online pre dataset *LPFS* (zdroj: vlastné spracovanie v softvéri R).

Podobne ako v prípade *PID-5_ domény*, sme najskôr otestovali normalitu dát pomocou *Shapirovho-Wilkovho testu*. Na základe výsledkov testu uvedených v tabuľke 2 bola hypotéza normality dát zamietnutá pri všetkých doménach. Teda na overenie štatistickej významnosti rozdielov medzi spôsobmi zberu dát sme použili *Wilcoxonov test*.

Výsledky testu (tabuľka 2) ukázali, že nulová hypotéza bola zamietnutá vo všetkých doménach. Rovnako ako v prípade datasetu *PID-5_ domény* sme údaje získané online z ďalších analýz vylúčili a vytvorili dataset pozostávajúci výlučne z dát získaných papierovou formou.

Tabuľka 2: P-hodnoty *Shapirovho-Wilkovho testu (SW)* a *Wilcoxonovho testu (W)*.

	Iden.	**Self.**	**Empa.**	**Inti.**
SW Online	0,002	0,0002	0,0002	$8,74 \times 10^{-6}$
SW Papier	$6,19 \times 10^{-13}$	$4,73 \times 10^{-16}$	$3,22 \times 10^{-14}$	$1,64 \times 10^{-14}$
W	$3,05 \times 10^{-9}$	$5,61 \times 10^{-5}$	0,002	0,0004

3.1.2 Porovnanie podľa pohlavia

V tejto časti sme porovnali rozdiely v našich dátach na základe premennej **gend.** Rozdiely sme skúmali osobitne v skupine nepacientov a pacientov, a to opäť len pre datasety *PID-5_ domény* a *LPFS*.

PID-5_ domény

Počet a rozdelenie respondentov podľa pohlavia môžeme pozorovať v tabuľke 3.

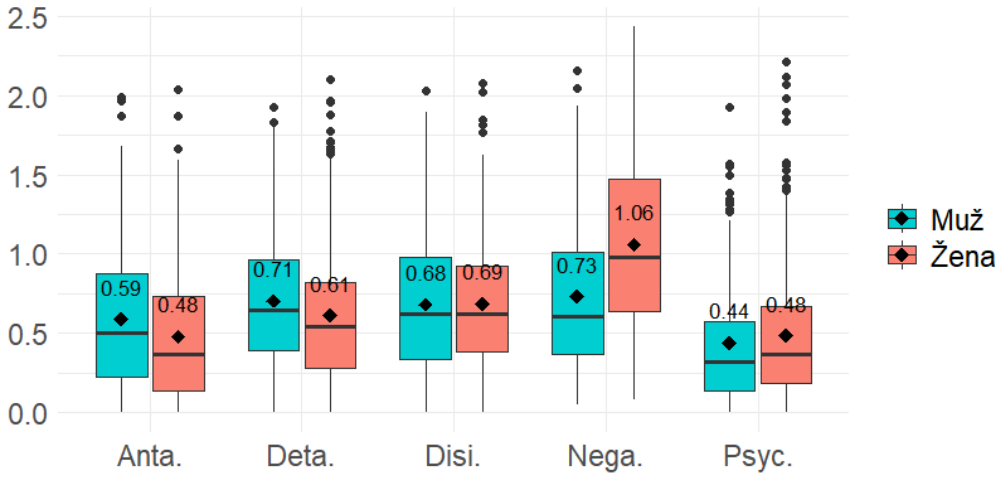
Tabuľka 3: Rozdelenie dát podľa pohlavia pre dataset *PID-5_ domény*.

Pohlavie	Počet nepacientov	Počet pacientov
Muž	190	289
Žena	366	129

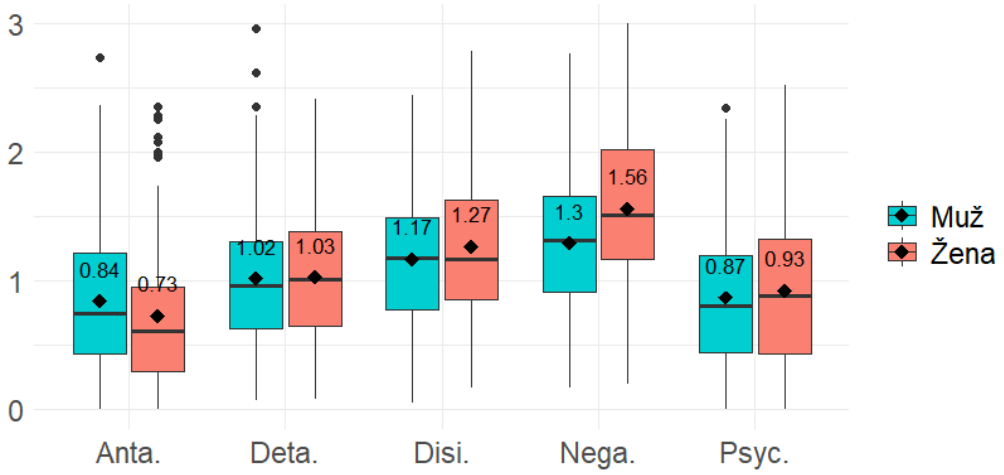
Ako prvé sme sa pozreli na základnú vizualizáciu rozdelenia domén medzi mužmi a ženami pre obe skupiny, samostatne pre skupinu nepacientov (Obr. 3a) a pacientov (Obr. 3b). Už na prvý pohľad je zrejmé, že sa v niektorých doménach muži a ženy líšia. V oboch skupinách ženy dosahujú výrazne vyššie skóre v doméne **Nega.**, čo je v súlade s tým, že ide o doménu súvisiacu s emočnou stránkou osobnosti. Naopak, muži dosahujú v oboch prípadoch vyššie skóre v doméne **Anta.**, ktorá je spojená s bezcitným a manipulatívnym správaním, čo môže súvisieť s tým, že tento typ správania sa u mužov vyskytuje častejšie než u žien. Na štatistické posúdenie týchto rozdielov sme najskôr overili normalitu dát pomocou *Shapirovho-Wilkovho testu*.

Z výsledkov v tabuľke 4 vyplýva, že domény v skupine nepacientov nevykazujú normálne rozdelenie. Z tohto dôvodu sme na overenie rozdielov medzi jednotlivými pohlaviami použili *Wilcoxonov test*. Štatisticky významné rozdiely medzi mužmi a ženami sa prejavili pri doménach **Anta.**, **Deta.** a **Nega.** Ostatné domény sa medzi skupinami štatisticky významne neodlišovali.

V prípade skupiny pacientov (tabuľka 5) nevykazovali normálne rozdelenie domény **Anta.** a **Psyc.** v oboch podskupinách (muži aj ženy). Rovnako sme zamietli normalitu dát aspoň v jednej z podskupín pri doménach **Deta.** a **Nega.** Z tohto dôvodu sme na porovnanie skupín použili *Wilcoxonov test*. V prípade domény **Disi.** obe podskupiny vykazovali normálne rozdelenie, a preto sme na testovanie rozdielov použili *dvojvýberový*



(a) Boxploty domén Muž vs. Žena pre *nepacientov*.



(b) Boxploty domén Muž vs. Žena pre *pacientov*.

Obr. 3: Boxploty domén Muž vs. Žena pre dataset *PID-5_ domény* (zdroj: vlastné spracovanie v softvéri R).

t-test s rovnakou nulovou hypotézou ako pri *Wilcoxonovom teste*.

Na základe výsledkov uvedených v tabuľke 5 sa ukázalo, že štatisticky významný rozdiel medzi mužmi a ženami sa prejavil pri doménach **Anta.** a **Nega.** Naopak, pri ostatných doménach sa rozdiel medzi pohlaviami štatisticky významný nepreukázal.

Tabuľka 4: P-hodnoty *Shapirovho-Wilkovho testu (SW)* a *Wilcoxonovho testu (W)* pre nepacientov.

	Anta.	**Disi.**	**Deta.**	**Nega.**	**Psyc.**
SW Muž	$1,69 \times 10^{-7}$	$7,49 \times 10^{-5}$	$3,22 \times 10^{-6}$	$4,64 \times 10^{-8}$	$8,77 \times 10^{-12}$
SW Žena	$1,56 \times 10^{-14}$	$7,13 \times 10^{-9}$	$7,61 \times 10^{-12}$	$1,97 \times 10^{-6}$	$4,56 \times 10^{-16}$
W	0,004	0,819	0,011	$3,16 \times 10^{-12}$	0,169

Tabuľka 5: P-hodnoty *Shapirovho-Wilkovho testu (SW)*, *Wilcoxonovho testu (W)* a *dvojvýberového t-testu (t-test)* pre pacientov.

	Anta.	**Disi.**	**Deta.**	**Nega.**	**Psyc.**
SW Muž	$2,69 \times 10^{-7}$	0,062	$1,04 \times 10^{-5}$	0,020	$9,22 \times 10^{-7}$
SW Žena	$4,15 \times 10^{-8}$	0,073	0,052	0,306	0,003
W / t-test	0,010	0,100	0,659	$6,10 \times 10^{-5}$	0,454

LPFS

V tejto časti sme sa zamerali na dataset *LPFS*, ktorého rozdelenie dát medzi pohlaviami je uvedené v tabuľke 6.

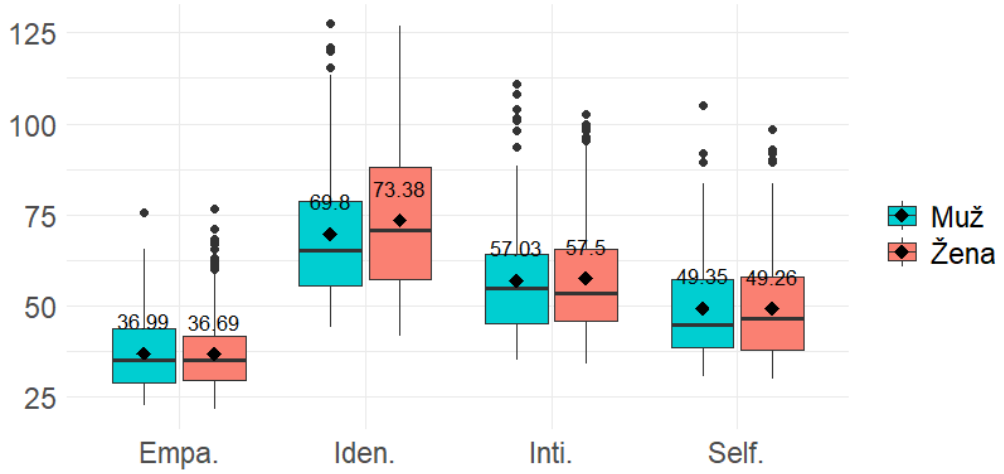
Tabuľka 6: Rozdelenie dát podľa pohlavia pre dataset *LPFS*.

Pohlavie	**Počet nepacientov**	**Počet pacientov**
Muž	172	254
Žena	324	118

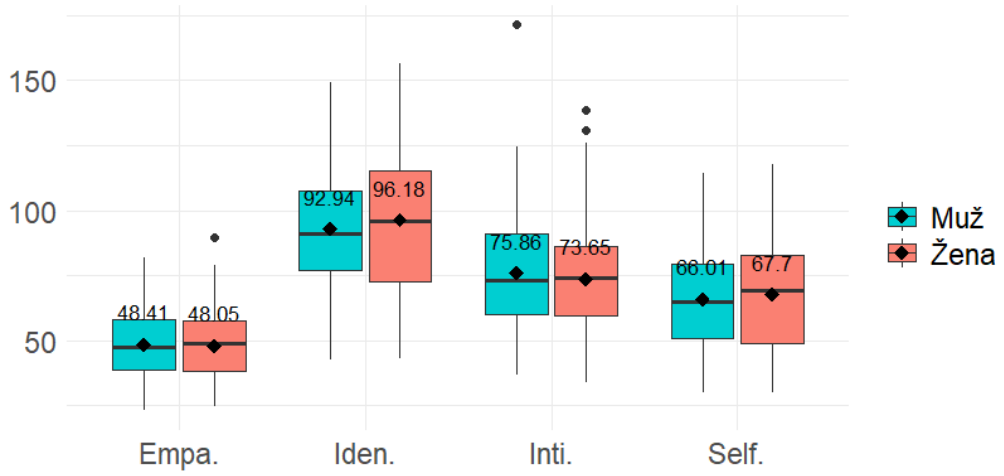
Na základe Obr. 4 možno pozorovať, že najvýraznejší rozdiel medzi skupinami sa prejavuje pri doméne **Iden.**, pričom ženy dosahujú v oboch prípadoch vyššie skóre. Pri ostatných doménach sa výraznejšie vizuálne rozdiely medzi pohlaviami neprejavujú.

Po overení normality dát pomocou *Shapirovho–Wilkovho testu* (tabuľka 7) pozorujeme v skupine nepacientov zamietnutie normality pri všetkých doménach. Následné testovanie rovnosti stredných hodnôt medzi mužmi a ženami pomocou *Wilcoxonovho testu* preukázalo štatisticky významný rozdiel len pri doméne **Iden.**, čo je v súlade s pozorovaniami z

boxplotov.



(a) Boxploty domén Muž vs. Žena pre *nepacientov*.



(b) Boxploty domén Muž vs. Žena pre *pacientov*.

Obr. 4: Boxploty domén Muž vs. Žena pre dataset *LPFS* (zdroj: vlastné spracovanie v softvéri R).

V skupine pacientov (tabuľka 8) sa normalita dát nepreukázala u mužov ani u žien pri doménach **Self.** a **Inti.** Pri doménach **Iden.** a **Empa.** bola hypotéza o normalite zamietnutá aspoň v jednej z podskupín. Na otestovanie rovnosti stredných hodnôt sme preto použili *Wilcoxonov test*, na základe ktorého nebol zistený štatisticky významný rozdiel medzi pohlaviami pri žiadnej z domén.

Tabuľka 7: P-hodnoty *Shapirovho-Wilkovho testu (SW)* a *Wilcoxonovho testu (W)* pre nepacientov.

	Iden.	**Self.**	**Empa.**	**Inti.**
SW Muž	$1,48 \times 10^{-8}$	$1,13 \times 10^{-9}$	$2,67 \times 10^{-6}$	$1,64 \times 10^{-8}$
SW Žena	$1,23 \times 10^{-8}$	$7,05 \times 10^{-12}$	$4,40 \times 10^{-12}$	$3,35 \times 10^{-11}$
W	0,034	0,623	0,740	0,680

Tabuľka 8: P-hodnoty *Shapirovho-Wilkovho testu (SW)* a *Wilcoxonovho testu (W)* pre pacientov.

	Iden.	**Self.**	**Empa.**	**Inti.**
SW Muž	0,140	0,001	0,011	$6,30 \times 10^{-5}$
SW Žena	0,040	0,005	0,062	0,007
W	0,415	0,537	0,870	0,292

3.2 Deskriptívna analýza dát

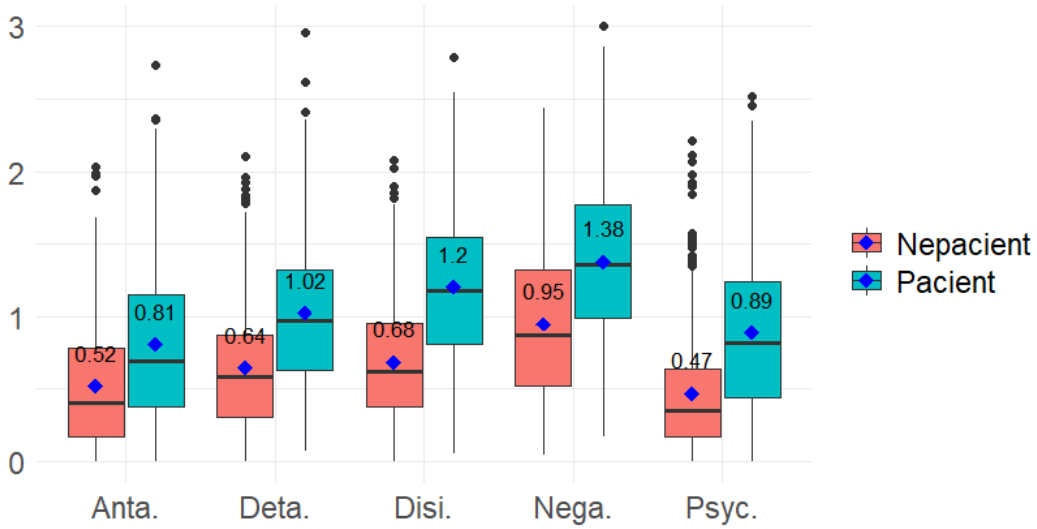
V časti *Deskriptívna analýza dát* sme sa pri každom datasete najskôr zamerali na rozdelenie jednotlivých domén medzi pacientov a nepacientov prostredníctvom boxplotov. Následne sme analyzovali základné charakteristiky jednotlivých datasetov pomocou histogramov. Konkrétny počet pozorovaní v datasetoch po ich úprave, ako aj zastúpenie respondentov v jednotlivých kategóriách, sú uvedené v tabuľke 9.

Tabuľka 9: Prehľad dát pre jednotlivé datasety.

Dataset	**Počet dát**	**Počet nepacientov**	**Počet pacientov**
PID-5_aspekty	969	553	416
PID-5_domény	974	556	418
LPFS	868	496	372

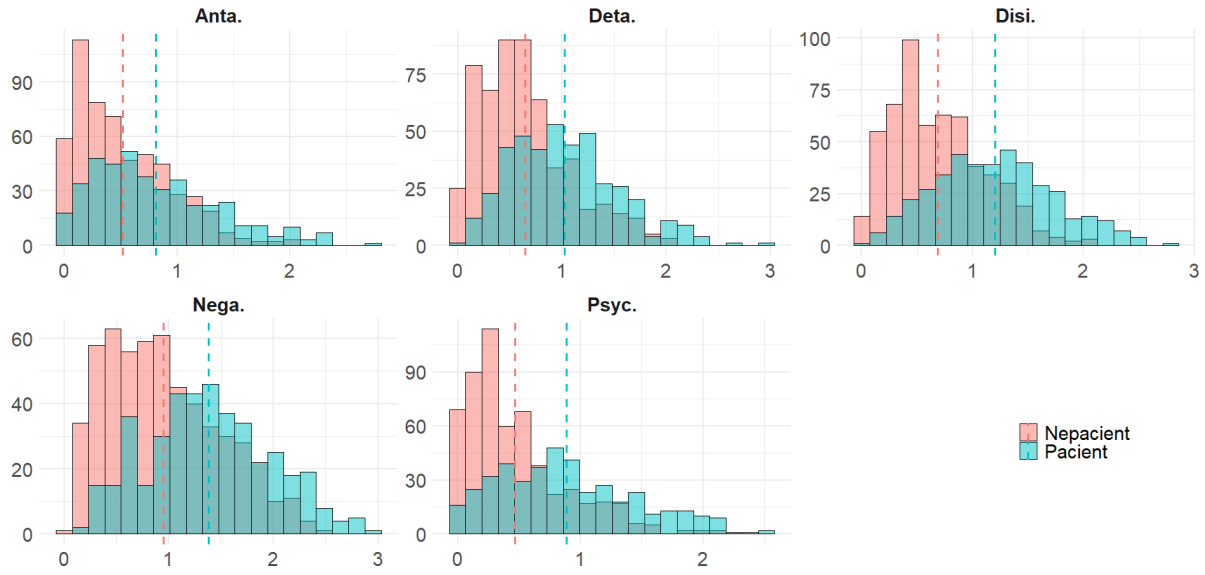
3.2.1 PID-5_domény

V prípade datasetu *PID-5_domény* sú na Obr. 5 zobrazené boxploty jednotlivých domén. Najvýraznejší rozdiel medzi skupinami pozorujeme pri doménach **Disi.** a **Psyc.** Tieto domény sa na základe vizuálnej analýzy javia ako potenciálne vhodné premenné na rozlišovanie medzi skupinou nepacientov a pacientov.



Obr. 5: Boxploty domén pre dataset *PID-5_domény* (zdroj: vlastné spracovanie v softvéri R).

Na základe Obr. 6 môžeme pozorovať, že väčšina domén vykazuje, najmä v skupine nepacientov, rozdelenie s pozitívnym sklonom, t. j. väčšina hodnôt sa sústreďuje v blízkosti hodnoty 0. Zároveň sa v skupine nepacientov prejavuje vyššie zastúpenie nižších hodnôt, zatiaľ čo v skupine pacientov pozorujeme vyššie zastúpenie vyšších hodnôt. V tomto prípade sa na základe distribúcie aj priemerných hodnôt (tabuľka 11) ukazuje, že najmenšie rozdiely medzi skupinami sa prejavujú pri doméne **Anta.** Na základe týchto výsledkov možno predpokladať, že táto doména bude mať nižšiu rozlišovaciu schopnosť medzi skupinami nepacientov a pacientov.



Obr. 6: Histogramy domén datasetu *PID-5_ domény* s vyznačenými priemernými hodnotami (zdroj: vlastné spracovanie v softvéri R).

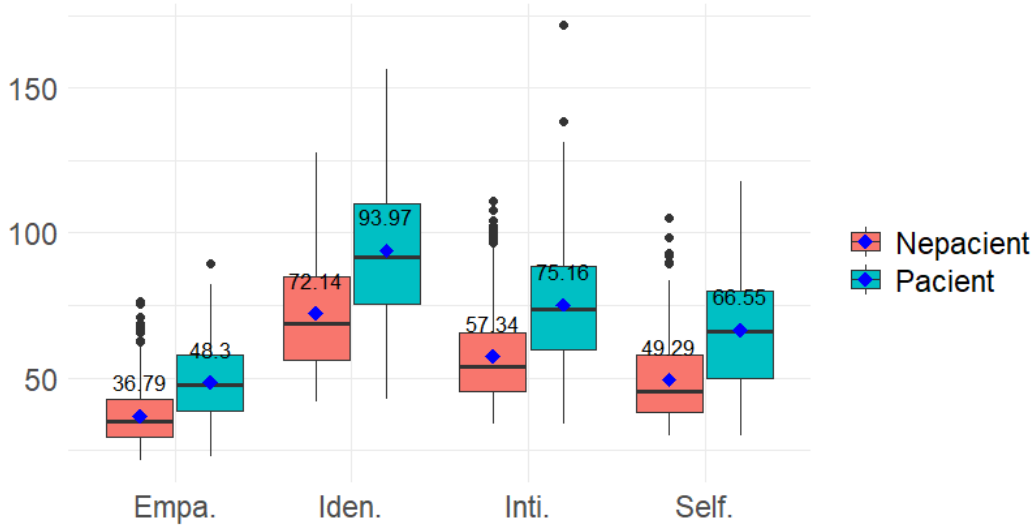
3.2.2 PID-5_aspekty

Na základe získaných priemerných hodnôt jednotlivých aspektov zvlášť pre skupinu nepacientov a pacientov (tabuľka 10), možno pozorovať, že pacienti dosahujú vyššie priemerné hodnoty vo všetkých sledovaných aspektoch v porovnaní s nepacientmi. Jednotlivé boxploty sú uvedené v prílohe A na Obr. 32 až Obr. 36 a histogramy v prílohe B na Obr. 37 až Obr. 41.

Najvýraznejší rozdiel medzi skupinami možno na základe priemerných hodnôt a vizualizácií pozorovať pri aspektoch **depr.**, **ecce.**, **irre.** a **sepa.** Rovnako ako v prípade datasetu *PID-5_ domény* možno predpokladať, že práve tieto aspekty by mohli zohrávať významnú úlohu pri klasifikácii respondentov. Naopak, najmenšie rozdiely medzi skupinami sa prejavujú pri aspektoch **atte., inti., rest., rigi.** a **subm.**

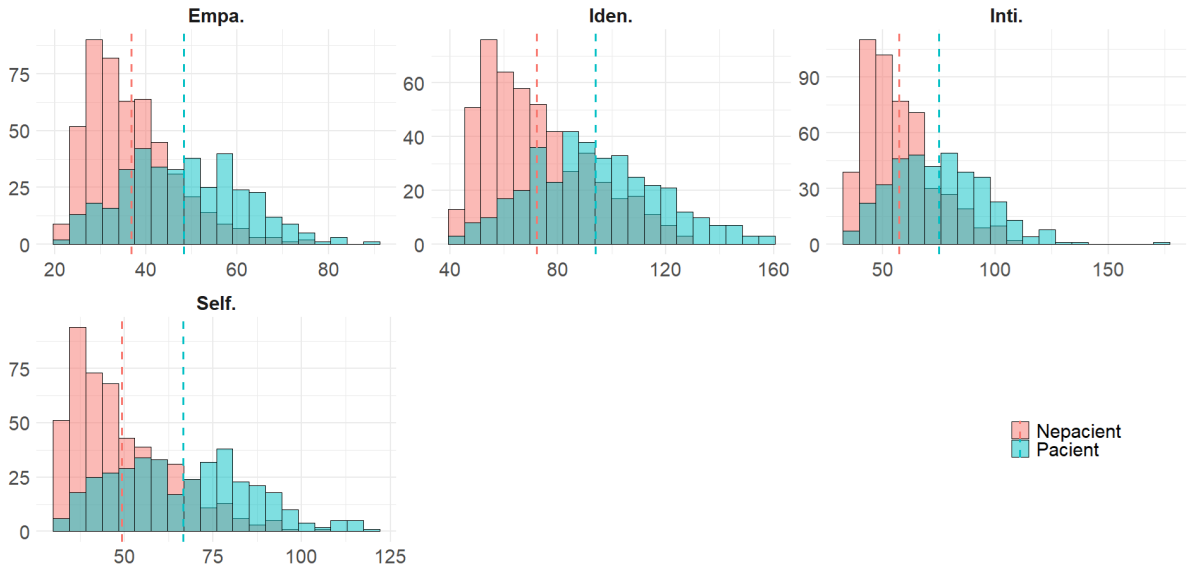
3.2.3 LPFS

V prípade datasetu *LPFS* sa na základe vizualizácie rozdelenia domén pomocou boxplotov (Obr. 7) javí doména **Iden.** ako tá s najväčším potenciálom rozlišovať medzi skupinami.



Obr. 7: Boxploty domén datasetu *LPFS* (zdroj: vlastné spracovanie v softvéri R).

Hoci na základe histogramov (Obr. 8) nie je možné jednoznačne určiť, pri ktorej doméne sa prejavuje výraznejší rozdiel medzi skupinami, pri porovnaní priemerných hodnôt jednotlivých domén v skupine nepacientov a pacientov (tabuľka 12) sa opäť ukazuje najväčší rozdiel pri doméne **Iden.** Naopak, najmenší rozdiel pozorujeme pri doméne **Empa.**



Obr. 8: Histogramy domén datasetu *LPFS* s vyznačenými priemernými hodnotami (zdroj: vlastné spracovanie v softvéri R).

3.3 Reliabilita dát

Ďalej sme sa zamerali na reliabilitu dát, konkrétne na vnútornú konzistenciu jednotlivých položiek, t. j. na mieru, do akej jednotlivé položky v rámci škály spolu súvisia. Na meranie vnútornej konzistentnosti dát sme použili Cronbachovu alfu (α). Kľúčová časť kódu je uvedená v prílohe H.

3.3.1 PID-5_aspekty

Na základe získaných hodnôt α, uvedených v tabuľke 10, môžeme u nepacientov pozorovať najsilnejšiu konzistenciu medzi položkami ($\alpha \geq 0{,}9$) pri aspekte **atte.**, **ecce.** a **depr.** Dobrú spoľahlivosť (α v rozmedzí 0,8 až 0,9) vykazujú aspekty **anxi., dist., host., perc., pers., rigi., risk.** a **soci.** Naopak, aspekty s relatívne nízkou konzistenciou ($\alpha < 0{,}7$) sú **impu.** a **rest.**, pričom najslabšiu spoľahlivosť vykazuje **susp.**

U pacientov najsilnejšiu konzistenciu pozorujeme rovnako ako u nepacientov pri **ecce.**, a zároveň pri **atte.** a **depr.** Najslabšiu mieru konzistencie opäť vykazuje aspekt **susp.** Výraznejšie zlepšenie miery konzistencie v skupine pacientov sa prejavuje pri aspektoch **call., dece., depr., gran., host., irre., risk., sepa.** a **unus.** v porovnaní s nepacientmi, zatiaľ čo zhoršenie pozorujeme pri aspektoch **pers.**, **rest.**, **rigi.** a **soci.**

Pre väčšinu aspektov (17 z 25) má klinická vzorka vyššiu hodnotu α. Celkovo sa však ukazuje, že väčšina aspektov vykazuje dobrú vnútornú konzistenciu medzi položkami.

Aspekty s najnižšou vnútornou konzistenciou sa v oboch skupinách ukazujú byť **susp.** a **rest.** Pri analýze problematických položiek sa v prípade aspektu **susp.** v oboch skupinách ukázalo, že problematickou položkou je otázka Q177: *„Zřídkakdy cítím, že mě lidé, které znám, zkouší využívat. “* Táto otázka najmenej zapadá do celkovej škály, keďže jej vynechaním by sa hodnota α zvýšila. Dôvodom môže byť aj to, že patrí do skupiny reverzne kódovaných položiek. Pre aspekt **rest.** ide v prípade nepacientov o otázku Q45: *„Když se něco stane, dotkne se mě to pouze na chvilku.“*, zatiaľ čo u pacientov by vynechanie akejkoľvek otázky hodnotu α znížilo.

Hodnoty α sme porovnali s výsledkami výskumu [14], uskutočneného na americkej populácii. Výskum bol realizovaný na vzorke rôznych respondentov, pričom sa zameriavame na vzorku 366 respondentov, ktorí vyhľadali psychologickú alebo psychiatrickú pomoc (v článku označovanú ako *selected sample*) a na vzorku 264 reprezentantov americkej bežnej

populácie (v článku označovanú ako *representative sample*) [14].

Rovnako ako v našom prípade sa ukázalo, že najsilnejšiu vnútornú konzistenciu majú v oboch skupinách populácie aspekty **ecce.** a **depr.** Výskum zároveň potvrdil, že problematické aspekty s nižšou vnútornou konzistenciou sú, podobne ako v našom prípade, **rest.** v oboch skupinách a **susp.**, najmä v skupine bežnej populácie. Aspekt **susp.** nevykazuje nižšiu konzistenciu iba v našom prípade či v prípade americkej populácie, ale systematicky sa javí ako problematický aj v iných výskumoch (napr. výskum na zdravej talianskej populácii [6] či výskum na klinickej populácii v Kanade [22]).

V porovnaní s našimi výsledkami sú hodnoty jednotlivých α uvádzané vo výskume [14] vo väčšine prípadov vyššie. Päť najvyšších hodnôt α sa pohybuje v rozmedzí 0,91 - 0,96 pre nepacientov a 0,91 - 0,95 pre pacientov, zatiaľ čo v našom prípade je to 0,87 - 0,93 pre nepacientov a 0,87 - 0,93 pre pacientov. Ak sa pozrieme na najnižšie hodnoty α, v prípade americkej populácie sa pohybujú v rozmedzí 0,72 - 0,80 pre nepacientov a 0,73 - 0,84 pre pacientov, zatiaľ čo v našom výskume sú tieto hodnoty nižšie 0,51 - 0,71 pre nepacientov a 0,52 - 0,73 pre pacientov. Najvýraznejší rozdiel v hodnote α medzi českou a americkou populáciou sa prejavuje v oboch skupinách pri **susp.**, kde je hodnota o 0,22 vyššia v prípade americkej bežnej populácie a o 0,28 vyššia v prípade klinickej populácii.

Tabuľka 10: Hodnoty priemeru **(P)**, smerodajnej odchýlky **(SO)** a Cronbachovej alfy (α) pre jednotlivé aspekty u nepacientov **(Nep.)** a pacientov **(Pac.)** v datasete *PID-5_ aspekty*.

Aspekt	**Nep. P(SO)**	**Pac. P(SO)**	**Nep.** α	**Pac.** α
anhe.	0,657 (0,492)	1,112 (0,620)	0,795	0,803
anxi.	1,035 (0,683)	1,443 (0,730)	0,893	0,873
atte.	0,752 (0,634)	0,931 (0,729)	0,897	0,899
call.	0,317 (0,308)	0,652 (0,495)	0,767	0,830
dece.	0,529 (0,420)	0,832 (0,601)	0,796	0,851
depr.	0,428 (0,454)	0,992 (0,663)	0,897	0,912
dist.	0,742 (0,587)	1,213 (0,676)	0,873	0,869
ecce.	0,613 (0,595)	1,128 (0,734)	0,930	0,931
emot.	0,965 (0,642)	1,366 (0,700)	0,791	0,788
gran.	0,449 (0,470)	0,730 (0,616)	0,765	0,784
host.	0,857 (0,498)	1,227 (0,600)	0,803	0,825
impu.	0,811 (0,511)	1,271 (0,604)	0,684	0,691
inti.	0,515 (0,549)	0,774 (0,667)	0,739	0,779
irre.	0,506 (0,431)	1,121 (0,640)	0,709	0,787
mani.	0,575 (0,547)	0,870 (0,694)	0,772	0,784
perc.	0,384 (0,394)	0,771 (0,541)	0,818	0,816
pers.	0,816 (0,533)	1,283 (0,515)	0,816	0,731
rest.	0,944 (0,502)	1,202 (0,522)	0,671	0,622
rigi.	1,089 (0,640)	1,320 (0,622)	0,876	0,838
risk.	1,076 (0,496)	1,418 (0,630)	0,818	0,868
sepa.	0,848 (0,601)	1,335 (0,731)	0,786	0,812
subm.	0,987 (0,620)	1,234 (0,710)	0,706	0,734
susp.	0,997 (0,435)	1,346 (0,474)	0,508	0,516
unus.	0,411 (0,470)	0,765 (0,627)	0,786	0,810
soci.	0,760 (0,548)	1,180 (0,591)	0,850	0,818

3.3.2 PID-5_ domény

V prípade datasetu *PID-5_ domény* môžeme na základe získaných hodnôt α (tabuľka 11) pozorovať, že všetkých päť domén vykazuje v oboch skupinách vysokú vnútornú konzistenciu (α v rozsahu 0,88 - 0,94). Rozdiely medzi hodnotami pre pacientov a nepacientov sú minimálne.

Ak porovnáme naše hodnoty α s výsledkami výskumu americkej populácie [14], môžeme pozorovať, že aj v tomto prípade dosahujú všetky domény vysokú vnútornú konzistenciu, pričom rozdiely medzi hodnotami našej vzorky a americkej populácie v oboch skupinách sú minimálne. Najväčší rozdiel možno pozorovať pri doménach **Anta.**, kde hodnoty v americkej štúdii dosahujú 0,95 pre nepacientov a 0,94 pre pacientov, a **Deta.**, kde tieto hodnoty dosahujú 0,96 pre nepacientov aj pacientov.

Tabuľka 11: Hodnoty priemeru, smerodajnej odchýlky a Cronbachovej alfy pre jednotlivé domény u nepacientov a pacientov v datasete *PID-5_ domény*.

Doména	Nep. P(SO)	Pac. P(SO)	Nep. α	Pac. α
Anta.	0,518 (0,415)	0,808 (0,553)	0,894	0,910
Disi.	0,685 (0,415)	1,201 (0,534)	0,883	0,896
Deta.	0,645 (0,433)	1,023 (0,509)	0,892	0,888
Nega.	0,947 (0,538)	1,378 (0,590)	0,916	0,908
Psyc.	0,469 (0,412)	0,887 (0,555)	0,933	0,940

3.3.3 LPFS

V závere tejto časti sme sa zamerali na dataset *LPFS*. Po výpočte hodnôt uvedených v tabuľke 12 sa ukázalo, že v skupine nepacientov vykazujú hodnotu $\alpha > 0{,}8$ domény **Iden.** a **Tota.**, zatiaľ čo v skupine pacientov túto hodnotu dosahujú domény **Iden.**, **Self.** a **Tota.** Najvyššiu vnútornú konzistenciu dosahuje v oboch skupinách doména **Tota.**, ktorá je tvorená kombináciou zvyšných štyroch domén.

Ako najslabšia časť dotazníka sa javí doména **Empa.** V skupine pacientov je hodnota α len mierne vyššia než hraničná hodnota ($\alpha = 0{,}7$), zatiaľ čo v skupine nepacientov je nižšia. Analýza problematických položiek v rámci domény **Empa.** ukázala, že položka

Q11: *„Dokážu oceniť názor druhých ľudí, aj keď s nimi nesúhlasím.“* najmenej zapadá do celkovej škály otázok merajúcich túto doménu.

Nižšiu mieru vnútornej konzistencie vykazuje aj doména **Inti.**, pričom medzi položky, ktoré negatívne korelujú s ostatnými položkami, patria Q26: *„Mám mnoho uspokojivých vztahů jak v soukromí, tak v práci.“*, Q37: *„Při jednání s druhými se snažím být flexibilní a spolupracující.“* a Q39: *„Pracuji na svých blízkých vztazích, protože jsou pro mě důležité.“* Dôvodom, prečo tieto položky najmenej zapadajú do celkovej škály, môže byť skutočnosť, že pri výpočte skóre danej domény majú zápornú váhu, a teda patria medzi reverzne kódované položky.

Hodnoty α sme porovnali s výsledkami výskumu [8] realizovaného na iránskej populácii. Výskum zahŕňal 471 vysokoškolských študentov (neklinická populácia) a 142 hospitalizovaných pacientov s diagnostikovanou poruchou osobnosti (klinická populácia).

Vo výskume [8] sú uvedené hodnoty α pre doménu **Iden.** (0,83 pre nepacientov / 0,81 pre pacientov), **Self.** (0,85 / 0,79), **Inti.** (0,75 / 0,77) a **Tota.** (0,94 / 0,93), pričom všetky tieto domény vykazujú dobrú vnútornú konzistenciu, podobne ako v našom výskume. Najnižšiu hodnotu α dosiahla doména **Empa.** (0,70 / 0,63), čo je taktiež v súlade s našimi zisteniami. Rozdiel spočíva v tom, že v našom výskume sú hodnoty α vo všetkých doménach vyššie v skupine pacientov, zatiaľ čo vo výskume [8] sú, s výnimkou domény **Inti.**, vyššie u nepacientov. Podobné výsledky, poukazujúce na nižšiu mieru vnútornej konzistencie v doméne **Empa.**, uvádza aj výskum realizovaný na portugalskej populácii [21].

Tabuľka 12: Hodnoty priemeru, smerodajnej odchýlky a Cronbachovej alfy pre jednotlivé domény u nepacientov a pacientov v datasete *LPFS*.

Doména	**Nep. P(SO)**	**Pac. P(SO)**	**Nep.** α	**Pac.** α
Iden.	72,14 (19,57)	93,97 (24,02)	0,841	0,860
Self.	49,29 (14,18)	66,55 (19,34)	0,784	0,802
Empa.	36,79 (9,77)	48,30 (12,86)	0,679	0,733
Inti.	57,34 (15,42)	75,16 (20,64)	0,706	0,739
Tota.	215,60 (52,74)	284,00 (68,28)	0,927	0,935

3.4 Validita dát

V záverečnej časti kapitoly *Analýza dát* sme skúmali validitu dát, t. j. či použité dotazníky skutočne merajú konštrukty, ktoré majú zachytávať. Na analýzu vhodnosti našich dát sme použili korelačnú analýzu, pri ktorej sme skúmali vzťahy medzi jednotlivými zložkami datasetov *PID-5_ aspekty*, *PID-5_ domény* a *LPFS*. Zároveň sme sa pozreli aj na vzájomné korelácie medzi týmito datasetmi.

Pri korelačnej analýze sme vo všetkých prípadoch používali *Pearsonov korelačný koeficient*. Aj keď dáta podľa histogramov nevykazovali normalitu, pre použitie tohto koeficientu sme sa rozhodli na základe veľkej vzorky dát (tabuľka 9).

Okrem toho sme overili vnútornú štruktúru dát pomocou konfirmačnej faktorovej analýzy (CFA; kód v prílohe I). Cieľom analýzy bolo otestovať predpokladanú štruktúru vzťahov medzi jednotlivými doménami dotazníkov *PID-5* a *LPFS-SR*, ako aj posúdiť, či český preklad položiek zachováva ich pôvodnú faktorovú štruktúru a adekvátne reprezentuje predpokladané psychologické konštrukty. Kvalitu modelu sme hodnotili na základe vybraných indexov zhody a miery vysvetlenej variability jednotlivých položiek (R^2).

3.4.1 PID-5_aspekty

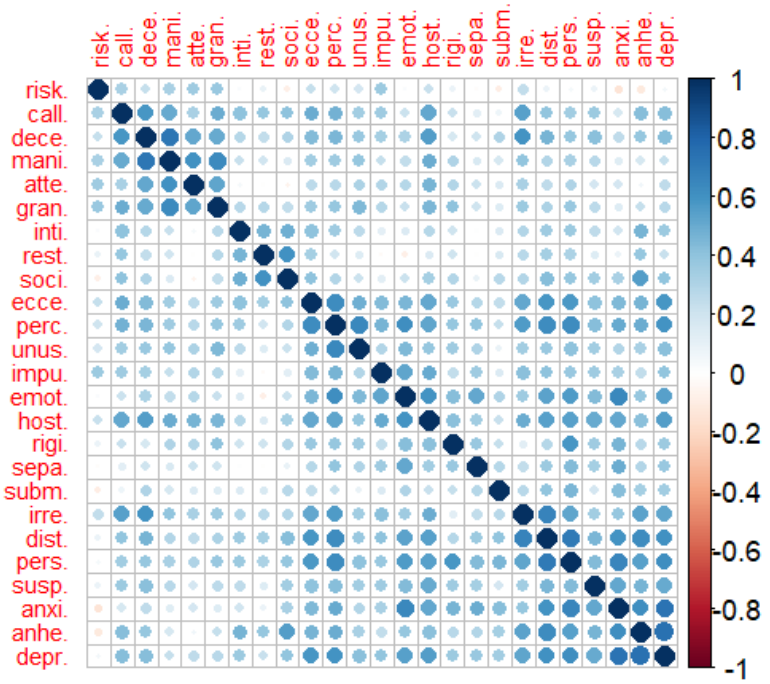
Korelačná analýza

Ako prvé sme sa v prípade datasetu *PID-5_ aspekty* zamerali na vzájomné korelácie medzi jednotlivými aspektmi v skupine nepacientov a pacientov. V oboch prípadoch väčšina aspektov medzi sebou pozitívne koreluje. Aspekty, ktoré spolu silne korelujú, sú zoskupené do zhlukov, čo uľahčuje identifikáciu vzťahov medzi nimi.

Pri skupine nepacientov (Obr. 9) možno identifikovať približne štyri zhluky aspektov: *prvý zhluk* – **call., dece., mani., atte., gran.**, *druhý zhluk* – **inti., rest., soci.**, *tretí zhluk* – **ecce., perc., unus., impu., emot., host.**, *štvrtý zhluk* – **irre., dist., pers., susp., anxi., anhe., depr.** Naopak, aspekty **risk., rigi., sepa., subm.** sa do žiadneho zhluku nezaraďujú.

Porovnanie identifikovaných zhlukov s doménami datasetu *PID-5_ domény* naznačuje čiastočnú zhodu medzi štruktúrou zhlukov a rozdelením aspektov do jednotlivých domén. Prvý zhluk obsahuje všetky tri aspekty domény **Anta.**, zatiaľ čo tretí zhluk zahŕňa všetky aspekty domény **Psyc.** Doména **Deta.** má dva aspekty zaradené v druhom zhluku a jeden

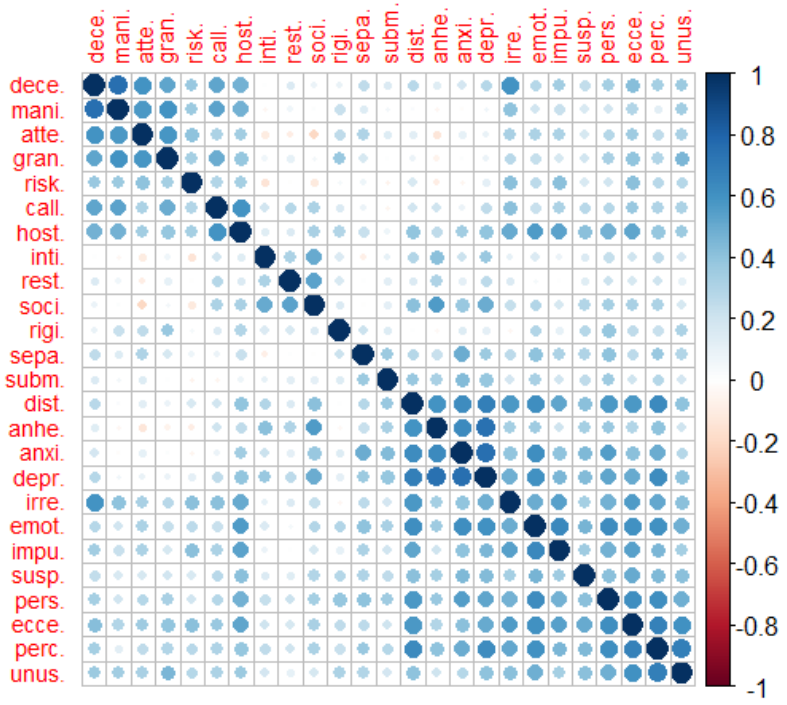
vo štvrtom, zatiaľ čo pri doméne **Disi.** má dva aspekty vo štvrtom zhluku a jeden v treťom. V prípade domény **Nega.** je každý z jej troch aspektov zaradený do iného zhluku, pričom napriek tomu vykazujú medzi sebou silnejšie korelácie.



Obr. 9: Korelačná matica pre *nepacientov* pre dataset *PID-5_ aspekty* (zdroj: vlastné spracovanie v softvéri R).

V prípade skupiny pacientov (Obr. 10) možno identifikovať dva väčšie zhluky: *prvý zhluk* – **dece., mani., atte., gran., risk., call., host.**, *druhý zhluk* – **dist., anhe., anxi., depr., irre., emot., impu., susp., pers., ecce., perc., unus.**, a menší zhluk tvorený aspektmi **inti., rest., soci.** Rovnako aj tu aspekty **rigi., sepa., subm.** nepatria do žiadneho zhluku.

Porovnaním oboch skupín sa ukazuje, že rovnaké aspekty majú tendenciu tvoriť spoločné zhluky, zatiaľ čo iné zostávajú mimo. Táto analýza naznačuje, že pri tvorbe modelov by mohlo postačovať vybrať reprezentantov z jednotlivých zhlukov, ktorí dobre vystihujú dané skupiny aspektov.

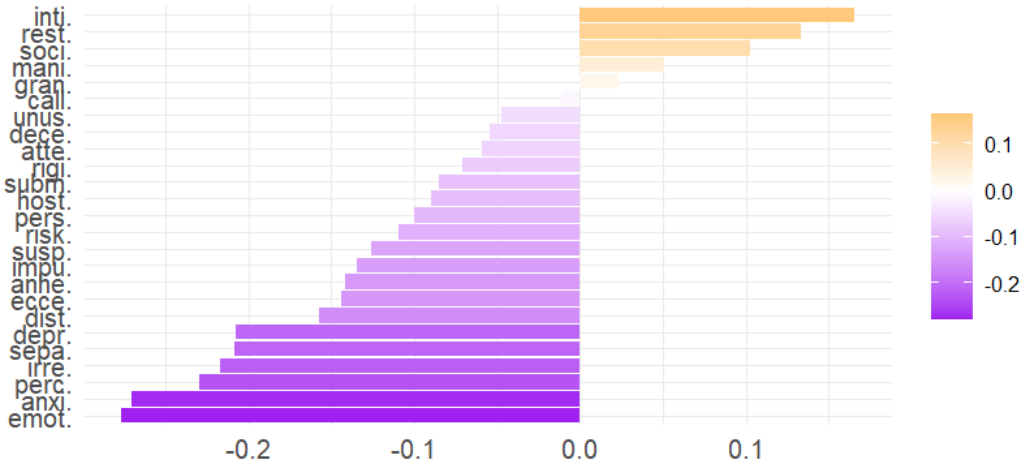


Obr. 10: Korelačná matica pre *pacientov* pre dataset *PID-5_ aspekty* (zdroj: vlastné spracovanie v softvéri R).

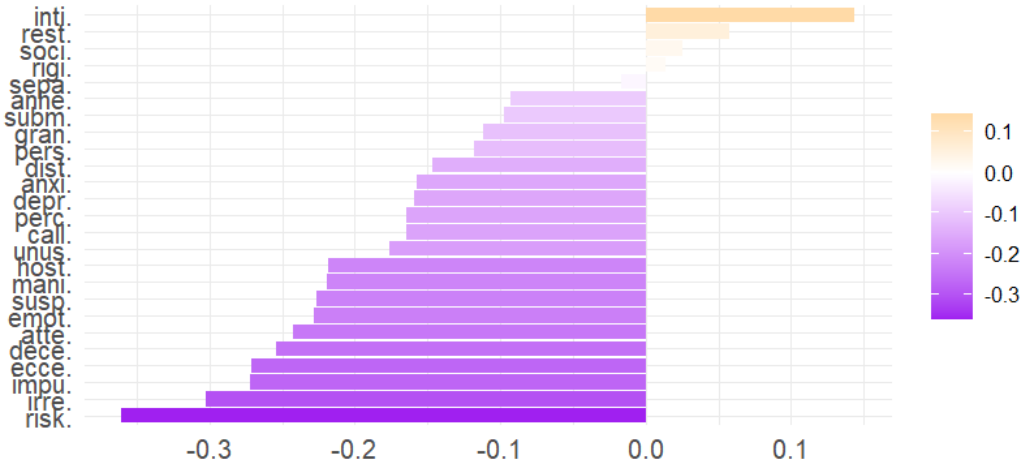
Obr. 11 a Obr. 12 znázorňujú korelácie aspektov s premennou **age.** V oboch skupinách majú starší respondenti vyššie skóre najmä v aspektoch súvisiacich s nižšou ochotou socializovať sa. Naopak, mladší respondenti dosahujú vyššie skóre v aspektoch súvisiacich s emočnou nestabilitou, ako aj v aspektoch odrážajúcich tendenciu k riskantnému správaniu.

CFA

V prípade dotazníka *PID-5* sa pri *CFA* zameriame na model vytvorený z domén datasetu *PID-5_ domény*, keďže v prípade datasetu *PID-5_ aspekty* by do modelu vstupovalo veľa parametrov vzhľadom na relatívne malý počet dát.



Obr. 11: Korelácie aspektov datasetu *PID-5_ aspekty* s premennou **age.** pre *nepacientov* (zdroj: vlastné spracovanie v softvéri R).



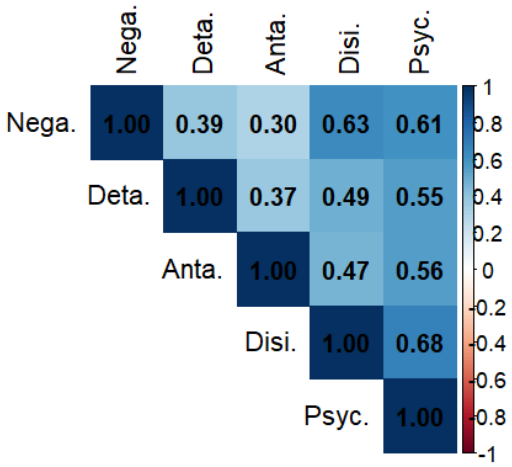
Obr. 12: Korelácie aspektov datasetu *PID-5_ aspekty* s premennou **age.** pre *pacientov* (zdroj: vlastné spracovanie v softvéri R).

3.4.2 PID-5_domény

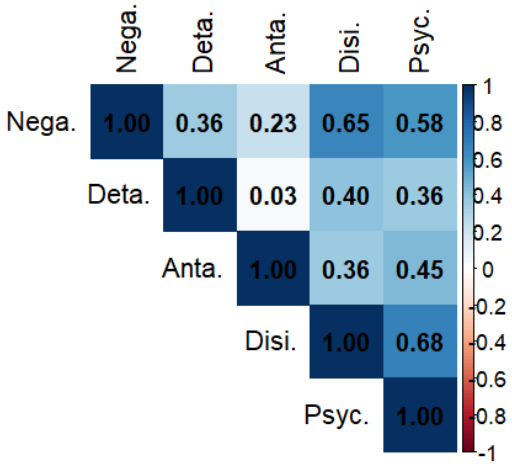
Korelačná analýza

Korelácie jednotlivých domén datasetu *PID-5_ domény* sú znázornené na Obr. 13 pre nepacientov a na Obr. 14 pre pacientov. V prípade korelácií jednotlivých domén s premennou

age. pozorujeme podobné závery ako pri datasete *PID-5_ aspekty*.



Obr. 13: Korelačná matica pre *nepacientov* pre dataset *PID-5_ domény* (zdroj: vlastné spracovanie v softvéri R).



Obr. 14: Korelačná matica pre *pacientov* pre dataset *PID-5_ domény* (zdroj: vlastné spracovanie v softvéri R).

CFA

Po úprave dát dotazníka *PID-5* sme dostali dataset s 988 respondentmi, z toho 567 nepacientov a 421 pacientov. Model, ktorý sme v rámci CFA skúmali, je uvedený v prílohe C.

Na posúdenie miery zhody modelu s pozorovanými dátami sme využili viaceré indexy uvedené v tabuľke 13, ktoré boli vybrané na základe časti 2.1.5. P-hodnota χ^2 testu naznačuje slabú celkovú zhodu modelu s dátami, pričom tento výsledok môže byť ovplyvnený veľkosťou výberového súboru. Podobne aj index *CFI* poukazuje na neuspokojivú zhodu modelu, keďže jeho hodnota je nižšia ako 0,80. Na druhej strane, index *RMSEA* indikuje dobrú zhodu modelu, keďže jeho hodnota je nižšia ako 0,05. Rovnaký záver podporuje aj index *SRMR*, podľa ktorého hodnoty v intervale 0,05 - 0,08 predstavujú prijateľnú mieru zhody modelu.

Parametre získané na našom datasete sme porovnali s výsledkami CFA analýzy iného výskumu uvedenými v článku [6]. Výsledky dotazníka *PID-5* boli získané na vzorke 710 talianskych dospelých dobrovoľníkov z bežnej populácie. CFA analýza dosiahla v rámci výskumu hodnotu χ^2=2154,57 s p-hodnotou menšou ako 0,001, $RMSEA = 0{,}11$ (90 % IS: 0,11; 0,12), $SRMR = 0{,}10$ a $CFI = 0{,}93$. Na základe porovnania týchto metrík možno konštatovať, že náš model dosahuje priaznivejšie hodnoty indexov *RMSEA* a *SRMR*, avšak menej priaznivú hodnotu indexu *CFI*.

Tabuľka 13: Indexy zhody modelu dotazníka *PID-5*.

χ^2 ***(p-hodnota)***	**RMSEA** ***(90 % IS)***	**SRMR**	**CFI**	**PNFI**
21172,307 (<0,001)	0,046 (0,045; 0,047)	0,073	0,702	0,596

Na detailnejšiu analýzu jednotlivých položiek sme sa zamerali na hodnoty ich R^2. Ako je uvedené v [22], kvalitu danej položky pre príslušný faktor hodnotíme nasledovne:

- ak je hodnota $R^2 \geq 0{,}20$, položka je primerane silno spojená s daným faktorom modelu,
- ak je hodnota $R^2 < 0{,}20$, položka neopisuje faktor dostatočne, keďže nízka hodnota R^2 poukazuje na veľmi vysokú mieru chyby.

Pri pohľade na hodnoty R^2 pre jednotlivé položky domén (tabuľka 25, príloha C) možno pozorovať, že v doméne **Nega.** sú problematickými položkami *Q12, Q57, Q64, Q96, Q102, Q165*, teda práve tieto položky najmenej reprezentujú danú doménu. Rovnaký počet problematických položiek má aj doména **Psyc.** (*Q37, Q44, Q77, Q143, Q150,*

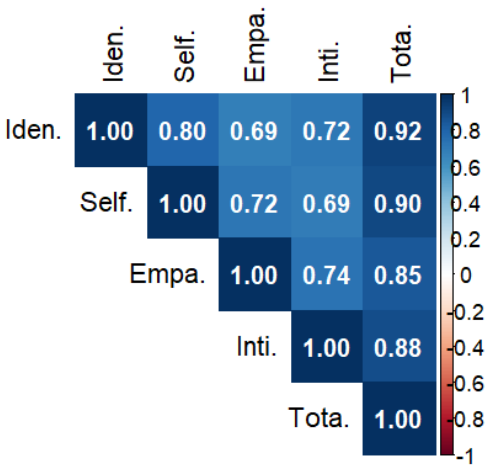
Q192). V prípade domény **Deta.** sa ako problematické ukázali položky *Q20, Q30, Q97, Q108, Q120*. Najmenej problematických položiek vykazujú domény **Anta.** (*Q142, Q179*) a **Disi.** (*Q58*).

V prípade položiek *Q58*: *„Obvykle myslím dříve, než jednám.“* a *Q142*: *„Snažím se říkat pravdu, i když bolí.“* sú hodnoty R^2 veľmi nízke, čo naznačuje, že tieto položky nemajú s danými doménami prakticky žiadnu súvislosť. Môže to súvisieť s tým, že obe položky sú reverzne kódované, a preto s ostatnými položkami nezapadajú optimálne do celkovej štruktúry danej domény.

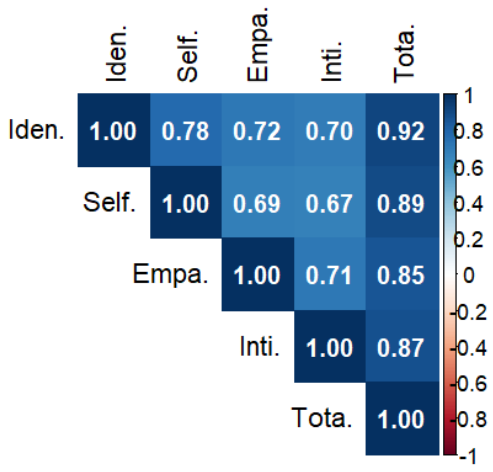
3.4.3 LPFS

Korelačná analýza

Táto časť sa zameriava na vzťahy medzi jednotlivými doménami datasetu *LPFS*. Obr. 15 a Obr. 16 znázorňujú korelačné matice pre nepacientov, resp. pacientov.

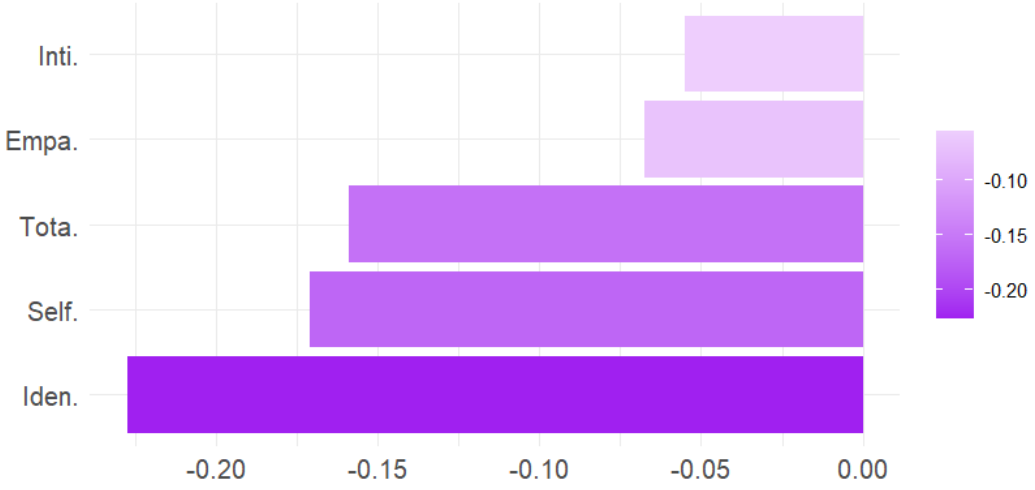


Obr. 15: Korelačná matica pre *nepacientov* pre dataset *LPFS* (zdroj: vlastné spracovanie v softvéri R).

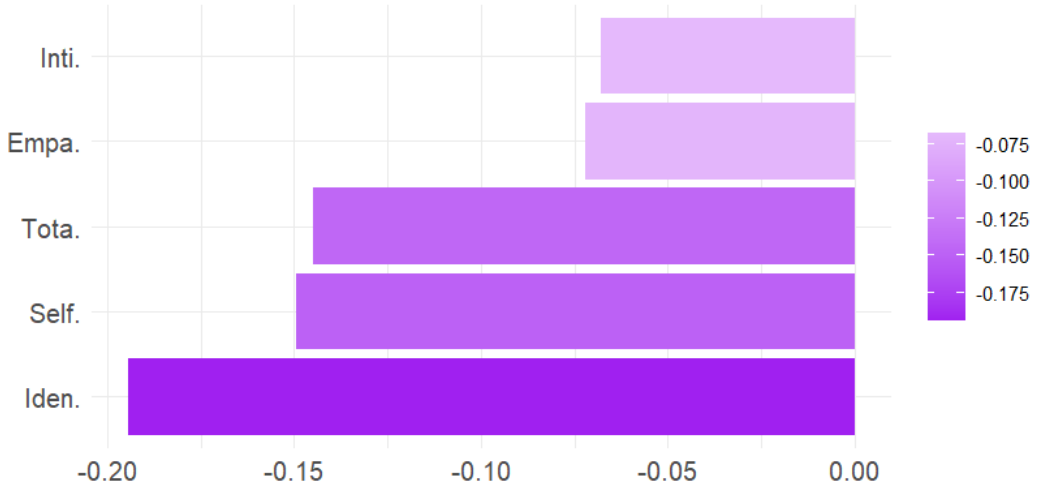


Obr. 16: Korelačná matica pre *pacientov* pre dataset *LPFS* (zdroj: vlastné spracovanie v softvéri R).

V prípade korelácií s premennou **age.** (Obr. 17 a Obr. 18) sa v oboch skupinách ukazuje, že všetky domény vykazujú s rastúcim vekom nižšie hodnoty, čo naznačuje výraznejšie prejavy u mladších respondentov. Tento trend môže súvisieť s ešte neustáleným vývinom osobnosti v mladšom veku. Najsilnejšia negatívna korelácia s vekom je pozorovaná pri doméne **Iden.**, ktorá predstavuje dysfunkciu vnímania sebahodnoty.



Obr. 17: Korelácie domén datasetu *LPFS* s premennou **age.** pre *nepacientov* (zdroj: vlastné spracovanie v softvéri R).



Obr. 18: Korelácie domén datasetu *LPFS* s premennou **age.** pre *pacientov* (zdroj: vlastné spracovanie v softvéri R).

CFA

Model CFA pre dotazník *LPFS-SR* je uvedený v prílohe D. Výsledky analýzy zhody modelu (tabuľka 14) sú konzistentné s výsledkami získanými pre dotazník *PID-5*. Podľa p-hodnoty χ^2 testu a indexu *CFI* model vykazuje slabšiu zhodu s dátami, zatiaľ čo hodnoty *RMSEA* a *SRMR* naznačujú prijateľnú až dobrú zhodu.

Výsledky získané na našej populácii porovnáme s výsledkami uvedenými vo výskume [8], ktorý bol realizovaný na iránskej vzorke respondentov. V prípade iránskej verzie dotazníka *LPFS-SR* boli zaznamenané hodnoty *CFI* = 0,975 a *RMSEA* = 0,056. V porovnaní s týmto výskumom sú naše indexy zhody modelu nižšie, pričom najvýraznejší rozdiel sa prejavuje pri parametri *CFI*.

Tabuľka 14: Indexy celkovej zhody modelu dotazníka *LPFS-SR*.

χ^2 ***(p-hodnota)***	**RMSEA** ***(90 % IS)***	**SRMR**	**CFI**	**PNFI**
10556,549 (<0,001)	0,053 (0,052; 0,054)	0,068	0,670	0,576

V doméne **Iden.** sa ako najproblematickejšie položky ukazujú *Q7, Q8, Q40, Q65, Q76* (tabuľka 26, príloha D). V doméne **Self.** možno ako najmenej reprezentatívne položky

označiť *Q1, Q14, Q22, Q34, Q51, Q64*. Najviac problematických položiek v prípade dotazníka *LPFS-SR* vykazujú domény **Empa.** (*Q2, Q11, Q17, Q20, Q38, Q44, Q61, Q68*) a **Inti.** (*Q12, Q13, Q27, Q37, Q39, Q43, Q45, Q69*). Ukazuje sa, že práve tieto dve domény predstavujú závažný problém dotazníka, keďže približne polovica všetkých položiek je problematická.

Veľmi nízke hodnoty R^2 sú zaznamenané pri položkách *Q40*: *„Jsem si jistý/á rozdílem mezi svými hodnotami a hodnotami, které ostatní chtějí abych měl/a.“*, *Q44*: *„Jsem si velmi dobře vědom/a, jak působím na druhé lidi.“* a *Q76*: *„Když mám pocit, že jsem něco udělal/a dobře, mám téměř vždy pravdu.“* Rovnako ako pri dotazníku *PID-5* to súvisí pravdepodobne s tým, že uvedené položky sú reverzne kódované.

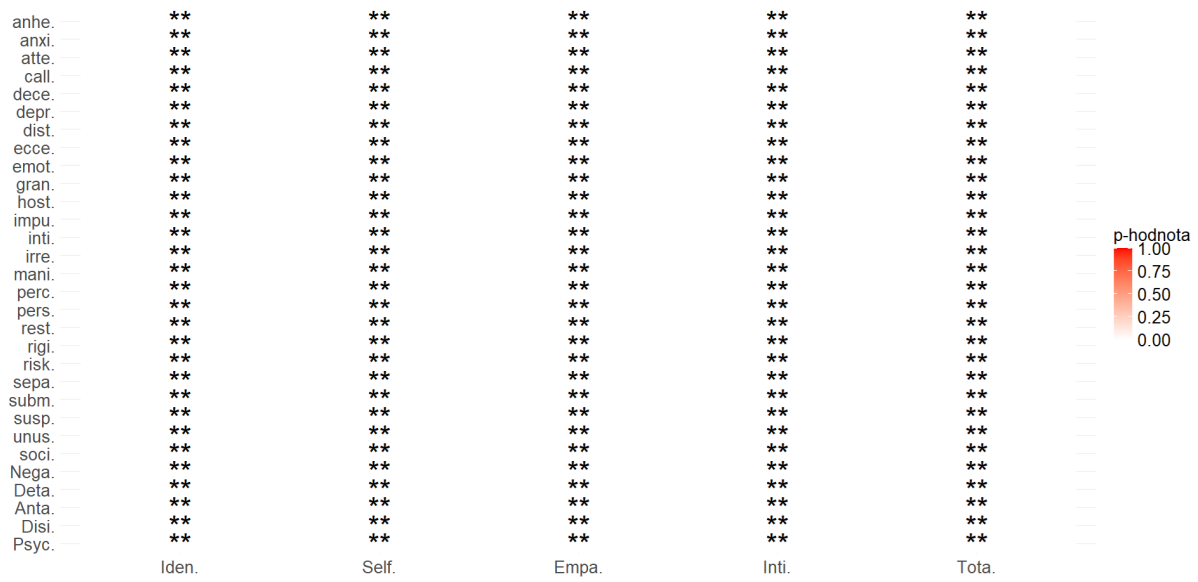
3.4.4 Vzťahy medzi dotazníkmi

Na záver analýzy validity dát sme sa zamerali na vzťahy medzi dotazníkmi *PID-5* a *LPFS-SR*. Vypočítali sme Pearsonove korelačné koeficienty medzi aspektmi a doménami dotazníka *PID-5* a doménami dotazníka *LPFS-SR*.

Pre lepšiu ilustráciu sme výsledky vizualizovali pomocou teplotnej mapy, pričom štatisticky významné korelácie boli označené hviezdičkami: jedna hviezdička pre p-hodnoty $< 0{,}05$ a dve hviezdičky pre p-hodnoty $< 0{,}01$. Pri pohľade na výsledky znázornené na Obr. 19 možno pozorovať, že všetky korelácie medzi dotazníkmi *PID-5* a *LPFS-SR* sú štatisticky významne odlišné od nuly. Podobné výsledky sme získali aj po rozdelení dát na skupinu nepacientov (Obr. 42, príloha F) a pacientov (Obr. 43, príloha F), pričom štatistická významnosť korelácií bola zachovaná vo väčšine domén v oboch skupinách.

Hodnoty korelácií získané z nášho dátového súboru sme porovnali s hodnotami korelácií uvedenými v článku [37], ktorý zahŕňal celkovo 44 štúdií. Z týchto štúdií bolo 27 vzoriek neklinických, 13 klinických a 7 zmiešaných (kombinácia klinických a neklinických respondentov), pričom veľkosti vzoriek sa pohybovali od $n = 88$ do $n = 3,019$. Používali sa rôzne verzie dotazníka *PID-5* a viacero typov dotazníkov *LPFS-SR* (najčastejšie *LPFS-BF* a *LPFS-SR*). Hodnoty jednotlivých korelácií boli definované ako priemerné hodnoty vypočítané z jednotlivých štúdií.

Pri pohľade na hodnoty korelácií medzi datasetmi *PID-5_ domény* a *LPFS* (tabuľka 15) sa ukázalo, že domény **Nega.**, **Disi.** a **Psyc.** vykazujú vo všetkých prípadoch korelácie



Obr. 19: Štatistická významnosť korelácií medzi dotazníkmi *PID-5* a *LPFS-SR* (zdroj: vlastné spracovanie v softvéri R).

vyššie alebo rovné hornej hranici 95 % intervalu spoľahlivosti uvádzaného v [37]. Hodnoty korelácií pre pacientov a nepacientov spolu sa pohybovali približne v rozmedzí 0,37 až 0,76, pričom väčšina korelácií presahovala hodnotu 0,50, s výnimkou domény **Anta.** Zároveň sa ukázalo, že vo väčšine prípadov sú korelácie vyššie v skupine nepacientov v porovnaní so skupinou pacientov. Vo výskume [37] sa priemerné hodnoty korelácií pohybovali v rozmedzí 0,44 až 0,64, pričom, podobne ako v našom prípade, najmenšie hodnoty dosiahla doména **Anta.**

V prípade aspektov *PID-5* sa ukázalo, že korelácie presahujúce hornú hranicu 95 % intervalu spoľahlivosti uvádzaného v [37] boli zaznamenané pri aspektoch **anxi.**, **depr.**, **dist.**, **ecce.**, **irre.** a **perc.** Práve tieto aspekty vstupujú do domén v datasete *PID-5_domény*, ktoré taktiež vykazovali zvýšené hodnoty korelácií v porovnaní s ostatnými. Hodnoty jednotlivých korelácií sa pohybovali v širšom intervale približne od 0,21 do 0,77, pričom opäť viac ako polovica z nich presahovala hodnotu 0,5. Opätovne sa potvrdilo, že vo väčšine prípadov boli korelácie vyššie v skupine nepacientov. V prípade výskumu [37] sa hodnoty pohybovali približne v rozmedzí 0,25 až 0,68, pričom vyššie hodnoty boli zaznamenané rovnako ako v našom prípade pri aspektoch **anxi.** a **depr.** Hodnoty korelácií a príslušné intervaly spoľahlivosti sú uvedené v prílohe E.

Tabuľka 15: Korelácie medzi doménami dotazníkov *PID-5* a *LPFS-SR*.

Doména		**Iden.**	**Self.**	**Empa.**	**Inti.**	**Tota.**
Nega.	Pac.	0,70	0,60	0,48	0,55	0,67
	Nep.	0,73	0,57	0,49	0,54	0,67
	Spolu	0,76	0,65	0,57	0,62	0,72
	95 % IS[*]	*[0,62; 0,70]*	*[0,48; 0,57]*	*[0,37; 0,44]*	*[0,39; 0,49]*	*[0,59; 0,67]*
Deta.	Pac.	0,39	0,50	0,42	0,51	0,51
	Nep.	0,53	0,56	0,56	0,62	0,63
	Spolu	0,55	0,60	0,57	0,63	0,65
	95 % IS[*]	*[0,50; 0,59]*	*[0,44; 0,52]*	*[0,41; 0,50]*	*[0,56; 0,65]*	*[0,60; 0,68]*
Anta.	Pac.	0,28	0,19	0,34	0,31	0,31
	Nep.	0,41	0,42	0,44	0,41	0,47
	Spolu	0,41	0,37	0,45	0,42	0,45
	95 % IS[*]	*[0,37; 0,52]*	*[0,29; 0,37]*	*[0,39; 0,46]*	*[0,29; 0,38]*	*[0,37; 0,50]*
Disi.	Pac.	0,64	0,63	0,50	0,52	0,65
	Nep.	0,69	0,73	0,58	0,55	0,72
	Spolu	0,73	0,75	0,64	0,63	0,76
	95 % IS[*]	*[0,44; 0,54]*	*[0,50; 0,60]*	*[0,37; 0,47]*	*[0,31; 0,42]*	*[0,48; 0,60]*
Psyc.	Pac.	0,64	0,53	0,57	0,54	0,64
	Nep.	0,63	0,57	0,63	0,54	0,66
	Spolu	0,69	0,62	0,67	0,62	0,71
	95 % IS[*]	*[0,45; 0,53]*	*[0,39; 0,47]*	*[0,42; 0,50]*	*[0,39; 0,47]*	*[0,52; 0,62]*

[*] Hodnoty 95 % intervalov spoľahlivosti sú uvedené podľa [37].

4 Štatistické modely

Kapitola *Štatistické modely* sa zaoberá aplikáciou rôznych štatistických modelov na získané dáta. V úvode sme sa zamerali na jednu z metód extrakcie premenných, konkrétne na *PCA*. Následne sme sa venovali metódam selekcie premenných, pomocou ktorých sme z datasetu *PID-5_ aspekty* vybrali súbor najrelevantnejších aspektov. Po získaní tohto redukovaného súboru sme na dáta aplikovali rôzne klasifikačné modely. Cieľom kapitoly bolo nielen porovnanie jednotlivých klasifikačných metód, ale aj identifikácia osobnostných čŕt, ktoré najviac prispievajú k rozlišovaniu medzi skupinami v rámci analyzovaných dát.

4.1 Metóda hlavných komponentov

Prvou metódou, ktorú sme aplikovali na získané dáta, bola metóda hlavných komponentov *(PCA)*. Naším cieľom bolo redukovať počet premenných pri zachovaní čo najväčšej variability dát a identifikovať hlavné faktory, ktoré sme neskôr použili ako vstup do modelu logistickej regresie. Okrem toho sme sa venovali aj vizualizácii dát v dvojrozmernom priestore, ktorá nám umožňuje lepšie pochopiť vzťahy medzi jednotlivými premennými.

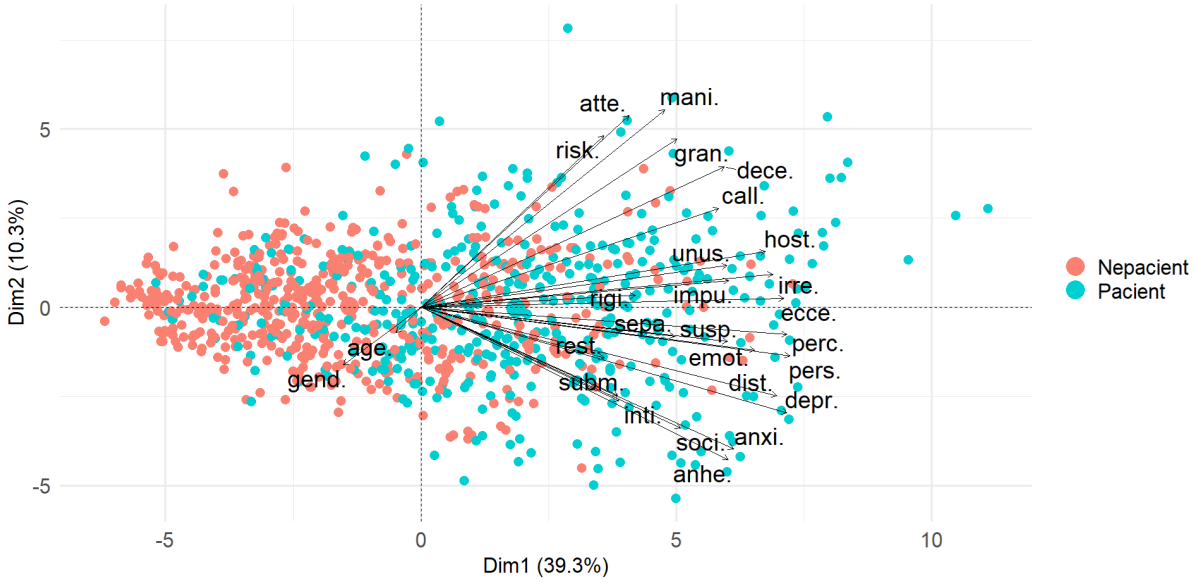
PID-5_aspekty

Ako prvé sme sa pozreli na aplikáciu PCA pre dataset *PID-5_ aspekty*. Na základe prvého hlavného komponentu (*PC1*) sa ukázalo, že všetky aspekty majú kladné váhy (v rozmedzí 0,12 až 0,25), zatiaľ čo premenné **gend.** a **age.** sú takmer nulové so záporným znamienkom. Tento komponent možno interpretovať ako celkovú klinickú záťaž, pričom rozdeľuje respondentov na tých, ktorí dosahujú zvýšené hodnoty jednotlivých aspektov, a na tých, ktorí ich majú nízke.

Pri druhom hlavnom komponente (*PC2*) bolo možné medzi jednotlivými aspektmi pozorovať rozdelenie váh na kladné a záporné. Výraznejšie kladné váhy vykazovali aspekty ako **atte.** (0,35), **call.** (0,19), **dece.** (0,27), **gran.** (0,33), **mani.** (0,37) a **risk.** (0,33). Práve tieto aspekty charakterizujú osobnostné črty spojené s uprednostňovaním vlastných záujmov a vyhľadávaním pozornosti na úkor empatie, pravdivosti a zodpovednosti voči druhým. Naopak, záporné váhy sa vyskytovali pri aspektoch **anhe.** (-0,27), **anxi.** (-0,28), **depr.** (-0,20), **dist.** (-0,17), **inti.** (-0,18), **subm.** (-0,17) a **soci.** (-0,23), čo je spojené s aspektmi reprezentujúcimi znížený schopnosť prežívať radosť sprevádzanú

úzkosťou, pesimizmom a tendenciou vyhýbať sa sociálnym väzbám. Vizualizáciu prvých dvoch hlavných komponentov môžeme pozorovať na Obr. 20.

Pri porovnaní hlavných komponentov s korelačnými maticami (Obr. 9 a Obr. 10) sa ukázalo, že všetky aspekty s kladnou váhou v *PC1* sú medzi sebou pozitívne korelované. Zároveň biplot potvrdzuje, že tie aspekty, ktoré sme v časti 3.4.1 identifikovali ako premenné vytvárajúce spoločné zhluky a vykazujúce silnú vzájomnú koreláciu, sa v biplote nachádzajú v tesnej blízkosti. Na základe rozdelení váh v *PC2* možno pozorovať, že aspekty s kladnou váhou medzi sebou vykazujú vyššiu mieru korelácie. Naopak, aspekty so zápornou váhou vykazujú len slabú koreláciu s aspektmi s kladnou váhou.



Obr. 20: Biplot dvoch hlavných komponentov pre dataset *PID-5_ aspekty* (zdroj: vlastné spracovanie v softvéri R).

PID-5_ domény

V prípade datasetu *PID-5_ domény* sme mohli na základe *PC1* opätovne pozorovať, že všetky klinické domény majú kladné váhy, zatiaľ čo premenné **gend.** a **age.** záporné.

Kladné váhy pri *PC2* dosiahli, okrem premennej **age.** (0,65), aj domény **Anta.** (0,12) a **Deta.** (0,21), ktoré sú spojené s egocentrickým správaním, nižšou mierou empatie a oslabenými sociálnymi väzbami. Záporné váhy nadobudli premenná **gend.** (-0,66) a domény **Disi.** (-0,05), **Nega.** (-0,27) a **Psyc.** (-0,05), ktoré poukazujú na impulzívne správanie, zvýšené prežívanie negatívnych emócií a tendenciu k neštandardnému vnímaniu reality.

Ak sa aj v tomto prípade pozrieme na vizualizáciu prvých dvoch hlavných komponentov (Obr. 21), v pravej časti grafu možno pozorovať skupinu respondentov so zvýšeným skóre jednotlivých domén, ktorú tvoria prevažne skutoční pacienti. Naopak, v ľavej časti sa nachádza skupina osôb s nižším skóre, ktorá je z väčšej časti zastúpená nepacientmi.

V hornej časti grafu, kde sú zastúpení pacienti aj nepacienti, možno pozorovať prevažne mužov vo vyššom veku, u ktorých sa prejavuje vyššia miera sociálnej izolácie (výraznejšia kladná váha pri **Deta.**). Naopak, v dolnej časti grafu, kde prevažujú nepacienti, ide prevažne o mladšie ženy so zvýšenou úzkosťou (výraznejšia záporná váha pri **Nega.**).



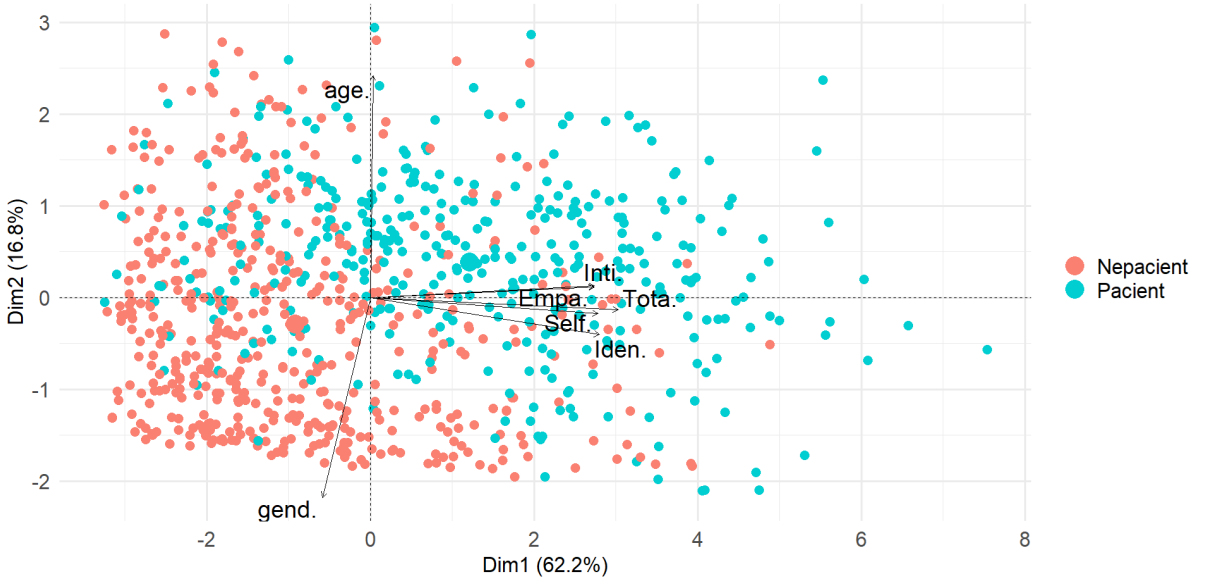
Obr. 21: Biplot dvoch hlavných komponentov pre dataset *PID-5_ domény* (zdroj: vlastné spracovanie v softvéri R).

LPFS

Aj v prípade datasetu *LPFS* sme mohli na základe prvého komponentu pozorovať, že všetky merané domény dosiahli kladné váhy. Rovnako aj na základe biplotu zobrazeného na Obr. 22 vidíme, že skutoční pacienti majú spravidla zvýšené skóre jednotlivých domén v porovnaní so skupinou nepacientov, čo je konzistentné s interpretáciou daného komponentu.

Na základe komponentu *PC2* dosiahli kladné váhy premenná **age.** (0,73) a domény **Empa.** (0,04) a **Inti.** (0,04), zatiaľ čo záporné váhy nadobudli **gend.** (-0,66) a domény **Iden.** (-0,12), **Self.** (-0,05) a **Tota.** (-0,04). Komponent *PC2* tak predovšetkým odráža

rozdiel medzi vekom a pohlavím respondentov, pričom príspevok ostatných domén je vzhľadom na nízke váhy len minimálny.



Obr. 22: Biplot dvoch hlavných komponentov pre dataset *LPFS* (zdroj: vlastné spracovanie v softvéri R).

4.2 Metódy selekcie premenných

V tejto časti sme sa zamerali na výber aspektov pre dataset *PID-5_ aspekty* prostredníctvom krokových metód a metódy LASSO. Naším cieľom bolo z 25 osobnostných aspektov identifikovať redukovaný súbor aspektov, ktorý dostatočne reprezentuje celý súbor aspektov. Ukážka kódu je uvedená v prílohe J.

Dáta sme rozdelili na trénovaciu (80 %) a testovaciu (20 %) vzorku. Počet dát v trénovacej a testovacej vzorke pre jednotlivé skupiny je uvedený v tabuľke 16.

Tabuľka 16: Počet dát v trénovacej a testovacej vzorke.

Vzorka	**Nepacienti / Pacienti**
Trénovacia	442 / 333
Testovacia	111 / 83

4.2.1 Krokové metódy

Z *krokových metód* sme použili metódy selekcie premenných založené na spätnej selekcii, pričom sme uplatnili dve informačné kritériá *AIC* a *BIC*. Na základe selektovaných premenných sme sa následne pokúsili vytvoriť model logistickej regresie, pričom sme jednotlivé modely medzi sebou porovnali.

Ako prvé sme aplikovali na trénovaciu vzorku dát spätnú selekciu pomocou kritéria *AIC*. Na základe tejto selekcie bolo vybraných **14 aspektov** spolu s premennými **gend.** a **age.**, pričom finálna hodnota kritéria *AIC* dosiahla **625,42**. V prípade selekcie pomocou kritéria *BIC* bolo vybraných len **5 aspektov** s hodnotou *BIC* **638,10**. Pre oba modely sme následne vytvorili predikcie modelu logistickej regresie na základe testovacej vzorky, pričom finálne modely sme vyhodnotili pomocou metrík uvedených v tabuľke 17.

Tabuľka 17: Porovnanie výkonnosti modelov vytvorených na základe spätnej selekcie.

Metrika	AIC	BIC
Presnosť	0,794	0,789
Senzitivita	0,829	0,793
Špecificita	0,747	0,783
AUC	0,890	0,870

Z tabuľky 17 vyplýva, že rozdiel vo výkonnosti medzi porovnávanými modelmi je minimálny. Na základe získaných výsledkov, ako aj vzhľadom na jednoduchosť a lepšiu interpretovateľnosť, by sme z týchto dvoch modelov uprednostnili aspekty vybrané spätnou selekciou na základe kritéria *BIC*.

4.2.2 LASSO

V tejto časti sme sa, podobne ako v časti 4.2.1, pokúsili skonštruovať model logistickej regresie využitím metódy LASSO. Pri výbere aspektov prostredníctvom tejto metódy sme na určenie optimálnej hodnoty parametra λ použili dva prístupy: parameter λ_{min}, ktorý zvyčajne vedie k presnejšiemu, avšak komplexnejšiemu modelu, a parameter λ_{1se}, ktorý má tendenciu vytvoriť jednoduchší a robustnejší model [20].

Pri voľbe parametra λ_{min} model ponechal až **18 aspektov** spolu s premennými **gend.** a **age.** V prípade parametra λ_{1se} bolo vyselektovaných **7 aspektov**. V oboch prípadoch sme po selekcii na vybraných aspektoch natrénovali model logistickej regresie. Výkonnosť oboch modelov sme následne vyhodnotili na základe metrík uvedených v tabuľke 18.

Tabuľka 18: Porovnanie výkonnosti modelov vytvorených na základe selekcie premenných metódou LASSO.

Metrika	**LASSO (λ_{min})**	**LASSO (λ_{1se})**
Presnosť	0,799	0,794
Senzitivita	0,834	0,802
Špecificita	0,747	0,783
AUC	0,888	0,874

Ak by sme si mali z týchto dvoch modelov vybrať, z dôvodu jednoduchosti a interpretovateľnosti by sme uprednostnili aspekty získané metódou LASSO s penalizačným parametrom λ_{1se}. Hoci je jeho výkonnosť mierne nižšia ako v prípade modelu s parametrom λ_{min}, model s takto zvolenými aspektmi predstavuje výhodný kompromis medzi výkonnosťou a zjednodušením modelu.

4.2.3 Finálny výber aspektov

Pri porovnaní modelov so selektovanými aspektmi (model spätnej selekcie založený na kritériu *BIC* a model LASSO s parametrom λ_{1se}) sme na základe mierne vyššej presnosti a diskriminačnej schopnosti zvolili redukovaný a zjednodušený súbor aspektov získaný metódou **LASSO s parametrom λ_{1se}**. Na základe selekcie premenných sme získali redukovaný súbor, ktorý budeme ďalej označovať ako *PID-5_ aspekty_ sel*, s nasledujúcimi aspektmi[2]: **gend.**, **age.**, **depr.**, **impu.**, **irre.**, **risk.**, **sepa.**, **susp.** a **soci.**

Ak sa zameriame na aspekty vybrané modelom, možno pozorovať, že model zvolil jednotlivé aspekty tak, aby reprezentovali jednotlivé zhluky identifikované v korelačných maticiach v časti 3.4.1. Zároveň sa ukázalo, že tieto aspekty boli vybrané aj ostatnými metódami selekcie premenných. To naznačuje, že práve tieto aspekty zohrávajú v rámci da-

[2] *Poznámka:* Vybrané aspekty získané pomocou všetkých metód selekcie sú uvedené v prílohe G.

tasetu *PID-5_ aspekty* kľúčovú úlohu pri klasifikácii. Z tohto dôvodu sme v nasledujúcich analýzach pracovali už iba so získaným súborom *PID-5_ aspekty_ sel* namiesto všetkých aspektov.

4.3 Logistická regresia

Ako prvý klasifikačný model sme vytvorili model *logistickej regresie*.

PID-5_aspekty_sel

Model logistickej regresie má tvar:

$$\ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 \cdot \text{gend.} + \beta_2 \cdot \text{age.} + \beta_3 \cdot \text{depr.} + \beta_4 \cdot \text{impu.} + \beta_5 \cdot \text{irre.} + \beta_6 \cdot \text{risk.} + \beta_7 \cdot \text{sepa.} + \beta_8 \cdot \text{susp.} + \beta_9 \cdot \text{soci.} \tag{28}$$

Pre robustné odhady parametrov a validáciu modelu sme použili metódu *bootstrap* s 1000 iteráciami. Do modelu logistickej regresie sme zaviedli váhy tak, aby zodpovedali skutočnému podielu pacientov a nepacientov v českej populácii. Váhy sme nastavili na základe článku [5], ktorý uvádza, že v českej populácii sa psychická porucha vyskytuje u *21,9 %* populácie, zatiaľ čo *78,1 %* populácie psychickou poruchou netrpí. Kľúčové časti kódu sú uvedené v prílohe K, pričom pri jeho písaní bola AI využitá len pri riešení čiastkových problémov, napríklad pri nastavovaní váh modelu.

Pre každú trénovaciu vzorku sme odhadli model logistickej regresie a na príslušnej testovacej vzorke sme vypočítali hodnoty výkonnostných metrík spolu s optimálnym prahom *(angl. threshold)*, ktorý predstavuje hranicu medzi klinickou a neklinickou populáciou. Tieto hodnoty sme zaznamenali pri každej iterácii a následne vypočítali ich priemer. Optimálny prah bol stanovený tak, aby maximalizoval súčet senzitivity a špecificity. Získané stredné hodnoty regresných koeficientov spolu s ich 95 % intervalmi spoľahlivosti sú uvedené v tabuľke 19, pričom spriemerované hodnoty jednotlivých metrík sú v tabuľke 20.

Následne sme zostrojili finálny model na celom datasete (bez delenia dát na trénovaciu a testovaciu vzorku), pričom sme upravili premennú **age.** tak, aby v zostavenom modeli referenčnou osobou bol 35-ročný muž s nulovým skóre jednotlivých osobnostných aspektov. Po exponenciálnej transformácii získaných regresných koeficientov sme získali hodnoty pomerov šancí, ktoré sú uvedené v tabuľke 19.

Tabuľka 19: Odhady regresných koeficientov s 95 % intervalmi spoľahlivosti a pomer šancí pre dataset *PID-5_ aspekty_ sel.*

Aspekt	Hodnota	2,5 % kvantil	97,5 % kvantil	Pomer šancí
(Intercept)	-7,271	-8,426	-6,162	0,164
gend.	-1,307	-1,789	-0,881	0,272
age.	0,063	0,047	0,082	1,070
depr.	0,715	0,271	1,171	1,973
impu.	0,339	-0,119	0,826	1,464
irre.	1,200	0,738	1,667	3,032
risk.	0,746	0,353	1,135	1,863
sepa.	0,773	0,474	1,105	2,021
susp.	0,442	-0,034	0,934	1,517
soci.	0,421	0,013	0,836	1,427

Na základe získaných hodnôt vykazujú v rámci modelu najvyššie odhadované pomery šancí aspekty **irre.** a **sepa.** Z toho vyplýva, že práve tieto aspekty by mohli významne prispievať k rozlišovaniu medzi skupinami. Toto pozorovanie je v súlade s našou intuíciou z časti 3.2.2.

Môžeme pozorovať, že do finálneho modelu boli vybrané aspekty, ktoré medzi sebou vykazujú slabé až stredne silné korelácie (Obr. 9 a Obr. 10). Následne sme medzi jednotlivými aspektmi overili mieru multikolinearity pomocou ukazovateľa *VIF*, ktorý vyjadruje mieru lineárnej závislosti vysvetľujúcej premennej od ostatných premenných v modeli, pričom hodnoty vyššie ako 5 poukazujú na potenciálny problém [10, s. 102]. Zistené hodnoty *VIF* sa pohybovali v intervale 1,19 až 1,80, čo nenaznačuje prítomnosť významnej multikolinearity.

Regresné koeficienty vo finálnom modeli sme overili pomocou *Waldovho testu*. Ukázalo sa, že aspekty **impu.** a **susp.** nevykazujú štatisticky významný prínos v rámci modelu. Následne sme využili funkciu *drop1* (Obr. 23), pomocou ktorej sme porovnali modely po postupnom vynechaní jednotlivých aspektov. Hodnota miery nekvality pre úplný model je 795,06 a hodnota AIC 815,06. Analýza pomocou tejto funkcie naznačila, že vynechanie

aspektov **impu.**, **susp.** a **soci.** vedie len k minimálnemu zhoršeniu modelu.

```
Model:
Patient ~ Gender + Age + Depressivity + Impulsivity + Irresponsibility +
    RiskTaking + SeparationInsecurity + Suspiciousness + SocialWithdrawal
                     Df Deviance    AIC    LRT  Pr(>Chi)
<none>                    795.06 815.06
Gender                1   841.80 859.80 46.738 8.113e-12 ***
Age                   1   876.14 894.14 81.082 < 2.2e-16 ***
Depressivity          1   805.05 823.05  9.993 0.0015714 **
Impulsivity           1   798.44 816.44  3.379 0.0660375 .
Irresponsibility      1   822.76 840.76 27.704 1.414e-07 ***
RiskTaking            1   806.61 824.61 11.550 0.0006776 ***
SeparationInsecurity  1   815.27 833.27 20.211 6.935e-06 ***
Suspiciousness        1   798.42 816.42  3.359 0.0668243 .
SocialWithdrawal      1   798.94 816.94  3.880 0.0488630 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Obr. 23: Výstup funkcie *drop1* pre dataset *PID-5_ aspekty_ sel* (zdroj: vlastné spracovanie v softvéri R).

Na základe predchádzajúcej analýzy sme otestovali submodel, v ktorom sme vynechali práve aspekty **impu.**, **susp.** a **soci.** Výsledná p-hodnota (0,004) ukázala, že ich spoločné vynechanie vedie k štatisticky významnému zhoršeniu modelu. Tento výsledok naznačuje, že hoci jednotlivé aspekty nemusia byť samostatne štatisticky významné, ich spoločný príspevok k modelu nie je zanedbateľný.

Okrem modelu s vybranými aspektmi sme vytvorili aj model obsahujúci všetkých 25 aspektov, pričom sme získali priemerné hodnoty jednotlivých metrík uvedené v tabuľke 20. Zároveň sme analyzovali model, v ktorom sme ako vstupné premenné použili hlavné komponenty získané metódou PCA. Na určenie vhodného počtu hlavných komponentov sme aplikovali *Kaiserovo pravidlo*, podľa ktorého sme do analýzy zahrnuli prvých 11 komponentov.

Po porovnaní všetkých modelov sa model s vybranými aspektmi ukázal ako najoptimalnejší. Dosiahol porovnateľnú výkonnosť ako ostatné modely, pričom jeho hlavnou výhodou je jednoduchosť a lepšia interpretovateľnosť.

Tabuľka 20: Porovnanie modelu logistickej regresie s vybranými aspektami *PID-5_ aspekty_ sel* (**Model 1**), modelu so všetkými aspektami *PID-5_ aspekty* (**Model 2**) a modelu založeného na komponentoch získaných metódou PCA (**Model 3**).

Metrika	Model 1	Model 2	Model 3
Presnosť	79,94 %	79,98 %	78,86 %
AUC	88,39 %	88,16 %	87,39 %
Senzitivita	78,91 %	77,62 %	75,48 %
Špecificita	80,70 %	81,75 %	81,39 %
Optimálny prah	0,230	0,235	0,252

PID-5_ domény

Po vytvorení modelu logistickej regresie pre dataset *PID-5_ domény* sme získali model, ktorý správne klasifikoval 77,19 % prípadov. Model zároveň dosiahol schopnosť rozlišovať medzi jednotlivými skupinami na úrovni 85,77 %, pričom správne identifikoval pacientov v 74,05 % prípadov a nepacientov 79,56 % prípadov.

Následne bol zostrojený model logistickej regresie na celom datasete. Na základe p-hodnôt jednotlivých regresných koeficientov sa ukázalo, že domény **Deta.** (0,311), **Anta.** (0,696) a **Psyc.** (0,331) nemajú štatisticky významný prínos pre model. Rovnaký záver potvrdil aj výstup funkcie *drop1* (Obr. 24).

```
Model:
Patient ~ Gender + Age + Negativeaffectivity + Detachment + Antagonism +
    Disinhibition + Psychoticism
                   Df Deviance    AIC    LRT  Pr(>Chi)
<none>                  884.24 900.24
Gender              1   965.04 979.04 80.803 < 2.2e-16 ***
Age                 1   957.56 971.56 73.327 < 2.2e-16 ***
Negativeaffectivity 1   899.86 913.86 15.628 7.709e-05 ***
Detachment          1   885.26 899.26  1.027    0.3109
Antagonism          1   884.39 898.39  0.152    0.6963
Disinhibition       1   923.66 937.66 39.423 3.412e-10 ***
Psychoticism        1   885.18 899.18  0.947    0.3306
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obr. 24: Výstup funkcie *drop1* pre dataset *PID-5_ domény* (zdroj: vlastné spracovanie v softvéri R).

Následne sme otestovali hypotézu, či je submodel s vynechanými doménami posta-

čujúci. Výsledná p-hodnota (0,396) ukázala, že zjednodušený model je dostatočný. Po natrénovaní nového modelu bez uvedených domén sme získali spriemerované hodnoty jednotlivých metrík uvedené v tabuľke 22. Vzhľadom na porovnateľnú výkonnosť zjednodušeného modelu a jeho väčšiu jednoduchosť bol ako finálny zvolený model:

$$\ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 \cdot \text{gend.} + \beta_2 \cdot \text{age.} + \beta_3 \cdot \text{Nega.} + \beta_4 \cdot \text{Disi.} \tag{29}$$

Odhadnuté hodnoty regresných koeficientov finálneho modelu sú uvedené v tabuľke 21.

Tabuľka 21: Odhady regresných koeficientov s 95 % intervalmi spoľahlivosti a pomer šancí pre dataset *PID-5_ domény* (finálny model).

Doména	**Hodnota**	**2,5 % kvantil**	**97,5 % kvantil**	**Pomer šancí**
(Intercept)	-5,410	-6,310	-4,578	0,088
gend.	-1,595	-1,986	-1,229	0,199
age.	0,055	0,040	0,072	1,064
Nega.	0,965	0,560	1,374	2,497
Disi.	1,910	1,497	2,362	6,801

Na základe získaných pomerov šancí (tabuľka 21) sa ukazuje, že najvýraznejší nárast šance pri zvýšení skóre sa prejavil pri doméne **Disi.** Na základe toho môžeme predpokladať, že práve táto doména má najväčší vplyv na pravdepodobnosť prítomnosti poruchy osobnosti. Tento záver zároveň podporuje našu intuíciu vytvorenú na základe vizualizácie jednotlivých domén v časti 3.2.1.

Za významnú možno považovať aj doménu **Nega.**, keďže ju spolu s doménou **Disi.** nie je možné z modelu vynechať. Hoci medzi doménami **Nega.** a **Disi.** bola zistená pomerne silná pozitívna korelácia (0,63 u nepacientov / 0,65 u pacientov), analýza multikolinearity (VIF < 2) nepreukázala jej významnú úroveň v modeli.

Rovnako ako v prípade datasetu *PID-5_ aspekty* sme vytvorili model založený na hlavných komponentoch, pričom podľa *Kaiserovho pravidla* boli do modelu zahrnuté prvé štyri komponenty. Výsledky naznačujú, že model využívajúci PCA nedosiahol vyššiu presnosť ani diskriminačnú schopnosť než model (29) (tabuľka 22).

Tabuľka 22: Porovnanie finálneho modelu logistickej regresie s doménami *PID-5_ domény* (**Model 1**) a modelu založeného na komponentoch získaných metódou PCA (**Model 2**).

Metrika	Model 1	Model 2
Presnosť	77,48 %	76,19 %
AUC	86,04 %	84,35 %
Senzitivita	73,29 %	75,80 %
Špecificita	80,61 %	76,49 %
Optimálny prah	0,247	0,223

LPFS

Na záver tejto časti sme zostrojili model logistickej regresie pre dataset *LPFS*, pričom doména **Tota.** bola z dôvodu redundancie vylúčená, keďže ide o agregovanú doménu pozostávajúcu zo všetkých štyroch domén. Takto vytvorený model dosiahol priemernú presnosť 77,17 %, diskriminačnú schopnosť 84,66 %, senzitivitu 75,80 % a špecificitu 78,19 %.

Analýza p-hodnôt a výstupu funkcie *drop1* (Obr. 25) ukázala, že domény **Empa.** a **Inti.** by bolo vhodné z modelu vynechať. Testovanie hypotézy o submodeli (p-hodnota = 0,058) neposkytlo štatisticky významný dôkaz o zhoršení modelu, čo podporuje použitie jednoduchšieho modelu. Na základe toho sme získali finálny model logistickej regresie:

$$\ln\left(\frac{p_x}{1-p_x}\right) = \beta_0 + \beta_1 \cdot \text{gend.} + \beta_2 \cdot \text{age.} + \beta_3 \cdot \text{Iden.} + \beta_4 \cdot \text{Self.} \tag{30}$$

```
Model:
Patient ~ Gender + Age + Identity + Selfdirection + Empathy +
    Intimacy
               Df Deviance    AIC    LRT  Pr(>Chi)
<none>              821.23 835.23
Gender          1   883.06 895.06 61.827 3.750e-15 ***
Age             1   874.83 886.83 53.599 2.458e-13 ***
Identity        1   834.26 846.26 13.033  0.000306 ***
Selfdirection   1   827.89 839.89  6.657  0.009878 **
Empathy         1   823.95 835.95  2.722  0.099001 .
Intimacy        1   822.13 834.13  0.900  0.342772
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Obr. 25: Výstup funkcie *drop1* pre dataset *LPFS* (zdroj: vlastné spracovanie v softvéri R).

Na takto vytvorený finálny model sme aplikovali metódu *bootstrap*, pričom odhadnuté hodnoty jednotlivých regresných koeficientov môžeme pozorovať v tabuľke 23 a sprieme-rované hodnoty metrík v tabuľke 24.

Tabuľka 23: Odhady regresných koeficientov s 95 % intervalmi spoľahlivosti a pomer šancí pre dataset *LPFS* (finálny model).

Doména	Hodnota	2,5 % kvantil	97,5 % kvantil	Pomer šancí
(Intercept)	-6,996	-8,156	-5,915	0,015
gend.	-1,358	-1,730	-1,018	0,250
age.	0,051	0,034	0,068	1,058
Iden.	0,034	0,022	0,047	1,034
Self.	0,030	0,013	0,047	1,031

Na základe finálneho modelu možno medzi najdôležitejšie domény pri klasifikácii da-tasetu *LPFS* zaradiť **Iden.** a **Self.**, keďže ich odstránenie z modelu by viedlo k zhoršeniu jeho kvality. Model logistickej regresie tak opätovne potvrdil zistenia uvedené v časti 3.2.3. Medzi týmito doménami bola zistená vysoká korelácia (0,80 u nepacientov / 0,78 u pacientov) a zároveň mierna úroveň multikolinearity (2,65 pre **Iden.** / 2,53 pre **Self.**). Pri interpretácii pomerov šancí je preto vhodné zohľadniť vzájomný vzťah týchto domén. Aj v prípade datasetu *LPFS* boli vypočítané metriky pre model založený na hlavných komponentoch, pričom boli použité prvé tri komponenty (tabuľka 24).

Tabuľka 24: Porovnanie finálneho modelu logistickej regresie s doménami *LPFS* (**Model 1**) a modelu založeného na komponentoch získaných metódou PCA (**Model 2**).

Metrika	Model 1	Model 2
Presnosť	76,90 %	77,53 %
AUC	84,69 %	84,69 %
Senzitivita	78,40 %	76,29 %
Špecificita	75,77 %	78,46 %
Optimálny prah	0,212	0,235

4.4 Lineárna diskriminačná analýza

Ďalšou metódou, ktorou sme sa zaoberali, je *lineárna diskriminačná analýza (LDA)*. Okrem vytvorenia samotného modelu a vyhodnotenia jeho výkonnosti bolo naším cieľom určiť, na základe váh získaných z modelu LDA dôležitosť jednotlivých domén pri klasifikácii.

PID-5_aspekty_sel

Pred aplikáciou samotnej metódy LDA sme otestovali predpoklady. Ako prvé sme sa zameriali na overenie normality pomocou *Shapirovho–Wilkovho testu* v oboch skupinách. Na základe výsledkov testu sa ukázalo, že aspekty, a to ako v skupine nepacientov, tak aj v skupine pacientov, nevykazujú normalitu. Napriek tomu sme sa túto metódu rozhodli aplikovať aj na naše dáta.

Ďalej sme sa pozreli na test homogenity kovariančných matíc pomocou *Boxovho M-testu*. Nízka p-hodnota poukázala na významné rozdiely v kovariančných maticiach medzi skupinami. Nakoniec sme pomocou testu rovnosti stredných hodnôt porovnali odhady priemerných hodnôt jednotlivých aspektov v oboch skupinách. Výsledky ukázali, že stredné hodnoty sa medzi skupinami štatisticky významne líšili.

Po otestovaní jednotlivých predpokladov sme na dáta aplikovali model LDA s 10-násobnou krížovou validáciou. V každej iterácii sme natrénovali model s apriórnymi pravdepodobnosťami, ktoré sme nastavili podľa článku [5] na hodnotu pravdepodobnosti patriť do skupiny nepacientov *0,781* a pravdepodobnosti patriť do skupiny pacientov *0,219*. Zároveň sme upravili hodnotu optimálneho prahu tak, aby namiesto predvolenej hodnoty 0,5 maximalizovala súčet senzitivity a špecificity. Pre každý natrénovaný model sme získali hodnoty jednotlivých metrík a následne vypočítali ich priemerné hodnoty.

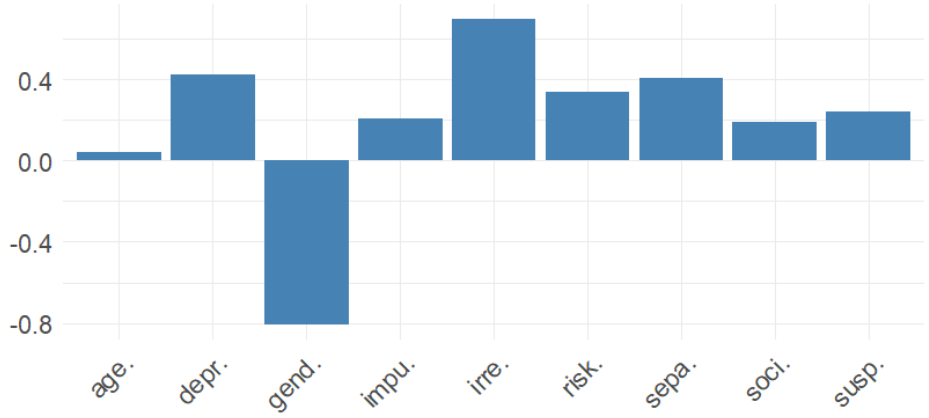
Výsledky krížovej validácie ukázali priemernú presnosť klasifikácie **79,87 %** a priemernú hodnotu AUC **88,47 %**. Zároveň hodnota senzitivity dosiahla **78,02 %**, špecificity **81,26 %** a hodnota optimálneho prahu **0,234**.

Následne sme vytvorili finálny LDA model na celom datasete. Po aplikovaní modelu nás zaujímali práve koeficienty prvého lineárneho diskriminačného komponentu (*LD1*), ktoré predstavujú váhy jednotlivých premenných v diskriminačnej funkcii. Váhy, ktoré

zároveň vyjadrujú dôležitosť jednotlivých aspektov pri klasifikácii, sú nasledovné:

$$\begin{aligned} \text{LD1} = &-0,807 \cdot \text{gend.} + 0,042 \cdot \text{age.} + 0,426 \cdot \text{depr.} + 0,207 \cdot \text{impu.} + 0,699 \cdot \text{irre.} \\ &+ 0,341 \cdot \text{risk.} + 0,407 \cdot \text{sepa.} + 0,239 \cdot \text{susp.} + 0,190 \cdot \text{soci.} \end{aligned} \tag{31}$$

Zo získaných váh vyplýva, že aspekt **irre.** má najsilnejší pozitívny vplyv na rozlíšenie medzi skupinami. Významnejšie k samotnej klasifikácii prispievajú aj **depr.** a **sepa.**, čo je v súlade s výsledkami modelu logistickej regresie. Výsledné váhy sú vizualizované pomocou stĺpcového grafu (Obr. 26).



Obr. 26: Váhy aspektov v *LD1* pre dataset *PID-5_ aspekty_ sel* (zdroj: vlastné spracovanie v softvéri R).

Po vytvorení LDA modelu s použitím všetkých aspektov datasetu *PID-5_ aspekty* sme získali model s priemernou presnosťou 81,42 %, hodnotou AUC 89,27 %, senzitivitou 79,42 %, špecificitou 82,85 % a optimálnym prahom 0,223. Ukazuje sa, že model s vybranými aspektmi dosiahol mierne nižšiu kvalitu klasifikácie v porovnaní s modelom so všetkými aspektmi. Napriek tomu predstavuje vhodnejšiu voľbu, keďže obsahuje menší počet aspektov.

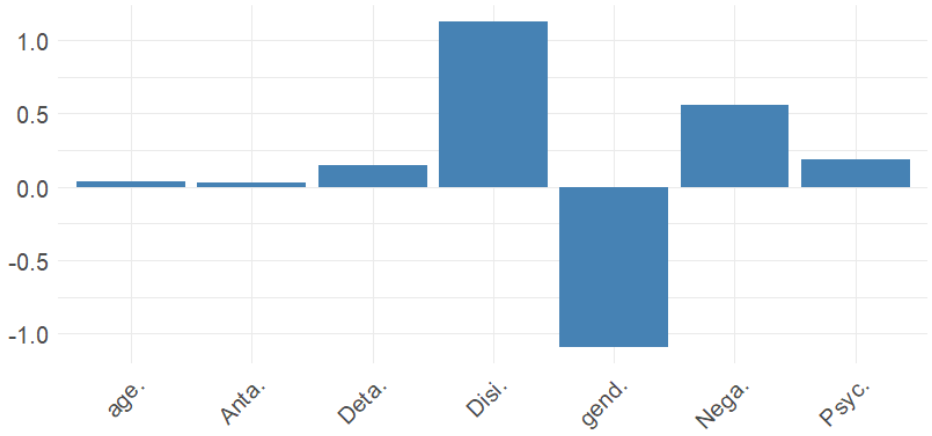
PID-5_ domény

Pred aplikáciou modelu LDA sme overili jednotlivé predpoklady, pričom sme získali rovnaké výsledky ako v prípade *PID-5_ aspekty_ sel*. Po aplikovaní LDA modelu pomocou 10-násobnej krížovej validácie sme získali model s priemernou presnosťou **77,42 %**, priemernou hodnotou AUC **86,16 %**, senzitivitou **76,21 %**, špecificitou **78,11 %** a optimálnym prahom **0,222**.

Po vytvorení finálneho modelu sme získali nasledujúcu diskriminačnú funkciu:

$$\begin{aligned} \text{LD1} = &-1,088 \cdot \text{gend.} + 0,041 \cdot \text{age.} + 0,556 \cdot \text{Nega.} + 0,145 \cdot \text{Deta.} + 0,032 \cdot \text{Anta.} \\ &+ 1,128 \cdot \text{Disi.} + 0,186 \cdot \text{Psyc.} \end{aligned} \tag{32}$$

Na základe jednotlivých váh (Obr. 27), sa ukazuje, že najvyššiu váhu v diskriminačnej funkcii má práve doména **Disi.** Táto doména sa potvrdzuje ako silný rozlišovací faktor aj na základe výsledkov uvedených v časti 4.3. Doména, ktorá naopak najmenej prispieva k rozlišovaniu medzi jednotlivými skupinami podľa váh v komponente *LD1*, je okrem premennej **age.** aj **Anta.**, čo je opäť v súlade s našou intuíciou vyplývajúcou z analýzy v časti 3.2.1.



Obr. 27: Váhy domén v *LD1* pre dataset *PID-5_ domény* (zdroj: vlastné spracovanie v softvéri R).

LPFS

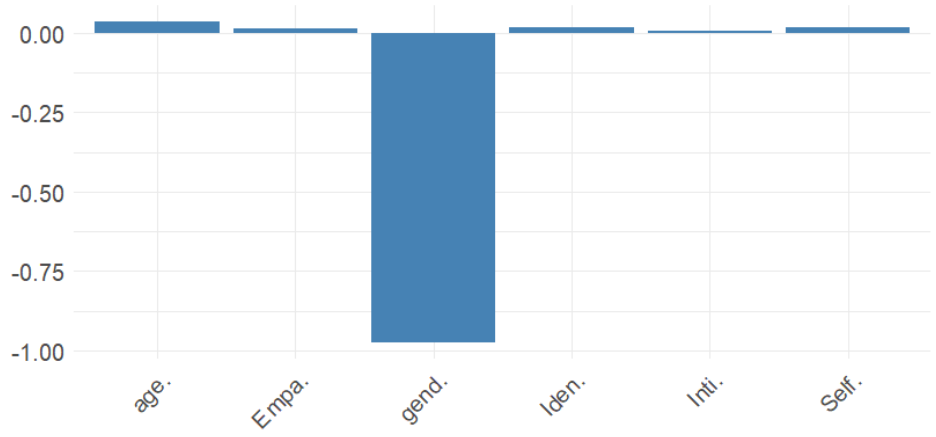
Po aplikovaní modelu LDA, z ktorého sme vynechali doménu **Tota.**, keďže je tvorená kombináciou ostatných domén a jej zaradenie by mohlo spôsobiť kolinearitu, sme získali model s priemernou presnosťou **77,19 %** a priemernou hodnotou AUC **84,93 %**. Senzitivita modelu dosiahla **76,22 %**, zatiaľ čo špecificita bola **77,44 %**. Optimálny prah dosiahol v priemere hodnotu **0,224**.

Následne sme vytvorili finálny model, v ktorom sme na základe komponentu *LD1*

získali diskriminačnú funkciu v tvare:

$$\begin{aligned} \text{LD1} = &-0,976 \cdot \text{gend.} + 0,037 \cdot \text{age.} + 0,019 \cdot \text{Iden.} + 0,016 \cdot \text{Self.} + 0,015 \cdot \text{Empa.} \\ &+ 0,005 \cdot \text{Inti.} \end{aligned} \tag{33}$$

Zo získaných váh jednotlivých domén (Obr. 28) vyplýva, že okrem premennej **gend.** neprispieva k rozlišovaniu medzi skupinami žiadna z domén datasetu *LPFS*. Na základe výsledkov modelu LDA sa neukázala žiadna doména ako významne dôležitá pre samotnú klasifikáciu. Zároveň sa podľa váh modelu ako najmenej prispievajúca ku klasifikácii javí doména **Inti.**



Obr. 28: Váhy domén v *LD1* pre dataset *LPFS* (zdroj: vlastné spracovanie v softvéri R).

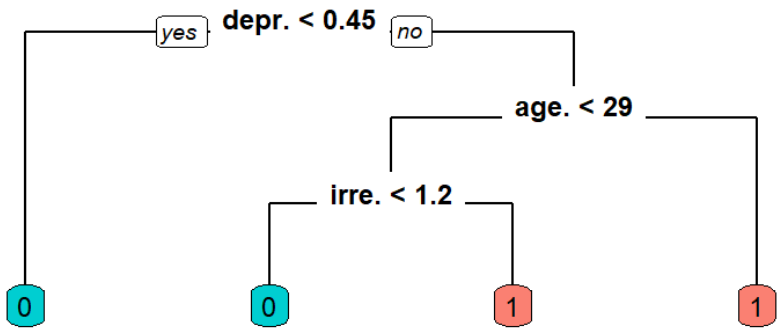
4.5 Klasifikačný strom

Poslednou štatistickou metódou, ktorej sme sa venovali, bol *klasifikačný strom*. Kód pre metódu klasifikačného stromu a metódu LDA je uvedený v prílohe L.

PID-5_aspekty_sel

Ako prvé sme vytvorili základný model, ktorý sme natrénovali na celom datasete. Model sme nastavili tak, aby sa pri delení uzlov používala *miera entropie* namiesto Giniho indexu. Na odhad chyby modelu sme aplikovali 100-násobnú krížovú validáciu. Pri analýze štruktúry stromu sa ukázalo, že ako optimálny sa javí strom s tromi uzlami. V koreňovom uzle dochádza k hlavnému rozdeleniu respondentov na základe aspektu **depr.** Jednotlivé rozhodovacie pravidlá klasifikačného stromu sú vizualizované na Obr. 29.

Ak sa zároveň zameriame na dôležitosť jednotlivých aspektov pri klasifikácii, ukazuje sa, že za najvýznamnejší aspekt bol označený **depr.** Ako ďalší významný aspekt bol identifikovaný **irre.** Uvedené aspekty sa zároveň potvrdili ako vhodné rozlišovacie faktory medzi skupinami aj pri aplikácii ostatných štatistických metód. Naopak, model označil aspekt **risk.** za najmenej významný z hľadiska klasifikácie.



Obr. 29: Klasifikačný strom pre dataset *PID-5_ aspekty_ sel* (zdroj: vlastné spracovanie v softvéri R).

Vzhľadom na známe skutočné rozdelenie skupín v českej populácii (článok [5]) sme okrem základného modelu zostrojili aj model klasifikačného stromu s apriórne nastavenými pravdepodobnosťami. Na vyhodnotenie kvality modelu sme na dáta aplikovali, podobne ako v prípade modelu LDA, 10-násobnú krížovú validáciu. Po vykonaní validácie sme získali model s priemernou presnosťou **79,36 %**, diskriminačnou schopnosťou **80,79 %** a optimálnym prahom **0,254**. Senzitivita dosiahla hodnotu **66,24 %** a špecificita **89,37 %**.

Pri porovnaní výsledkov s modelom, v ktorom bola validácia vykonaná pomocou všetkých aspektov datasetu *PID-5_ aspekty*, sa ukázalo, že presnosť modelu sa mierne znížila na 78,43 %. Rovnako sa znížila aj hodnota AUC (79,23 %) a senzitivity (65,23 %), špecificity (88,47 %) a optimálneho prahu (0,236). Na základe získaných metrík vyplýva, že model využívajúci všetky aspekty nevykazuje lepšiu výkonnosť v porovnaní s modelom s vybranými aspektmi.

PID-5_ domény

Po vytvorení základného modelu aplikovaného na celý dataset *PID-5_ domény* sa ako

optimálne ukázalo rozdelenie stromu na 12 uzlov. Hlavné rozdelenie v koreňovom uzle bolo založené na doméne **Disi.** Klasifikačný strom je znázornený na Obr. 30.

Model označil ako najdôležitejšie domény pri klasifikácii **Disi.** a **Nega.** Ich význam potvrdzujú aj výsledky logistickej regresie (časť 4.3), ako aj koeficienty diskriminačnej funkcie v modeli LDA (časť 4.4). Naopak, najmenej významnou doménou je **Anta.**, čo je opätovne v súlade s výsledkami ostatných klasifikačných metód.

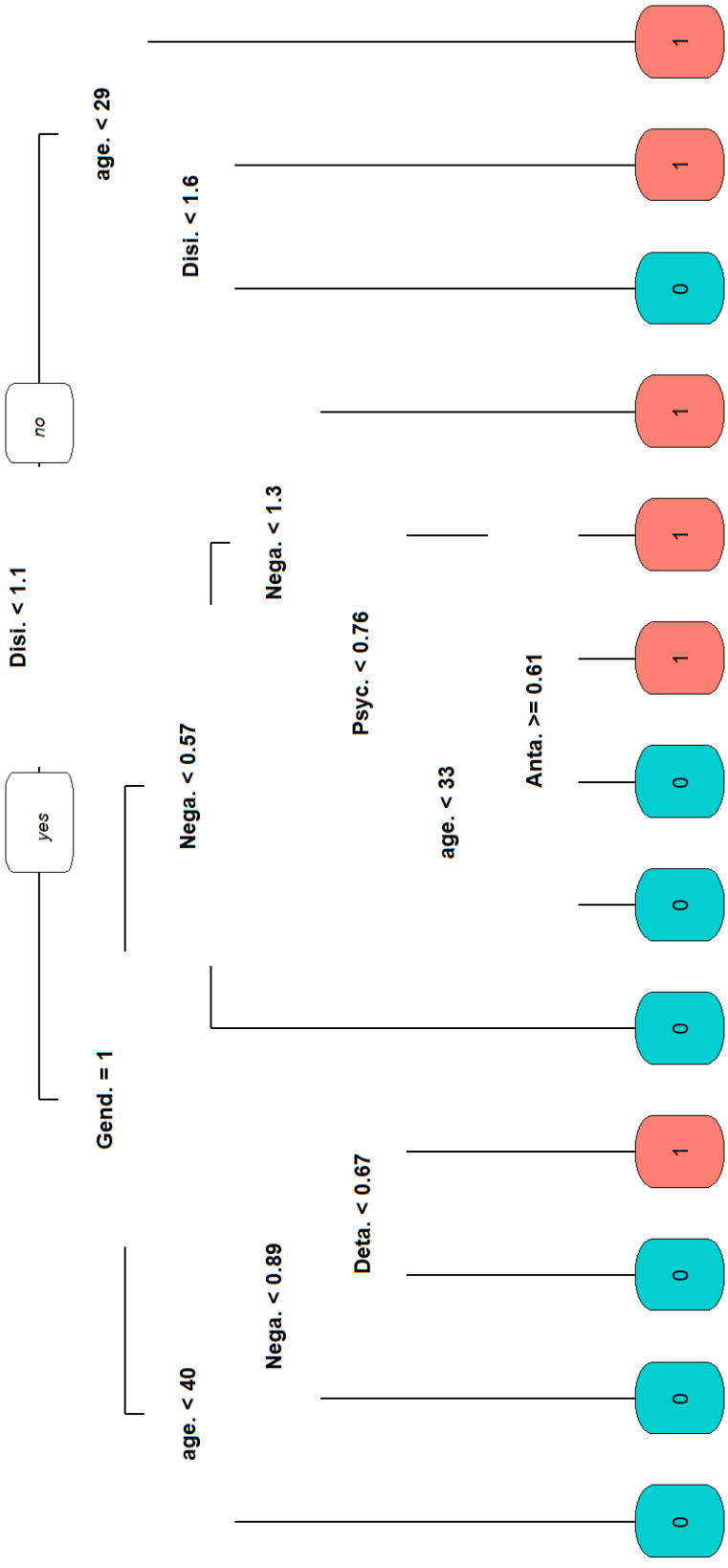
Rovnako ako v prípade datasetu *PID-5_aspekty_sel* sme vytvorili model klasifikačného stromu s apriórnymi pravdepodobnosťami. Po aplikovaní krížovej validácie model dosiahol presnosť **75,97 %**, hodnotu AUC **79,29 %**, senzitivitu **62,23 %**, špecificitu **86,60 %** a optimálny prah **0,220**.

LPFS

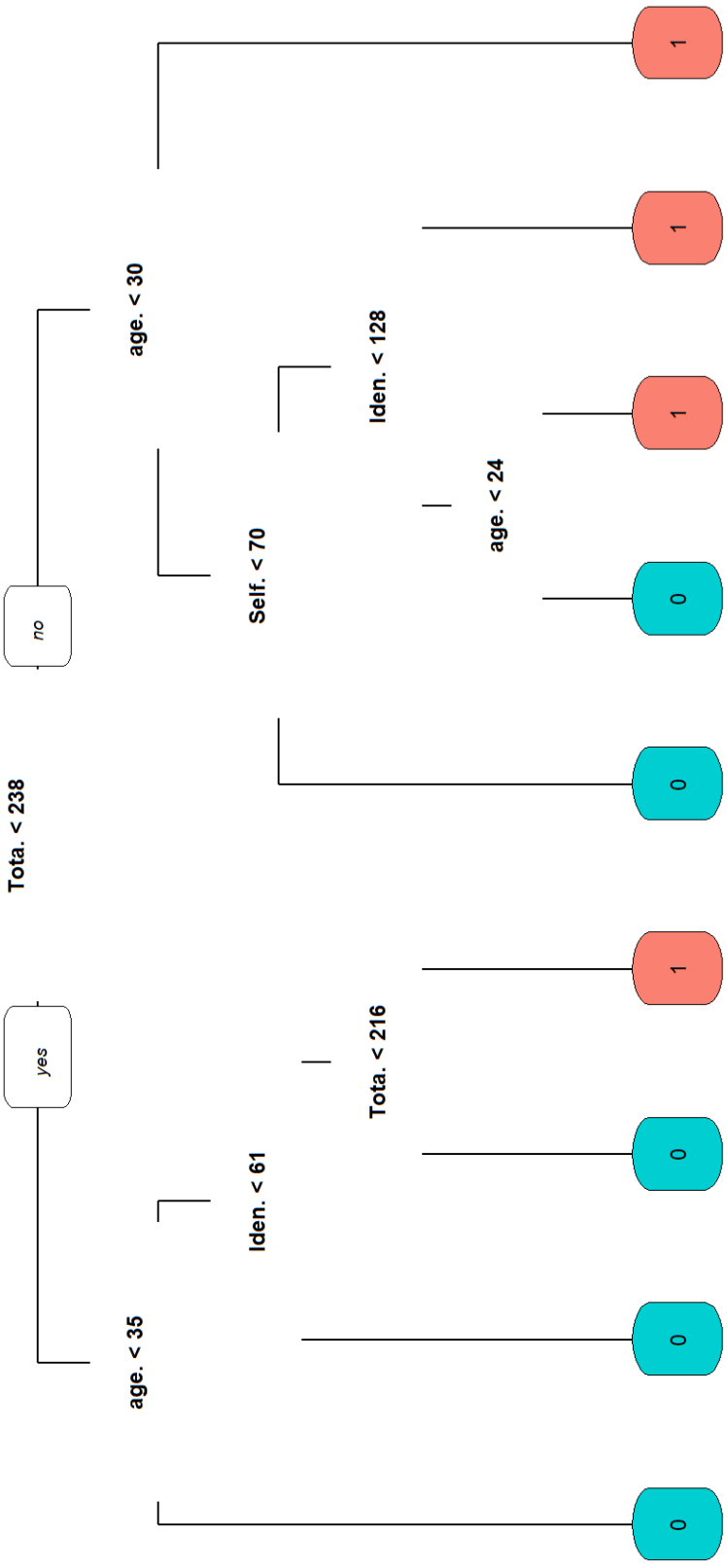
Pre dataset *LPFS* sme vytvorili základný model, ktorý obsahoval 8 uzlov, pričom hlavné rozdelenie v koreňovom uzle bolo založené na doméne **Tota.** (Obr. 31).

V prípade datasetu *LPFS* sa ako domény, ktoré dokážu dobre rozlišovať medzi skupinami, ukázali na základe vytvoreného modelu domény **Tota.**, **Iden.** a **Self.** Dôležitosť domén **Iden.** a **Self.** potvrdil aj model logistickej regresie. Model označil za najmenej dôležitú doménu **Empa.**, čo je opäť v súlade s našimi predchádzajúcimi pozorovaniami.

Po aplikovaní krížovej validácie sme získali model s presnosťou **76,50 %**, schopnosťou rozlišovať medzi skupinami **78,44 %** a optimálnym prahom **0,222**. Model dokázal správne klasifikovať skutočných pacientov na úrovni **59,56 %** a skutočných nepacientov na úrovni **89,10 %**.



Obr. 30: Klasifikačný strom pre dataset *PID-5_ domény* (zdroj: vlastné spracovanie v softvéri R).



Obr. 31: Klasifikačný strom pre dataset *LPFS* (zdroj: vlastné spracovanie v softvéri R).

4.6 Porovnanie modelov

Na záver kapitoly sme porovnali jednotlivé klasifikačné metódy a zároveň zhrnuli dôležitosť jednotlivých domén, t. j. ktoré domény sa ukázali ako najdôležitejšie pri rozlišovaní medzi skupinami.

PID-5_aspekty

Pri porovnaní aplikovaných štatistických metód sa ukázalo, že všetky modely dosiahli porovnateľnú presnosť klasifikácie. Najnižšiu schopnosť rozlišovať medzi skupinami na základe hodnoty AUC dosiahol klasifikačný strom. Model logistickej regresie a LDA vykazovali relatívne vyvážené hodnoty senzitivity a špecificity, pričom v oboch prípadoch bola špecificita mierne vyššia. Tento výsledok súvisí so spôsobom konštrukcie modelov, pri ktorom bol optimálny prah stanovený tak, aby maximalizoval súčet hodnôt senzitivity a špecificity. Napriek tomu klasifikačný strom dosiahol výrazne nižšiu senzitivitu v porovnaní so špecificitou než ostatné modely.

Na základe porovnania jednotlivých metód sa klasifikačný strom zo štatistického hľadiska javí ako najmenej vhodný. Hoci jeho významnou výhodou je jednoduchá a intuitívna interpretácia, v analyzovaných dátach dosahuje nízku hodnotu senzitivity v porovnaní s ostatnými metódami. Nevýhodou tejto metódy je zároveň jej nestabilita, keďže aj malé zmeny v trénovacej množine môžu viesť k výrazne odlišnej štruktúre výsledného stromu (časť 2.2.6).

Metóda LDA vykazovala výkonnosť porovnateľnú s logistickou regresiou, avšak v našom súbore dát došlo k porušeniu jej predpokladov. Ako bolo uvedené v časti 2.2.5, porušenie predpokladu rovnosti kovariančných matíc môže ovplyvniť lineárny charakter rozhodovacích hraníc, ktorý predstavuje kľúčovú vlastnosť tejto metódy. Z tohto dôvodu by bolo vhodné zvážiť aplikáciu metódy *kvadratickej diskriminačnej analýzy (QDA)* na naše dáta.

Pri porovnateľnej výkonnosti oboch prístupov sa preto ako vhodnejšia javí *logistická regresia*, a to najmä vzhľadom na jej menšiu citlivosť na porušenie predpokladov a jednoduchú interpretáciu parametrov prostredníctvom *pomerov šancí*. Zároveň, ako už bolo uvedené v časti 2.2.4, logistická regresia patrí medzi najčastejšie využívané viacrozmerné modely pri analýze binárnych výsledkov v medicínskom výskume.

Takto špecifikovaný model môže predstavovať efektívny nástroj na klasifikáciu nových

pozorovaní, pričom využitie optimálneho rozhodovacieho prahu umožňuje zaradiť respondentov do skupiny s prítomnosťou poruchy osobnosti na základe predikovanej pravdepodobnosti premennej **Pati.** Zvolené nastavenie prahu zodpovedá vyváženému pomeru senzitivity a špecificity, čo však otvára otázku, či je takýto kompromis v klinickej praxi optimálny, alebo či by bolo vhodnejšie uprednostniť zvýšenie jednej z týchto charakteristík na úkor druhej.

Všetky použité metódy označili aspekt **irre.** ako najdôležitejší pri klasifikácii. Významnú úlohu zohrávali aj aspekty **depr.** a **sepa.**, zatiaľ čo aspekty **impu.**, **risk.**, **susp.** a **soci.** prispievali k rozlišovaniu medzi skupinami najmenej.

Identifikácia aspektov, ktoré najefektívnejšie odlišujú klinickú a neklinickú populáciu, môže byť prínosná pri interpretácii jednotlivých skóre. Zvýšené hodnoty v týchto aspektoch môžu naznačovať vyššiu pravdepodobnosť zaradenia do skupiny pacientov, čo môže byť užitočné pre klinickú prax.

PID-5_domény

Rovnako ako v prípade datasetu *PID-5_aspekty_sel* vykazovali všetky modely porovnateľnú celkovú presnosť, pričom model klasifikačného stromu opäť dosiahol najnižšiu hodnotu AUC. Modely logistickej regresie a LDA dosahovali vyváženejšiu schopnosť identifikovať skutočných pacientov aj nepacientov v porovnaní s klasifikačným stromom. Hoci logistická regresia a LDA dosiahli porovnateľnú výkonnosť, zo štatistického hľadiska sa aj pre dataset *PID-5_domény* javí ako najvhodnejší model *logistickej regresie*.

Jednotlivé metódy identifikovali doménu **Disi.** ako najdôležitejšiu pri rozlišovaní medzi skupinami, pričom významnú úlohu zohráva aj doména **Nega.** Naopak, domény **Anta.**, **Deta.** a **Psyc.** mali na klasifikáciu najnižší prínos.

LPFS

V prípade datasetu *LPFS* dosiahli všetky modely z hľadiska celkovej presnosti opäť porovnateľný výkon. Podobne ako v predchádzajúcich dvoch datasetoch dosiahli logistická regresia a LDA vyššiu diskriminačnú schopnosť a vyváženejšie hodnoty senzitivity a špecificity než klasifikačný strom. Preto sa aj v tomto prípade javí *logistická regresia* ako vhodný model, keďže pri porovnateľnom výkone s ostatnými metódami zároveň poskytuje vyššiu mieru interpretovateľnosti.

Kľúčovými doménami pre klasifikáciu v tomto datasete boli **Iden.** a **Self.**, ktoré boli

identifikované ako dôležité logistickou regresiou aj klasifikačným stromom. Naopak, domény **Empa.** a **Inti.** sa ukázali ako najmenej významné, pričom ich bolo možné z modelu logistickej regresie vylúčiť bez výraznej straty výkonnosti.

Záver

Cieľom diplomovej práce bolo aplikovať klasifikačné metódy na dáta získané z českých verzií sebahodnotiacich dotazníkov a identifikovať tie domény a aspekty, ktoré najvýraznejšie prispievajú k rozlišovaniu medzi klinickou a neklinickou populáciou. Ďalšími čiastkovými cieľmi bolo porovnanie rozdielov medzi jednotlivými skupinami respondentov a skúmanie reliability a validity získaných dát.

V teoretickej časti práce boli predstavené dva typy sebahodnotiacich dotazníkov používaných pri diagnostike porúch osobnosti, konkrétne *PID-5* (pôvodná verzia [15]) a *LPFS-SR* (pôvodná verzia [18]). Okrem ich štruktúry bol popísaný aj spôsob výpočtu jednotlivých skóre, ktoré vstupovali do následných analýz. Súčasťou tejto časti bol aj prehľad aspektov a domén jednotlivých dotazníkov spolu s ich stručnou charakteristikou založenou na článku [1].

Nasledujúca časť práce bola venovaná teoretickému pozadiu štatistických metód vhodných na analýzu viacrozmerných osobnostných dát, so zameraním najmä na metódy redukcie dimenzie, overovanie reliability a validity a klasifikačné prístupy.

V praktickej časti práce boli najskôr analyzované rozdiely medzi respondentmi podľa spôsobu zberu dát a pohlavia. Na základe výsledkov v časti 3.1.1 sa ukázalo, že dáta získané online formou sa štatisticky významne líšia od dát získaných papierovou formou, a preto boli z ďalších analýz vylúčené. Následne boli pomocou deskriptívnej štatistiky a vizualizácií preskúmané základné charakteristiky jednotlivých domén, čo umožnilo získať prvotnú predstavu o ich diskriminačnej schopnosti.

V časti 3.3 bola overená reliabilita dát pomocou *Cronbachovej alfy*, pričom väčšina hodnotených domén dosahovala uspokojivú úroveň vnútornej konzistencie. Validita dát bola analyzovaná v časti 3.4 pomocou *konfirmačnej faktorovej analýzy*, na základe ktorej boli identifikované položky, ktoré najmenej reprezentovali príslušné domény. Korelačná analýza preukázala štatisticky významné vzťahy medzi doménami dotazníkov *PID-5* a *LPFS-SR*. Na základe týchto výsledkov možno oba dotazníky považovať za spoľahlivé a validné nástroje aj v rámci analyzovaného súboru dát.

V záverečnej časti práce boli aplikované klasifikačné modely s cieľom rozlíšiť klinickú a neklinickú populáciu. Pre dataset *PID-5_ aspekty* boli najskôr použité metódy selekcie premenných, na základe ktorých bola identifikovaná redukovaná množina aspektov dosta-

točne reprezentujúca pôvodný súbor aspektov. Porovnaním logistickej regresie, lineárnej diskriminačnej analýzy a klasifikačného stromu sa ako najvhodnejší model z hľadiska výkonnosti a interpretovateľnosti ukázala *logistická regresia*. Za kľúčové aspekty pri klasifikácii boli označené **irre.**, **depr.** a **sepa.**

Aj v prípade datasetov *PID-5_ domény* a *LPFS* dosiahla *logistická regresia* najlepší kompromis medzi presnosťou klasifikácie a interpretovateľnosťou modelu. V datasete *PID-5_ domény* sa ako najvýznamnejšie ukázali domény **Disi.** a **Nega.**, zatiaľ čo v datasete *LPFS* boli za kľúčové označené **Iden.** a **Self.**

Za hlavný prínos tejto práce možno považovať aplikáciu viacrozmerných štatistických metód na analýzu dát zo sebahodnotiacich dotazníkov, ktorá prispieva k efektívnejšiemu rozlíšeniu klinickej a neklinickej populácie. Identifikácia domén a aspektov, ktoré najvýraznejšie prispievajú k samotnej klasifikácii, môže mať praktický význam pri interpretácii výsledkov dotazníkov a pri odhadovaní rizika prítomnosti poruchy osobnosti. Navrhnutý model *logistickej regresie* predstavuje dobre interpretovateľný a presný nástroj, ktorý môže prispieť k efektívnejšiemu predklinickému skríningu a podporiť rozhodovanie v klinickej praxi.

Výber použitých metód bol motivovaný ich interpretovateľnosťou a praktickou využiteľnosťou. Možným rozšírením tejto práce je aplikácia sofistikovanejších štatistických prístupov, ktoré by mohli viesť k zlepšeniu presnosti klasifikácie a hlbšiemu pochopeniu štruktúry analyzovaných dát.

Zoznam použitej literatúry

[1] „Alternative DSM-5 Model for Personality Disorders,“ *Focus*, roč. 11, č. 2, s. 189–203, 2013. URL: https://psychiatryonline.org/toc/foc/11/2, (Dostupné na internete, 04.05.2026).

[2] J. Andel, „Základy matematické statistiky,“ *Matfyzpress, Praha*, 2005.

[3] S. Balakrishnama a A. Ganapathiraju, „Linear discriminant analysis-a brief tutorial,“ *Institute for Signal and information Processing*, roč. 18, č. 1998, s. 1–8, 1998. URL: https://datajobstest.com/data-science-repo/LDA-Primer-%5BBalakrishnama-and-Ganapathiraju%5D.pdf, (Dostupné na internete, 04.05.2026).

[4] R. Bonniga a D. A. Saraswathi, „Literature review of cronbachalphacoefficient and mcdonald's omega coefficient,“ *European Journal of Molecular & Clinical Medicine*, roč. 7, č. 06, 2020. URL: https://ssrn.com/abstract=4443362, (Dostupné na internete, 04.05.2026).

[5] T. Formánek, A. Kagström, P. Cermakova, L. Csémy, K. Mladá a P. Winkler, „Prevalence of mental disorders and associated disability: Results from the cross-sectional CZEch mental health Study (CZEMS),“ *European psychiatry*, roč. 60, s. 1–6, 2020. DOI: https://doi.org/10.1016/j.eurpsy.2019.05.001, (Dostupné na internete, 04.05.2026).

[6] A. Fossati, R. F. Krueger, K. E. Markon, S. Borroni a C. Maffei, „Reliability and validity of the Personality Inventory for DSM-5 (PID-5) predicting DSM-IV personality disorders and psychopathy in community-dwelling Italian adults,“ *Assessment*, roč. 20, č. 6, s. 689–708, 2013. DOI: https://doi.org/10.1177/1073191113504984, (Dostupné na internete, 04.05.2026).

[7] R. Harman, „Multivariate Statistical Analysis,“ FMFI UK, tech. spr., 2020, Učebné materiály. URL: http://www.iam.fmph.uniba.sk/ospm/Harman/VSAp.pdf, (Dostupné na internete, 04.05.2026).

[8] A. Hemmati, L. C. Morey, M. N. McCredie, F. Rezaei, A. Nazari a F. Rahmani, „Validation of the Persian translation of the Level of Personality Functioning Scale—Self-Report (LPFS-SR): Comparison of college students and patients with personality

disorders,“ *Journal of Psychopathology and Behavioral Assessment*, roč. 42, č. 3, s. 546–559, 2020. DOI: https://doi.org/10.1007/s10862-019-09775-6, (Dostupné na internete, 04.05.2026).

[9] D. C. Hooper a J. Coughlan, „Structural equation modelling: guidelines for determining model fit,“ *Electronic Journal of Business Research Methods*, roč. 6, č. 1, s. 53–60, 2008. URL: https://academic-publishing.org/index.php/ejbrm/article/view/1224/1187, (Dostupné na internete, 04.05.2026).

[10] G. James, D. Witten, T. Hastie a R. Tibshirani, *An introduction to statistical learning: with applications in R*. Springer, 2013, zv. 103.

[11] I. Jolliffe, „Principal component analysis,“ *Encyclopedia of statistics in behavioral science*, 2005. DOI: https://doi.org/10.1002/0470013192.bsa501, (Dostupné na internete, 04.05.2026).

[12] I. Kennedy, „Sample size determination in test-retest and Cronbach alpha reliability estimates,“ *British Journal of Contemporary Education*, roč. 2, č. 1, s. 17–29, 2022. DOI: https://doi.org/10.52589/BJCE-FY266HK9, (Dostupné na internete, 04.05.2026).

[13] C. L. Kimberlin a A. G. Winterstein, „Validity and reliability of measurement instruments used in research,“ *American journal of health-system pharmacy*, roč. 65, č. 23, s. 2276–2284, 2008. DOI: https://doi.org/10.2146/ajhp070364, (Dostupné na internete, 04.05.2026).

[14] R. F. Krueger, J. Derringer, K. E. Markon, D. Watson a A. E. Skodol, „Initial construction of a maladaptive personality trait model and inventory for DSM-5,“ *Psychological medicine*, roč. 42, č. 9, s. 1879–1890, 2012. DOI: https://doi.org/10.1017/S0033291711002674, (Dostupné na internete, 04.05.2026).

[15] R. F. Krueger, J. Derringer, K. E. Markon, D. Watson a A. E. Skodol, *Personality Inventory for DSM-5 (PID-5): Adult Form*, Copyright © 2013 American Psychiatric Association. Material may be reproduced for research and clinical use., American Psychiatric Association, 2013. URL: https://www.psychiatry.org/File%20Library/Psychiatrists/Practice/DSM/APA_DSM5_The-Personality-

Inventory - For - DSM - 5 - Full - Version - Adult . pdf, (Dostupné na internete, 04.05.2026).

[16] M. P. LaValley, „Logistic regression,“ *Circulation*, roč. 117, č. 18, s. 2395–2399, 2008. DOI: https://doi.org/10.1161/CIRCULATIONAHA.106.682658, (Dostupné na internete, 04.05.2026).

[17] „Chapter 3 - Linear mixed-effects models,“ in *Methods and Applications of Longitudinal Data Analysis*, X. Liu, ed., Oxford: Academic Press, 2016, s. 61–94, ISBN: 978-0-12-801342-7. DOI: https://doi.org/10.1016/B978-0-12-801342-7.00003-4, (Dostupné na internete, 04.05.2026).

[18] L. C. Morey, *Level of Personality Functioning Scale – Self-Report (LPFS-SR)*, PDF questionnaire, 2016. URL: https://supp.apa.org/psycarticles/supplemental/pas0000450/LevelQuesForm.pdf, (Dostupné na internete, 04.05.2026).

[19] T. G. Nick a K. M. Campbell, „Logistic regression,“ *Topics in biostatistics*, s. 273–301, 2007. DOI: https://doi.org/10.1007/978-1-59745-530-5_14, (Dostupné na internete, 04.05.2026).

[20] A. M. Olusegun, H. G. Dikko a S. U. Gulumbe, „Identifying the limitation of stepwise selection for variable selection in regression analysis,“ *American Journal of Theoretical and Applied Statistics*, roč. 4, č. 5, s. 414–419, 2015. DOI: https://doi.org/10.11648/j.ajtas.20150405.22, (Dostupné na internete, 04.05.2026).

[21] R. Pires et al., „The Portuguese version of the self-report form of the DSM-5 Level of Personality Functioning Scale (LPFS-SR) in a community and clinical sample,“ *Plos one*, roč. 19, č. 6, e0300706, 2024. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0300706, (Dostupné na internete, 04.05.2026).

[22] L. C. Quilty, L. Ayearst, M. Chmielewski, B. G. Pollock a R. M. Bagby, „The psychometric properties of the personality inventory for DSM-5 in an APA DSM-5 field trial sample,“ *Assessment*, roč. 20, č. 3, s. 362–369, 2013. DOI: https://doi.org/10.1177/1073191113486183, (Dostupné na internete, 04.05.2026).

[23] R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2023.

[24] J. Ranstam a J. A. Cook, „LASSO regression,“ *Journal of British Surgery*, roč. 105, č. 10, s. 1348–1348, 2018. DOI: https://doi.org/10.1002/bjs.10895, (Dostupné na internete, 04.05.2026).

[25] N. M. Razali, Y. B. Wah et al., „Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests,“ *Journal of statistical modeling and analytics*, roč. 2, č. 1, s. 21–33, 2011. URL: https://www.nbi.dk/~petersen/Teaching/Stat2017/Power_Comparisons_of_Shapiro-Wilk_Kolmogorov-Smirn.pdf, (Dostupné na internete, 04.05.2026).

[26] K. Riegel, M. Preiss, A. Ksinan, D. Šamánková a P. Harsa, „Psychometric properties of the Czech version of the personality inventory for DSM-5: Internal consistency, validity and discrimination capacity of the measure,“ *Ceskoslovenska Psychologie*, roč. 61, s. 128–143, 2017.

[27] L. Rokach a O. Maimon, „Classification trees,“ in *Data mining and knowledge discovery handbook*, Springer, 2010, s. 149–174. DOI: https://doi.org/10.1007/978-0-387-09823-4_9, (Dostupné na internete, 04.05.2026).

[28] S. Rosa, *Redukcia dimenzie dát: poznámky k prednáškam*, Interné materiály k predmetu, Nepublikované, 2025.

[29] S. K. Sarkar, „Principal component analysis,“ *Statistical Procedures for Analyzing Agricultural Data using R*, s. 139–156, 2023. URL: https://icar-iirr.org/books/chapters/Statistical-Procedures_ch10.pdf, (Dostupné na internete, 04.05.2026).

[30] R. P. Sarmento a V. Costa, „Confirmatory factor analysis–a case study,“ *arXiv preprint arXiv:1905.05598*, 2019. URL: https://arxiv.org/abs/1905.05598, (Dostupné na internete, 04.05.2026).

[31] L. Schlosserova, L. Rosova, T. S. Zbornik a K. D. Riegel, „Self-Reported Personality Functioning in Patients With Substance Use Disorders: Similarities and Specifics With Respect to the Czech Community Sample,“ *Alpha Psychiatry*, roč. 26, č. 4, s. 46 058, 2025. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC12416040/, (Dostupné na internete, 04.05.2026).

[32] Y.-Y. Song a Y. Lu, „Decision tree methods: applications for classification and prediction,“ *Shanghai archives of psychiatry*, 2015. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4466856/, (Dostupné na internete, 04.05.2026).

[33] S. Sperandei, „Understanding logistic regression analysis,“ *Biochemia medica*, roč. 24, č. 1, s. 12–18, 2014. DOI: https://doi.org/10.11613/BM.2014.003, (Dostupné na internete, 04.05.2026).

[34] C. D. Stapleton, „Basic Concepts and Procedures of Confirmatory Factor Analysis.,“ 1997. URL: https://eric.ed.gov/?id=ED407416, (Dostupné na internete, 04.05.2026).

[35] M. Tavakol a R. Dennick, „Making sense of Cronbach's alpha,“ *International journal of medical education*, roč. 2, s. 53, 2011. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4205511/, (Dostupné na internete, 04.05.2026).

[36] A. Tharwat, T. Gaber, A. Ibrahim a A. E. Hassanien, „Linear discriminant analysis: A detailed tutorial,“ *AI communications*, roč. 30, č. 2, s. 169–190, 2017. DOI: https://doi.org/10.3233/AIC-170729, (Dostupné na internete, 04.05.2026).

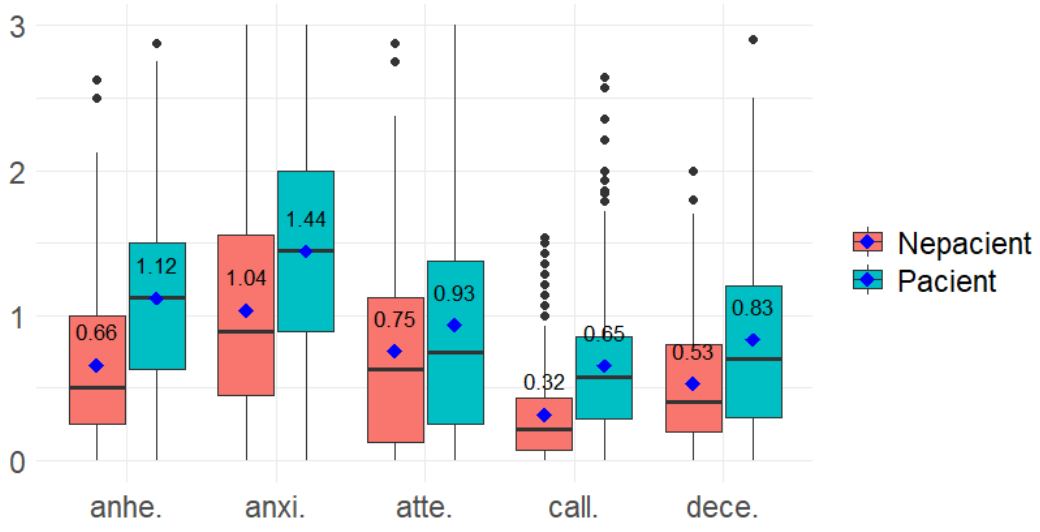
[37] J. Thimm, „The associations between personality functioning (Criterion A) and pathological personality traits (Criterion B) in the alternative model for personality disorders in DSM-5 (DSM-5 AMPD): A meta-analysis,“ *Frontiers in Psychiatry*, roč. 16, s. 1 673 139, 2025. DOI: https://doi.org/10.3389/fpsyt.2025.1673139, (Dostupné na internete, 04.05.2026).

[38] M. Tranmer a M. Elliot, „Binary logistic regression,“ *Cathie Marsh for census and survey research, paper*, roč. 20, s. 90 033–90 039, 2008. URL: https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2008/2008-20-binary-logistic-regression.pdf, (Dostupné na internete, 04.05.2026).

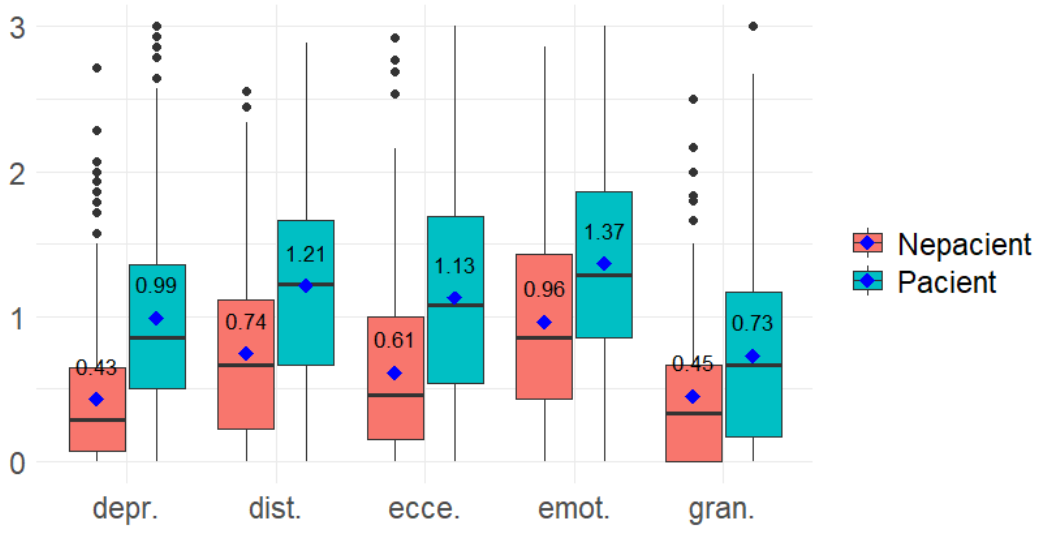
[39] B. W. Yap a C. H. Sim, „Comparisons of various types of normality tests,“ *Journal of Statistical Computation and Simulation*, roč. 81, č. 12, s. 2141–2155, 2011. DOI: https://doi.org/10.1080/00949655.2010.520163, (Dostupné na internete, 04.05.2026).

Príloha A

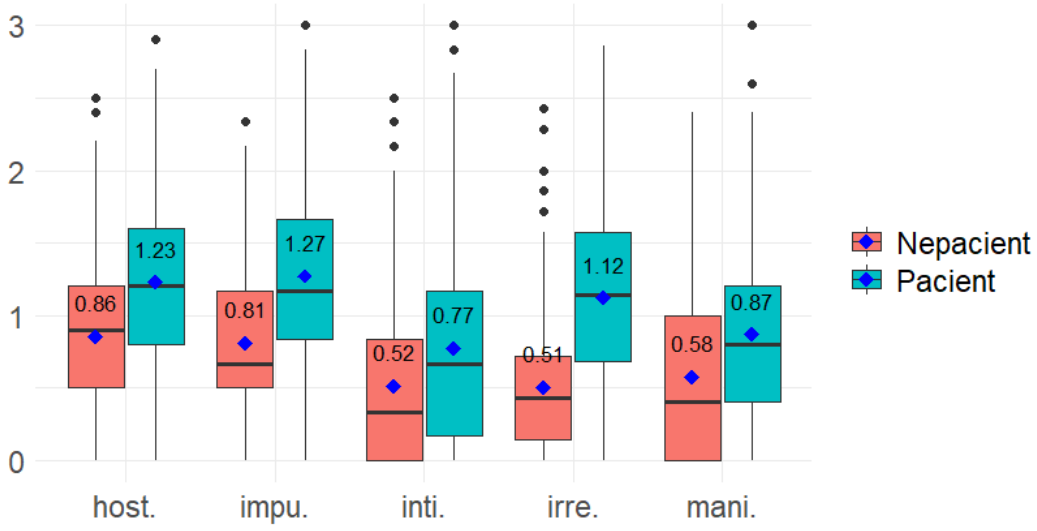
Boxploty aspektov datasetu *PID-5_ aspekty*



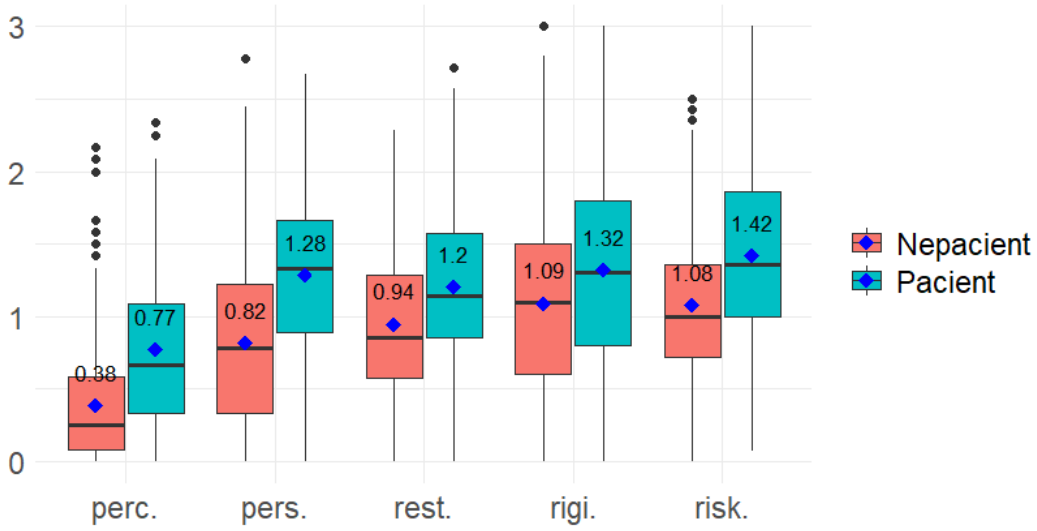
Obr. 32: Boxploty aspektov datasetu *PID-5_ aspekty* (zdroj: vlastné spracovanie v softvéri R).



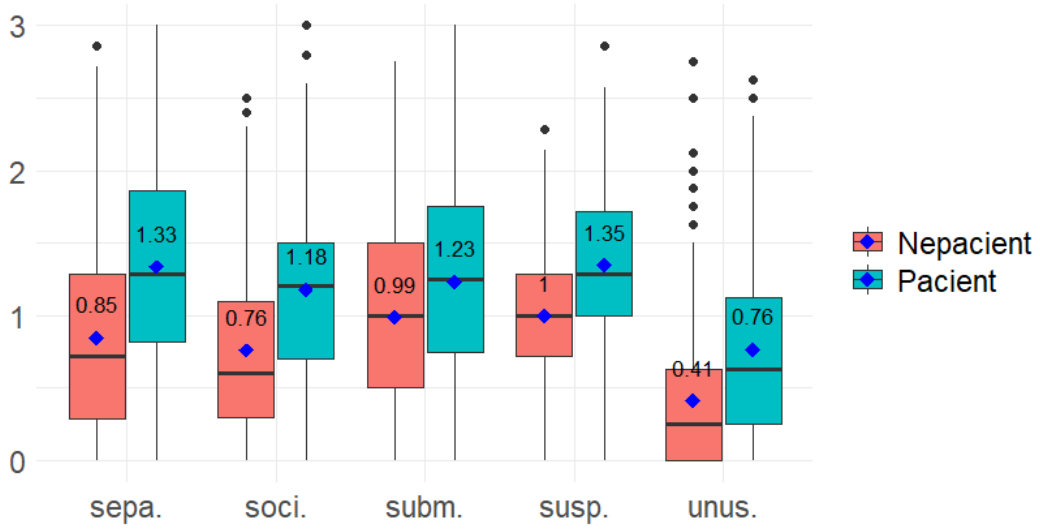
Obr. 33: Boxploty aspektov datasetu *PID-5_ aspekty* (zdroj: vlastné spracovanie v softvéri R).



Obr. 34: Boxploty aspektov datasetu *PID-5_ aspekty* (zdroj: vlastné spracovanie v softvéri R).



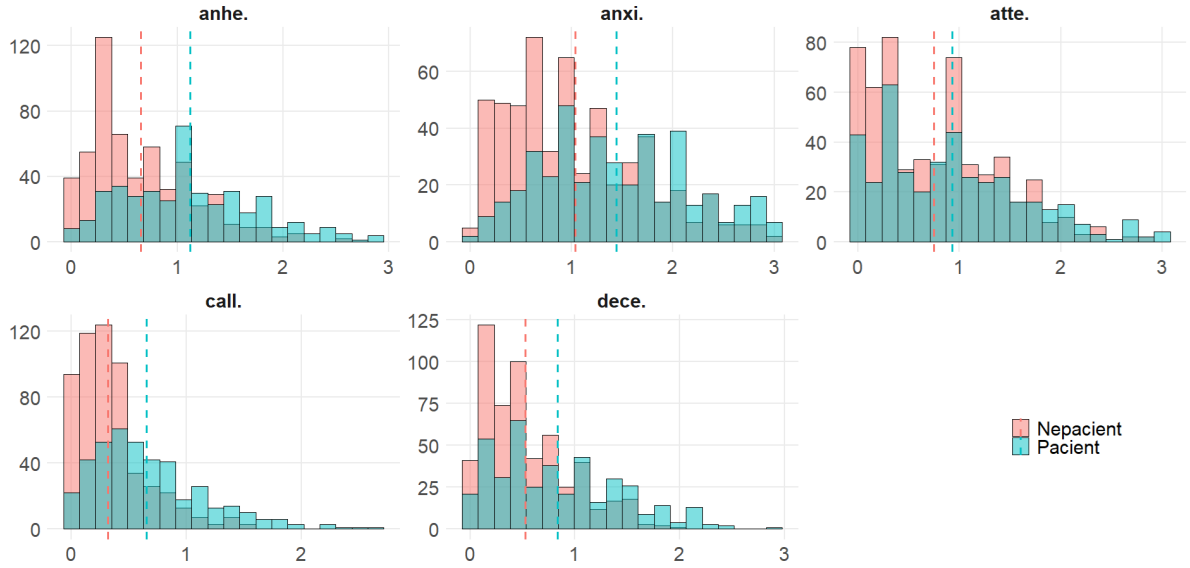
Obr. 35: Boxploty aspektov datasetu *PID-5_ aspekty* (zdroj: vlastné spracovanie v softvéri R).



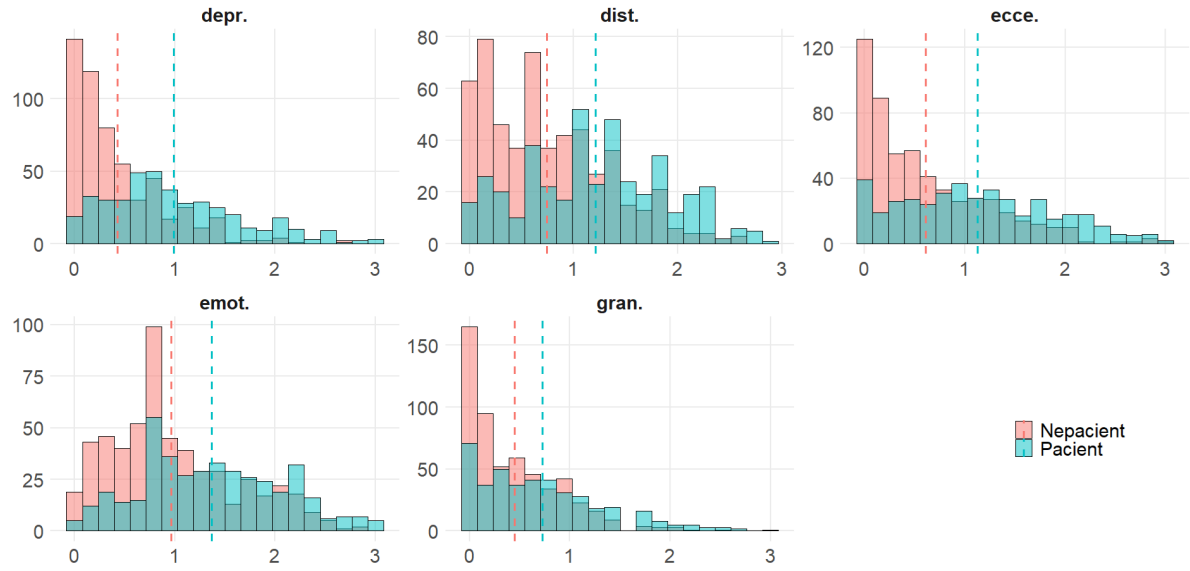
Obr. 36: Boxploty aspektov datasetu *PID-5_ aspekty* (zdroj: vlastné spracovanie v softvéri R).

Príloha B

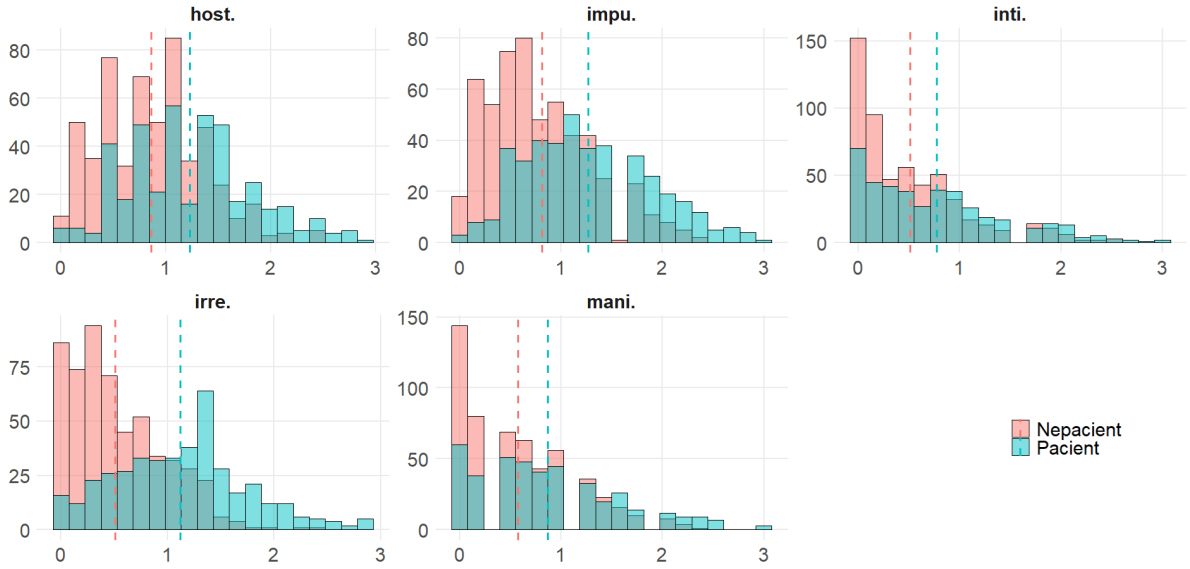
Histogramy aspektov datasetu *PID-5_ aspekty*



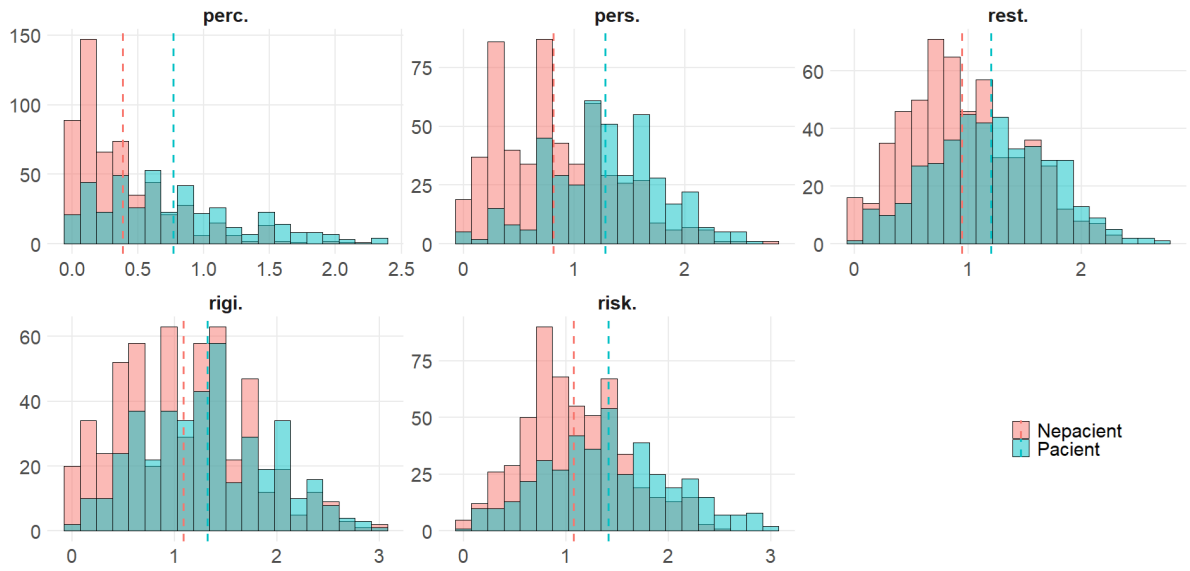
Obr. 37: Histogramy aspektov datasetu *PID-5_ aspekty* s vyznačenými priemernými hodnotami (zdroj: vlastné spracovanie v softvéri R).



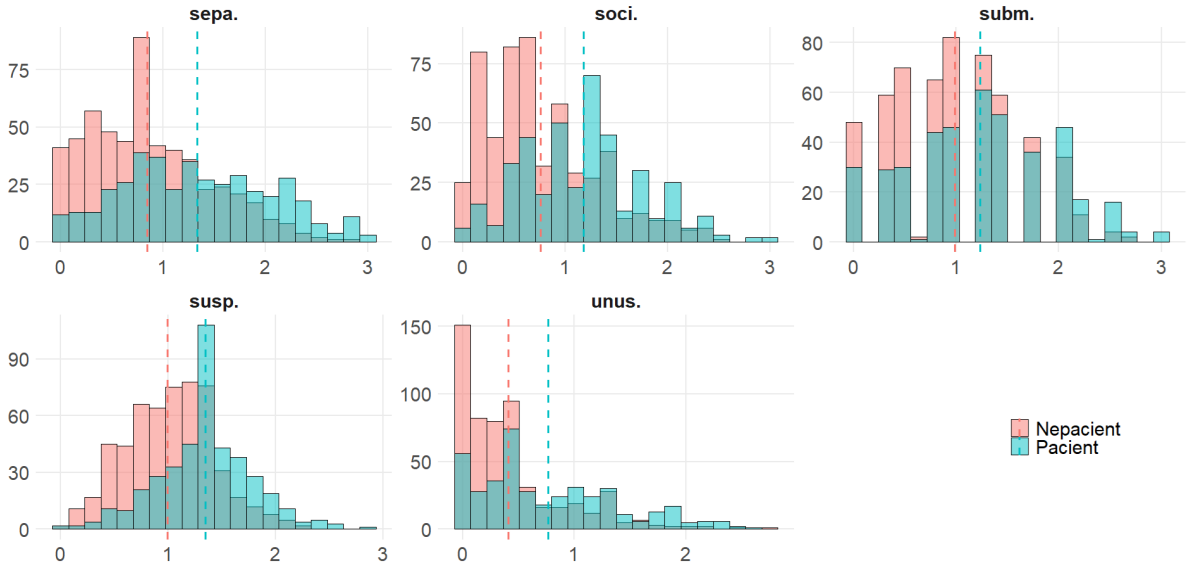
Obr. 38: Histogramy aspektov datasetu *PID-5_ aspekty* s vyznačenými priemernými hodnotami (zdroj: vlastné spracovanie v softvéri R).



Obr. 39: Histogramy aspektov datasetu *PID-5_ aspekty* s vyznačenými priemernými hodnotami (zdroj: vlastné spracovanie v softvéri R).



Obr. 40: Histogramy aspektov datasetu *PID-5_ aspekty* s vyznačenými priemernými hodnotami (zdroj: vlastné spracovanie v softvéri R).



Obr. 41: Histogramy aspektov datasetu *PID-5_ aspekty* s vyznačenými priemernými hodnotami (zdroj: vlastné spracovanie v softvéri R).

Príloha C

Model CFA a hodnoty R^2 jednotlivých položiek pre dotazník *PID-5*

$$
\begin{aligned}
\text{Nega.} =& Q12 + Q18 + Q50 + Q57 + Q62 + Q64 + Q79 + Q93 + Q95 + Q96 + Q102 \\
&+ Q109 + Q110 + Q122 + Q127 + Q130 + Q138 + Q141 + Q149 + Q165 \\
&+ Q174 + Q175 + Q181, \\
\text{Deta.} =& Q1 + Q10 + Q20 + Q23 + Q26 + Q30 + Q75 + Q82 + Q89 + Q97 + Q108 \\
&+ Q120 + Q124 + Q136 + Q145 + Q146 + Q147 + Q155 + Q157 + Q161 \\
&+ Q182 + Q186 + Q189 + Q203, \\
\text{Anta.} =& Q40 + Q41 + Q53 + Q56 + Q65 + Q76 + Q107 + Q114 + Q125 + Q126 \\
&+ Q134 + Q142 + Q162 + Q179 + Q180 + Q187 + Q197 + Q206 + Q214 \\
&+ Q218 + Q219, \\
\text{Disi.} =& Q4 + Q6 + Q16 + Q17 + Q22 + Q29 + Q31 + Q47 + Q58 + Q68 + Q88 \\
&+ Q118 + Q129 + Q132 + Q144 + Q156 + Q160 + Q171 + Q199 + Q201 \\
&+ Q204 + Q210, \\
\text{Psyc.} =& Q5 + Q21 + Q24 + Q25 + Q33 + Q36 + Q37 + Q42 + Q44 + Q52 + Q55 \\
&+ Q59 + Q70 + Q71 + Q77 + Q83 + Q94 + Q99 + Q106 + Q139 + Q143 \\
&+ Q150 + Q152 + Q154 + Q172 + Q185 + Q192 + Q193 + Q194 + Q205 \\
&+ Q209 + Q213 + Q217.
\end{aligned}
$$

Tabuľka 25: Hodnoty R^2 jednotlivých položiek modelu pre dotazník *PID-5*.

Nega.	R^2	**Deta.**	R^2	**Anta.**	R^2	**Disi.**	R^2	**Psyc.**	R^2
Q12	0,150	**Q1**	0,260	**Q40**	0,302	**Q4**	0,283	**Q5**	0,384
Q18	0,410	**Q10**	0,328	**Q41**	0,212	**Q6**	0,284	**Q21**	0,478
Q50	0,389	**Q20**	0,132	**Q53**	0,325	**Q16**	0,361	**Q24**	0,476
Q57	0,122	**Q23**	0,372	**Q56**	0,374	**Q17**	0,415	**Q25**	0,460
Q62	0,372	**Q26**	0,395	**Q65**	0,229	**Q22**	0,321	**Q33**	0,521
Q64	0,174	**Q30**	0,127	**Q76**	0,454	**Q29**	0,411	**Q36**	0,325
Q79	0,436	**Q75**	0,330	**Q107**	0,322	**Q31**	0,439	**Q37**	0,142
Q93	0,487	**Q82**	0,423	**Q114**	0,322	**Q47**	0,300	**Q42**	0,227
Q95	0,441	**Q89**	0,222	**Q125**	0,466	**Q58**	0,003	**Q44**	0,158
Q96	0,062	**Q97**	0,121	**Q126**	0,337	**Q68**	0,512	**Q52**	0,572
Q102	0,191	**Q108**	0,171	**Q134**	0,443	**Q88**	0,538	**Q55**	0,469
Q109	0,577	**Q120**	0,196	**Q142**	0,002	**Q118**	0,487	**Q59**	0,227
Q110	0,512	**Q124**	0,392	**Q162**	0,486	**Q129**	0,492	**Q70**	0,545
Q122	0,352	**Q136**	0,398	**Q179**	0,128	**Q132**	0,453	**Q71**	0,608
Q127	0,312	**Q145**	0,204	**Q180**	0,261	**Q144**	0,468	**Q77**	0,174
Q130	0,538	**Q146**	0,255	**Q187**	0,231	**Q156**	0,202	**Q83**	0,472
Q138	0,502	**Q147**	0,315	**Q197**	0,343	**Q160**	0,296	**Q94**	0,220
Q141	0,408	**Q155**	0,262	**Q206**	0,586	**Q171**	0,283	**Q99**	0,210
Q149	0,208	**Q157**	0,420	**Q214**	0,427	**Q199**	0,406	**Q106**	0,248
Q165	0,174	**Q161**	0,415	**Q218**	0,527	**Q201**	0,322	**Q139**	0,295
Q174	0,640	**Q182**	0,235	**Q219**	0,627	**Q204**	0,308	**Q143**	0,093
Q175	0,229	**Q186**	0,484			**Q210**	0,252	**Q150**	0,135
Q181	0,409	**Q189**	0,410					**Q152**	0,603
		Q203	0,261					**Q154**	0,206
								Q172	0,502
								Q185	0,403
								Q192	0,198
								Q193	0,351
								Q194	0,351
								Q205	0,591
								Q209	0,368
								Q213	0,307
								Q217	0,433

Príloha D

Model CFA a hodnoty R^2 jednotlivých položiek pre dotazník *LPFS-SR*

$$
\begin{aligned}
\text{Iden.} =& Q4 + Q7 + Q8 + Q15 + Q23 + Q33 + Q35 + Q40 + Q47 + Q48 + Q50 + Q55 \\
&+ Q56 + Q57 + Q62 + Q65 + Q66 + Q71 + Q72 + Q74 + Q76 + Q77 + Q79, \\
\text{Self.} =& Q1 + Q14 + Q16 + Q22 + Q24 + Q25 + Q28 + Q29 + Q30 + Q34 + Q36 + Q41 \\
&+ Q42 + Q51 + Q53 + Q58 + Q59 + Q64 + Q70 + Q78 + Q80, \\
\text{Empa.} =& Q2 + Q11 + Q17 + Q18 + Q20 + Q21 + Q31 + Q32 + Q38 + Q44 + Q49 + Q54 \\
&+ Q60 + Q61 + Q68 + Q75, \\
\text{Inti.} =& Q3 + Q5 + Q6 + Q9 + Q10 + Q12 + Q13 + Q19 + Q26 + Q27 + Q37 + Q39 \\
&+ Q43 + Q45 + Q46 + Q52 + Q63 + Q67 + Q69 + Q73, \\
\text{Tota.} =& \text{Iden.} + \text{Self.} + \text{Empa.} + \text{Inti.}
\end{aligned}
$$

Tabuľka 26: Hodnoty R^2 jednotlivých položiek modelu pre dotazník *LPFS-SR*.

Iden.	R^2	**Self.**	R^2	**Empa.**	R^2	**Inti.**	R^2
Q4	0,213	**Q1**	0,130	**Q2**	0,170	**Q3**	0,452
Q7	0,059	**Q14**	0,067	**Q11**	0,056	**Q5**	0,398
Q8	0,119	**Q16**	0,400	**Q17**	0,160	**Q6**	0,340
Q15	0,301	**Q22**	0,043	**Q18**	0,259	**Q9**	0,218
Q23	0,196	**Q24**	0,431	**Q20**	0,110	**Q10**	0,311
Q33	0,507	**Q25**	0,450	**Q21**	0,382	**Q12**	0,183
Q35	0,479	**Q28**	0,508	**Q31**	0,265	**Q13**	0,148
Q40	0,001	**Q29**	0,431	**Q32**	0,272	**Q19**	0,410
Q47	0,446	**Q30**	0,314	**Q38**	0,096	**Q26**	0,250
Q48	0,498	**Q34**	0,176	**Q44**	0,001	**Q27**	0,093
Q50	0,450	**Q36**	0,309	**Q49**	0,430	**Q37**	0,077
Q55	0,327	**Q41**	0,436	**Q54**	0,275	**Q39**	0,101
Q56	0,427	**Q42**	0,499	**Q60**	0,258	**Q43**	0,165
Q57	0,349	**Q51**	0,112	**Q61**	0,033	**Q45**	0,076
Q62	0,209	**Q53**	0,310	**Q68**	0,131	**Q46**	0,423
Q65	0,060	**Q58**	0,338	**Q75**	0,303	**Q52**	0,249
Q66	0,313	**Q59**	0,416			**Q63**	0,427
Q71	0,410	**Q64**	0,057			**Q67**	0,211
Q72	0,456	**Q70**	0,260			**Q69**	0,147
Q74	0,391	**Q78**	0,199			**Q73**	0,280
Q76	0,006	**Q80**	0,342				
Q77	0,329						
Q79	0,287						

Príloha E

Korelácie medzi aspektmi dotazníka *PID-5* a doménami *LPFS-SR*

Tabuľka 27: Korelácie medzi aspektmi dotazníka *PID-5* a doménami *LPFS-SR*.

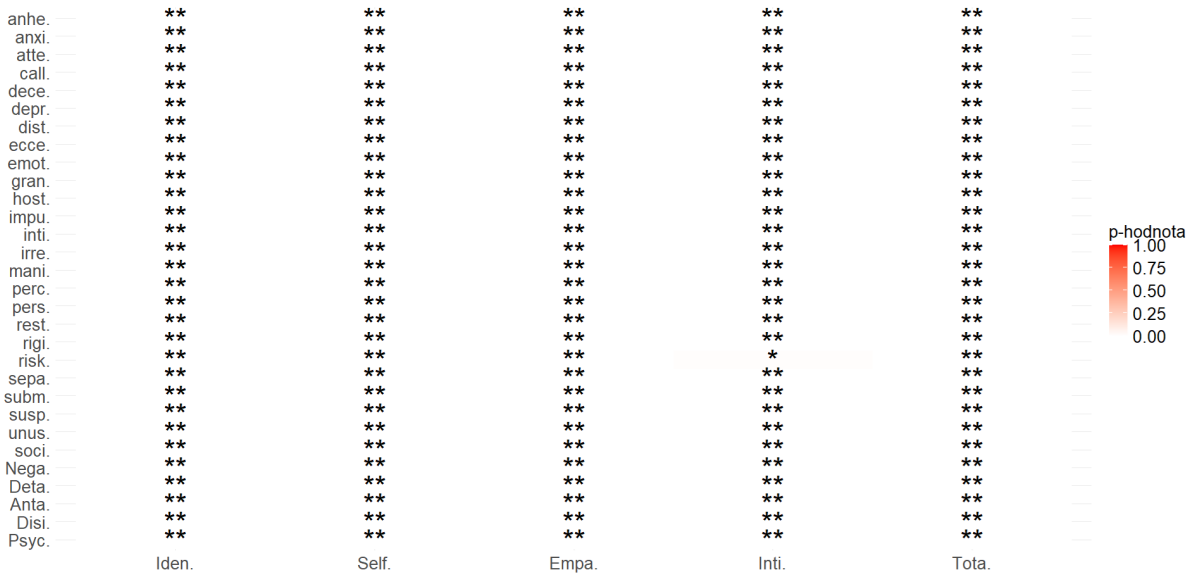
Aspekt		**Iden.**	**Self.**	**Empa.**	**Inti.**	**Tota.**
anne.	Pac.	0,46	0,59	0,37	0,43	0,53
	Nep.	0,63	0,64	0,55	0,60	0,69
	Spolu	0,62	0,68	0,55	0,59	0,67
	*95 % IS**	*[0,48; 0,66]*	*[0,47; 0,65]*	*[0,35; 0,52]*	*[0,42; 0,59]*	*[0,55; 0,71]*
anxi.	Pac.	0,66	0,60	0,43	0,51	0,63
	Nep.	0,70	0,58	0,44	0,49	0,64
	Spolu	0,71	0,63	0,50	0,55	0,67
	*95 % IS**	*[0,53; 0,65]*	*[0,37; 0,53]*	*[0,27; 0,43]*	*[0,32; 0,53]*	*[0,51; 0,63]*
atte.	Pac.	0,28	0,14	0,26	0,19	0,24
	Nep.	0,31	0,27	0,23	0,22	0,29
	Spolu	0,31	0,23	0,27	0,23	0,29
	*95 % IS**	*[0,07; 0,38]*	*[0,14; 0,38]*	*[0,10; 0,37]*	*[0,00; 0,37]*	*[0,14; 0,36]*
call.	Pac.	0,27	0,25	0,37	0,41	0,36
	Nep.	0,41	0,49	0,52	0,49	0,52
	Spolu	0,43	0,45	0,52	0,53	0,52
	*95 % IS**	*[0,10; 0,53]*	*[0,12; 0,59]*	*[0,44; 0,69]*	*[0,33; 0,65]*	*[0,25; 0,60]*
dece.	Pac.	0,34	0,29	0,36	0,31	0,36
	Nep.	0,47	0,51	0,46	0,47	0,53
	Spolu	0,46	0,44	0,46	0,44	0,49
	*95 % IS**	*[0,25; 0,47]*	*[0,26; 0,50]*	*[0,30; 0,53]*	*[0,24; 0,50]*	*[0,33; 0,51]*
depr.	Pac.	0,65	0,67	0,50	0,57	0,68
	Nep.	0,71	0,69	0,58	0,59	0,73
	Spolu	0,74	0,75	0,62	0,66	0,77
	*95 % IS**	*[0,56; 0,74]*	*[0,53; 0,69]*	*[0,38; 0,56]*	*[0,48; 0,60]*	*[0,62; 0,73]*
dist.	Pac.	0,64	0,67	0,51	0,48	0,66
	Nep.	0,69	0,71	0,56	0,52	0,71
	Spolu	0,71	0,73	0,60	0,58	0,72
	*95 % IS**	*[0,44; 0,61]*	*[0,42; 0,61]*	*[0,29; 0,48]*	*[0,25; 0,48]*	*[0,49; 0,64]*
ecce.	Pac.	0,60	0,48	0,51	0,50	0,60
	Nep.	0,59	0,52	0,56	0,47	0,60
	Spolu	0,65	0,58	0,60	0,57	0,66
	*95 % IS**	*[0,34; 0,58]*	*[0,30; 0,51]*	*[0,36; 0,53]*	*[0,29; 0,56]*	*[0,45; 0,64]*
emot.	Pac.	0,62	0,52	0,41	0,48	0,59
	Nep.	0,66	0,49	0,44	0,49	0,60
	Spolu	0,68	0,56	0,49	0,54	0,64
	*95 % IS**	*[0,54; 0,68]*	*[0,41; 0,59]*	*[0,31; 0,51]*	*[0,33; 0,56]*	*[0,53; 0,68]*
gran.	Pac.	0,21	0,11	0,31	0,30	0,25
	Nep.	0,30	0,32	0,38	0,34	0,37
	Spolu	0,32	0,28	0,40	0,38	0,37
	*95 % IS**	*[0,10; 0,30]*	*[0,15; 0,31]*	*[0,31; 0,48]*	*[0,22; 0,40]*	*[0,23; 0,38]*
host.	Pac.	0,46	0,32	0,40	0,42	0,45
	Nep.	0,61	0,53	0,56	0,54	0,63
	Spolu	0,59	0,50	0,54	0,55	0,60
	*95 % IS**	*[0,42; 0,63]*	*[0,35; 0,57]*	*[0,39; 0,59]*	*[0,36; 0,59]*	*[0,50; 0,67]*
impu.	Pac.	0,50	0,41	0,34	0,38	0,47
	Nep.	0,44	0,42	0,41	0,36	0,46
	Spolu	0,56	0,52	0,48	0,48	0,56

Aspekt		Iden.	Self.	Empa.	Inti.	Tota.
	*95 % IS**	*[0,32; 0,56]*	*[0,42; 0,61]*	*[0,31; 0,53]*	*[0,26; 0,48]*	*[0,42; 0,61]*
inti.	Pac.	0,19	0,28	0,27	0,37	0,31
	Nep.	0,31	0,36	0,37	0,42	0,40
	Spolu	0,31	0,37	0,38	0,44	0,41
	*95 % IS**	*[0,18; 0,39]*	*[0,11; 0,43]*	*[0,13; 0,45]*	*[0,27; 0,49]*	*[0,26; 0,49]*
irre.	Pac.	0,47	0,51	0,40	0,44	0,51
	Nep.	0,55	0,66	0,44	0,45	0,59
	Spolu	0,61	0,67	0,55	0,56	0,66
	*95 % IS**	*[0,26; 0,57]*	*[0,39; 0,63]*	*[0,33; 0,56]*	*[0,28; 0,54]*	*[0,42; 0,61]*
mani.	Pac.	0,19	0,10	0,23	0,22	0,21
	Nep.	0,31	0,31	0,33	0,29	0,34
	Spolu	0,31	0,26	0,33	0,31	0,33
	*95 % IS**	*[0,10; 0,30]*	*[0,23; 0,30]*	*[0,21; 0,33]*	*[0,15; 0,32]*	*[0,21; 0,31]*
perc.	Pac.	0,61	0,58	0,54	0,52	0,64
	Nep.	0,61	0,58	0,58	0,53	0,65
	Spolu	0,68	0,65	0,63	0,61	0,70
	*95 % IS**	*[0,24; 0,56]*	*[0,26; 0,59]*	*[0,27; 0,58]*	*[0,14; 0,52]*	*[0,35; 0,63]*
pers.	Pac.	0,59	0,51	0,46	0,48	0,58
	Nep.	0,63	0,59	0,48	0,50	0,63
	Spolu	0,67	0,63	0,56	0,57	0,68
	*95 % IS**	*[0,47; 0,71]*	*[0,47; 0,68]*	*[0,38; 0,60]*	*[0,34; 0,63]*	*[0,61; 0,70]*
rest.	Pac.	0,19	0,17	0,28	0,26	0,25
	Nep.	0,17	0,27	0,30	0,29	0,27
	Spolu	0,27	0,30	0,36	0,35	0,34
	*95 % IS**	*[0,01; 0,29]*	*[0,08; 0,36]*	*[0,20; 0,44]*	*[0,19; 0,40]*	*[0,18; 0,37]*
rigi.	Pac.	0,26	0,05	0,24	0,28	0,23
	Nep.	0,39	0,26	0,34	0,36	0,38
	Spolu	0,36	0,21	0,33	0,35	0,35
	*95 % IS**	*[0,24; 0,44]*	*[0,12; 0,34]*	*[0,14; 0,39]*	*[0,19; 0,43]*	*[0,29; 0,46]*
risk.	Pac.	0,20	0,15	0,13	0,17	0,19
	Nep.	0,10	0,12	0,13	0,06	0,11
	Spolu	0,26	0,25	0,25	0,23	0,27
	*95 % IS**	*[0,06; 0,44]*	*[0,21; 0,46]*	*[0,20; 0,47]*	*[0,14; 0,44]*	*[0,19; 0,49]*
sepa.	Pac.	0,45	0,36	0,34	0,37	0,44
	Nep.	0,46	0,34	0,35	0,38	0,44
	Spolu	0,54	0,46	0,45	0,47	0,53
	*95 % IS**	*[0,42; 0,56]*	*[0,34; 0,50]*	*[0,26; 0,45]*	*[0,28; 0,45]*	*[0,40; 0,53]*
subm.	Pac.	0,45	0,38	0,27	0,23	0,39
	Nep.	0,42	0,39	0,22	0,26	0,38
	Spolu	0,46	0,42	0,30	0,29	0,41
	*95 % IS**	*[0,45; 0,48]*	*[0,37; 0,43]*	*[0,17; 0,35]*	*[0,24; 0,32]*	*[0,39; 0,48]*
susp.	Pac.	0,51	0,42	0,49	0,51	0,54
	Nep.	0,53	0,53	0,56	0,57	0,61
	Spolu	0,60	0,55	0,60	0,61	0,65
	*95 % IS**	*[0,46; 0,67]*	*[0,35; 0,58]*	*[0,44; 0,68]*	*[0,50; 0,72]*	*[0,55; 0,71]*
unus.	Pac.	0,47	0,35	0,46	0,41	0,47
	Nep.	0,39	0,34	0,46	0,37	0,43
	Spolu	0,50	0,43	0,53	0,47	0,53
	*95 % IS**	*[0,24; 0,50]*	*[0,23; 0,48]*	*[0,34; 0,49]*	*[0,20; 0,50]*	*[0,35; 0,55]*
soci.	Pac.	0,32	0,34	0,38	0,45	0,41
	Nep.	0,39	0,39	0,46	0,51	0,48
	Spolu	0,45	0,46	0,50	0,56	0,53
	*95 % IS**	*[0,44; 0,52]*	*[0,38; 0,47]*	*[0,39; 0,51]*	*[0,46; 0,68]*	*[0,51; 0,60]*

* Hodnoty 95 % intervalov spoľahlivosti sú uvedené podľa [37].

Príloha F

Štatistická významnosť korelácií medzi dotazníkmi *PID-5* a *LPFS-SR* pre jednotlivé podskupiny

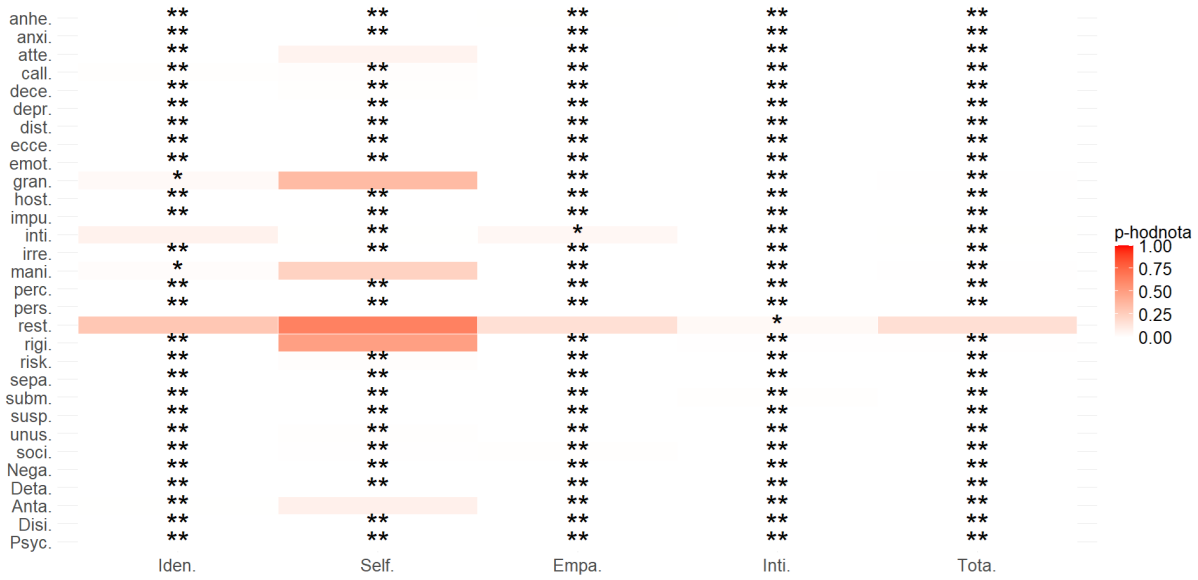


(a) Štatistická významnosť korelácií pre **nepacientov (ženy)**.

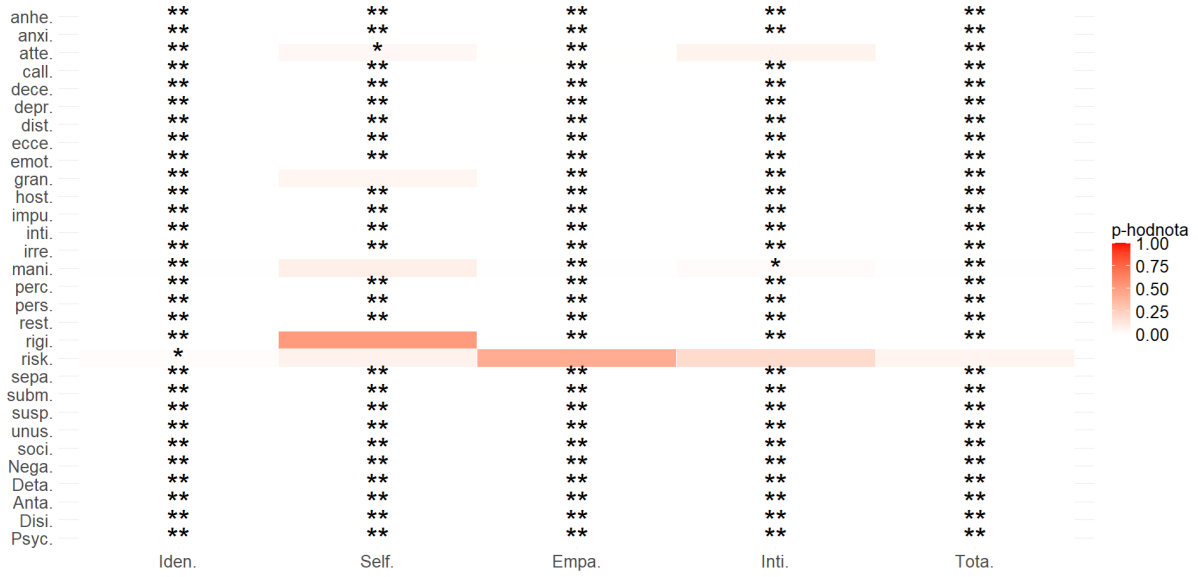


(b) Štatistická významnosť korelácií pre **nepacientov (muži)**.

Obr. 42: Štatistická významnosť korelácií medzi dotazníkmi *PID-5* a *LPFS-SR* pre *nepacientov* (zdroj: vlastné spracovanie v softvéri R).



(a) Štatistická významnosť korelácií pre **pacientov (ženy)**.



(b) Štatistická významnosť korelácií pre **pacientov (muži)**.

Obr. 43: Štatistická významnosť korelácií medzi dotazníkmi *PID-5* a *LPFS-SR* pre *pacientov* (zdroj: vlastné spracovanie v softvéri R).

Príloha G

Vybrané aspekty datasetu *PID-5_ aspekty* podľa jednotlivých metód selekcie

Tabuľka 28: Vybrané aspekty podľa jednotlivých metód selekcie.

Spätná selekcia (AIC)	**Spätná selekcia (BIC)**	**Lasso ($\lambda_{\mathbf{min}}$)**	**Lasso ($\lambda_{\mathbf{1se}}$)**
gend.	gend.	gend.	gend.
age.	age.	age.	age.
atte.	atte.	anhe.	depr.
dece.	depr.	atte.	impu.
depr.	irre.	call.	irre.
dist.	risk.	dece.	risk.
host.	sepa.	depr.	sepa.
impu.		dist.	susp.
inti.		host.	soci.
irre.		impu.	
mani.		inti.	
rest.		irre.	
risk.		mani.	
sepa.		perc.	
susp.		rest.	
soci.		rigi.	
		risk.	
		sepa.	
		susp.	
		soci.	

Príloha H

Kód: Cronbachová alfa

```
# Rozdelenie dat na pacientov a nepacientov
data_DSM_pac <- data_DSM[data_DSM$patient == 1, ]
data_DSM_nep <- data_DSM[data_DSM$patient == 0, ]

# Premenne -> PACIENT
Antagonism <- data_DSM_pac[, c("Q107", "Q125", "Q162", "Q180", "Q219", "Q41", "Q53", "
    Q56", "Q76", "Q126",
                           "Q134", "Q142", "Q206", "Q214", "Q218", "Q40", "Q65", "Q114
    ", "Q179", "Q187",
                           "Q197")]
### Kronbachova alfa
alpha(Antagonism, check.keys = FALSE)$total$raw_alpha

# Premenne -> NEPACIENT
Antagonism <- data_DSM_nep[, c("Q107", "Q125", "Q162", "Q180", "Q219", "Q41", "Q53", "
    Q56", "Q76", "Q126",
                             "Q134", "Q142", "Q206", "Q214", "Q218", "Q40", "Q65", "
    Q114", "Q179", "Q187",
                             "Q197")]
### Kronbachova alfa
alpha(Antagonism, check.keys = FALSE)$total$raw_alpha
```

Príloha I

Kód: CFA

```
# CFA model
model <- '
  Negativeaffectivity =~ Q18 + Q62 + Q102 + Q122 + Q138 + Q165 + Q181 + Q79 + Q93 +
    Q95 + Q96 + Q109 + Q110 + Q130 + Q141 + Q174 + Q12 + Q50 + Q57 + Q64 + Q127 + Q149
     + Q175
  Detachment =~ Q10 + Q20 + Q75 + Q82 + Q136 + Q146 + Q147 + Q161 + Q182 + Q186 + Q1 +
     Q23 + Q26 + Q30 + Q124 + Q155 + Q157 + Q189 + Q89 + Q97 + Q108 + Q120 + Q145 +
    Q203
  Antagonism =~ Q107 + Q125 + Q162 + Q180 + Q219 + Q41 + Q53 + Q56 + Q76 + Q126 + Q134
     + Q142 + Q206 + Q214 + Q218 + Q40 + Q65 + Q114 + Q179 + Q187 + Q197
  Desinhibition =~ Q31 + Q129 + Q156 + Q160 + Q171 + Q201 + Q210 + Q4 + Q16 + Q17 +
    Q22 + Q58 + Q204 + Q6 + Q29 + Q47 + Q68 + Q88 + Q118 + Q132 + Q144 + Q199
  Psychoticism =~ Q94 + Q99 + Q106 + Q139 + Q143 + Q150 + Q194 + Q209 + Q5 + Q21 + Q24
     + Q25 + Q33 + Q52 + Q55 + Q70 + Q71 + Q152 + Q172 + Q185 + Q205 + Q36 + Q37 + Q42
     + Q44 + Q59 + Q77 + Q83 + Q154 + Q192 + Q193 + Q213 + Q217
'
cfa_result <- cfa(model, data = data_DSM, estimator = "MLR", std.lv = TRUE)
summary(cfa_result, standardized = TRUE, ci=TRUE, fit.measures = TRUE, rsquare = TRUE)
```

Príloha J

Kód: Metódy selekcie premenných

```
# Rozdelenie na train (80%) a test (20%)
set.seed(123456)
n <- nrow(data_DSM_1)
index <- sample(1:n)
ind.train <- index[1:floor(0.8*n)]
ind.test <- index[(floor(0.8*n) + 1):n]
train <- data_DSM_1[ind.train,]
test <- data_DSM_1[ind.test,]

## Krokove metody
n.train <- nrow(train)
mena <- colnames(data_DSM_1[, -1])
m.full <- glm(Patient ~ ., data=train, family = binomial(link = "logit"))

# BACKWARD AIC (rovnaky postup aj pre BACKWARD BIC)
backward.AIC <- step(m.full, direction="backward", k=2)
backward.AIC$coef

# Metriky
var_bAIC <- names(backward.AIC$coefficients)[-c(1)]
var_bAIC <- c("Patient", var_bAIC)
test_bAIC <- test[, var_bAIC, drop = FALSE]
y.bAIC <- predict(backward.AIC, test_bAIC, type="response")

roc_train <- roc(train$Patient, predict(backward.AIC, train, type="response"))
optimal_threshold_bAIC <- coords(roc_train, "best")$threshold
roc_bAIC <- roc(test_bAIC$Patient, y.bAIC)
auc_bAIC <-  auc(roc_bAIC)
predicted_classes_bAIC <- ifelse(y.bAIC >= optimal_threshold_bAIC, 1, 0)
confusionMatrix(as.factor(predicted_classes_bAIC), as.factor(test_bAIC$Patient))

## Lasso s lambda.min (rovnaky postup aj pre LASSO S LAMBDA.1SE)
x <- as.matrix(train[, -1])
y <- train$Patient
cv_lasso <- cv.glmnet(x, y, alpha = 1, family = "binomial", type.measure = "deviance",
    nfolds = 100)
lambda_opt <- cv_lasso$lambda.min
coef_lasso <- coef(cv_lasso, s = lambda_opt)
selected_components <- rownames(coef_lasso)[which(coef_lasso != 0)][-1]

# Logisticka regresia
train_selected <- train[, c(selected_components, "Patient")]
test_selected <- test[, c(selected_components, "Patient")]
log_model <- glm(Patient ~ ., data = train_selected, family = binomial(link = "logit")
   )

predictions <- predict(log_model,test_selected, type="response")
log_roc <- roc(response = test_selected$Patient, predictor = predictions)
auc(log_roc)
train_predictions <- predict(log_model, train_selected, type = "response")
train_roc <- roc(train_selected$Patient, train_predictions)
optimal_threshold <- coords(train_roc, "best")$threshold[1]
predicted_classes <- ifelse(predictions >= optimal_threshold, 1, 0)
confusionMatrix(as.factor(predicted_classes), as.factor(test_selected$Patient))
```

Príloha K

Kód: Logistická regresia

```
# Bootstrap
R <- 1000
n <- nrow(data_DSM_1_sel)
log_model <- glm(Patient ~ ., data = data_DSM_1_sel, family = binomial(link = "logit")
    )
k <- length(coef(log_model))
A <- matrix(nrow = R, ncol = k, dimnames = list(paste("Sample",1:R), names(coef(log_
    model))))
B <- matrix(nrow = R, ncol = k, dimnames = list(paste("Sample",1:R), names(coef(log_
    model))))
C <- matrix(nrow = R, ncol = 5, dimnames = list(paste('Sample',1:R), c("auc","
    threshold", "accuracy","sensitivity","specificity")))
true_prop_0 <- 0.781
true_prop_1 <- 0.219

set.seed(111)
for(i in 1:R){
  train_idx <- sample(1:n, size = n, replace = TRUE)
  oob_idx   <- setdiff(1:n, unique(train_idx))
  train_data <- data_DSM_1_sel[train_idx, ]
  test_data  <- data_DSM_1_sel[oob_idx,  ]
  # Pocet dat
  n_total <- nrow(train_data)
  n_one <- sum(train_data$Patient==1)
  n_zero <- sum(train_data$Patient==0)
  # Vahy
  weight_0 <- true_prop_0 / (n_zero / n_total)
  weight_1 <- true_prop_1 / (n_one / n_total)
  train_data$weights <- ifelse(train_data$Patient == 0, weight_0, weight_1)
  # Model
  log_model <- glm(Patient ~ Gender + Age + Depressivity + Impulsivity +
    Irresponsibility + RiskTaking + SeparationInsecurity + Suspiciousness +
    SocialWithdrawal, data = train_data, family = binomial(link = "logit"), weights =
    weights)
  summary_model <- summary(log_model)
  B[i,] <- coef(log_model)
  A[i,] <- exp(coef(log_model))
  # Optimalny prah
  train_predicted_probs <- predict(log_model, newdata = train_data, type = "response")
  roc_train <- roc(train_data$Patient, train_predicted_probs)
  optimal_threshold <- coords(roc_train, "best")$threshold[1]
  # Predikcie
  predicted_probs <- predict(log_model, newdata = test_data,type = "response")
  # ROC krivka
  roc <- roc(test_data$Patient, predicted_probs)
  auc <- auc(roc)
  predicted_classes <- ifelse(predicted_probs >= optimal_threshold, 1, 0)
  # Confusion matrix
  confusion_matrix <- table(Predicted = factor(predicted_classes, levels = c(0,1)),
    Actual = factor(test_data$Patient, levels = c(0,1)))
  accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
  # Senzitivita / Specificita
  TP <- confusion_matrix["1","1"]
  TN <- confusion_matrix["0","0"]
```

```
  FP <- confusion_matrix["1","0"]
  FN <- confusion_matrix["0","1"]
  sensitivity <- TP / (TP + FN)
  specificity <- TN / (TN + FP)
  C[i,1] <- auc
  C[i,2] <- optimal_threshold
  C[i,3] <- accuracy
  C[i,4] <- sensitivity
  C[i,5] <- specificity
}

# Finalny model
data_DSM_1_sel$Age <- data_DSM_1_sel$Age - 35
log_final <- glm(Patient ~ ., data = data_DSM_1_sel, family = binomial(link = "logit")
    )
summary(log_final)
coef(log_final)
exp(coef(log_final))
vif(log_final)
drop1(log_final, test = "Chisq")

# Test o submodeli
log_SUBMODEL <- glm(Patient ~ Gender + Age + Depressivity  + Irresponsibility +
    RiskTaking + SeparationInsecurity, data = data_DSM_1_sel, family = binomial(link =
     "logit"))
residual.devianceSUBMODEL <- log_SUBMODEL$deviance
residual.deviance <- log_final$deviance
1-pchisq(residual.devianceSUBMODEL-residual.deviance, df=10-7)
```

Príloha L

Kód: LDA a klasifikačný strom

```
### LDA
k <- 10
folds <- sample(cut(seq(1,nrow(data_DSM_1_sel)), breaks=k, labels=FALSE))
lda_acc <- rep(0, k)
lda_auc <- rep(0, k)
lda_sensitivity <- rep(0, k)
lda_specificity <- rep(0, k)
lda_threshold <- rep(0,k)
for (i in 1:k) {
  test_ind <- which(folds == i)
  train_cv <- data_DSM_1_sel[-test_ind,]
  test_cv <- data_DSM_1_sel[test_ind,]
  # Model
  lda_model <- lda(Patient ~ ., data = train_cv, prior = c(0.781, 0.219))
  # Optimalny prah
  train_predictions <- predict(lda_model, train_cv)
  train_roc <- roc(response = train_cv$Patient, predictor = train_predictions$
    posterior[,2])
  lda_threshold[i] <- coords(train_roc, "best")$threshold[1]
  # AUC
  predictions <- predict(lda_model, test_cv)
  lda_roc <- roc(response = test_cv$Patient, predictor = predictions$posterior[,2])
  lda_auc[i] <- auc(lda_roc)
  predicted_classes <- ifelse(predictions$posterior[,2] >= lda_threshold[i], 1, 0)
  # Presnost
  confusion_matrix <- table(Predicted = predicted_classes, Actual = test_cv$Patient)
  lda_acc[i] <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
  # Senzitivita / Specificita
  TN <- confusion_matrix[1, 1]
  FP <- confusion_matrix[2, 1]
  FN <- confusion_matrix[1, 2]
  TP <- confusion_matrix[2, 2]
  lda_sensitivity[i] <- TP / (TP + FN)
  lda_specificity[i] <- TN / (TN + FP)
}
# Finalny Model
lda_model <- lda(Patient ~ ., data = data_DSM_1_sel, prior = c(0.781, 0.219))
lda_weights <- as.data.frame(lda_model$scaling)
colnames(lda_weights) <- "LD1_Weight"

### Rozhodovacie stromy
set.seed(123)
k <- 10
folds <- sample(rep(1:k, length.out = nrow(data_DSM_1_sel)))
ks_acc <- numeric(k)
ks_sensitivity <- numeric(k)
ks_specificity <- numeric(k)
ks_auc <- numeric(k)
ks_threshold <- numeric(k)
for (i in 1:k) {
  test_ind <- which(folds == i)
  train_cv <- data_DSM_1_sel[-test_ind, ]
  test_cv  <- data_DSM_1_sel[test_ind, ]
  # Model
```

```
  rpart_model <- rpart(
    Patient ~ .,
    data = train_cv,
    method = "class",
    parms = list(prior = c(.781,.219), split = "information"))
  # Optimalny prah
  train_preds <- predict(rpart_model, train_cv, type = "prob")
  train_roc <- roc(response = train_cv$Patient, predictor = train_preds[,2])
  ks_threshold[i] <- coords(train_roc, "best")$threshold[1]
  # Predikcie
  preds <- predict(rpart_model, test_cv, type = "prob")
  pred_class <- ifelse(preds[,2] >= ks_threshold[i], 1, 0)
  # Confusion matrix
  confusion_matrix <- table(
    Predicted = pred_class,
    Actual = as.numeric(test_cv$Patient))
  TN <- confusion_matrix[1, 1]
  FP <- confusion_matrix[2, 1]
  FN <- confusion_matrix[1, 2]
  TP <- confusion_matrix[2, 2]
  # Presnost
  total <- TN + FP + FN + TP
  ks_acc[i] <- (TP + TN) / total
  # Senzitivita / Specificita
  ks_sensitivity[i] <- TP / (TP + FN)
  ks_specificity[i] <- TN / (TN + FP)
  # AUC
  roc <- roc(response = test_cv$Patient, predictor = preds[,2])
  ks_auc[i] <- auc(roc)
}
# Vykreslenie klasifikacneho stromu
patient_tree <- rpart(pati.~., data = data_DSM_1_sel,
                      method = "class", parms = list(split = "information"),
                      control = rpart.control(xval = 100))
summary(patient_tree)
rpart.plot(patient_tree,main="")
```