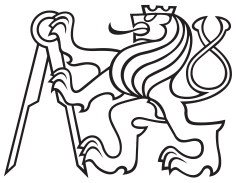
Diplomová práce



České
vysoké
učení technické
v Praze

F3

Fakulta elektrotechnická
Katedra teorie obvodů

Predikce funkčního vyústění schizofrenie z multimodálních dat

Marie Turnovcová

Vedoucí: Ing. Eduard Bakštein, Ph.D.
Studijní program: Lékařská elektronika a bioinformatika
Specializace: Zpracování signálů
Květen 2024

Poděkování

Ráda bych poděkovala vedoucímu mé diplomové práce Ing. Eduardu Bakšteinovi, Ph.D. za jeho čas, ochotu i podporu při konzultacích této práce. Děkuji také mé rodině a Petrovi za jejich podporu během celého studia.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracovala samostatně, a že jsem uvedla veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze, 24. května 2024

Abstrakt

Schizofrenie je závažné duševní onemocnění s různými klinickými průběhy, jehož neurobiologická podstata není známa a neexistuje univerzální léčba. Predikce funkčního vyústění může zlepšit prognózu a kvalitu života pacientů, zejména v časných stádiích nemoci. Tato diplomová práce zkoumá možnosti predikce funkčního vyústění z multimodálních dat a porovnává několik metod regrese pro data s vysokou mírou kolinearity. Použitý multimodální dataset pochází z Národního ústavu duševního zdraví z longitudinální studie pacientů s první epizodou schizofrenie (Early-Stage Schizophrenia Outcome). Data byla analyzována pomocí metod supervidovaných hlavních komponent (SPCA) dle Baira a Barshana. Byly použity modely lineární regrese a elasticNet regrese, přičemž hyperparametry modelů byly nastaveny pomocí 5-fold krosvalidace. Výsledky ukazují, že Bairova metoda SPCA nebyla užitečná, zatímco Barshanovy metody SPCA a Dual SPCA mohou být přínosné, pokud jsou doplněny robustními modely. Nejvyšší přesnosti predikce bylo dosaženo u psychologické domény kvality života s modelem lineární regrese a u součtu negativních příznaků s modelem elasticNet regrese, obě predikce po agregaci dat metodou Dual Barshan SPCA. Modely vysvětlily na testovací sadě 29 % variability dat.

Klíčová slova: schizofrenie, predikce, SPCA

Vedoucí: Ing. Eduard Bakštein, Ph.D.

Abstract

Schizophrenia is a serious mental illness with a variety of clinical courses, the neurobiological basis of which is unknown and there is no universal treatment. Predicting the functional outcome can improve the prognosis and quality of life of patients, especially in the early stages of the disease. This thesis explores the possibilities of predicting functional outcome from multimodal data and compares several regression methods for data with a high degree of collinearity. The multimodal dataset used is from the National Institute of Mental Health longitudinal study of patients with the first episode of schizophrenia (Early-Stage Schizophrenia Outcome). Data were analyzed using supervised principal component analysis (SPCA) methods according to Bair and Barshan. Linear regression and ElasticNet regression models were used, and the hyperparameters of the models were adjusted using 5-fold cross-validation. The results show that Bair's SPCA method was not useful, while Barshan's SPCA and Dual SPCA methods can be beneficial when complemented with robust models. The highest prediction accuracy was achieved for the psychological quality of life domain with the linear regression model and for the sum of negative symptoms with the elasticNet regression model. Both predictions were made using data aggregation with the Dual Barshan SPCA method. The models explained 29 % of the variability in the data in the test set.

Keywords: schizophrenia, prediction, SPCA

Title translation: Clinical outcome prediction of schizophrenia from multimodal data

Obsah

Obrázky

Tabulky

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Turnovcová** Jméno: **Marie** Osobní číslo: **483053**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra teorie obvodů**

Studijní program: **Lékařská elektronika a bioinformatika**

Specializace: **Zpracování signálů**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Predikce funkčního vyústění schizofrenie z multimodálních dat

Název diplomové práce anglicky:

Clinical Outcome Prediction in Schizophrenia Using Multimodal Data

Pokyny pro vypracování:

Schizofrenie je závažné psychiatrické onemocnění, charakterizované m.j. vysokou variabilitou klinických průběhů - četnost klinických epizod i míra progrese se mezi pacienty výrazně liší. Standardně probíhá diagnostika pomocí posouzení klinických příznaků a strukturovaného rozhovoru s psychiatrem.
Cílem projektu je predikce funknčního vyústění schizofrenie z multimodálních dat, zahrnujících kromě klinických a demografických parametrů také již předzpracovaná neurozobrazovací data. Projekt využívá předzpracovaná tabulková data ze studie ESO, sledující uvedené parametry pacientů při výskytu první epizody a následně po 1 a 3 letech. Dataset zahrnuje několik domén: klinická data, demografii, anamnézu, psychiatrické škály a rosáhlá tabulková data, získaná zpracováním několika modalit magnetické rezonance.
Úkoly:
1. Nastudujte problematiku hodnocení funkčního vyústění u schizofrenie (viz také Meehan 2022) a parametrů, sledovaných ve studii ESO.
2. Proveďte přípravu datasetu a následnou explorační analýzu, zaměřenou také na souvislosti mezi proměnnými napříč různými doménami.
3. Navrhněte vhodné metody pro agregaci dat a selekci příznaků (např SPCA dle Barshan et al. a Bair et al.).
4. Vytvořte model pro predikci funkčního vyústění schizofrenie ze zadaných dat, využívající agregace z předchozího bodu.
5. Pro nastavení hyper/parametrů modelů zvolte vhodné krosvaliační schéma, viz Chekroud 2024.

Seznam doporučené literatury:

[1] Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by supervised principal components. Journal of the American Statistical Association, 101(473), 119–137. https://doi.org/10.1198/016214505000000628
[2] Barshan, E., Ghodsi, A., Azimifar, Z., & Zolghadri Jahromi, M. (2011). Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. Pattern Recognition, 44(7), 1357–1371. https://doi.org/10.1016/j.patcog.2010.12.015
[3] Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P. R., Koutsouleris, N., Krumholz, H. M., Krystal, J. H., & Paulus, M. (2024). Illusory generalizability of clinical prediction models. Science, 383(6679), 164–167. https://doi.org/10.1126/science.adg8538
[4] Meehan, A. J., Lewis, S. J., Fazel, S., Fusar-Poli, P., Steyerberg, E. W., Stahl, D., & Danese, A. (2022). Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. Molecular Psychiatry, 27(6), 2700–2708. https://doi.org/10.1038/s41380-022-01528-4

Jméno a pracoviště vedoucí(ho) diplomové práce:

Ing. Eduard Bakštein, Ph.D. Analýza a interpretace biomedicínských dat FEL

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

Datum zadání diplomové práce: **15.02.2024** Termín odevzdání diplomové práce: **24.05.2024**

Platnost zadání diplomové práce: **21.09.2025**

Ing. Eduard Bakštein, Ph.D.	doc. Ing. Radoslav Bortel, Ph.D.	prof. Mgr. Petr Páta, Ph.D.
podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

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

. | Datum převzetí zadání

Podpis studentky

Kapitola 1

Úvod

Schizofrenie je závažné duševní onemocnění, které se projevuje různými klinickými průběhy a ovlivňuje myšlení, vnímání, emoce či chování jedince. Postihuje až 1% populace a má značný dopad na jedince i na celou společnost. Bohužel neurobiologická podstata tohoto onemocnění není známa, proto by schopnost predikce funkčního vyústění mohla představovat klíčovou oblast výzkumu a napomoci tak zlepšit prognózu a kvalitu života pacientů, a to zejména u časných stádií nemoci (tj. při první epizodě). Funkční vyústění schizofrenie zahrnuje škálu možných výsledků, od plného zotavení po úplné funkční omezení, kdy pacient v důsledku svého zdravotního stavu ztrácí schopnost vykonávat běžné denní aktivity. Použití multimodálního datasetu umožňuje, díky kombinaci dat z různých zdrojů, analyzovat schizofrenii komplexněji a poskytuje tak do budoucna naději na hlubší porozumění příčin i průběhu této nemoci.

Tato diplomová práce zkoumá možnosti predikce funkčního vyústění z multimodálních dat a porovnáná několik metod přístupu k řešení úlohy regrese z velkého počtu příznaků s vysokou mírou kolinearity. Multimodální dataset použitý v této práci, pochází z Národního ústavu duševního zdraví, z longitudinální studie pacientů s první epizodou schizofrenie (ESO, Early-Stage Schizophrenia Outcome).

Cíle této práce jsou:

1. nastudovat problematiku hodnocení funkčního vyústění u schizofrenie a parametrů sledovaných ve studii ESO,

2. provést přípravu datasetu a explorační analýzu datasetu,
3. navrhnout vhodné metody pro výběr příznaků a pro agregaci dat,
4. vytvořit model pro predikci funkčního vyústění schizofrenie, využívající navržené agregace,
5. zvolit vhodné krosvalidační schéma pro nastavení hyperparametrů modelu.

V kapitole 2 Úvod do problematiky je popsána schizofrenie, její symptomy a léčba, dále pak problematika funkčního vyústění. V další kapitole 3, věnované metodám použitým v této práci, je popsána jedna z metod vhodných pro agregaci dat: supervidovaná analýza hlavních komponent (Supervised principal component analysis, SPCA), dále regresní modely pro predikci a použité metriky pro evaluaci těchto modelů. Následující kapitola 4 zahrnuje popis multimodálního datasetu z ESO studie, podkapitolu tykající se parametrů, jež jsou v datasetu zahrnuty, podkapitolu věnovanou explorační analýze zaměřené také na souvislosti mezi proměnnými a na závěr podkapitolu popisující předzpracování datasetu. V předposlední kapitole 5 jsou uvedeny výsledky použitých modelů a následná diskuze. Závěrečná kapitola 6 tuto diplomovou práci shrnuje a uzavírá.

Kapitola 2

Úvod do problematiky

V této kapitole je nejprve v sekci 2.1 představeno onemocnění schizofrenie, včetně její základní charakteristiky (2.1.1), symptomů, jimiž se projevuje (2.1.2) a dostupné léčby (2.1.3). V sekci 2.2 je pak popsáno funkční vyústění a škály, které se k jeho hodnocení používají.

2.1 Schizofrenie

Schizofrenie je soubor symptomů zatím neznámé etiologie, převážně definovaný pozorovanými příznaky psychózy. [1] Její průběh i nástup je variabilní. Schizofrenie s časným nástupem je obvykle spojena s horší prognózou, zatímco při pozdějším nástupu nemoci jsou afektivní a sociální funkce s větší pravděpodobností zachovány. [2]

2.1.1 Základní popis

Schizofrenie je soubor neurovývojových poruch, které zahrnují změny v mozkových okruzích. [1] Onemocnění se u pacientů projevuje velmi individuálně. Nejde jen o rozličné příznaky, ale také o rychlost rozvinutí nemoci a závažnost nemoci. Někteří jedinci mohou mít ojedinělé epizody schizofrenie, jiní mohou trpět stálým chronickým stavem a u některých se mohou střídat relapsy s

remisemi. [3] Psychóza se většinou objevuje typicky už v pozdní adolescenci nebo rané dospělosti (mezi 18. a 25. rokem života). [1] Schizofrenie s plnou symptomatikou se před pubertou objevuje jen velmi zřídka. [2]

K rozvoji nemoci mohou přispět různé vlivy, například psychosociální faktory a užívání drog. [4] Mezi další popsané vlivy, které se mohou podílet na rozvoji schizofrenie, patří: podvýživa matky během těhotenství, infekce ve druhém trimestru těhotenství, perinatální poranění a expozice cytokinům. [1] Z dosavadních studií vyplývá, že schizofrenie má také genetický základ. Jedinci, kteří mají příbuzné s touto poruchou, mají vyšší riziko výskytu schizofrenie. Například podle studie od vědců Wray a Gottesman, kteří použili údaje z dánských národních registrů, má jedinec 67% predispozici k rozvoji schizofrenie, pokud alespoň jeden z jeho rodičů touto nemocí trpí. [5]

Protože neurobiologická podstata schizofrenie zatím není známa, předpokladem pro rozvoj této nemoci je tak interakce genetické informace s rizikovými faktory prostředí a životním stylem. [6]

Podle GBD (Global Burden of Disease) je odhadovaná absolutní hodnota globální prevalence schizofrenie 0,29%, avšak jiné zdroje uvádějí hodnoty obvykle vyšší. [7] Například podle článku z roku 2015 byl mediánový odhad celoživotní prevalence schizofrenie v letech 1990-2013 0,48 %. [8] Celosvětově se počet lidí s diagnózou schizofrenie zvýšil z 13,1 milionu v roce 1990 na 20,9 milionu v roce 2016. [9]

2.1.2 Symptomy

Schizofrenie je charakterizována poruchami v duševních modalitách, včetně myšlení, vnímání, sebeprožívání, poznávání či chování. Poruchy v myšlení se projevují například přetrvávajícími bludy či dezorganizovaným myšlením, poruchy ve vnímání halucinacemi. [2] Mezi příznaky schizofrenie se dále řadí ztráta sebeuvědomění, která se projevuje narušenou schopností rozlišit vlastní a cizí jednání [10], pocit, že myšlenky nebo chování jsou pod kontrolou vnější síly. Poruchy v poznávání mohou zahrnovat zhoršenou pozornost či verbální paměť [11]. Dalšími projevy schizofrenie může být ztráta motivace, otupené emoční projevy, bezúčelné chování, nepředvídatelné nebo nepřiměřené emoční reakce. Přítomny mohou být také psychomotorické poruchy, včetně katatonie. [2] Jedním z dalších symptomů mohou být poruchy řeči a další. [1]

Pro diagnózu schizofrenie musí podle ICD WHO alespoň 2 z následujících

příznaků přetrvávat po dobu nejméně jednoho měsíce a nesmějí být projevem jiného zdravotního stavu (např. nádoru mozku) ani nemohou být způsobeny účinkem látky nebo léku na centrální nervový systém, ani abstinenčními příznaky. Zároveň musí být splněn alespoň 1 příznak z následujícího výčtu od a) do d): [2]

a) Přetrvávající bludy,

b) Trvalé halucinace,

c) Dezorganizované myšlení (např. volné asociace, irelevantní řeč, neologismy),

d) Zážitky ovlivňování, pasivity nebo kontroly.

e) Negativní příznaky, jako je například afektivní zploštění, asociálnost nebo chudost řeči,

f) Hrubě dezorganizované chování bránící cílené činnosti (např. bizarně nebo bezúčelně vypadající chování, nepředvídatelné nebo nepřiměřené emoční reakce),

g) Psychomotorické poruchy (např. katatonický neklid či stupor).

Positive and Negative Syndrome Scale

Positive and Negative Syndrome Scale (PANSS) je škála k hodnocení dimenzí symptomů schizofrenie, která obsahuje 30 položek. Původně byly tyto položky seskupeny pouze do tří skupin: pozitivní a negativní symptomy a obecná psychopatologie. Avšak studie naznačují, že strukturu PANSS u osob se schizofrenií lépe zachycuje model pětifaktorový, obsahující opět skupinu pozitivních a negativních symptomů, dále dezorganizované myšlení, příznaky rozrušení a depresivní příznaky. Oproti trojfaktorovému modelu poskytuje pětifaktorový model empiricky podložené rozlišení jiných dimenzí symptomů. Položky PANSS škály jsou na základě strukturovaného klinického rozhovoru ohodnoceny na stupnici od 1 (bez příznaků) do 7 (extrémně symptomatické). [12]

Mezi pozitivní symptomy patří takové symptomy, které jsou abnormálně přítomné, příkladem jsou halucinace, iluze či grandiozita. Negativní symptomy abnormálně chybí, jde například o ztrátu či snížení schopnosti vyjadřovat emoce nebo je prožívat či nedostatek spontánnosti. [13] Podle International

Statistical Classification of Diseases (ICD WHO) mají pozitivní příznaky tendenci časem přirozeně slábnout, zatímco negativní příznaky často přetrvávají a jsou úzce spojeny s horší prognózou. [2] Dezorganizované myšlení zahrnuje slabou pozornost či potíže s abstraktním myšlením. Příznakem rozrušení může být nepřátelskost či nespolupráce a mezi depresivní příznaky se řadí například úzkost, deprese či pocity viny. [12]

2.1.3 Léčba

Psychogenetické testy ani specifické biomarkery, které by dokázaly určit vhodnou formu léčby schizofrenie, bohužel navzdory intenzivnímu výzkumu zatím neexistují. Léčba je indikována podle symptomů konkrétního pacienta. Mnohdy tak dochází ke změnám medikace, než je nalezena vyhovující. [14] Medikace zahrnuje podávání antipsychotik, která tlumí symptomy nemoci na základě mechanismu blokády dopaminového receptoru D2. [15]

Antipsychotika někdy poskytují dramatickou symptomatickou úlevu od halucinací a bludů a zlepšení u dezorganizovaných myšlenek i chování. Bohužel jsou ale spojeny s množstvím nežádoucích účinků, z nichž některé jsou medicínsky závažné, mnohé z nich pak také negativně ovlivňují postoje pacientů k léčbě. Někteří odborníci a směrnice dokonce doporučují výběr antipsychotik ne za základě účinnosti, jež je v mnoha případech podobná, ale na základě profilů vedlejších účinků, které jsou rozmanité. [16]

Dále mohou být při léčbě vhodné různé psychoterapeutické techniky. Psychologická léčba, například kognitivně behaviorální terapie (KBT) nebo podpůrná psychoterapie, může mírnit příznaky nemoci, pomáhá rozvíjet sociální dovednosti, získat soběstačnost, identifikovat včasné varovné příznaky relapsu a prodloužit období remise. [13]

Pro pacienty trpící schizofrenií bývá často typická také nutná hospitalizace. Některým jedincům užívání antipsychotik výrazně zlepší kvalitu života a sníží projevy nemoci, někteří jedinci bohužel na medikaci téměř nereagují.[3]

2.2 Hodnocení funkčního vyústění v dostupné literatuře

Funkční vyústění lze definovat nebo měřit například jako kvalitu života, zaměstnání, schopnost samostatného života nebo schopnost plánovat a měnit základní denní činnosti. V používaných definicích funkčního vyústění avšak chybí shoda a vyvstává tak potřeba její definici standardizovat. [17]

V tabulce 2.1 jsou uvedeny vybrané studie, ve kterých byla řešena úloha spojená s funkčním vyústěním u schizofrenie. Funkční vyústění je v těchto studiích hodnoceno pomocí různých škál. Tyto škály jsou blíže popsány v následujícím seznamu:

- SLOF: Specific Level of Functioning Scale je hybridní škála, která hodnotí různé oblasti fungování. Obsahuje 43 položek seskupených do 6 následujících domén: fyzické fungování, dovednosti osobní péče, mezilidské vztahy, sociální přijímání, dovednosti každodenního života a pracovní dovednosti. Hodnocení fungování pacientů je provedeno klíčovým pečovatelem [18]

- GAF: Global Assessment of Functioning neboli globální hodnocení funkčnosti je systém pro hodnocení psychologického, sociálního a pracovního fungování. Tato škála byla odvozena od škály GAS a nyní, stejně jako u GAS, je její rozsah od 0 do 100, kde vyšší hodnota znamená zdravějšího jedince. Tato škála je měřítkem celkového poškození způsobeného duševními faktory. [19] Podle některých studií je GAF psychometricky problematický, protože je vázán spíše s psychiatrickými symptomy pacienta, než s jeho skutečným fungováním: není schopen odlišit psychiatrické symptomy od vztahovéno, sociálního a pracovního fungování. [20, 21]

- SOFAS: Social and Occupational Functioning Assessment Scale je hodnotící škála sociálního a pracovního fungování. Oproti GAFu, který zahrnuje do hodnocení pouze duševní poškození, tato škála do hodnocení zahrnuje také dysfunkce související s obecným zdravotním postižením. Používá se k hodnocení úrovně fungování u pacientů se schizofrenií. [20, 19]

- RFS: Role Functioning Scale je škála pro hodnocení fungování jedinců ve specifikovaných oblastech každodenního života. Konkrétně jde o následující oblasti: produktivitu v práci, samostatné bydlení a péči o sebe, vztahy v bezprostřední sociální síti a vztahy v širší sociální síti. Hodnoty v každé z oblastí se pohybují na škále od 1 (minimální úroveň fungování v dané roli) po 7. Celkové skóre ze všech čtyř oblastí představuje globální index fungování s hodnotami v rozmezí 4 až 28. [22]

- GF: Social and Role je škála globálního fungování zaměřená na sociální fungování a fungování v rolích. Byla vyvinuta speciálně pro použití u jednotlivců s vysokým rizikem psychózy. Škála zohledňuje věk, fázi onemocnění, změny v průběhu času a odděluje domény sociálního fungování a fungování v rolích. Pro obě domény je škála od 0 do 10. Tyto škály nejsou ovlivněny závažností symptomů a posuzují specifické domény fungování. [21]

- SF-36: 36-Item Short Form Health Survey je dotazník, který měří 8 oblastí: fyzické fungování, fyzickou roli, tělesnou bolest, celkové zdraví, vitalitu, sociální fungování, emocionální role a duševní zdraví. Těchto 8 oblastí je možné seskupit do 2 konceptů, a to do fyzické a mentální dimenze. [23]

- PAS: Premorbid Adjustment Scale neboli Škála premorbidního přizpůsobení je škála, která hodnotí úroveň fungování v oblastech sociální přístupnosti, vztahy s vrstevníky, schopnosti fungovat mimo nukleární rodinu a schopnost vytvářet intimní sociosexuální vazby, a to před nástupem schizofrenie. [24]

- SASS: Social Adaptation Self-evaluation Scale (sebehodnotící škála sociální adaptace) je škála 21 položek, které zkoumají oblasti práce a volného času, rodinných i mimorodinných vztahů, intelektuálních zájmů, spokojenosti v rolích a vnímání své schopnosti řídit a kontrolovat své prostředí. [25]

- SIB-R: Scales of Independent Behavior-Revised je revidovaná škála nezávislého chování. Tato škála je komplexní hodnocení adaptivního a maladaptivního chování sloužící k určení úrovně fungování člověka v klíčových oblastech chování. Hodnocení je vedeno formou strukturovaného rozhovoru nebo pomocí kontrolního seznamu. [26]

- GAS: Global Assessment Scale neboli škála globálního hodnocení je měřítko celkové závažnosti poruchy. Hodnoty skóre této škály se pohybují od 1 (nejvíce nemocný) do 100 (bez symptomů). Stupnice je rozdělena na 10 rovnoměrných intervalů, každý interval má maximální skóre 10. [27] Speciální forma této škály přizpůsobená dětem se jmenuje Children's Global Assessment Scale neboli dětská škála globálního hodnocení. Ta je měřítkem celkové závažnosti poruchy pro děti, jejíž skóre se opět pohybuje od 1 (největší postižení) do 100 (bez symptomů). Do číselného skóre je otisknuto sociální i psychiatrické fungování. [28]

- PSP: Personal and Social Performance Scale je polostrukturovaný rozhovor, který zkoumá následující oblasti: péči o sebe, společensky užitečné činnosti, osobní a sociální vztahy a rušivé a agresivní chování. Každá oblast je hodnocena od 0 do 5 bodů, kde 5 je velmi závažné postižení. [29]

Z dalších škál používaných pro hodnocení funkčnosti je možné zmínit Defensive Functioning Scale a Global Assessment of Relational Functioning, dále například Social Functioning Scale (SFS), Disability Assessment Scale (DAS II), Strauss-Carpenter Level of Functioning scale (SC-LOF), Social Adjustment Scale-II (SAS-II) nebo Social Behaviour Schedule (SBS). [19, 30]

Je zřejmé, že používaných škál je mnoho. Tyto škály hodnotí různé aspekty funkčnosti pacienta, většina z nich využívá k vyhodnocení bodového hodnocení. Některé škály jsou určeny pro specifické cílové skupiny (např. GF: Social a GF: Role pro jednotlivce s vysokým rizikem psychózy), zatímco některé jsou obecné (např. SF-36). Dalším zásadním rozdílem mezi používanými škálami jsou různé metody hodnocení: některé škály jsou sebehodnotící (např. SASS), jiné jsou hodnoceny odborníky, klíčovými pracovníky. Deficity ve funkčnosti jsou jedny z určujících znaků schizofrenie. Sociální funkčnost zahrnuje pracovní oblast, interpersonální vztahy a péči o sebe. Rozlišit deficit v sociální funkčnosti od přetrvávajících negativních symptomů není snadné. [30]

Studie	Řešená úloha	Vstupní data	Funkční vyústění
Giuliani [18]	Vztahy mezi faktory a FO u osob se SZ a u jejich příbuzných prvního stupně	PANSS, BNSS, CDSS, SHRS a další	SLOF
Walther [31]	Predikce FO a FC pomocí gestikulace a neverbální sociální precepce	PANSS, SOFAS, TULIA, PONS	GAF, SOFAS
Agid[32]	Predikce klinického a funkčního zlepšení u pac. se schizofrenií léčených antipsychotiky	demografie, dávkování antipsychotik, změny v klinických symptomech(BPRS skóre), kogn. postižení v poč. fázi, pohybové poruchy, GAF	GAF (zlepšení GAF skóre alespoň o 50% po dobu 6 měsíců)
Kimmy[33]	Role zpracování emocí u pac. se schizofrenií a její vztah k psychosoc. funkčnímu výsledku	demografie, délka nemoci,diagnostika, užívaná antipsychotika, výsledky psycholog. a psychiatrických měření	RFS
Ricardo[34]	Identifikace prediktorů dlouhodobého funkčního vyústění u osob s klinickým vysoce rizikovým stavem pro psychózu	kognitivní, klinická a demografická data	GF:Social, GF:Role
Rebecca[35]	Predikce a hodnocení funkčního výsledku u schizofrenie	PANSS skóre, sociodemograf. informace, klinické hodnoty	GAF, SOFAS a SF-36 skóre přesahující určené hodnoty
Kravariti[36]	Jaké faktory ovlivňují psychosociálnímu vyústění SZ u pacientů s EOS	premorbidní funkce, věk při nástupu SZ, PANSS, DUP, závažnost symptomů	GAF, PAS, SASS
Cervellione [37]	Vztah mezi neurokognitivními funkcemi adolescentů s EOS a jejich FO	Neuropsychologické testy	SIB-R, CGAS
Giordano [29]	Identifikace faktorů pro FO u osob trpících SZ	demografie, klinické charakteristiky, neurokognitivní testy, symptomatika SZ (BNSS, PANSS)	PSP (celkové PSP skóre, oblast společensky užitečných aktivit a oblast osobních a sociálních vztahů)

SZ: schizofrenie, DUP: doba neléčené psychózy, EOS: raný nástup schozofrenie, FO: funkční vyústění, FC: funkční kapacita, TULIA: Test of Upper Limb Apraxia (hodnocení gestikulace), PONS: Profile of Nonverbal Sensitivity (hodnocení neverbálního sociálního vnímání)

Tabulka 2.1: Vybrané studie

Kapitola 3

Použité metody

V této kapitole budou nejprve představeny vybrané metody pro agregaci dat (3.1), konkrétně jde o metody analýzy hlavních komponent pod dohledem (SPCA), a jejich aplikace v této diplomové práci (3.1.3). Následuje popis vybraných regresních modelů (3.2), lineární regrese (3.2.1) a elasticNet regrese (3.2.2) a jejich aplikace při predikci funkčního vyústění schizofrenie (3.2.3). Poslední sekce této kapitoly je věnována evaluaci modelu (3.3). Procesy agregace, predikce i veškeré související zpracování bylo realizováno v Pythonu ve vývojovém prostředí PyCharm.

3.1 Vybrané metody pro agregaci dat

Vysokorozměrná data často obsahují mnoho nadbytečných informací, včetně korelovaných či duplicitních faktorů. V tomto případě hraje roli redukce dimenzionality, díky které je vytvořen nízkodimenzionální prostor příznaků. Tímto způsobem by měly být účinky irelevantních informací omezeny. [38]

Protože mechanismus nemoci schizofrenie zatím není znám, neznáme ani konkrétní parametry, ze kterých její funkční či klinické vyústění predikovat. Proto byly pro tuto práci vybrány metody supervised principal components, které podmnožinu prediktorů vybírají na základě asociace s outcomem (vyústěním). Konkrétně jde o metodu SPCA podle Baira, popsanou v následující sekci 3.1.1, a podle Barshan, která je popsána v sekci 3.1.2.

3.1.1 SPCA Bair

Jednou z nejrozšířenejších metod supervidovaných hlavních komponent je metoda publikovaná Bairem a spoluautory [39], která se zaměřuje na metodu predikce výsledné proměnné Y na základě souboru prediktivních proměnných $X_1, X_2, ..., X_p$ naměřených u N jedinců. Tento přístup je vhodný použít v případě, že počet p výrazně převyšuje počet jedinců N. Technika řízených hlavních komponent umožňuje automaticky odhalit smysluplnou strukturu v datech a větší váhu přikládá prediktivním proměnným silně korelovaným s výslednou proměnnou Y. Vyhledává tedy hlavní komponenty s maximální závislostí na výsledné proměnné, čímž se liší od klasické PCA (Principal Component Analysis).

Popis metody

X je matice dat o rozměrech $p \times n$, kde p je počet proměnných a n je počet pozorování, Y je vektor výsledné proměnné o rozměru n. Cílem úlohy je najít ortogonální projekci $U^T X$ tak, aby závislost mezi $U^T X$ a Y byla maximální.

Postup metody je následující:

1. Výpočet standardizovaných regresních koeficientů s_j, které měří univerzální efekt na výslednou proměnnou Y každé proměnné X_j samostatně:

$$s_j = \frac{x_j^T y}{\|x_j\|}, \tag{3.1}$$

   kde $\|x_j\| = \sqrt{x_j^T x_j}$.

2. Vytvoření redukované datové matice X_θ obsahující pouze proměnné, jejichž jednorozměrný regresní koeficient překračuje prahovou hodnotu θ. Tedy platí, že C_θ je soubor indexů, pro které $\mid s_j \mid > \theta$. X_θ se pak skládá ze sloupců matice X odpovídajících C_θ. Parametr θ je odhadnut křížovou validací.

3. Výpočet první (nebo prvních několik) hlavních složek redukované matice dat pomocí singulárního rozkladu (SVD):

$$X_\theta = U_\theta D_\theta V_\theta^T, \tag{3.2}$$

   kde U_θ, D_θ a V_θ jsou příslušné matice singulárního rozkladu.

$U_\theta = (u_{\theta,1}, u_{\theta,2}, ..., u_{\theta,m})$, kde $u_{\theta,1}$ je první řízená hlavní komponenta.

4. Použití získaných hlavních komponent v regresním modelu k predikci výsledné proměnné Y. Univariátní regresní model s odpovědí y a prediktorem $u_{\theta,1}$:

$$\hat{y}^{spc,\theta} = \bar{y} + \hat{\gamma} \cdot u_{\theta,1}. \tag{3.3}$$

Matici U_θ můžeme vyjádřit následovně:

$$U_\theta = X_\theta V_\theta D_\theta^{-1} = X_\theta W_\theta \tag{3.4}$$

Tedy například $u_{\theta,1}$ je lineární kombinací sloupců matice X_θ: $u_{\theta,1} = X_\theta w_{\theta,1}$. Díky tomu je možno tuto predikci lineárního regresního modelu chápat jako omezený odhad lineárního modelu, který využívá všechny prediktory v X_θ:

$$\hat{y}^{\mathrm{spc},\theta} = \bar{y} + \hat{\gamma} \cdot X w_{\theta,1} = \bar{y} + X_\theta \hat{\beta}_\theta \tag{3.5}$$

kde $\hat{\beta}_\theta = \hat{\gamma} w_{\theta,1}$.

Predikce z regresního modelu je při testování vektoru příznaků x^* provedena následovně:

1. Vycentrování každé složky x^* pomocí průměrů odvozených z trénovacích dat, tj. $x_j^* \leftarrow x_j^* - \bar{x}_j$.

2. $$\hat{y}^* = \bar{y} + \hat{\gamma} \cdot x_\theta^{*T} w_{\theta,1} = \bar{y} + x_\theta^{*T} \hat{\beta}_\theta, \tag{3.6}$$

   kde x_θ^* je příslušný podvektor x^*.

3.1.2 SPCA Barshan

Odlišnou metodou supervidovaných hlavních komponent navrhla Barshan a spoluautoři v článku [40] z roku 2011.

Pro sadu n datových bodů $\{\mathbf{x}_i\}_{i=1}^{n}$, kde každý bod obsahuje p rysů, které jsou uspořádány do matice X o rozměrech $p \times n$, máme Y: matici o rozměru n obsahující výslednou proměnnou. Hledáme podprostor $U^T X$, ve kterém je maximalizována závislost mezi projekcí dat $U^T X$ a výsledkem Y. Pro měření závislosti mezi $U^T X$ a výstupní proměnnou Y používá tato metoda Hilbertovo-Schmidtovo kritérium nezávislosti, blíže popsané v následující podkapitole s názvem Empirické HSIC (3.1.2).

Potřebujeme tedy maximalizovat $\mathbf{tr}(HKHL)$, kde:

- K je jádro $U^T X$,
- L je jádro Y,
- $H_{ij} = I - n^{-1}\mathbf{e}\mathbf{e}^T$.

Tuto úlohu můžeme formulovat následovně:

$$\mathbf{tr}(HKHL) = \mathbf{tr}(HX^T UU^T XHL) = \mathbf{tr}(U^T XHLH^T X^T U)$$

Hledáme ortogonální transformační matici U, která mapuje data do prostoru, kde jsou prvky nekorelované. Optimalizační problém tak má následující podobu:

$$\arg\max_U \operatorname{tr}(U^T XHLH^T X^T U) \tag{3.7}$$

$$\text{za podmínky} \quad U^T U = I \tag{3.8}$$

Tento optimalizační problém lze řešit v uzavřeném tvaru. Pokud má symetrická reálná matice $Q = XHLH^T X^T$ vlastní hodnoty $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$ s odpovídajícími vlastními vektory $v_1, \ldots, v_p$, pak maximální hodnota funkce, která splňuje podmínku, je $\lambda_p + \lambda_{p-1} + \ldots + \lambda_{p-d+1}$ a optimální řešení je $U = [v_p, v_{p-1}, \ldots, v_{p-d+1}]$, kde d označuje dimenzi výstupního prostoru S.

Algoritmus SPCA Barshan je popsán v sekci s názvem Algoritmus Supervised PCA (3.1.2).

Mnohdy je dimenzionalita p matice dat X mnohem větší, než počet pozorování. Z tohoto důvodu byla vytvořena metoda Dual SPCA, která je méně závislá na dimenzi p. Její algoritmus je popsán v sekci s názvem Algoritmus Dual Supervised PCA (3.1.2).

Empirické HSIC

Empirické kritérium HSIC je odhad závislosti mezi dvěma náhodnými proměnnými X a Y na základě empirických dat, jehož rovnice výpočtu vypadá následovně:

$$HSIC_{emp}(Z, F, G) = (n-1)^{-2}\mathbf{tr}(KHLH), \tag{3.9}$$

kde:

- $Z = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ je soubor n nezávislých pozorování z rozdělení $P_{X,Y}$,
- K a L jsou matice jádrových funkcí pro X a Y odpovídající prvkům souboru Z,
- n je počet pozorování,
- H je matice centrování definovaná jako $H_{ij} = I - n^{-1}\mathbf{e}\mathbf{e}^T$, kde I je identita a $\mathbf{e}$ je vektor jedniček.

Algoritmus Supervised PCA

Algoritmus pro SPCA s použitím HSIC je dán následovně:

1. Inicializace $H_{ij} \leftarrow I - n^{-1}\mathbf{e}\mathbf{e}^T$.
2. Výpočet $Q \leftarrow XHLHX^T$.
3. Výpočet bází: U jsou vlastní vektory Q korespondující nejvyšším d vlastním hodnotám.
4. Kódování trénovacích data: $Z \leftarrow U^T X$.
5. Kódování testovacího příkladu: $\mathbf{z} \leftarrow U^T \mathbf{x}$.

Algoritmus Dual Supervised PCA

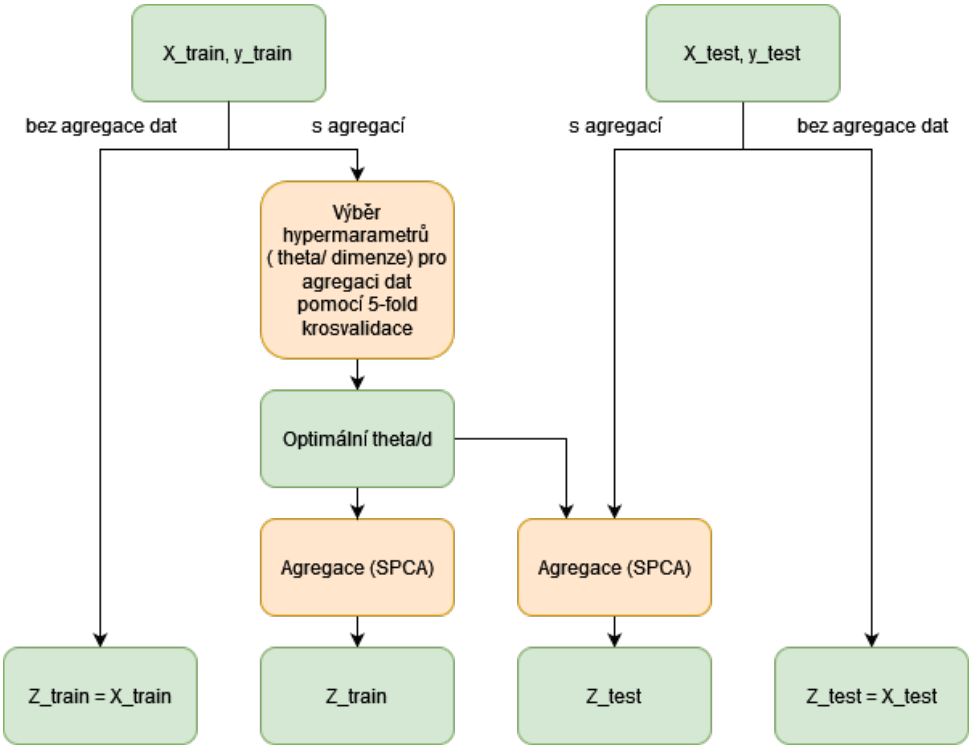
Algoritmus Dual Supervised PCA probíhá v následujících krocích:

1. Dekompozice L, $L = D^T D$.
2. $H_{ij} \leftarrow I - n^{-1}\mathbf{e}\mathbf{e}^T$
3. $\Psi \leftarrow XH\Delta^T$
4. Výpočet báze:
   - $V \leftarrow$ vlastní vektory $\Psi^T\Psi = \Delta H[X^T X]H\Delta^T$ odpovídající nejvyšším d vlastním hodnotám.
   - $\varepsilon \leftarrow$ diagonální matice odmocnin nejvyšších d vlastních hodnot $\Psi^T\Psi$.
   - $U \leftarrow \Psi V\varepsilon^{-1}$
5. Kódování trénovacích dat: $Z \leftarrow U^T X = \varepsilon^{-1}V^T\Delta H[X^T X]$
6. Kódování testovacího příkladu: $\mathbf{z} \leftarrow U^T\mathbf{x} = \varepsilon^{-1}V^T\Delta H[X^T\mathbf{x}]$

3.1.3 Agregace dat

Na obrázku 3.1 je zobrazeno schéma pro agregaci dat použité v této práci. Připravená data pro trénovací (_train) i testovací (_test) sadu jsou tvořena maticí X, která obsahuje vybrané příznaky pacientů z první vizity, a vybranou výstupní proměnnou y z druhé vizity. Příprava dat před agregací je blíže popsána v části 4.6.

V případě, že agregace dat není požadována, jsou výstupní matice Z shodné s vstupními maticemi X. V případě agregace dat je na trénovací sadě nejprve pomocí pětinásobné krosvalidace vybrán vhodný parametr pro agregaci. Pokud je zvolená metoda agregace Bair SPCA, je tímto parametrem theta, tedy prahová hodnota pro regresní koeficienty. Pokud je zvolená metoda Barshan SPCA, je hledaným parametrem dimenze dat. Tyto nalezené parametry poté figurují v agregaci jak trénovacích, tak testovacích dat. Agregace je realizována pomocí tříd v Pythonu, jejichž pseudokódy jsou uvedeny v příloze B.1. Výstupem jsou agregované matice dat Z pro trénovací a testovací sadu.



Obrázek 3.1: Schéma pro agregaci

3.2 Vybrané regresní modely

V definici remise u schizofrenie chybí shoda a jednotnost. Celkový koncept zahrnuje zlepšení a klinický výsledek se obvykle definuje buď pomocí procentuálního snížení celkového počtu symptomů nebo pomocí hraničního kritéria pro konkrétní symptomy. [17] Z tohoto důvodu byla v této práci zvolena regresní predikční úloha, protože nalezení vhodného prahu pro konkrétní proměnné je náročné.

3.2.1 Lineární regrese

Lineární regrese je jednoduchý model pro predikci. Lineární regrese jediné prediktorové proměnné X je popsána v následující části s názvem Jednoduchá lineární regrese. V další části je pak popsána metoda lineární regrese, která zahrnuje více prediktorů.

Jednoduchá lineární regrese

Jednoduchá lineární regrese je lineární přístup k predikci kvantitativní odezvy Y na základě jediné prediktorové proměnné X za předpokladu, že existuje přibližný následující lineární vztah mezi X a Y:

$$Y \approx \beta_0 + \beta_1 X,$$

kde β_0 a β_1 jsou parametry modelu představující intercept a sklon v lineárním modelu, díky nimž můžeme predikovat $\hat{y}$: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$,

Aby výsledná přímka byla co nejblíže datovým bodům, přístup nejmenších čtverců volí $\hat{\beta}_0$ a $\hat{\beta}_1$ tak, aby minimalizoval RSS (Residual Sum of Squares).

$$RSS = e_1^2 + e_2^2 + \ldots + e_n^2,$$

kde $e_i = y_i - \hat{y}_i$ představuje i-tý reziduální rozdíl.

Pak:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{3.10}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{3.11}$$

kde $\bar{x}$ a $\bar{y}$ jsou průměry hodnot X a Y.

Vícenásobná lineární regrese

Vícenásobná lineární regrese je rozšířením jednoduché lineární regrese tak, aby mohla zahrnovat více prediktorů. Pro p různých prediktorů má vícenásobný regresní model následující tvar:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

kde X_j představuje j-tý prediktor a β_j kvantifikuje spojení mezi touto proměnnou a odezvou.

Parametry jsou, stejně jako u jednoduché lineární regrese, odhadovány pomocí přístupu nejmenších čtverců. β_0, β_1, ..., β_p jsou tedy voleny tak, aby minimalizovaly RSS:

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip}\right)^2.$$

3.2.2 ElasticNet regrese

ElasticNet regrese je metoda kombinující výhody lasso (L_1 penalizace) a ridge (L_2 penalizace) regrese. Při pevně daných nezáporných hodnotách λ_1 a λ_2 se minimalizuje kritérium elasticNet regrese, které je váženou kombinací L_1 a L_2 penalizace. Pro $\alpha = 1$ je elasticNet shodná s ridge regresí, pro $\alpha = 0$ s lasso regresí. [41] Tato kombinace umožňuje učení řídkého modelu, kde má jen málo vah nenulovou hodnotu a zároveň zachovává regularizační vlastnosti ridge. L_1- a L_2-norma jsou regulovány pomocí parametru l_1_ratio (ρ).

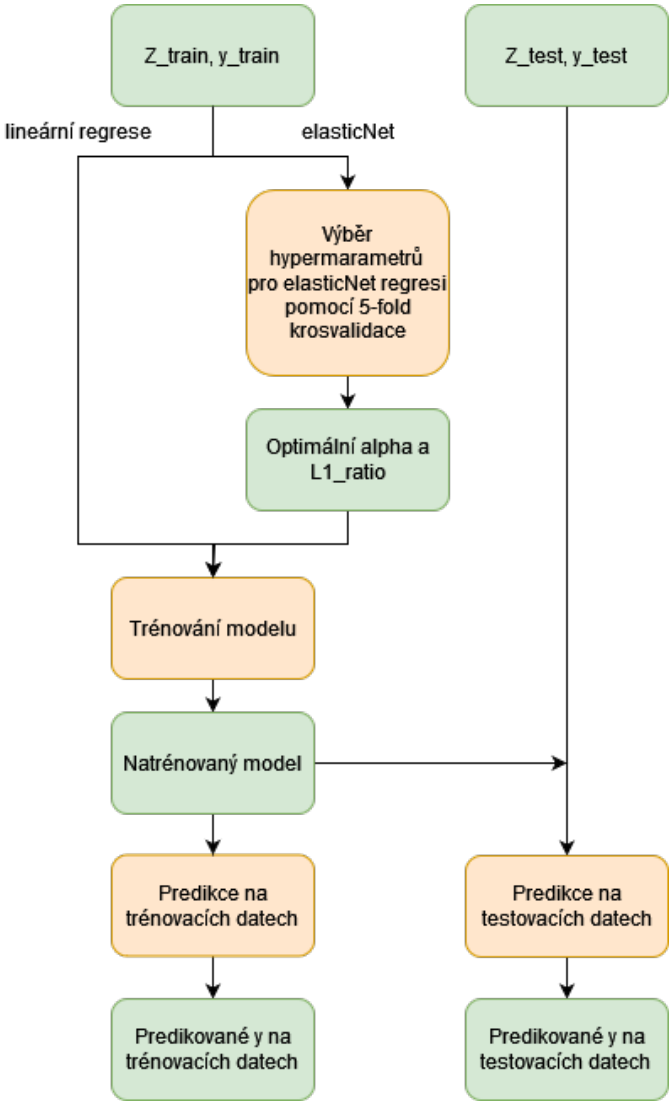
Účelová funkce vypadá následovně:

$$\min_{\omega} \frac{1}{2n_{samples}}\|X\omega - y\|_2^2 + \alpha\rho\|\omega\|_1 + \frac{\alpha(1-\rho)}{2}\|\omega\|_2^2 \qquad (3.12)$$

ElasticNet model je vhodný, pokud jsou některé proměnné mezi sebou korelované. [42]

3.2.3 Predikce

Na obrázku 3.2 je zobrazeno schéma predikce. Vstupními daty jsou Z_train a y_train pro trénování modelu a Z_test a y_test pro testování modelu. (Získání těchto dat je blíže popsáno v sekci 3.1.3 obrázkem 3.1.) Pokud je zvoleným modelem elasticNet regrese, jsou nejprve hledány optimální hyperparametry α a $L_1 ratio$ pomocí 5-fold krosvalidace na trénovacích datech. Tyto parametry pak vstupují do modelu, který je použit k predikci jak na trénovacích, tak na testovacích datech.



Obrázek 3.2: Schéma predikce

3.3 Evaluace modelu

3.3.1 Metriky pro evaluaci regresních modelů

RMSE a MAE jsou standardní metriky používané při evaluaci modelu. [43] Pro vzorek n pozorování y a odpovídající modelové predikce $\hat{y}$ je RMSE (Root Mean Squared Error) neboli odmocnina ze střední kvadratické chyby definována vzorcem 3.13 a MAE (Mean Absolute Error) neboli střední absolutní

chyba vzorcem 3.14.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3.13}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \tag{3.14}$$

R^2 (R-squared) neboli koeficient determinace, který udává podíl variace odpovědí vysvětlené dostupnými prediktory, je uveden v následujícím vzorci:

$$R^2 = 1 - \frac{SSE(X)}{SST}, \tag{3.15}$$

kde SSE 3.16 (Residual Sum of Squares) zohledňuje variabilitu odpovědí, která není vysvětlena dostupnými prediktory a SST 3.17 (Total Sum of Squares) zohledňuje celkovou variabilitu odpovědí.

$$SSE(X) = \sum_{i=1}^{n}(y_i - \hat{y}_i(X))^2 \tag{3.16}$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.17}$$

Tento koeficient R^2 je dobře definován pro lineární regresní modely a v praxi se používá jako míra kvality přizpůsobení modelů.

3.3.2 Zvolené schéma krosvalidace

Multifaktoriální a heterogenní povaha schizofrenie spolu s větším subjektivním a nepřímým měřením symptomů přispívá k pomalejší progresi predikční vědy v této oblasti ve srovnání s jinými klinicými specializacemi. I přes tyto

výzvy se ale tzv. precizní psychiatrie rychle vyvíjí a přináší naději na zlepšení diagnózy, prognózy i predikci léčebné odpovědi. [44] Prediktivní výsledky léčby u schizofrenie bývají silně závislé na kontextu a mohou proto mít omezenou obecnou platnost. Pro to, aby precizní medicína zlepšila výsledky v klinické praxi, musí být vyvíjené modely robustní: schopné předpovídat výsledky také pro pacienty, na kterých nebyly trénovány. To se v praxi většinou neděje, protože data pro takovéto studie jsou vzácná. Vědci proto obvykle rozdělí účastníky studie do určitého počtu náhodných skupin. Model je pak sestaven s využitím dat z jedné skupiny (trénovací skupina) a testován na druhé skupině. [45]

Přestože validace predikčních modelů na různých klinických vzorcích vede obvykle k nižšímu výkonům, je zásadním krokem pro vývoj těchto modelů a poskytuje věrnější hodnocení potenciálu statistických modelů zlepšit klinickou praxi.

Při výběru schématu krosvalidace byl v této diplomové práci následován Adam M. Chakroud a spoluautoři, kteří ve svém článku [45] používají opakovanou desetinásobnou cross-validaci, při níž datový soubor rozdělí do 10 náhodných skupin a 9 z nich použijí pro trénování, zbylou skupinu použijí pro testování. Tento proces je opakován desekrát tak, aby každá ze skupin byla jednou použita pro testování. Tento postup byl v této práci následován jak při výběru hyperparametrů u SPCA (theta pro Bair SPCA a d (dimenze) pro Barshan SPCA), tak při výběru parametrů pro elasticNet regresi (alfa a L_1 ratio), avšak s tím rozdílem, že byla z důvodu zkrácení doby výpočtu použita pouze pětinásobná krosvalidace.

Kapitola 4

Dataset

4.1 Základní popis

Dataset byl poskytnut Národním ústavem duševního zdraví (NÚDZ), kde byla data převážně naměřena. Část měření byla uskutečněna také v Institutu klinické a experimentální medicíny (IKEM). Sběr dat proběhl v rámci ESO studie (Early-Stage Schizophrenia Outcome). ESO studie je longitudinální studie pacientů s první epizodou schizofrenie, kteří věkem spadají do rozmezí 18 a 60 let, mají diagnózu schizofrenie, akutní polymorfní psychotické poruchy, akutní psychotické poruchy podobné schizofrenii nebo schizoafektivní poruchy. Zároveň pro zařazení pacienta do studie nesmí být psychóza přítomna více než 24 měsíců a v době hodnocení pacient užívá antipsychotické léky.

Celkově 543 pacientům byla naměřena první vizita (V1), která by podle plánu měla proběhnout vždy po první schizofrenií zapříčiněné hospitalizaci pacienta. Z těchto 543 pacientů 315 absolvovalo i druhou vizitu (V2), datovanou rok od vizity první. Pouze 140 pacientů absolvovalo i třetí vizitu (V3), která byla naplánována 4 roky od první vizity.

Počty pacientů v jednotlivých vizitách a jejich vybrané charakteristiky jsou uvedeny v tabulce 4.1. V této tabulce můžeme například vidět, že průměrný věk pacientů při první vizitě je 28 let, což je více, než jaký je obvyklý průměrný věk při objevení psychózy. [1] V tabulce 4.2 jsou pak uvedeny rozdíly naplánovaných časových odstupů vizit od skutečných časových odstupů naměřených vizit. Obě vizity (V2 A V3) byly naměřeny přibližně o půl roku

později oproti původnímu plánu. Na výkon predikce pak může mít vliv vysoká směrodatná odchylka blížící se jednomu roku.

Vizita	Počet pacientů	Mužů	Žen	Průměrný věk	Průměrný DUP [měsíc]
1.	543	331	212	28 ± 8	4.72 ± 9.50
2.	315	188	127	30 ± 8	4.39 ± 8.91
3.	140	83	57	34 ± 8	4.87 ± 11.73

Tabulka 4.1: Počet pacientů v jednotlivých vizitách a vybrané charakteristiky

Doba v letech	mezi 1. a 2. vizitou	mezi 1. a 3. vizitou
Naplánovaná	1	4
Průměrná realizovaná	$1,4 \pm 0,9$	$4,6 \pm 0,9$

Tabulka 4.2: Průměrná doba mezi vizitami

4.2 Zahrnuté parametry

Dataset obsahuje 2454 sloupců, v nichž můžeme najít:

- Základní informace o pacientovi (rok narození, pohlaví, počet sourozenců, datum vizity, pořadí vizity, místo měření,...),
- Demografické údaje (národnost, vzdělání, zaměstnání, rodinný stav a bydlení, informace o dětech, rodičích i sourozencích, a informace o užívání návykových látek (cigarety, alkohol, drogy)),
- Klinické údaje (pacientova diagnóza, první příznaky nemoci, zahájení lečby a její délka, komorbidity, vývoj pozitivních příznaků, informace ze specifických lékařských testů a diagnóz (ADHD, dyslexie apod.), fyzické parametry (hmotnost, výška, BMI), krevní tlak, informace o menstruačním cyklu),
- Informace o medikaci (o antipsychotické léčbě a somatické medikaci (glukokortikoidy, hormonální léky, inzulin, antidiabetika, statiny, antihypertenziva, antihistaminika)),
- Škály (hodnocení klinické závažnosti (CGI), dodržování léčby (Compliance), globální funkční hodnocení (GAF), hodnocení Mini-Mental stavu (MINI)),

- PANSS škála (pozitivní (P1-P7) a negativní (N1-N7) symptomy, specifické skupiny symptomů (G1-G16), jejich celkové hodnocení (suma) a dále pětifaktorový model PANSS (Positive, Negative, Excited, Disorganized, Depressed)), viz 2.1.2,
- WHOQoL škála (bydlení a životní prostředí, hmotné zabezpečení, zdravotní stav, obecné kvalita života (G1-G4), fyzicka kvalita života (F1-F24) a duševní kvalita života (D1-D4)),
- Osobní anamnézy pacientů,
- Laboratorní screening (základní krevní testy, testy jater a žlučových cest, hormonální testy, krevní obraz,
- Cytokiny z laboratoře v Plzni a z Fyziologického ústavu Akademie věd,
- Kynureniny (dopamin, serotonin, glutamát, tryptofan, ...),
- Zpracované hodnoty z magnetické rezonance a související výpočty (strukturální věk šedé/bílé hmoty, degenerace, a další). Dataset obsahuje T1 vážené snímky předzpracované pomocí optimalizovaného výpočetního anatomického toolboxu (Computational Anatomical Toolbox, CAT 12), blíže popsaného v článku Gasera a spoluautorů ([46]), normalizované a segmentované na šedou a bílou hmotu a na mozkomíšní mok; odhadnuté parametry kvality obrazu a výsledné hodnocení jsou v datasetu sloučeny do váženého průměru hodnocení kvality obrazu (Image Quality Rating, IQR); podle AAL atlasu ([47]) odhadnuté průměrné objemy tkání v *ml* jsou také součástí datasetu.

Dataset je neúplný. Například některé z výše jmenovaných informací se začaly zaznamenávat až s postupem času, a u mnoha pacientů tak tyto příznaky chybí. Tento problém se týká například některých skupin laboratorních dat.

4.3 Explorační analýza

Na základě výsledků z ESO studie [48] byly vybrány příznaky GAF a kvalita života v psychologické doméně (QoL D2) jako predikované proměnné pro funkční vyústění schizofrenie. Přidány byly dále také proměnné PANSS Factor Negative (F-Negative) a součet PANSS Negative (Negative) jako predikované proměnné pro klinické vyústění schizofrenie. Tabulka 4.3 zobrazuje průměrné hodnoty těchto vybraných proměnných v jednotlivých vizitách. Můžeme vidět, že hodnoty pro funkční vyústění se průměrně s vizitami zvyšují, zatímco symptomatika (negativní příznaky) se naopak snižuje.

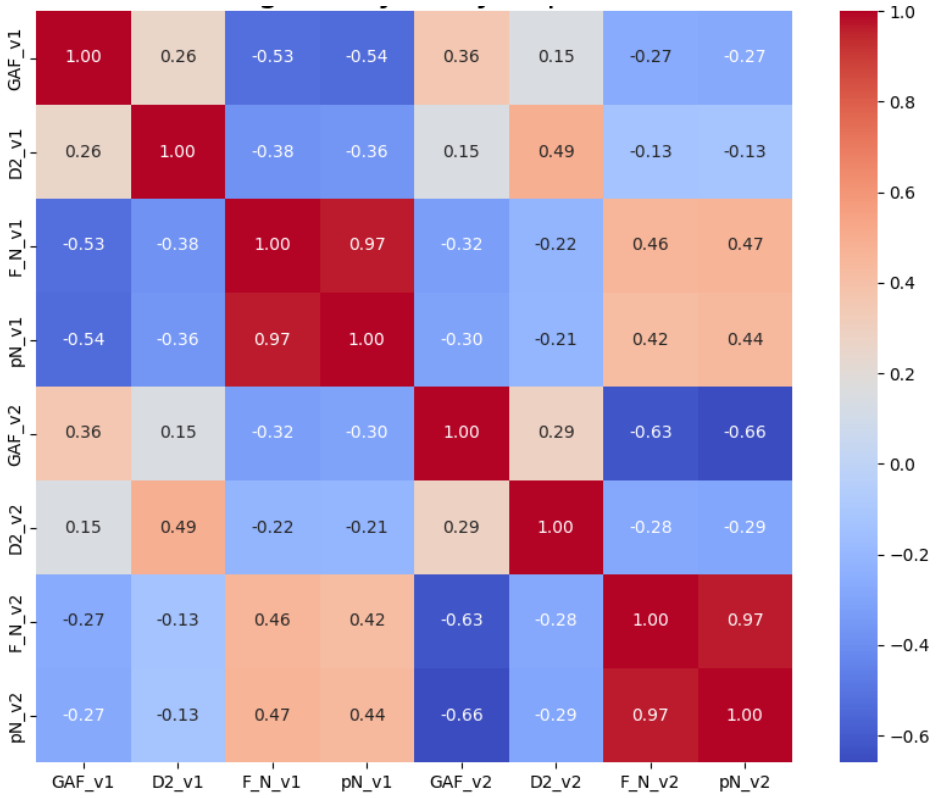
Velikost pearsonových korelačních koeficientů mezi těmito příznaky vykreslují obrázky 4.1 a 4.2. Obrázek 4.1 zobrazuje korelace mezi proměnnými v první a druhé vizitě, obrázek 4.2 zobrazuje korelace mezi proměnnými ve vizitě první a třetí. Heatmapy jsem vykreslila takto zvlášť, protože třetí vizitu absolvovalo méně pacientů. Proto ani korelace pro proměnné v první vizitě nejsou na obrázcích heatmap shodné.

Z obrázků můžeme vidět, že existují korelace mezi hodnotami proměnných naměřených při první vizitě a hodnotami těchto stejných proměnných naměřených jak při vizitě druhé, tak při vizitě třetí. Z nich nejvýznamnější jsou korelace mezi hodnotami psychologické domény, PANSS Factor Negative, a mezi hodnotami součtů PANSS Negative. Dále můžeme pozorovat korelaci mezi proměnnými PANSS Factor Negative a součet PANSS Negative, která je významná nejen pro hodnoty proměnných ve stejné vizitě (0,97 pro vizitu 1 i 2, 0,95 pro vizitu 3), ale také v různých vizitách (>0,4). V příloze jsou uvedeny vykreslené heatmapy pro větší množství vybraných příznaků, viz B.1, B.2.

Odhady hustoty pravděpodobnosti vybraných příznaků, vypočítané pomocí metody Kernel Density Estimation (KDE), jsou vykresleny na obrázku 4.3. Jednotlivé odhady jsou vypočítány pro konkrétní proměnné ze všech naměřených záznamů v dané vizitě. Pro první vizitu je tedy hustota pravděpodobnosti odhadnuta z největšího množství pacientů, na druhé straně pro třetí vizitu z nejmenšího, protože s vizitami se počet pacientů snižuje. Můžeme vidět, že rozdělení nejsou Gaussovská, a také že stav pacientů se v čase zlepšuje.

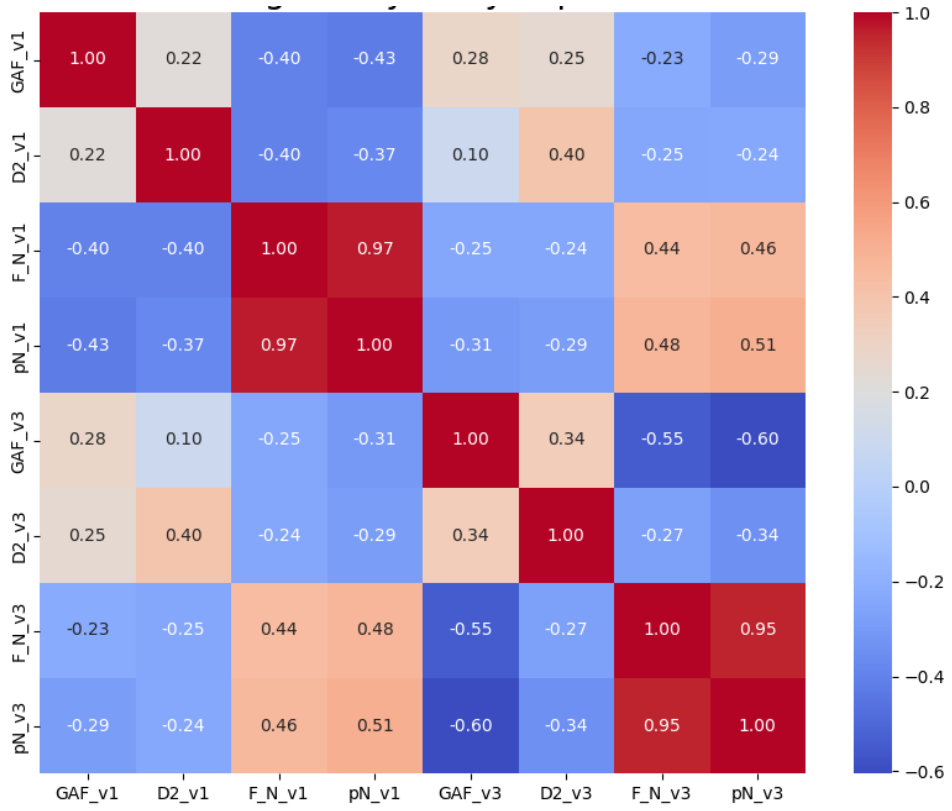
	GAF	QoL D2	F-Negative	Negative
V1	$67,0 \pm 16,0$	$13,7 \pm 2,8$	$2,3 \pm 0,9$	$15,6 \pm 5,9$
V2	$76,2 \pm 14,6$	$14,3 \pm 2,8$	$2 \pm 0,9$	$14,1 \pm 5,7$
V3	$77,7 \pm 14,4$	$14,5 \pm 3$	$1,8 \pm 0,8$	$12,4 \pm 5,2$

Tabulka 4.3: Průměrné hodnoty vybraných proměnných v jednotlivých vizitách



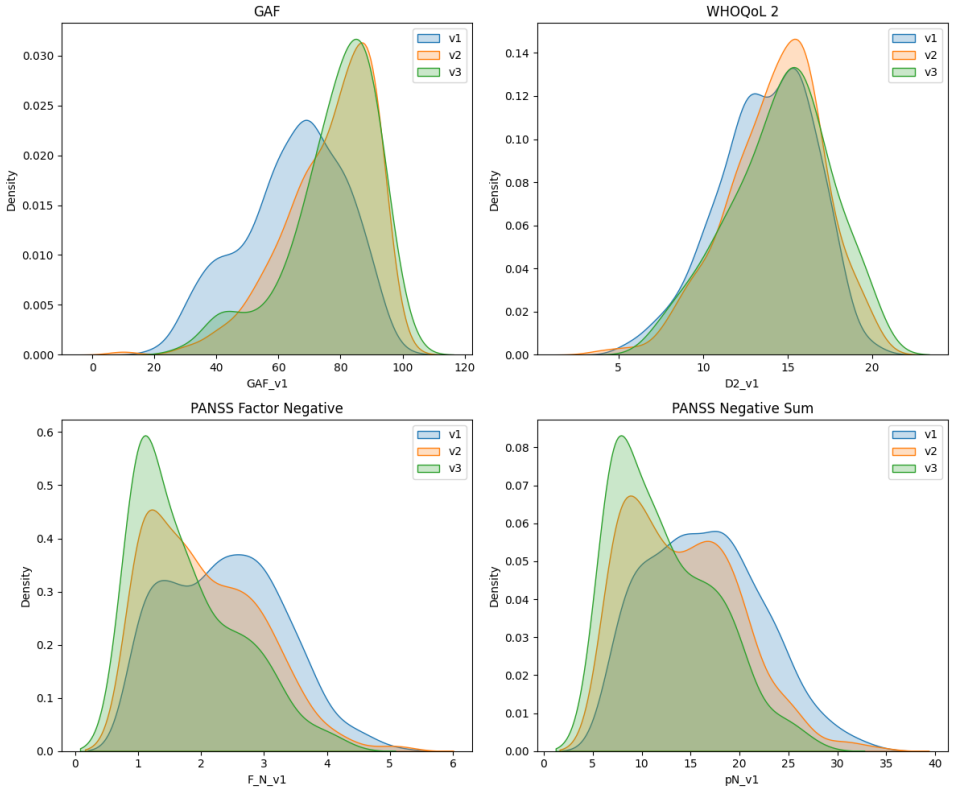
GAF, D2: QoL psychologická doména 2, F_N: PANSS Factor Negative, pN: součet PANSS Negative)

Obrázek 4.1: Heatmapa vybraných příznaků pro první (_v1) a druhou (_v2) vizitu



GAF, D2: QoL psychologická doména 2, F_N: PANSS Factor Negative, pN: součet PANSS Negative)

Obrázek 4.2: Heatmapa vybraných příznaků pro první (_v1) a třetí (_v3) vizitu



GAF, WHOQoL 2: QoL psychologická doména 2, PANSS Factor Negative, PANSS Negative Sum

Obrázek 4.3: Odhady hustoty pravděpodobnosti vybraných příznaků pro první (v1), druhou (v2) a třetí (v3) vizitu pro vybrané proměnné

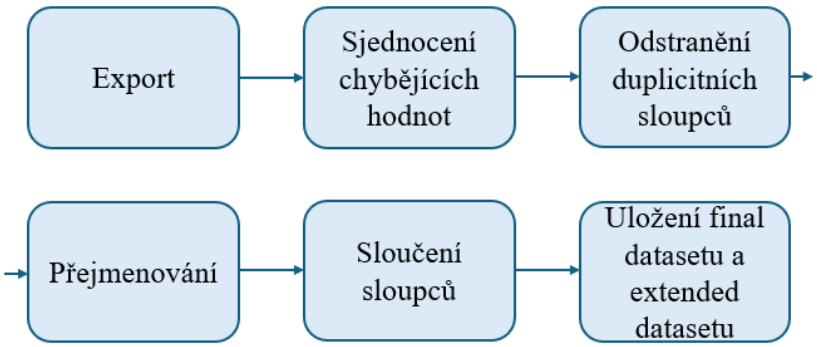
4.4 Předzpracování datasetu

Protože původní dataset nebylo kvůli jeho nekonzistentnosti možné přímo programově zpracovat, naimplementovala jsem nejdříve funkci pro předzpracování datasetu, jejímž výstupem je final dataset a extended dataset. Tato funkce byla napsána na základě požadavků specifikovaných v cols.csv - souboru, který byl vypracován odborníky, kteří spolupracují na ESO studii. Výsledný dataset je přehlednější, vhodný pro zpracování a funkce pro předzpracování ulehčuje případné změny ve výběru proměnných.

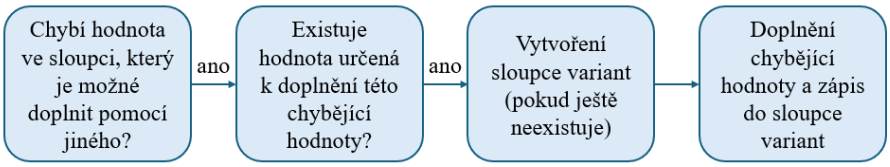
Cols.csv zahrnuje následující sloupce:

- colNameExport: jména sloupců v původním datasetu,
- colNameNew: nová jména sloupců,
- ColInFinal: zahrnutí sloupců do final datasetu (0/1),
- ColInExt: zahrnutí sloupců do extended datasetu (0/1),
- mergeWith: jméno sloupce, k jehož doplnění má být sloupec v colNameExport použit.

Pro práci s původním datasetem bylo potřeba nejprve rozpoznat a sjednotit chybějící hodnoty. Dále byly odstraněny duplicitní sloupce. Sloupce byly přejmenovány z českých názvů do anglických. Následně ve funkci dochází k sloučení sloupců, respektive k doplnění chybějících hodnot sloupců definovaných ve sloupci mergeWith pomocí vybraných sloupců na odpovídající řádce ve sloupci colNameExport. Po doplnění hodnot jsou vymazány sloupce, ze kterých byly chybějící hodnoty doplněny. Poté jsou podle ColInFinal a ColInExt vymazány sloupce, které do datasetů nemají být zahrnuty, a výsledné datasety jsou uloženy. Schéma zobrazující postup při slučování sloupců je zobrazen na obrázku 4.5 a schéma předzpracování datasetu na obrázku 4.4



Obrázek 4.4: Předzpracování datasetu



Obrázek 4.5: Slučování sloupců

Příklad procesu sloučení sloupců je zobrazen na obrázcích 4.4, 4.5 a 4.6. Na obrázku 4.4 můžeme vidět, že pro pacienty s osobním kódem 79 a 115 nebyly zaznamenány informace v1 o užívané medikaci antipsychotik a jejich dávkovém ekvivalentu (chlpz). Obrázek 4.5 je příkladem Cols.csv, tedy souboru, jenž obsahuje pravidla pro sloučení a přejmenování sloupců. Z tohoto obrázku je zřejmé, že zmíněné v1 informace je možné doplnit pomocí informací v2. Po předzpracování datasetu je poté ve výsledném uloženém datasetu, jehož příklad je uveden na obrázku 4.6, uvedena ve sloupci medikace i ve sloupci dávkového ekvivalentu doplněná hodnota v2 a informace o této variantě je zaznamenána ve sloupci Variant.

osobni_kod	medikace_v1	chlpz_v1	medikace_v2	chlpz_v2
79	x	x	0	0
115	N/A	N/A	risperidon 2.5mg	208

Tabulka 4.4: Původní dataset - příklad

colNameExport	colNameNew	ColInFinal	ColInExt	mergeWith
osobni_kod	id	1	1	
medikace_v1	Antypsychotics	1	1	
chlpz_v1	chlpz	1	1	
medikace_v2		0	0	medikace_v1
chlpz_v2		0	0	chlpz_v2

Tabulka 4.5: Cols csv - příklad

id	Antypsychotics	AntypsychoticsVariant	chlpz	chlpzVariant
79	0	medikace_v2	0	chlpz_v2
115	risperidon 2.5mg	medikace_v2	208	chlpz_v2

Tabulka 4.6: Final csv - příklad

V následující tabulce je uveden počet sloupců původního datasetu a počet sloupců ve final a extended datasetu uložených při předzpracování. Počet řádek se předzpracováním nezměnil.

Dataset	Počet sloupců
Původní	2453
Final	312
Extended	425

Tabulka 4.7: Počet sloupců v jednotlivých datasetech

4.5 Výběr příznaků

Z původního datasetu byly vytvořeny 3 sady příznaků popsané níže. Výsledné počty vybraných příznaků v jednotlivých sadách jsou shrnuty v tabulce 4.8. Výsledné počty pacientů pro jednotlivé sady uvedeny v tabulce 4.9.

Initial Set: První sada příznaků byla vybrána na základě připravovaného článku Kudelky a spoluautorů ([48]) a obsahuje 14 příznaků, které v uvedené studii při predikci proměnných ve 3. vizitě dosáhly významnějších predikčních výsledků.

Extended Set: Druhá sada byla vytvořena na základě výběru 170 příznaků do rozšířeného datasetu odborníky. Předzpracováním datasetu (viz 4.4) byly tyto příznaky rozšířeny o 24 sloupců variant. Následně byly některé příznaky převedeny z textových řetězců na číselné hodnoty. (Konkrétně šlo o sloupce: pohlaví, zařazení pacienta, centrum a všechny sloupce variant.) Z těchto příznaků byly následně vybrány pouze takové, po jejichž přidaní do seznamu příznaků se počet pacientů v trénovací sadě nesnížil na méně než 95% oproti počtu pacientů v trénovací sadě při použití seznamu příznaků bez přidaného příznaku. Výsledný extended set je tvořen 72 vybranými příznaky.

Comprehensive Set: Třetí sada zahrnuje nejvíce příznaků. Při její tvorbě bylo nejdříve vybráno 618 příznaků jak numerických, tak i nenumerických, avšak vhodných k převodu na numerické. Předzpracováním byly tyto příznaky rozšířeny o 150 sloupců variant. Příznaky v textovém tvaru byly převedeny na numerické (pohlaví, škály MINI, sloupce osobní anamnézy, sloupce variant, zařazení pacienta, centrum a klinická diagnóza). Z těchto příznaků byly následně vybrány pouze takové, po jejichž přidaní do seznamu příznaků se počet pacientů v trénovací sadě nesnížil na méně než 95% oproti počtu pacientů v trénovací sadě při použití seznamu příznaků bez tohoto přidaného příznaku. Výsledný comprehensive set je pak tvořen 385 příznaky.

Sada	Počet příznaků
Initial Set	14
Extended Set	72
Comprehensive Set	385

Tabulka 4.8: Počet příznaků v jednotlivých sadách

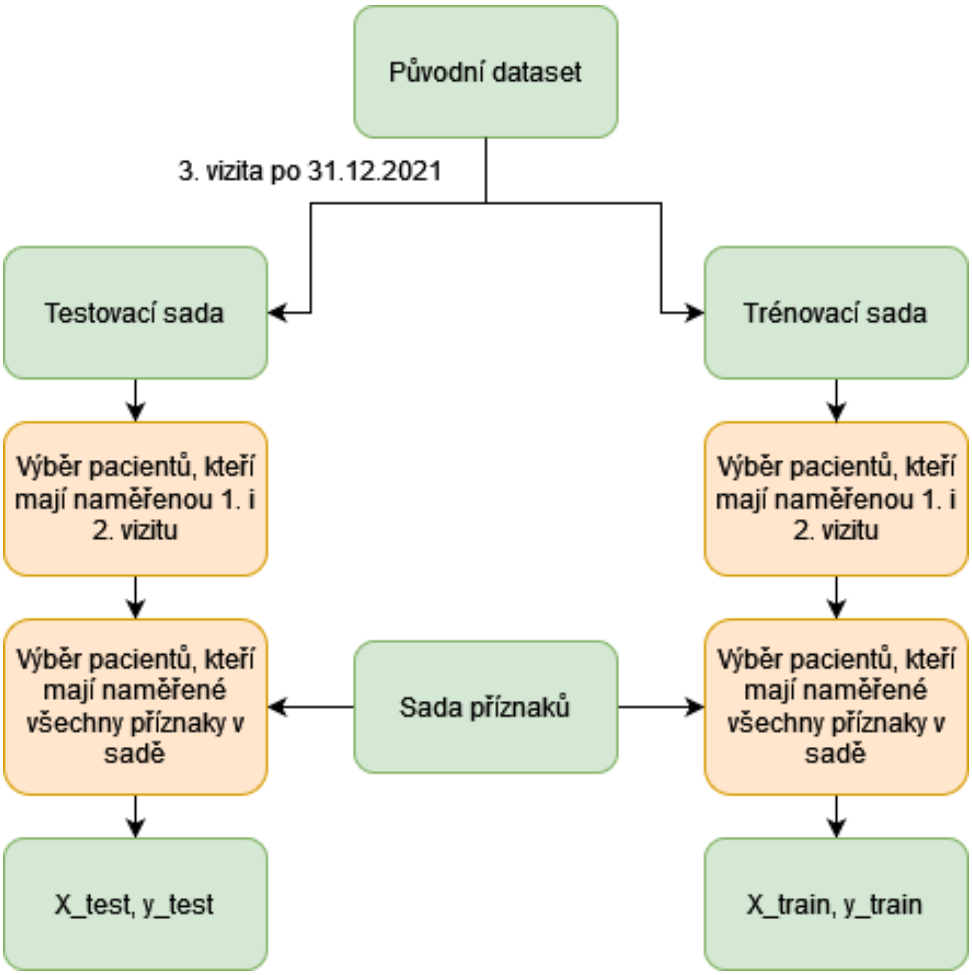
	Initial Set	Extended Set	Comprehensive Set
Trénovací sada	239	223	211
Testovací sada	41	39	37

Tabulka 4.9: Počet pacientů v jednotlivých sadách

4.6 Příprava dat

Před agregací dat a samotnou predikcí vybraných proměnných 2. vizity bylo nutné data připravit. Schéma přípravy dat je zobrazno na obrázku 4.6. Původní dataset byl nejprve rozdělen na testovací a trénovací, a to na základě podmínky týkající se data třetí vizity. Pokud třetí vizita pacienta proběhla po 31.12.2021, pak byl pacient zařazen do testovací sady. Trénovací sada tedy obsahuje pacienty, u kterých třetí vizita proběhla buď před tímto datem, případně pokud třetí vizitu vůbec nemají.

Dále byly z testovací i trénovací sady vybrání pouze ti pacienti, kteří mají naměřenou jednak první jednak druhou vizitu. Dále bylo zkontrolováno, že vybraní pacienti mají naměřené vybrané příznaky ze zvolené sady příznaků, pokud ne, byly vyřazeni. Posledním krokem při přípravě bylo vytvoření sady X, která pro jednotlivé pacienty obsahuje příznaky z vybrané sady naměřené ve vizitě 1, a y, která pro stejné pacienty obsahuje prediktivní proměnnou naměřenou ve vizitě 2.



Obrázek 4.6: Schéma přípravy dat

Kapitola 5

Výsledky

Nad třemi různými datasety byly regresními modely predikovány proměnné GAF, kvalita života v psychologické doméně, součet negativních symptomů na PANSS škále a negativní symptomy na pětifaktorové PANSS škále. Před samotnou predikcí byla data agregována pomocí Bair SPCA nebo Barshan SPCA, Dual Barshan SPCA, popřípadě byla použita data bez agregování.

Tato kapitola nejprve uvádí výsledky použitých modelů nad jednotlivými sety příznaků, dále následuje diskuze výsledků, uvedené limitace této práce a možnosti pro její rozšíření.

Initial Set

Initial sada příznaků obsahuje 14 prediktivních proměnných a 239 pacientů v trénovací sadě, 41 pacientů v sadě testovací. Tabulka 5.1 zobrazuje výsledky predikce pomocí lineární regrese při použití Initial Setu. V tabulce 5.2 jsou uvedeny výsledky predikce se stejnou sadou příznaků při použití elasticNet regresního modelu. V obou případech měla Bairova metoda SPCA na testovacích datech vždy horší výsledky než samotná lineární regrese.

U lineární regrese při predikci GAFu (Global Assessment of Functioning) dosáhl nejlepšího výsledku model s agregací dat pomocí metody Barshan ($R^2 = 0,21$). Stejného výsledku dosáhl také model s agregací dat pomocí metody Dual Barshan a model bez agregace dat pro predikci PANSS Factor

Negative (F-Negative). Při predikci sumy negativních příznaků (Negative) dosáhl nejlepších výsledků model bez agregace dat.

Celkově nejlepší výsledek z Initial Setu příznaků vykázala predikce psychologické domény kvality života (QoL D2) pomocí lineární regrese s metodou agragace dat Dual SPCA podle Barshan ($R^2 = 0,29$), což je také nejlepší výsledek i v kontextu ostatních sad příznaků.

U regresního modelu elasticNet bylo dosaženo stejných nejvyšších hodnot R^2 pro predikci GAFu, F-Negative i Negative, jako u lineární regrese. U QoL D2 bylo nejvyšší hodnoty dosaženo díky modelu bez agregace dat, s R^2 0,28.

Output	SPCA	Hyper	Set	RMSE	MAE	R2
GAF	Bair	1168	Train	14.55	11.41	0.04
			Test	19.58	14.48	-0.89
	Barshan	7	Train	13.55	10.60	0.16
			Test	12.66	10.71	0.21
	Dual	7	Train	13.39	10.51	0.18
			Test	12.79	10.70	0.19
	None	-	Train	13.13	10.22	0.22
			Test	13.07	10.47	0.16
QoL D2	Bair	216	Train	2.39	1.89	0.25
			Test	4.10	3.27	-0.71
	Barshan	9	Train	2.43	1.90	0.23
			Test	2.80	2.15	0.20
	Dual	9	Train	2.38	1.89	0.26
			Test	2.63	1.97	0.29
	None	-	Train	2.38	1.88	0.26
			Test	2.66	1.98	0.28
F-Negative	Bair	31	Train	0.78	0.63	0.21
			Test	1.05	0.91	-0.15
	Barshan	7	Train	0.79	0.65	0.18
			Test	0.89	0.73	0.18
	Dual	12	Train	0.77	0.62	0.23
			Test	0.87	0.73	0.21
	None	-	Train	0.77	0.62	0.23
			Test	0.87	0.73	0.21
Negative	Bair	213	Train	5.02	4.03	0.21
			Test	6.58	5.73	-0.33
	Barshan	11	Train	5.30	4.37	0.12
			Test	5.07	4.33	0.21
	Dual	6	Train	5.25	4.37	0.14
			Test	5.35	4.54	0.12
	None	-	Train	4.87	3.88	0.26
			Test	4.98	4.12	0.24

Žlutou barvou jsou zobrazeny nejlepší výsledky na testovací sadě pro jednotlivé proměnné, oranžová barva značí nejlepší výsledek pro model celkově

Tabulka 5.1: Predikce 2. vizity pomocí lineární regrese za použití Initial sady příznaků

Output	SPCA	Hyper	Alpha	L1 ratio	Set	RMSE	MAE	R2
GAF	Bair	1162	10.0	0.0	Train	13.61	10.71	0.16
					Test	15.35	11.72	-0.16
	Barshan	9	10.0	0.0	Train	13.64	10.73	0.15
					Test	12.61	10.73	0.21
	Dual	6	10.0	0.0	Train	13.53	10.63	0.17
					Test	12.61	10.66	0.21
	None	-	1.0	0.38	Train	13.32	10.42	0.19
					Test	12.66	10.55	0.21
QoL D2	Bair	202	1.0	0.14	Train	2.39	1.89	0.25
					Test	3.16	2.56	-0.02
	Barshan	9	0.1	0.99	Train	2.45	1.94	0.21
					Test	2.82	2.16	0.19
	Dual	7	1.0	0.12	Train	2.39	1.89	0.25
					Test	2.68	2.02	0.26
	None	-	1.0	0.19	Train	2.39	1.89	0.25
					Test	2.66	2.01	0.28
F-Negative	Bair	30	1e-05	0.0	Train	0.78	0.63	0.21
					Test	4.82	4.7	-23.3
	Barshan	12	0.1	0.06	Train	0.79	0.65	0.18
					Test	0.88	0.72	0.2
	Dual	12	1e-05	0.0	Train	0.77	0.62	0.23
					Test	0.87	0.73	0.21
	None	-	0.1	0.38	Train	0.77	0.63	0.22
					Test	0.87	0.73	0.21
Negative	Bair	217	1e-05	0.99	Train	5.03	4.03	0.21
					Test	8.29	7.07	-1.11
	Barshan	12	0.1	0.0	Train	5.27	4.36	0.14
					Test	5.01	4.29	0.23
	Dual	12	1e-05	0.0	Train	4.87	3.88	0.26
					Test	5.01	4.15	0.23
	None	-	0.1	0.99	Train	4.88	3.9	0.26
					Test	4.96	4.15	0.24

Žlutou barvou jsou zobrazeny nejlepší výsledky na testovací sadě pro jednotlivé proměnné, oranžová barva značí nejlepší výsledek pro model celkově

Tabulka 5.2: Predikce 2. vizity pomocí elasticNet regrese za použití Initial sady příznaků

Extended Set

Extended sada příznaků obsahuje 72 prediktivních proměnných a 223 pacientů v trénovací sadě, 39 pacientů v sadě testovací. Tabulka 5.3 zobrazuje výsledky lineární regrese při použití Extended Setu. V tabulce 5.4 jsou uvedeny výsledky predikce pro stejnou sadu příznaků při použití elasticNet regresního modelu. I pro tuto sadu příznaků Bairova metoda SPCA, stejně jako u sady Initial, na testovacích datech měla výsledky horší než průměrování.

U lineární regrese při predikci GAFu dosáhl nejlepšího výsledku model bez

agregace dat s $R^2 = 0,24$. Při predikci F-Negative dosáhly stejného R^2 (0,23) model s agregací dat pomocí metody Dual Barshan a model bez agregace.

Celkově nejlepší výsledek z Extended Setu příznaků vykázala predikce QoL D2 při použití regresního modelu elasticNet bez agregace dat, s $R^2 = 0,25$. Predikce proměnné Negative s agregací dat metodou Dual Barshan dosáhla podobného výsledku ($R^2 = 0,24$).

Output	SPCA	Hyper	Set	RMSE	MAE	R2
GAF	Bair	1142	Train	13.45	10.83	0.07
			Test	439.2	438.96	-935.4
	Barshan	18	Train	12.13	9.67	0.24
			Test	12.95	10.64	0.19
	Dual	18	Train	12.08	9.63	0.25
			Test	13.06	10.81	0.17
	None	-	Train	10.57	8.23	0.42
			Test	12.54	10.45	0.24
QoL D2	Bair	208	Train	2.4	1.91	0.24
			Test	17.71	17.08	-29.73
	Barshan	18	Train	2.72	2.16	0.03
			Test	3.15	2.51	0.03
	Dual	24	Train	2.31	1.81	0.3
			Test	2.82	2.1	0.22
	None	-	Train	2.09	1.64	0.42
			Test	2.86	2.24	0.2
F-Negative	Bair	29	Train	0.76	0.6	0.25
			Test	81.78	81.7	-7337.14
	Barshan	28	Train	0.77	0.62	0.24
			Test	0.88	0.73	0.14
	Dual	49	Train	0.69	0.54	0.38
			Test	0.84	0.65	0.23
	None	-	Train	0.67	0.53	0.41
			Test	0.84	0.67	0.23
Negative	Bair	206	Train	4.94	3.94	0.24
			Test	991.81	990.75	-31037.75
	Barshan	22	Train	5.09	4.16	0.2
			Test	5.04	4.17	0.2
	Dual	15	Train	4.96	3.98	0.24
			Test	5.07	4.14	0.19
	None	-	Train	4.32	3.34	0.42
			Test	5.09	3.93	0.18

Žlutou barvou jsou zobrazeny nejlepší výsledky na testovací sadě pro jednotlivé proměnné, oranžová barva značí nejlepší výsledek pro model celkově

Tabulka 5.3: Predikce 2. vizity pomocí lineární regrese za použití Extended sady příznaků

Output	SPCA	Hyper	Alpha	L1 ratio	Set	RMSE	MAE	R2
GAF	Bair	432	10.0	0.36	Train	12.55	10.08	0.19
					Test	16.44	14.4	-0.31
	Barshan	36	10.0	0.0	Train	12.37	9.95	0.21
					Test	13.04	10.9	0.18
	Dual	71	1.0	0.99	Train	11.64	9.37	0.3
					Test	12.7	10.5	0.22
	None	-	10.0	0.0	Train	12.17	9.74	0.24
					Test	12.8	10.69	0.2
QoL D2	Bair	36	1.0	0.37	Train	2.4	1.9	0.25
					Test	9.6	6.87	-8.03
	Barshan	17	100.0	0.99	Train	2.76	2.21	0.0
					Test	3.19	2.53	-0.0
	Dual	14	0.1	0.99	Train	2.4	1.9	0.25
					Test	2.86	2.19	0.2
	None	-	1.0	0.57	Train	2.4	1.9	0.25
					Test	2.77	2.12	0.25
F-Negative	Bair	7	1.0	0.0	Train	0.72	0.58	0.32
					Test	3.45	3.2	-12.05
	Barshan	30	1.0	0.0	Train	0.79	0.64	0.19
					Test	0.92	0.77	0.07
	Dual	29	0.1	0.44	Train	0.73	0.58	0.31
					Test	0.85	0.69	0.2
	None	-	0.1	0.52	Train	0.74	0.6	0.29
					Test	0.85	0.7	0.21
Negative	Bair	204	0.01	0.87	Train	4.94	3.93	0.24
					Test	895.73	895.0	-25315.59
	Barshan	54	10.0	0.16	Train	5.4	4.44	0.1
					Test	5.73	4.68	-0.04
	Dual	39	1.0	0.49	Train	4.8	3.89	0.29
					Test	4.91	4.07	0.24
	None	-	1.0	0.37	Train	4.88	3.93	0.26
					Test	4.98	4.1	0.22

Žlutou barvou jsou zobrazeny nejlepší výsledky na testovací sadě pro jednotlivé proměnné, oranžová barva značí nejlepší výsledek pro model celkově

Tabulka 5.4: Predikce 2. vizity pomocí elasticNet regrese za použití Extended sady příznaků

Comprehensive Set

Comprehensive sada příznaků obsahuje 385 prediktivních proměnných a 211 pacientů v trénovací sadě, 37 pacientů v sadě testovací. Tabulka 5.5 zobrazuje výsledky lineární regrese při použití Comprehensive Setu, který obsahuje 385 příznaků. V tabulce 5.6 jsou uvedeny výsledky predikce pro stejnou sadu příznaků při použití elasticNet regresního modelu. V tabulkách již není uvedena Bairova metoda SPCA, která ani pro menší množství vybraných příznaků nedokázala dosáhnout lepších výsledků, než průměrování.

Tabulka 5.5 nejlépe demonstruje přínos použití metod SPCA a Dual SPCA podle Barshan. U lineární regrese bez agregace dat dojde při použití většího množství příznaků k přeučení. Proto při predikci všech predikovaných pro-

měnných dosahuje tento model na trénovací sadě $R2 = 1$ a na testovací sadě selhává. Oproti tomu při použití agregace dat je pro predikovanou psychologickou doménu dosažené R^2 rovno 0,23 a pro predikovaný GAF je hodnota R^2 dokonce nejvyšší hodnotou R^2 pro predikci GAFu (společně s predikcí GAFu při použití modelu lineární regrese bez agregace z příznaků Extended Set).

ElasticNet není tak jako lineární regrese náchylná na přeučení a dosahuje při použití 385 prediktorů výsledků shodných nebo lepších v porovnání s modelem lineární regrese, a to i bez agregace dat. Avšak i zde poskytuje agragace dat pro některé predikované proměnné (GAF, F-Negative a Negative) lepší výsledek, než elasticNet regrese bez agregace dat.

Output	SPCA	Hyper	Set	RMSE	MAE	R2
GAF	Barshan	23	Train	11.82	9.3	0.27
			Test	12.64	10.11	0.24
	Dual	14	Train	12.0	9.46	0.25
			Test	13.25	10.87	0.17
	None	-	Train	0.0	0.0	1.0
			Test	36.52	27.01	-5.33
QoL D2	Barshan	9	Train	2.37	1.91	0.23
			Test	2.93	2.27	0.2
	Dual	11	Train	2.32	1.83	0.26
			Test	2.88	2.2	0.23
	None	-	Train	0.0	0.0	1.0
			Test	7.38	5.96	-4.09
F-Negative	Barshan	8	Train	0.78	0.65	0.15
			Test	0.95	0.79	0.05
	Dual	16	Train	0.75	0.6	0.22
			Test	0.89	0.73	0.17
	None	-	Train	0.0	0.0	1.0
			Test	2.65	2.09	-6.4
Negative	Barshan	7	Train	4.95	4.04	0.18
			Test	5.34	4.38	0.12
	Dual	5	Train	5.04	4.18	0.15
			Test	5.39	4.46	0.11
	None	-	Train	0.0	0.0	1.0
			Test	17.88	14.2	-8.86

Žlutou barvou jsou zobrazeny nejlepší výsledky na testovací sadě pro jednotlivé proměnné, oranžová barva značí nejlepší výsledek pro model celkově

Tabulka 5.5: Predikce 2. vizity pomocí lineární regrese za použití Comprehensive sady příznaků

Output	SPCA	Hyper	Alpha	L1 ratio	Set	RMSE	MAE	R2
GAF	Barshan	58	1.0	0.0	Train	11.47	9.12	0.32
					Test	12.63	10.47	0.24
	Dual	26	10.0	0.18	Train	12.11	9.61	0.24
					Test	12.9	10.77	0.21
	None	-	10.0	0.0	Train	11.9	9.45	0.26
					Test	12.85	10.79	0.22
QoL D2	Barshan	85	1.0	0.21	Train	2.38	1.91	0.23
					Test	2.89	2.27	0.22
	Dual	27	1.0	0.41	Train	2.32	1.84	0.26
					Test	2.88	2.23	0.22
	None	-	1.0	0.82	Train	2.37	1.89	0.23
					Test	2.79	2.19	0.28
F-Negative	Barshan	35	0.1	0.43	Train	0.72	0.59	0.27
					Test	0.87	0.71	0.21
	Dual	37	0.1	0.62	Train	0.68	0.54	0.35
					Test	0.83	0.68	0.28
	None	-	0.1	0.8	Train	0.73	0.59	0.25
					Test	0.87	0.71	0.21
Negative	Barshan	73	1.0	0.49	Train	4.89	4.0	0.21
					Test	5.3	4.36	0.13
	Dual	51	1.0	0.33	Train	4.45	3.63	0.34
					Test	4.78	3.84	0.29
	None	-	1.0	0.51	Train	4.78	3.88	0.24
					Test	5.05	4.13	0.21

Žlutou barvou jsou zobrazeny nejlepší výsledky na testovací sadě pro jednotlivé proměnné, oranžová barva značí nejlepší výsledek pro model celkově

Tabulka 5.6: Predikce 2. vizity pomocí elasticNet regrese za použití Comprehensive sady příznaků

5.1 Diskuze

Funkční a klinické vyústění bylo predikováno regresními modely (lineární regresí a elasticNet regresí) na sadách příznaků Initial, Extended a Comprehensive, které se liší zejména v počtu příznaků. Před samotnou predikcí byla data agregována pomocí metod Bair SPCA, Barshan SPCA a Dual Barshan SPCA. Predikce byla evaluována jak na trénovacích datech, které byly použity k trénování zvoleného modelu, tak na testovacích datech, se kterými se model při trénování nesetkal.

Nejvyšší přesnosti predikce bylo dosaženo při predikci psychologické domény kvality života s modelem lineární regrese při použití sady příznaků Initial a při predikci součtu negativních příznaků s modelem regrese elasticNet při použití sady příznaků Comprehensive. Těmto predikcím předcházela agregace dat pomocí Dual Barshan SPCA metody a při predikci je modelem vysvětleno

29% variability dat.

Větší množství příznaků by mohlo do modelu přinést další informace. Avšak zpracování většího množství příznaků přináší úskalí a jednoduché modely, jako je lineární regrese, selhávají, protože u nich dochází k přeučení. K tomuto jevu došlo v případě použití Comprehensive sady, která obsahuje 385 příznaků, ale pouze 211 pacientů. Lineární regrese na trénovací sadě má pro tuto sadu hodnoty R^2 rovné 1, zatímco na testovací sadě její výsledky nejsou lepší než průměrování. Tento problém řeší různé metody agregace nebo sofistikovanější modely. Použití metod Barshan SPCA i Dual Barshan SPCA zabraňuje přeučení, stejně tak použití regresního modelu elasticNet. Větší množství použitých příznaků avšak obsahuje také více šumových a korelovaných komponent, což ztěžuje a zpomaluje učení modelu. Tento problém řeší buď výběr příznaků anebo metody SPCA.

Pro lineární regresi s předchozí agregací dat Barshan SPCA a Dual Barshan SPCA na sadě Comprehensive v průměru nedosáhl model lepších výsledků oproti čisté lineární regresi na sadě Initial. V případě použití lineární regrese tedy můžeme říci, že je lepší použít Initial sadu proměnných oproti použití velkého počtu prediktorů v sadě Comprehensive s předchozí agregací.

Pro elasticNet regresi s agregací dat metodou Barshan na sadě Comprehensive model stejně jako u lineární regrese nedosáhl v průměru lepších výsledků než pouhá elasticNet regrese na Initial sadě. Oproti tomu použití elasticNet regrese s agregací dat pomocí metody DualBarshan měla v průměru lepší výsledky než použití modelu elasticNet na sadě příznaků Initial. Regrese elasticNet bez agregace na sadě Comprehensive v průměru nepřinesla lepší výsledky než na sadě Initial. Z toho je možné vyvodit, že použití elasticNet regrese současně s agregací dat pomocí metody Dual Barshan SPCA v tomto případě přináší lepší výsledky oproti ostatním vyzkoušeným kombinacím.

Porozumění a výběr vhodných prediktorů zůstává přesto zásadní v úloze predikce, protože díky výběru kandidátních prediktorů na základě existujících výzkumných důkazů a klinických znalostí je minimalizováno riziko zkreslení vedoucí k přeučení nebo k nalezení vztahů, které nejsou obecně platné. [44]

Funkční vyústění schizofrenie bylo predikováno pomocí psychologické domény kvality života a globálního hodnocení funkčnosti, u kterých bylo nejvyšší dosažené R^2 0,29, resp. 0,24. Klinické vyústění schizofrenie bylo predikováno pomocí součtu negativních příznaků na stupnici PANSS, kde nejvyšší R^2 dosáhlo hodnoty 0,29, a pomocí proměnné Factor Negative z pětifaktorové PANSS stupnice, u které je nejvyšší dosažená hodnota R^2 0,28. Na základě dosažených výsledků je možné říci, že do jisté míry lze tyto parametry, repre-

zentující funkční a klinické vyústění schizofrenie, predikovat.

Z vybraných a použitých metod SPCA je užitečná metoda Barshan SPCA a především metoda Dual Barshan SPCA. Oproti tomu metoda Bair SPCA nepřinesla žádný přínos. Obecně se dá říci, že metody SPCA napomáhají při použití velkých multimodálních datasetů pro predikci, nicméně samy o sobě si neumí dostatečně dobře poradit s přidáváním vyššího počtu korelovaných či šumových příznaků. Proto je zapotřebí kombinovat jejich použití s dostatečně robustními modely.

Nejlepší model této práce vysvětluje pouze 29% variability dat a jeho schopnost predikovat funkční i klinické vyústění schizofrenie je omezená. Potenciání klinické využití predikce může být v systému počítačem asistovaného rozhodování, budovaném v Národním ústavu duševního zdraví.

Tato predikční úloha byla zpracována jako regresní úloha, jejímž cílem bylo predikovat konkrétní hodnoty cílové proměnné. Nicméně by bylo možné k úloze přistoupit jako k úloze klasifikační, jejímž cílem by byla predikce příslušnosti k určité kategorii, vyhodnocena například pomocí váhované přesnosti (balanced accuracy). V případě tohoto přístupu se kategorie stanovují například pro remisi schizofrenie a bylo by tedy nutné vhodně zvolit práh pro odlišení kategorií.

Limitace

Protože některé příznaky jsou zaznamenány pouze u některých pacientů, větší sady příznaků znamenají méně pacientů v trénovacích i testovacích sadách. Ačkoliv příznaky byly do sady Extended i Comprehensive přidávány pouze za podmínky, že po přidání příznaku se počet pacientů v sadě nesníží na méně než 95%, sada Extended, resp. Comprehensive obsahuje o 7%, resp. 12% pacientů méně než sada Initial. Tento rozdíl může mít vliv na výsledky modelů.

Použitá data do jisté míry porušují podmínky pro použití lineární i elasticNet regrese (např. není známé, zda vztah mezi prediktory a závislou proměnnou je lineární). Dále u sad příznaků Extended a Comprehensive není splněna podmínka pro doporučené minimální množství pozorovaní (pacientů) pro regresní model při použití daného počtu prediktorů.

Možná rozšíření

Pro predikci na tomto datasetu by mohly být užitečné nelineární metody, případně voting či agregace více modelů nad jednotlivými modalitami, v budoucnu by proto bylo vhodné tyto metody predikce vyzkoušet. Další možností je doplnění chybějících hodnot datasetu za pomoci imputačních metod.

Literatura uvádí pravděpodobnou existenci podtypů schizofrenie. Toho by bylo možné využít a použít diferencované modely na různé subpopulace. K tomu by ale pravděpodobně byl zapotřebí větší dataset.

Nadále zůstává potřeba identifikace faktorů a signifikantních proměnných důležitých pro úspěšnou predikci vyústění schizofrenie.

Kapitola 6

Závěr

Hlavními cíli této práce bylo udělat rešerši literatury týkající se problematiky hodnocení funkčního vyústění u schizofrenie a problematiky parametrů sledovaných ve studii ESO, dále provedení přípravy datasetu a jeho explorační analýzy, navržení vhodných metod pro výběr příznaků a pro agregaci dat a predikce funkčního vyústění schizofrenie, která využívá těchto navržených metod agregace.

Po provedení rešerše byl připraven dataset, pro nějž byla následně udělána explorační analýza. Výběrem příznaků vznikly 3 rozdílně velké sady, z nichž jedna byla vytvořena na základě výsledků připravované studie, druhá zahrnovala příznaky vybrané odborníky a třetí sada byla navržena tak, aby obsahovala co největší množství příznaků z různých zdrojů. Data byla agregována metodami supervidovaných hlavních komponent podle metodiky publikované Bairem a spoluautory a podle metodiky publikované Barshan a spoluautory. Z přístupů Barshan byly konkrétně vybrány SPCA a Dual SPCA. S použitím modelů lineární regrese a elasticNet regrese byly z vybraných příznaků v první vizitě predikovány vybrané závislé proměnné v druhé vizitě. Pro nastavení hyperparametrů modelu byla použita 5-fold krosvalidace.

Bairova metoda SPCA se na testovacích dat neukázala být užitečnou. Metody Barshan SPCA a Dual Barshan SPCA oproti tomu mohou být přínosné, avšak jejich použití musí být doplněno dostatečně robustním modelem.

Nejvyšší přesnosti predikce bylo dosaženo při predikci psychologické domény kvality života s modelem lineární regrese při použití sady příznaků Initial a při predikci součtu negativních příznaků s modelem regrese elasticNet při použití

sady příznaků Comprehensive. Oběma těmto predikcím předcházela agregace dat pomocí Dual Barshan SPCA metody. Natrénované modely vysvětlily na testovací sadě 29% variability dat.

Potenciání klinické využití predikce může být v systému počítačem asistovaného rozhodování, budovaném v Národním ústavu duševního zdraví.

Příloha A

Literatura

[1] Thomas R. Insel. Rethinking schizophrenia. *Nature*, 468:187–193, 11 2010.

[2] World Health Organization. ICD WHO. `https://icd.who.int/browse11/l-m/en#/http%3a%2f%2fid.who.int%2ficd%2fentity%2f1683919430` Viděno 2024-01-30.

[3] William S. Kremen, Larry J. Seidman, Stephen V. Faraone, Rosemary Toomey, and Ming T. Tsuang. Heterogeneity of schizophrenia: A study of individual neuropsychological profiles. *Schizophrenia Research*, 71:307–321, 12 2004.

[4] World Health Organization. Fact sheet: Schizophrenia. `https://www.who.int/news-room/fact-sheets/detail/schizophrenia` Viděno 2024-01-30.

[5] Naomi R. Wray and Irving I. Gottesman. Using summary data from the danish national registers to estimate heritabilities for schizophrenia, bipolar disorder, and major depressive disorder. *Frontiers in Genetics*, 3, 2012.

[6] John H Gilmore. Understanding what causes schizophrenia: a developmental perspective. *American Journal of Psychiatry*, 167(1):8–10, 2010.

[7] Marco Solmi, Georgios Seitidis, Dimitris Mavridis, Christoph U. Correll, Elena Dragioti, Synthia Guimond, Lauri Tuominen, Aroldo Dargél, Andre F. Carvalho, Michele Fornaro, Michael Maes, Francesco Monaco, Minjin Song, Jae Il Shin, and Samuele Cortese. Incidence, prevalence,

and global burden of schizophrenia - data, with critical appraisal, from the global burden of disease (gbd) 2019. *Molecular Psychiatry*, 2023.

[8] Jason C. Simeone, Alexandra J. Ward, Philip Rotella, Jenna Collins, and Ricarda Windisch. An evaluation of variation in published estimates of schizophrenia prevalence from 1990-2013: A systematic literature review. *BMC Psychiatry*, 15, 8 2015.

[9] Fiona J. Charlson, Alize J. Ferrari, Damian F. Santomauro, Sandra Diminic, Emily Stockings, James G. Scott, John J. McGrath, and Harvey A. Whiteford. Global epidemiology and burden of schizophrenia: Findings from the global burden of disease study 2016. *Schizophrenia Bulletin*, 44:1195–1203, 10 2018.

[10] Louis A Sass and Josef Parnas. Schizophrenia, consciousness, and the self. *Schizophrenia bulletin*, 29(3):427–444, 2003.

[11] Andreas Rosén Rasmussen, Maja Zandersen, Julie Nordgaard, Karl Erik Sandsten, and Josef Parnas. Pseudoneurotic symptoms in the schizophrenia spectrum: An empirical study. *Schizophrenia Research*, 250:164–171, 12 2022.

[12] R. S. Wallwork, R. Fortgang, R. Hashimoto, D. R. Weinberger, and D. Dickinson. Searching for a consensus five-factor model of the positive and negative syndrome scale for schizophrenia. *Schizophrenia Research*, 137:246–250, 5 2012.

[13] American Psychiatric Association. What is schizophrenia. `https://www.psychiatry.org/patients-families/schizophrenia/what-is-schizophrenia` Viděno 2024-01-30.

[14] John Lally and James H. MacCabe. Antipsychotic medication in schizophrenia: a review. *British Medical Bulletin*, 114:169–179, 6 2015.

[15] Robert A McCutcheon, Tiago Reis Marques, and Oliver D Howes. Schizophrenia—an overview. *JAMA psychiatry*, 77(2):201–210, 2020.

[16] T Scott Stroup and Neil Gray. Management of common adverse effects of antipsychotic medications. *World Psychiatry*, 17(3):341–356, 2018.

[17] Martin Lepage, Michael D. Bodnar, and Christopher R. Bowie. Neurocognition: Clinical and functional outcomes in schizophrenia. *The Canadian Journal of Psychiatry*, 59:12 – 5, 2014.

[18] Luigi Giuliani, Giulia Maria Giordano, Paola Bucci, Pasquale Pezzella, Francesco Brando, and Silvana Galderisi. Improving knowledge on pathways to functional outcome in schizophrenia: main results from the italian network for research on psychoses. *Frontiers in Psychiatry*, 12:791117, 2021.

[19] Geir Pedersen and Sigmund Karterud. The symptom and function dimensions of the global assessment of functioning (gaf) scale. *Comprehensive Psychiatry*, 53(3):292–298, 2012.

[20] Myrto T Samara, Rolf R Engel, Aurelie Millier, Julia Kandenwein, Mondher Toumi, and Stefan Leucht. Equipercentile linking of scales measuring functioning and symptoms: examining the gaf, sofas, cgi-s, and panss. *European Neuropsychopharmacology*, 24(11):1767–1772, 2014.

[21] Ricardo E Carrión, Andrea M Auther, Danielle McLaughlin, Ruth Olsen, Jean Addington, Carrie E Bearden, Kristin S Cadenhead, Tyrone D Cannon, Daniel H Mathalon, Thomas H McGlashan, et al. The global functioning: social and role scales—further validation in a large sample of adolescents and young adults at clinical high risk for psychosis. *Schizophrenia bulletin*, 45(4):763–772, 2019.

[22] Sherryl H Goodman, Daniel R Sewell, Eileen L Cooley, and Naomi Leavitt. Assessing levels of adaptive functioning: the role functioning scale. *Community mental health journal*, 29(2):119–131, 1993.

[23] Liliane Lins and Fernando Martins Carvalho. Sf-36 total score as a single measure of health-related quality of life: Scoping review. *SAGE open medicine*, 4:2050312116671725, 2016.

[24] H Eleanor Cannon-Spoor, Steven G Potkin, and Richard Jed Wyatt. Premorbid adjustment scale. *Schizophrenia Bulletin*, 1982.

[25] M. Bosc, A. Dubini, and V. Polin. Development and validation of a social functioning scale, the social adaptation self-evaluation scale. *European Neuropsychopharmacology*, 7(1, Supplement 1):S57–S70, 1997.

[26] NELSON. Series - Psychoeducational Assessments SIBR - Nelson. `https://school.nelson.com/psych-ed-assessment/SIBR`.

[27] Richard CW Hall. Global assessment of functioning: a modified scale. *Psychosomatics*, 36(3):267–275, 1995.

[28] Hector R Bird, Glorisa Canino, Maritza Rubio-Stipec, and Julio C Ribera. Further measures of the psychometric properties of the children's global assessment scale. *Archives of General Psychiatry*, 44(9):821–824, 1987.

[29] Giulia M Giordano, Pasquale Pezzella, Armida Mucci, Stephen F Austin, Andreas Erfurth, Birte Glenthøj, Alex Hofer, Jan Hubenak, Jan Libiger, Ingrid Melle, et al. Negative symptoms and social cognition as mediators of the relationship between neurocognition and functional outcome in schizophrenia. *Frontiers in Psychiatry*, 15:1333711, 2024.

[30] T Burns and D Patrick. Social functioning as an outcome measure in schizophrenia studies. *Acta Psychiatrica Scandinavica*, 116(6):403–418, 2007.

[31] Sebastian Walther, Sarah Eisenhardt, Stephan Bohlhalter, Tim Vanbellingen, René Müri, Werner Strik, and Katharina Stegmayer. Gesture performance in schizophrenia predicts functional outcome after 6 months. *Schizophrenia bulletin*, 42(6):1326–1333, 2016.

[32] Ofer Agid, Cynthia O. Siu, Elizabeth Pappadopulos, Douglas Vanderburg, and Gary Remington. Early prediction of clinical and functional outcome in schizophrenia. *European Neuropsychopharmacology*, 23:842–851, 8 2013.

[33] Kimmy S. Kee, Michael F. Green, Jim Mintz, and John S. Brekke. Is emotion processing a predictor of functional outcome in schizophrenia? *Schizophrenia bulletin*, 29 3:487–97, 2003.

[34] Ricardo E. Carrión, Danielle McLaughlin, Terry E. Goldberg, Andrea M. Auther, Ruth H. Olsen, Doreen M. Olvet, Christoph U. Correll, and Barbara A. Cornblatt. Prediction of functional outcome in individuals at clinical high risk for psychosis. *JAMA Psychiatry*, 70:1133–1142, 2013.

[35] Rebecca Schennach-Wolff, Markus Jäger, Florian Seemüller, Michael Obermeier, Thomas Messer, Gerd Laux, Herbert Pfeiffer, Dieter Naber, Lutz G. Schmidt, Wolfgang Gaebel, Wolfgang Huff, Isabella Heuser, Wolfgang Maier, Matthias R. Lemke, Eckart Rüther, Gerd Buchkremer, Markus Gastpar, Hans Jürgen Möller, and Michael Riedel. Defining and predicting functional outcome in schizophrenia and schizophrenia spectrum disorders. *Schizophrenia Research*, 113:210–217, 9 2009.

[36] Eugenia Kravariti, Robin G Morris, Sophia Rabe-Hesketh, Robin M Murray, and Sophia Frangou. The maudsley early-onset schizophrenia study: cognitive function in adolescent-onset schizophrenia. *Schizophrenia Research*, 65(2):95–103, 2003.

[37] Kelly L. Cervellione, Katherine E. Burdick, John G. Cottone, Joseph P. Rhinewine, and Sanjiv Kumra. Neurocognitive deficits in adolescents with schizophrenia: longitudinal stability and predictive utility for short-term functional outcome. *Journal of the American Academy of Child and Adolescent Psychiatry*, 46 7:867–78, 2007.

[38] Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou. Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, 8(3):2663–2693, 2022.

[39] Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101:119–137, 3 2006.

[40] Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, 2011.

[41] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

[42] scikit-learn. 1.1. linear models — scikit-learn 1.5.0 documentation. `https://scikit-learn.org/stable/modules/linear_model.html#elastic-net` Viděno 2024-05-20.

[43] T. O. Hodson. Root-mean-square error (rmse) or mean absolute error (mae): when to use them or not. *Geoscientific Model Development*, 15(14):5481–5487, 2022.

[44] Alan J Meehan, Stephanie J Lewis, Seena Fazel, Paolo Fusar-Poli, Ewout W Steyerberg, Daniel Stahl, and Andrea Danese. Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Molecular psychiatry*, 27(6):2700–2708, 2022.

[45] Adam M Chekroud, Matt Hawrilenko, Hieronimus Loho, Julia Bondar, Ralitza Gueorguieva, Alkomiet Hasan, Joseph Kambeitz, Philip R Corlett, Nikolaos Koutsouleris, Harlan M Krumholz, et al. Illusory generalizability of clinical prediction models. *Science*, 383(6679):164–167, 2024.

[46] Christian Gaser, Robert Dahnke, Paul M Thompson, Florian Kurth, Eileen Luders, and Alzheimer's Disease Neuroimaging Initiative. Cat–a computational anatomy toolbox for the analysis of structural mri data. *biorxiv*, pages 2022–06, 2022.

[47] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Octave Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

[48] Jan Kudelka, Eduard Bakštein, Andrea Slováková, and Filip Spaniel. Unlocking the future: Medium-term prediction of clinically-relevant outcomes in first-episode schizophrenia patients. 2024. Článek v přípravě.

Příloha B

B.1 Implementované třídy

SpcaBair class

V této třídě je naimplementována redukce dimenze pomocí Bairovy metody (viz 3.1.1). Obsahuje následující metody:

- *___init___(self)* metoda inicializuje třídu a nastavuje proměnné `used_cols` na `None`, `U` na `None`, `theta` na 0 a `coefficients` na `None`.
- *find_optimal_hyper_param(self, num_folds, X, y, model)* metoda hledá optimální hyperparametr `theta` pro model využívající Bairovu SPCA. Výsledná optimální `theta` je vyhodnocena na základě křížové validace.
- *fit(self, X_train, y_train, theta=None)* metoda slouží k přizpůsobení modelu na trénovacích datech. Pokud je specifikována `theta`, použije se jako parametr, jinak je použita `theta` získána pomocí metody *find_optimal_hyper_param*. Metoda vrací transformační matici `U`, dekódovací matici `Z` a vlastní čísla `eigvals`.
- *transform(self, X_test)* metoda transformuje testovací data `X_test` pomocí transformační matice `U` a vrací dekódovaná data.
- *standardize_coefficients(X, y)* statická metoda, která vrací pole standardizovaných koeficientů.

SpcaBarshan class

V této třídě je naimplementována redukce dimenze pomocí metody Barshan (viz 3.1.2). Obsahuje následující metody:

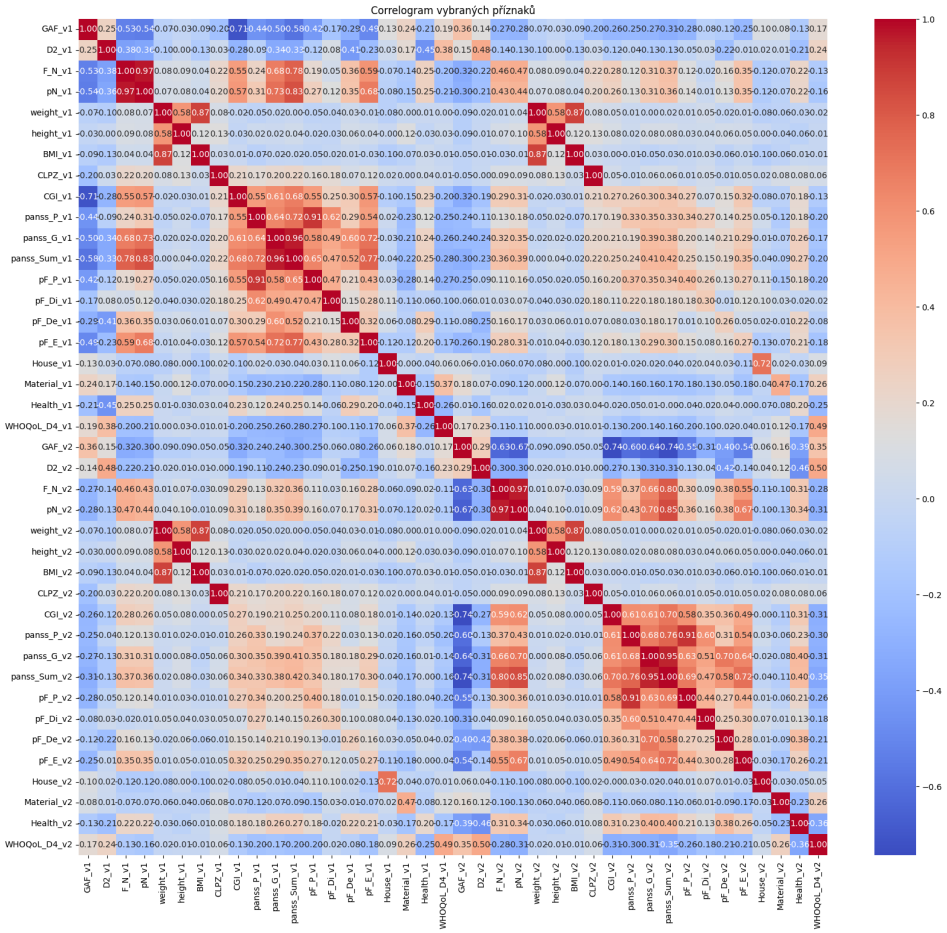
- *___init___(self)* metoda inicializuje třídu a nastavuje proměnné `U` na `None` a `d` na 0.
- *find_optimal_hyper_param(self, num_folds, X, y, model)* metoda slouží k nalezení optimálního hyperparametru `d` pro zvolený model. Optimální hodnota `d` je vyhodnocena pomocí křížové validace.
- *fit(self, X_train, y_train, d=None)* metoda slouží k přizpůsobení modelu na trénovacích datech. Pokud je specifikováno `d`, použije se toto `d`, jinak je použito optimální `d` nalezené pomocí metody předchodí. Tato metoda vrací optimalizovanou transformační matici `U`, dekódovací matici `Z` a vlastní čísla `eigvals`.
- *transform(self, X_test)* metoda transformuje testovací data `X_test` pomocí transformační matice `U` a vrací dekódovaná data.

SpcaDualBarshan class

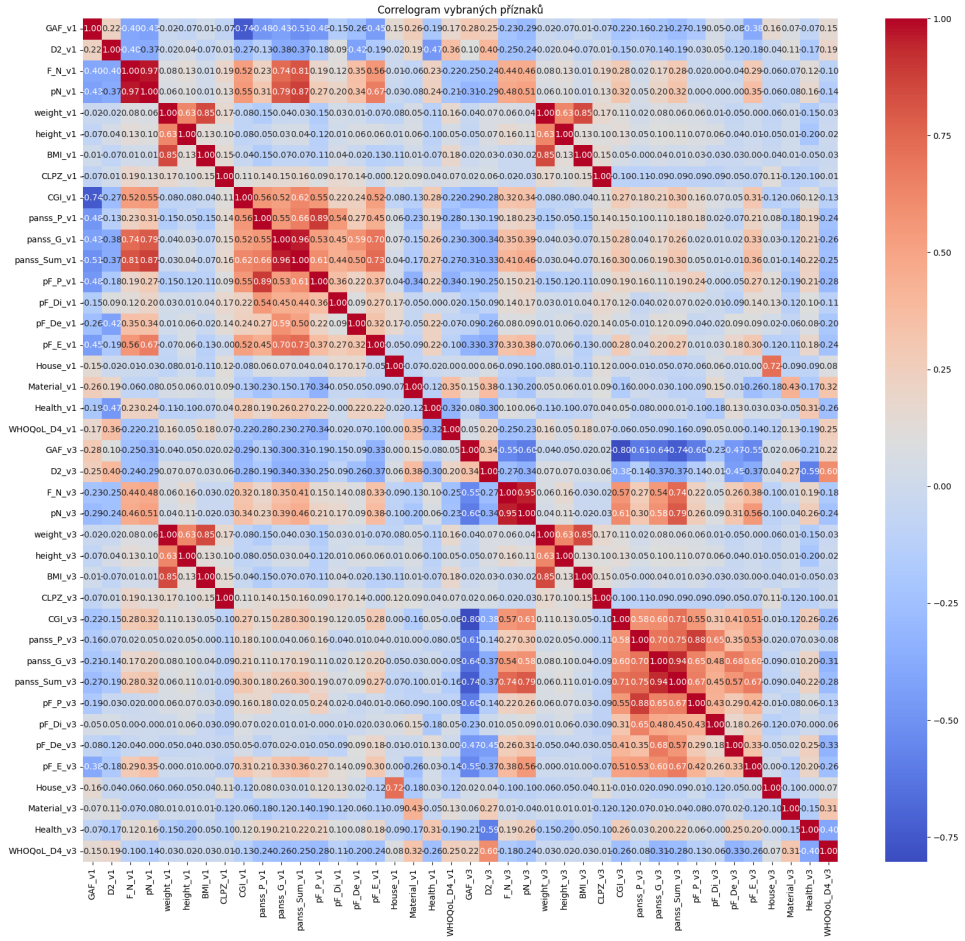
V této třídě je naimplementována redukce dimenze pomocí metody Barshan Dual SPCA (viz 3.1.2). Obsahuje následující metody:

- *___init___(self)* metoda inicializuje třídu a nastavuje proměnné `U` na `None` a `d` na 0.
- *find_optimal_hyper_param(self, num_folds, X, y, model)* metoda slouží k nalezení optimálního hyperparametru `d` pro zvolený model. Optimální hodnota `d` je vyhodnocena pomocí křížové validace.
- *fit(self, X_train, y_train, d=None)* metoda slouží k přizpůsobení modelu na trénovacích datech. Pokud je specifikováno `d`, použije se jako parametr, jinak je použito optimální `d` nalezené pomocí metody předchozí. Tato metoda vrací optimalizovanou transformační matici `U`, dekódovací matici `Z` a vlastní čísla `eigvals`.
- *transform(self, X_test)* metoda transformuje testovací data `X_test` pomocí naučené transformační matice `U` a vrací dekódovaná data.

B.2 Heatmapy vybraných příznaků



Obrázek B.1: Correlogram vybraných příznaků pro první (v1) a druhou (v2) vizitu



Obrázek B.2: Correlogram vybraných příznaků pro první (v1) a třetí (v3) vizitu