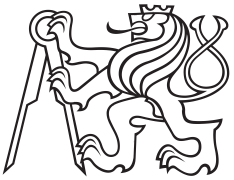
Diplomová práce



České
vysoké
učení technické
v Praze

F3

Fakulta elektrotechnická
Katedra kybernetiky

Individualizovaná detekce relapsu schizofrenních pacientů v programu ITAREPS

Bc. Alisa Housková
1802T002 - Biomedicínská informatika

Leden 2017

Vedoucí práce: Ing. Eduard Bakštein

Poděkování / Prohlášení

Děkuji všem, kteří přispěli ke vzniku této práci, jakýmkoliv činem. Zároveň děkuji i těm, kteří mi umožnili mi věnovat se studiu na vysoké škole, které mě naplňuje. V neposlední řadě děkuji svému vedoucímu za trpělivost a konzultace a za možnost spolupracovat na tak zajímavém projektu, jakým ITAREPS bez pochyby je.

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 dne 9. 1. 016

.....

Abstrakt / Abstract

Cílem práce je prozkoumat dostupná data z programu ITAREPS, který pomáhá přecházet hospitalizaci schizofrenních pacientů. Pacienti i jejich rodinní příslušníci během programu ITAREPS zasílají SMS zprávy s odpověďmi na 10 otázek, které se zaměřují na typické příznaky u pacientů trpících schizofrenií. Na jejich základě ITAREPS vyhodnocuje, zda se pacient zhoršuje a je tak podezřelý z hospitalizace. Práce se zaměřuje zejména na SMS posílané pacienty a rodinnými příslušníky. Díky pozorování posílaných SMS v čase je možné vypozorovat charakteristická období pro hospitalizované pacienty. Dále je zkoumána závislost jednotlivých otázek a také, zda existují shluky mezi pacienty dle zasílaných hodnot. Jedna z mála informací, které jsou známy o pacientovi při vstupu do programu, jsou jeho symptomy, proto se práce zaměřuje i na ně. Práce zkoumá závislosti mezi symptomy samotnými. Neméně zajímavou částí je hledání podobností mezi pacienty dle jejich vstupních symptomů. Nalezené shluky totiž pomáhají nalézt více individualizované řešení. Druhou částí práce je vyhodnocení úspěšnosti současného systému, který nastavuje globální prahy pro všechny pacienty stejné, a na základě znalosti dat navrhnout alternativní řešení, které více zohledňuje různorodost pacientů. Současný systém je vyhodnocen pro původní prahy, ale i pro nové nastavení optimálních prahů. Alternativní řešení k současnému systému využívá rozdělení pacientů do shluků dle symptomů a poté je použito lineárního klasifikátoru SVM. Klasifikátor je učen na několik sad dat tak, aby co nejvíce simuloval původní systém. Klasifikátor detekuje podezřelé pacienty z hospitalizace úspěšněji než současný systém s optimálními prahy. Práce nabízí jinou možnost, jak detekovat pacienty, kterým se zhoršuje stav a může jim hrozit hospitalizace.

Klíčová slova: ITAREPS, schizofrenie, individualizovaná detekce, explorativní analýza, strojové učení, klasifikace pacientů

The goal of this research is to analyze the data made available by the ITAREPS program. This program helps to prevent the need for hospitalisation of patients with schisophrenia. Here is how it works. On regular basis, the patients and their family members send SMSs with answers to 10 questions related to symptoms typical for patients with schisophrenia. Based on the answers provided, ITAREPS evaluates the patient's condition and anticipates their need for hospitalisation. This thesis is particularly focused on analyzing the SMSs sent by the patients and their family members. Thanks to the analysis, it is possible to identify characteristic periods for hospitalised patients. Additionally, all the questions are being examined to find any possible dependencies on each other. Last, but not least, all the answers are being analysed to determine any groups of patients with similar answer values. The most important factors when a patient is joining the program are their symptoms. Hence the focus of exploration is also directed to them. This research is looking for any dependencies on each other as well as dependencies between symptoms and number of hospitalisations before joining the ITAREPS program and during the ITAREPS program. It also helps us find dependencies between symptoms and number of hospitalisations before and during the ITAREPS program. We are trying to reveal any similarities between patients and their symptoms. Hereby identified groups could help to find more individualised solution. The next part of the thesis will focus on evaluating the effectiveness of the current system. The system is actually based on global thresholds which are the same for all patients. The goal here is to find a more individualized method which will take into account the diversity of patients. Once the current data has been evaluated, we can recommend new optimal thresholds. Another solution would be to sort the patients into groups based on their symptoms and then use the linear classifier SVM. This classifier works by using several sets of data in order to simulate the original system as much as possible. The SVM classifier identifies a patient's need for hospitalisation more effectively than the original system with optimal thresholds. In conclusion, this thesis offers alternative options to detect patient's deteriorating conditions in order to better prevent the need for their hospitalisation.

Keywords: ITAREPS, schisophrenia, individualized detection, explorative analysis, machine learning, classification of patients

Obsah /

Tabulky / Obrázky

Kapitola 1
Úvod

Pacienti trpící schizofrenií mohou zažívat opětovné propuknutí choroby nebo jejich příznaků, tzv. relaps, během něhož pacient není schopen se sám o sebe postarat a potřebuje akutní lékařskou pomoc. Doktoři své pacienty nevídají příliš často, v některých případech pouze jednou měsíčně. To může vést k přehlédnutí příznaků nadcházejícího relapsu a k jeho pozdní detekci.

Monitorovací program ITAREPS (Information Technology Aided Relapse Prevention Program in Schizophrenia) [1–2] vznikl, aby pomohl včas tyto příznaky detekovat. Program pomocí telemedicíny sbírá a uchovává data zasílaná pacienty a na jejich základě pomáhá detekovat možný relaps. Telemedicína, neboli medicína na dálku, je v současné době velmi populární a její využití je dnes již standardem. Jedná se nejčastěji o technologie umožňující sběr dat v pohodlí, například z domova či z nemocničního lůžka. Sbíraná data jsou odesílána přes komunikační síť do systému a ten je následně ukládá a vyhodnocuje. V programu ITAREPS je používán komunikační kanál SMS (Short Message Service). V současném systému ale není detekce relapsu individualizována, což znamená, že pro všechny pacienty jsou uplatňována stejná pravidla při vyhodnocování rizika relapsu. Z tohoto důvodu v mnohých případech systém relaps u pacientů nedetekuje, ale ve skutečnosti k němu dojde.

Tato práce se soustředí na nalezení alternativy k současné detekci relapsů prováděné systémem programu ITAREPS. Cílem práce je návrh metody, která bude detekovat příznaky více individuálně a u které bude k chybným upozorněním na relaps docházet pokud možno jen v minimální míře. Výsledný návrh může spočívat v dílčích úpravách aktuálně používaného systému a způsobu detekce relapsů anebo v kompletním nahrazení současného systému. Jako hlavním vstupem pro analýzu je mimo jiné velké množství dat nashromážděných během provozu programu ITAREPS.

1.1 Schizofrenie

Schizofrenie je mentální onemocnění. Odhaduje se, že jím trpí téměř 30 milionů lidí, z nichž 20 milionů žije ve vyspělých zemích. Celoživotní výskyt onemocnění

v populaci je 0,5 - 1 % a výskyt onemocnění je poměrně rovnoměrný po celém světě [3–4]. Jedná se psychotické onemocnění, které narušuje vnímání reálného světa okolo a tím dochází ke změnám osobnosti. Onemocnění doprovází mnoho různých příznaků, které se často dělí do následujících skupin: pozitivní, negativní, kognitivní, afektivní, katatonní a dezorganizace.

Často ale není snadné příznaky detekovat, mohou být totiž doprovodnými jevy jiných nemocní a vyskytují se i u zdravých jedinců. Důležité je, jaký vliv příznaky mají na chování nemocného.

Pozitivní příznaky jsou takové, kdy osoba ztrácí přehled o realitě. Jedná se o halucinace a bludy, přičemž bludy bývají často paranoidní. Dále se jedná o poruchy pohybu a poruchy myšlenkové, kdy pacient nedokáže logicky propojovat jednotlivé myšlenky.

Negativní příznaky ovlivňují pacientův sociální život a často bývají zaměňovány s leností a laxností. Není tedy snadné je vypozorovat. Například oploštění emocí způsobuje ztrátu některých emocí; gestikulace i mimika se stávají méně aktivní.

Kognitivní příznaky se též hůře pozorují, neboť se jedná o poruchy pozornosti, paměti. Jsou považovány za přímý projev primární patologie nemoci. K jejich detekci slouží specializované neuropsychologické nástroje a testy.

Mezi afektivní příznaky je řazena například deprese nebo emoce neúměrné k dané situaci.

Dezorganizace myšlení a chování způsobuje u nemocných nezvyklé chování, často podivínské.

Narušená psychomotorika signalizuje katatonní příznak - člověk nevykonává žádnou spontánní aktivitu, na povely ale reaguje příslušnými pohyby [5].

Výše zmíněné příznaky pomáhají dělit schizofrenii do nejrůznějších tříd [6], používá se ale i dělení podle převládajících příznaků pozitivních či negativních na typ I a typ II, nebo pozitivní, negativní a smíšený typ.

Přesné příčiny onemocnění jsou stále neznámé, současná léčba se snaží především eliminovat příznaky onemocnění pomocí antipsychotik, různě zaměřených rehabilitací, nebo různými kurzy zvládání nemoci pro rodinu a pacienta.

1.2 ITAREPS

Program ITAREPS (akronym Information Technology Aided Relapse Prevention Programme in Schizophrenia) využívá moderních technologií jako je telemedicína za použití mobilních telefonů a krátké textové zprávy (dále jen SMS) pro prevenci relapsů u schizofrenních pacientů.

Pacienti trpící psychotickou poruchou, jakou je schizofrenie, kteří se dobrovolně zapojili do programu, týdně odpovídají na 10 otázek dotazníku Early Warning Signs Questionnaire (dále jen EWSQ), viz tabulka 1.2. Účast rodinného příslušníka není povinná, byť je velmi doporučena, neboť hodnocení z pohledu pozorovatele může být méně subjektivní, než hodnocení sebe sama. Systém ITAREPS všem zúčastněným každý týden ve stejný den i čas posílá připomínku s žádostí o vyplnění EWSQ dotazníku. Otázky určené pacientovi se liší od otázek kladených rodinným příslušníkům.

Skóre	Význam
0	Beze změny či zlepšení stavu
1	Mírná změna k horšímu
2	Střední změna k horšímu
3	Výrazná změna k horšímu
4	Extrémní změna k horšímu

Tabulka 1.1. Skóre odpovědí dotazníku

Dotazník byl navržen tak, aby pozoroval změnu stavu (případné zlepšení, nebo zhoršení) oproti poslednímu týdnu, kdy byl dotazník vyplněn. Každá otázka nabývá hodnot 0 až 4, kde 0 znamená bez změny stavu, případně značí zlepšení. Na druhé straně škály stojí hodnota 4, která oznamuje extrémní zhoršení stavu tázaného.

V případě, že skóre EWSQ dotazníku překročí stanovený práh, systém vygeneruje ALERT (upozornění) a zašle upozorňující email příslušnému psychiatrovi s kódovým označením pacienta. Psychiatr na základě Early Intervention Algorithm (EIA) [7] zvýší dávku antipsychotik o 20 % pro příštích 24 hodin. Tento přístup byl prokázán jako efektivní.

Po vygenerovaném varovném signálu nastupuje třítýdenní alertní období, ve kterém je žádost o vyplnění dotazníku posílána dvakrát týdně. Jestliže práh není během ALERT PERIOD (alertní období) překročen, pak je psychiatr upozorněn, že se pacientův stav nezhoršuje a lze mu upravit medikaci na původní hladinu. Jestliže je práh překročen, tj. stav pacienta se zhoršuje, systém vygeneruje pohotovostní e-mail a alertní perioda je prodloužena o další tři týdny.

Otázky EWSQ10 určené pacientovi a rodinným příslušníkům jsou uvedeny v Tabulce 1.2.

Č.	Otázky pro pacienta (EWSQ-10P)	Otázky pro rodinné příslušníky (EWSQ-10FM)
1	Zhoršil se u vás od posledního hodnocení spánek?	Změna charakteru spánku.
2	Zhoršila se u vás od posledního hodnocení chuť k jídlu?	Nápadná změna chování.
3	Zhoršilo se u vás od posledního hodnocení soustředění, například při čtení či sledování televize?	Sociální stažení.
4	Zpozoroval/a/ jste u sebe od posledního hodnocení strach, obavy či jiné nepříjemné pocity z ostatních lidí?	Zhoršené fungování v každodenních činnostech.
5	Zpozoroval/a/ jste u sebe od posledního hodnocení zvýšený neklid nebo podrážděnost?	Zhoršení v oblasti osobní hygieny.
6	Zpozoroval/a/ jste u sebe od posledního hodnocení, že se bezprostředně kolem vás dějí věci, kterým nerozumíte?	Ztráta iniciativy, motivace.
7	Zpozoroval/a/ jste u sebe od posledního hodnocení ztrátu energie a zájmu?	Nápadné obsahy myšlení, nápadné zaujetí zvláštními myšlenkami.
8	Zhoršila se u vás od posledního hodnocení schopnost řešit každodenní problémy?	Nápadná chudost v řeči a myšlení.
9	Slyšel/a/ jste od posledního hodnocení hlasy, i když nikdo v té chvíli nebyl ve vašem okolí?	Podrážděnost, neklid, agresivita.
10	Zpozoroval/a/ jste od posledního hodnocení jiný časný varovný příznak typický pro vás?	Jiná nápadná změna v porovnání s předchozím stavem.

Tabulka 1.2. Otázky EWSQ10 pro pacienta a rodinné příslušníky

1.3 Motivace

Současný systém nastavuje práh pro všechny pacienty stejný. Každý pacient ale svůj stav vnímá odlišně, hodnocení je subjektivní. Motivací této práce najít způsoby, které mohou pomoci rozhodovat o nástupu kritického stavu (nastal varovný signál a po něm alertní perioda) více individuálně a zavčas podat zvýšenou dávku antipsychotik a předejít tak dalšímu zhoršování a případnému relapsu. Ke splnění cíle budou užity statistické metody, které hloubkově prozkoumají data, jejich charakter a závislosti, a metody strojového učení pro aplikaci individualizovaného klasifikátoru.

Kapitola 2
Použité metody

Kapitola krátce uvede některé použité metody v této práci. Nejedná se o úplný popis teorie, spíše krátké seznámení s některými úskalími, zejména u metod strojového učení. Pro výpočty a implementaci použitých algoritmů byl použit software $MATLAB^{®}$. Kromě standardních knihoven práce používá i knihovnu LibSVM [8].

2.1 Statistické metody

Neparametrické metody statistických testů předpokládají normální rozdělení v datech, zatímco parametrické předpokládají data s jinou než normální distribucí. Vícenásobné testy nad stejnými daty mohou vést k falešně pozitivním výsledkům. Tento problém řeší Bonferronniho korekce, která spočívá ve vydělení kritické testové hodnoty α počtem provedených testů nad těmiž daty m.

2.2 Principal Components Analysis

Principal Component Analysis (PCA) metoda se snaží najít skryté závislosti a redukuje dimenzi vstupních dat, která jsou namapována do nového prostoru při zachování maximální variance. Tato transformace se snaží data promítnout do hlavních komponent, kdy první komponenta obsahuje co možno nejvíce variability a maximalizuje varianci promítnutí všech hlavních komponent do nového prostoru. Metoda je často využívána k nalezení skrytých redundancí a trendů v objemných datech. Jednotlivé komponenty popisují, na kolik jsou sycena danými pozorováními. Lze tak pozorovat, zda některá pozorování nepřevažují, či zda jsou rovnoměrně zastoupena.

2.3 Hierarchické shlukování

Hierarchické shlukování hledá hierarchii mezi shluky v datech. Pro nalezení hierarchie je potřeba zohlednit vybranou metriku pro výpočet vzdálenosti mezi jednotlivými klastry a linkage (spojovací) kritérium, jež klastry poté propojí.

Grafickou reprezentaci označujeme jako dendrogram, kde mnohem snáze pozorujeme, zda vznikly nějaké shluky či nikoliv a na kolik jsou si blízké. Při sestavování lze použít mnoho výpočtů vzdáleností, například vážená průměrná vzdálenost (WPGMA), nevážená průměrná vzdálenost (UPGMA), největší a nejkratší vzdálenost. Tyto jednotlivé metody počítají vzdálenosti na základě nejběžnějších metrik jako euklidovská, Minkowskiho metrika, a další.

2.4 Nehierarchické shlukování

Algoritmus K-means je velmi používaným algoritmem pro nehierarchické shlukování. Není deterministický a ne vždy najde globální minimum, je důležité před jeho použitím zjistit optimální počet klastrů. Toho lze dosáhnout například použitím koeficientů silhouette. Díky celkové průměrné nejvyšší hodnotě můžeme najít optimální počet shluků. Silhouette graf znázorňuje, jak moc jsou body v konkrétním shluku blízko (v rozsahu -1, 1) k bodům v sousedních shlucích, tzv. silhouette koeficienty. Silhoutte koeficient udává míru podobnosti bodu ve svém shluku v porovnání s podobností k jiným shlukům. Čím je koeficient větší, tím více je daný bod podobný bodům ve svém shluku a tedy méně podobný ostatním bodům v jiných shlucích. Jestliže většina bodů má nízké či záporné hodnoty, pak je počet shluků zvolen špatně.

2.5 Sensitivita a specificita

Ohodnotit binární klasifikační úlohy pomáhají metriky sensitivita a specificita. Jedná se o statistické metriky. K jejich spočtení je potřeba znát zastoupení kladně pozitivních, falešně pozitivních, kladně negativních a falešně negativních klasifikací.

Kladně pozitivní (TP) jsou vzorky, které spadají do pozitivní třídy a byly tak skutečně klasifikovány. Falešně pozitivní (FP) jsou naopak vzorky, které patří do pozitivní, ale byly klasifikovány do třídy negativních. Obdobně pak pro třídu negativních, kdy negativní vzorky misklasifikované do pozitivní třídy se značí jako falešně negativní (FN) a negativní vzorky zařazené do správné třídy negativních jako kladně negativní (TN).

Na základě těchto hodnot lze určit následující metriky.

$$Sensitivita = TP/(TP + FN)$$

$$Specificita = TN/(TN + FP)$$

$$Přesnost = (TP + TN)/(TP + TN + FP + FN)$$

$$F1\ skore = 2 \cdot TP/(2 \cdot TP + FP + FN)$$

$$Youden's\ J\ index = sensitvita + specificita - 1$$

2.6 Učení s učitelem

Algoritmu jsou předložena trénovací data s jejich předem známou příslušností do konkrétní třídy. Validace a ověření výsledků pak probíhá nad testovací/validační množinou taktéž s předem známou příslušností do konkrétní třídy.

Trénovací data by měla představovat co nejpopisnější vzorek z reálného světa (celá množina dostupných dat). Jak vhodně vybrat trénovací a validační/testovací množinu a problémy, které s tím souvisí, popisuje mnoho článků. Jedná se například o vyvážení zaujetí a variance [9].

Klasifikátor, který se velmi přesně naučí na trénovací množinu, tedy má velké zaujetí a malá variance, bude nejspíše špatně detekovat testovací data. Zatímco klasifikátor s malým zaujetím a velkou variancí může klasifikovat již trénovací množinu nepřesně.

Dalším problémem je nevyvážené učení [10], kdy zastoupení vzorků jedné třídy je mnohonásobně menší než třídy druhé. Ideální zastoupení v obou třídách by mělo být stejné. Ve skutečnosti však taková data často nejsou k dispozici. Zejména ve světě biologie, medicíny se často jedná o poměry 1:10, ale i 1:1000 [11].

Existuje hned několik způsobů řešení nevyváženosti tříd vstupních dat:

Metoda podvzorkování spočívá v tom, že některé vzorky početnější třídy se ze vstupních dat odstraní. Vstupní množina dat je tedy sestavena z maximálního počtu vhodných vzorků n z méně početné třídy a z $k \cdot n$ náhodně vybraných vzorků z početnější třídy. Ideální k je rovno 1, nicméně takto silně náhodně podvzorkovaná data mohou natolik zkreslit charakter vstupu, že výsledek trénování je velmi zaujatý. Volba většího k ovšem může zase vést k tomu, že klasifikátor přehlédne vzorky z menší množiny a natrénuje se na množinu početnější.

Metoda nadvzorkování rozkopíruje minoritní třídu tak, aby počet vzorků odpovídal velikosti majoritní množiny. Tato metoda může přinášet lepší výsledky,

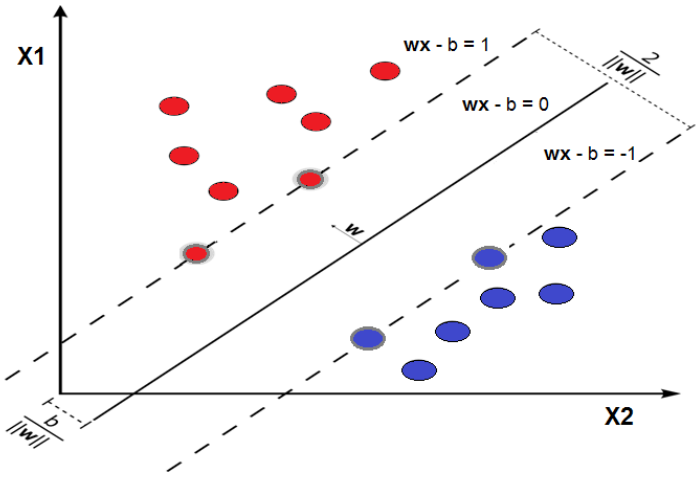
jelikož není omezena variabilita dat početnější třídy. V této práci je použito i metody nadvzorkování.

Neméně důležitým faktorem je i vhodný výběr poměru trénovací a testovací množiny. Obvykle platí, že trénovací množina by měla být co největší, zároveň ale je potřeba, aby testovací data obsahovala velkou varianci dat pro testování naučeného modelu. Častým poměrem trénovacích ku testovacích dat bývá 70:30 či 80:20. Validační množina se vytváří z dat trénovacích v podobném poměru a slouží k validování zvolených parametrů modelu. V práci je použita metoda křížové validace Monte-Carlo, kde rozdělení na trénovací a validační množinu proběhne náhodně nkrát. Pro všech n běhů je spočten medián výsledných statistik a jsou vybrány nejlepší paramtry modelu. Testovací množina funguje jako nová data a slouží k ověření vybraného modelu.

2.7 Support Vector Machine

Původní lineární algoritmus Support Vector Machine (SVM) navrhnul Vladimir N. Vapnik a Alexey Ya Cherbonenkis už v roce 1963, poté v roce 1992 V. N. Vapnik spolu s dalšími badateli navrhnuli rozšíření na nelineární klasifikátor pomocí aplikace kernelu. Standard využívající soft margin byl publikován v roce 1995.

Jedná se o metodu strojového učení s učitelem. Vstupem pro SVM je množina dvojic (x, y) z X, Y. SVM je trénováno na množině pozorování, kde pro každé pozorování X, $x(i), i = 1 : n \in R^n$ existuje štítek Y, $y(i), i = 1 : n \in -1, 1$. Cílem je najít vhodnou třídu $y \in Y$ pro již pozorované $x \in X$, tj klasifikátor $y = f(x, \alpha)$, kde α jsou vstupní parametry klasifikátoru.



Obrázek 2.1. SVM lineární klasifikátor. *Obrázek ukazuje dvě nadroviny oddělující dvě klasifikované třídy, v jejich středu lezí super nadrovina.*

Jakákoliv nadrovina R^n lze popsat $w \cdot x - b = 0$, kde w je normálnový vektor hyperroviny. Parametr $\frac{b}{||w||}$ je odsazení od počátku hyperroviny.

Jestliže jsou data lineárně separabilní, klasifikátor rozdělí data na dvě nadroviny tak, aby odstup mezi nimi byl co největší, $\frac{2}{||w||}$. Hyperroviny lze popsat rovnicemi $w \cdot x - b = 1$ a $w \cdot x - b = -1$.

Přesně v polovině odstupu mezi nalezenými hyperrovinami leží maximální hyperrovina. Úloha spočívá v nalezení maximálního odstupu mezi nadrovinami pomocí minimalizace $||w||$. Zároveň chceme zajistit to, aby vzorky spadaly pouze vně a na hranici nadrovin, k tomu slouží pevné nastavení hranic, které je popsáno rovnicemi níže.

$w \cdot x(i) - b >= 1$ pro $y(i) = 1$ a $w \cdot x(i) - b <= -1$ pro $y(i) = -1$. Pro X v prostoru R^n tak vznikne klasifikátor $x \rightarrow sign(w \cdot x - b)$, vektor x je tvořen body, které leží přesně na hranici dvou hyperrovin, tzv. support vektory.

V případě, že pracujeme s nelineárně separabilními daty, použijeme měkký odstup (soft margin) a tzv. hinge ztrátovou funkci $max(0, 1 - y(i)(w \cdot x(i) - b))$, která říká, že neleží-li $x(i)$ na správné straně (ve své nadrovině), pak je výsledek funkce roven vzdálenosti od hranice správné nadroviny, v opačném případě, leží-li $x(i)$ na správné straně, pak je výsledek roven 0.

Je potřeba zavést parametr λ, který upravuje vzdálenost $||w||^2$, a tak lze určit optimální poměr mezi velkým odstupem hyperrovin od sebe a počtem správně klasifikovaných $x(i)$.

$$C \cdot \sum_{i=1}^{n} max(0, 1 - y(i)(w \cdot x(i) - b)) + \lambda \cdot ||w||^2$$

Penalizační váhy C pro misklasifikaci do jedné, či druhé třídy posouvají hranice hyperrovin a umožňují tak klasifikátoru se zaměřit na misklasifikaci, která je v dané úloze považována za kritičtější.

Kapitola 3
Analýza dat a otázek

Jedná se o dotazníková data, a proto je potřeba se i zaměřit na podobnosti a souvislosti jednotlivých otázek mezi sebou. Lze předpokládat, že některé otázky spolu souvisí, jelikož mohou poukazovat na podobné příznaky schizofrenie. Stejně tak mohou být otázky na sobě závislé. Pomocí PCA metody a klastrovacích metod lze tyto souvislosti a skryté závislosti nalézt.

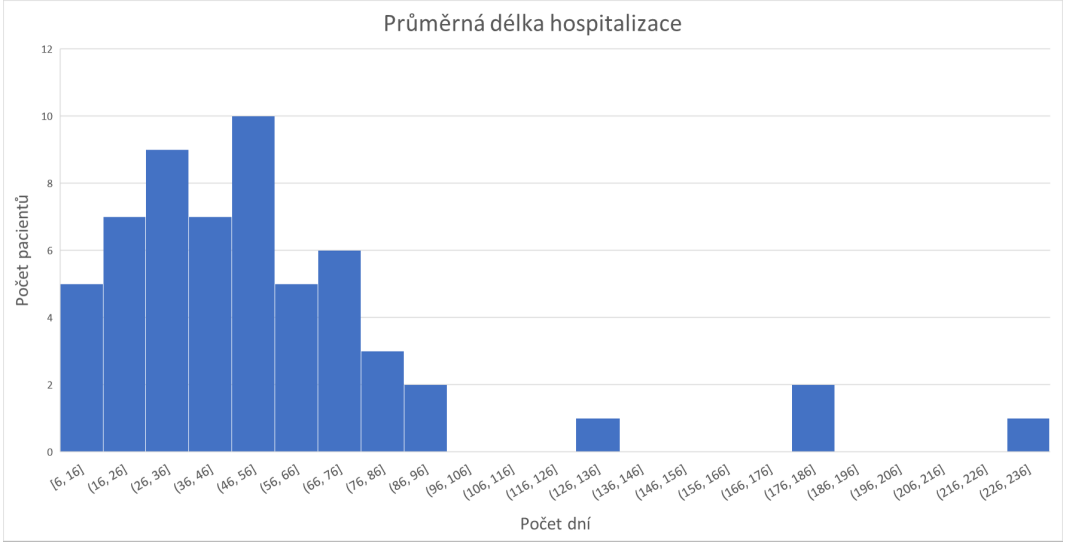
Explorativní analýza nám poskytuje především přehled o rozdělení výběru odlehlých hodnot, jako jsou kvantily, či průměr. Získané hodnoty slouží k testování hypotéz pomocí testové statistiky. Pro kvantilové rozložení souboru se často používá krabicový graf, zejména proto, že ukazuje odhad mediánu, symetrii v oblasti kvartilů (0,25 a 0,75) a také zobrazuje odlehlé hodnoty pozorování. Nejen výše zmíněné poznatky umožňují pozorovat charakter a vlastnosti dat, díky nimž lze lépe vyvozovat závěry a vybrat vhodněji další metody zpracování.

Tyto poznatky pomohou porozumět datům a mohou vést k nalezení skrytých spojitostí, které lze využít pro vytvoření klasifikátoru.

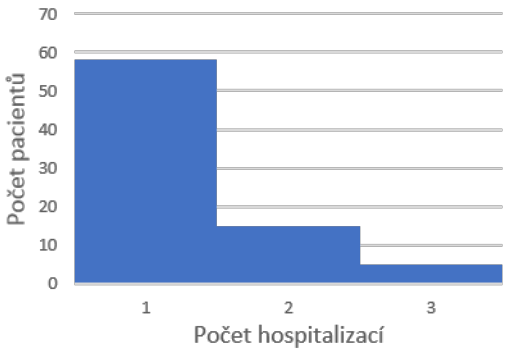
3.1 Základní přehled o hospitalizacích

Údaje o počátku a o konci jednotlivých hospitalizací zadávají lékaři do systému ITAREPS ručně.

Programu ITAREPS se zúčastnilo celkem 280 pacientů, 146 z nich pak pouze v rámci krátkodobé studie. Během programu ITAREPS bylo hospitalizováno 58 pacientů alespoň jedenkrát, 15 pacientů alespoň dvakrát a 5 pacientů bylo hospitalizováno třikrát. Průměrně hospitalizace trvala 58 dní. Základní histogramy hospitalizací a průměrné délky hospitalizací jsou zobrazeny na grafech obrázek 3.1.



Obrázek 3.1. Průměrná délka hospitalizace



Obrázek 3.2. Počet hospitalizací během ITAREPS

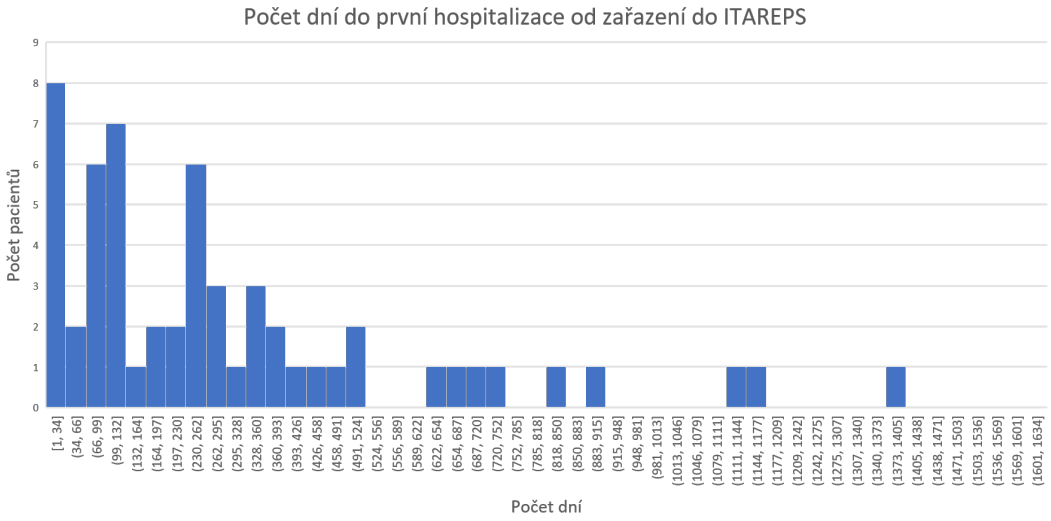
Tabulka 3.1 popisuje průměrné délky hospitalizací a počty dní od konce jedné hospitalizace do počátku druhé. Pro 4. hospitalizaci nejsou data uvedena, jelikož se týkala pouze úzké skupiny pacientů v programu ITAREPS.

Hospitalizace	Délka	Do další hostpitalizace
1.	41	304
2.	55	180
3.	44	-

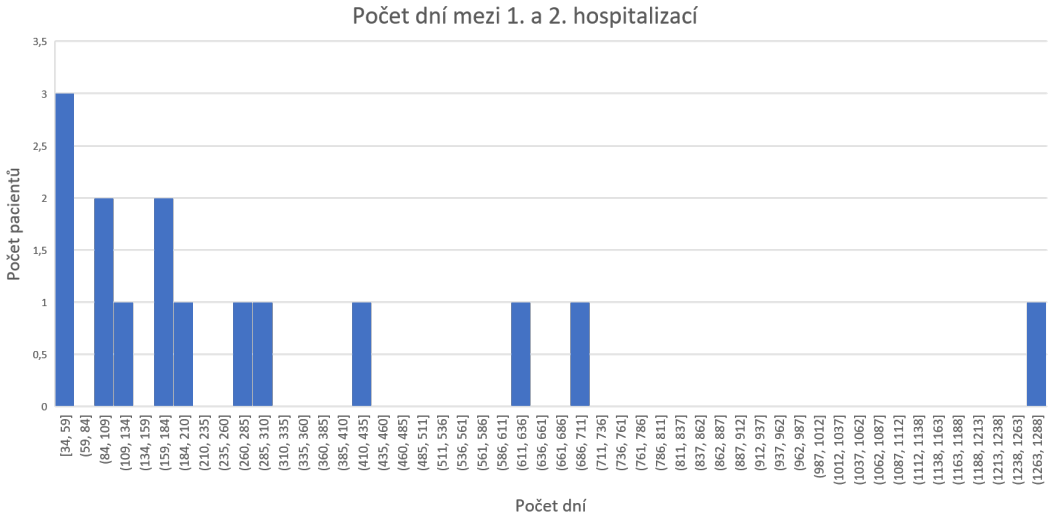
Tabulka 3.1. Hospitalizace, jejich délka ve dnech a období mezi hospitalizacemi. *Celkový stav pacienta po hospitalizaci se může zhoršovat a relapsy se mohou objevovat v kratších intervalech.*

V uvedených datech lze vypozorovat, že každá hospitalizace pacienta negativně poznamená. Vzhledem k tomu, že se rozestupy mezi jednotlivými hospitalizacemi zkracují, je možné usoudit, že pacientův stav se zhoršuje. Není ani výjimkou, že

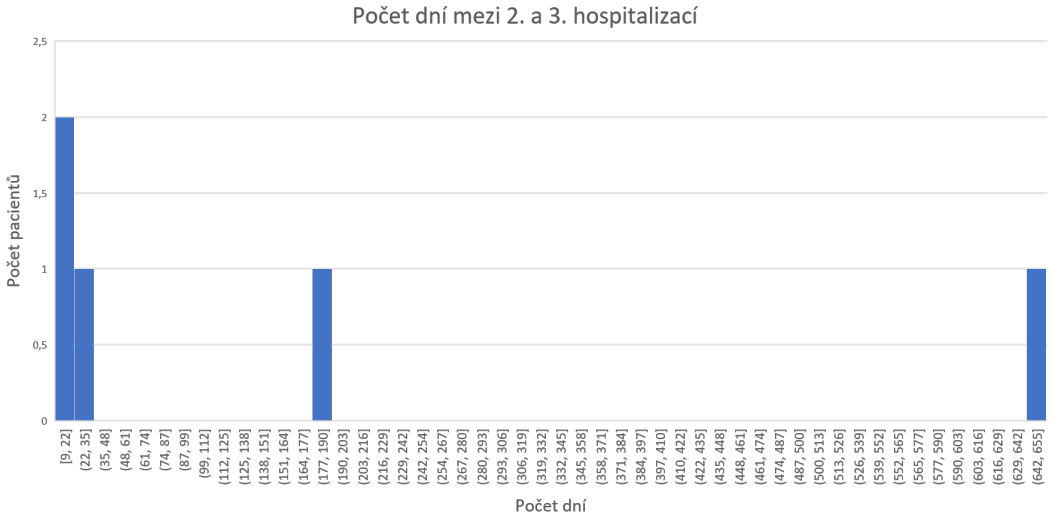
délka následujících hospitalizací se prodlužuje. Celkový stav pacienta po hospitalizaci se může zhoršovat a relapsy se mohou objevovat v kratších intervalech – což je v souladu s očekáváními a s tím, co je známo o schizofrenii.



Obrázek 3.3. Počet dní do první hospitalizace od zařazení do ITAREPS



Obrázek 3.4. Počet dní mezi 1. a 2. hospitalizací



Obrázek 3.5. Počet dní mezi 2. a 3. hospitalizací

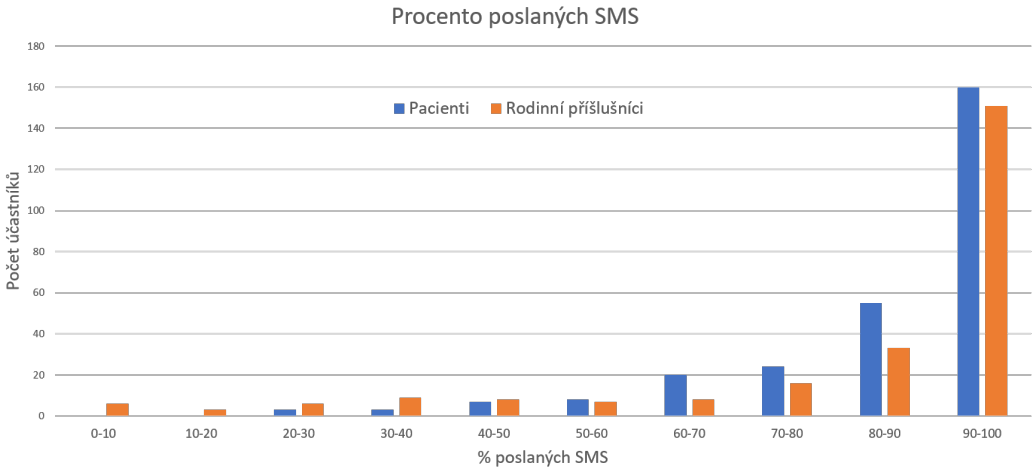
Obrázek výše ukazuje absolutní délky hospitalizací pro kompletní sadu pacientů. Je patrné, že absolutní počty dní hospitalizací se velmi liší. Může se ale jednat o odlehlá pozorování, což je těžké potvrdit vzhledem k malému počtu dat, zejména u třetí hospitalizace.

3.2 Základní přehled o poslaných SMS

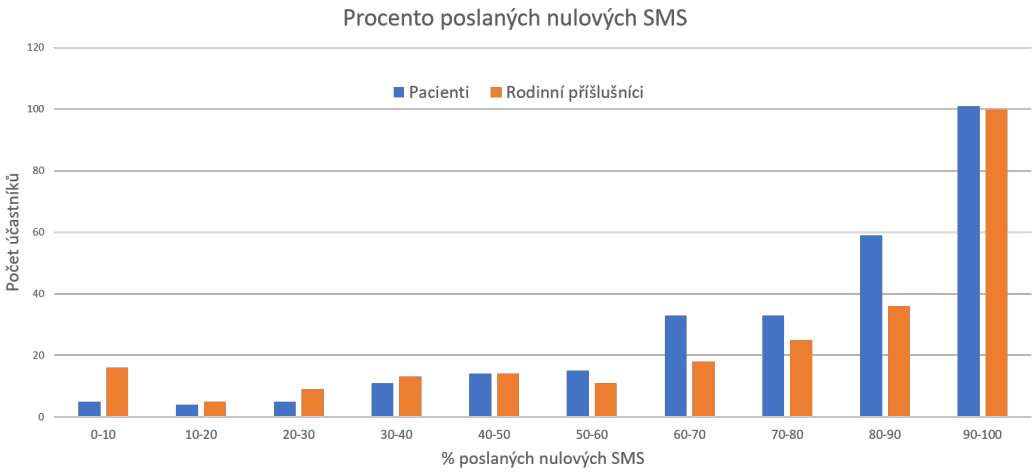
Histogramy níže ukazují procento poslaných SMS z celkového počtu SMS, které měly být dle programu ITAREPS poslány (každý týden jedna SMS). Další sada histogramů zobrazuje procento poslaných SMS s nulovým obsahem (celková suma číselných odpovědí v SMS je rovna nule). Ve všech histogramech jsou znázorněna procenta SMS posílaných pacienty a jejich rodinnými příslušníky.

Pokud se rodinný příslušník nezúčastnil, procento poslaných SMS je pro něj nulové. U hospitalizovaných pacientů je ze záznamů vyloučeno období hospitalizace, během kterého se SMS nezasílají.

Více než tři čtvrtiny pacientů (79,29 %) a zhruba dvě třetiny (63,80 %) rodinných příslušníků posílá alespoň 70 % vyžádaných SMS. Na první pohled by se mohlo zdát, že rodinní příslušníci jsou méně aktivní v posílání SMS než pacienti. Je to způsobeno tím, že v grafech jsou zaznamenány absolutní počty účastníků, ale ne všichni rodinní příslušníci se zúčastnili programu a tudíž je jejich počet menší. Trend patrný v grafech ukazuje, že rodinní příslušníci jsou podobně aktivní, jako pacienti.

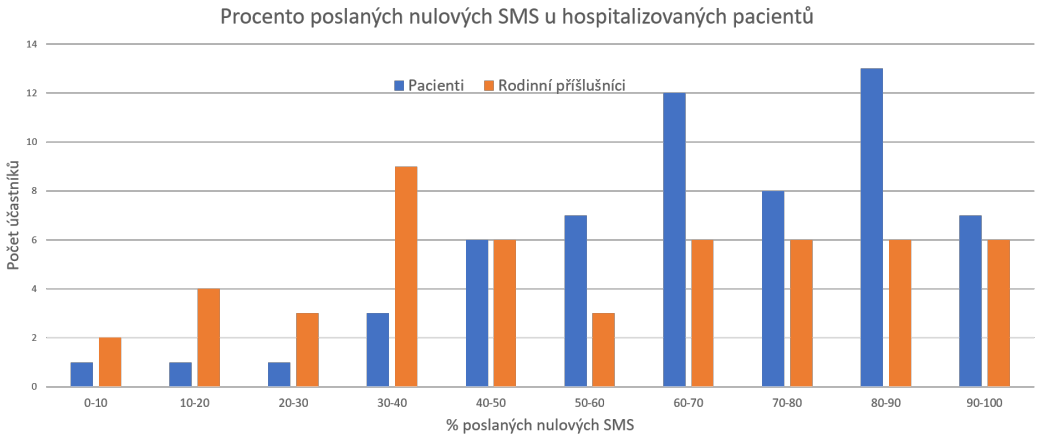


Obrázek 3.6. Procento poslaných SMS u všech pacientů a rodinných příslušníků

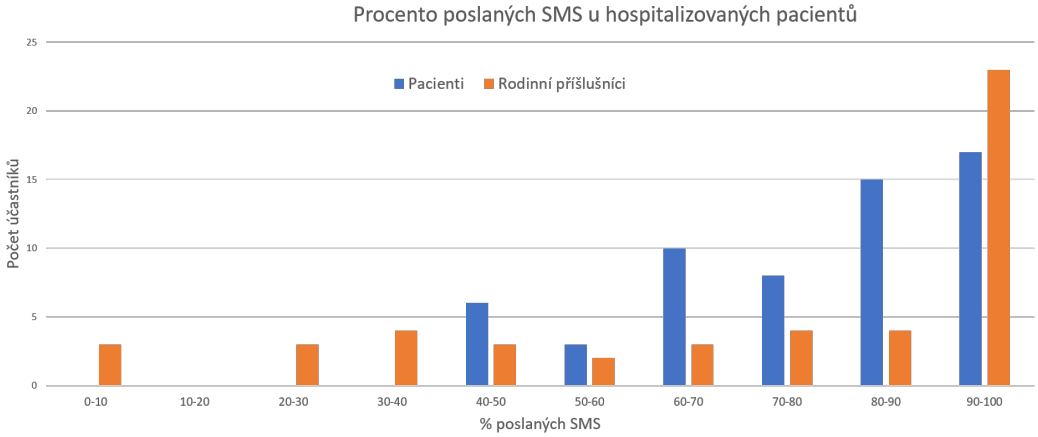


Obrázek 3.7. Procento poslaných nulových SMS u všech pacientů a rodinných příslušníků

Z obrázku je patrné, že velké procento poslaných SMS tvoří SMS s hodnotou nula. Pro nehospitalizované pacienty histogramy vypadají velmi podobně, zatímco pro hospitalizované pacienty během programu ITAREPS jsou výsledky poměrně odlišené. Majoritní skupina se ze 70 % poslaných SMS posouvá až na 50 % poslaných SMS, což může být způsobené tím, že pacienti trpí negativními příznaky (bude rozebráno dále) a jsou tak méně aktivní, případně zapomětliví. Musíme si uvědomit, že pacientů, kteří prošli hospitalizací během ITAREPS, je málo a jejich rodinných příslušníků ještě méně.



Obrázek 3.8. Procento poslaných SMS u hospitalizovaných pacientů a rodinných příslušníků



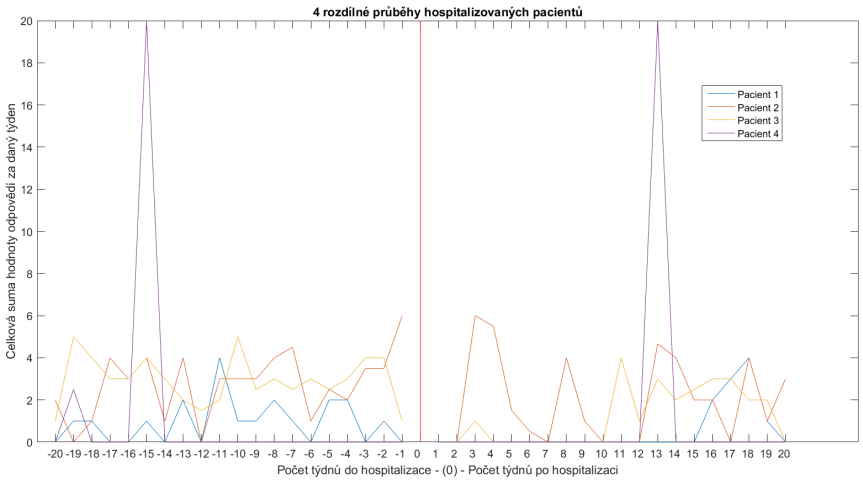
Obrázek 3.9. Procento poslaných nulových SMS u hospitalizovaných pacientů a rodinných příslušníků

Po prozkoumání dat o zaslaných SMS bylo vypozorováno, že mnoho pacientů vynechává posílání (ať už úmyslně, nebo například protože trpí negativními příznaky). Neposlané SMS působí problém při vyhodnocování, jelikož pro některé týdny záznamy chybí. Pro zefektivnění práce s daty, byly chybějící týdny nahrazeny hodnotami Not a Number (NaN), se kterými si lze později snadno poradit a využít je při procesu vyhodnocení. Nahrazení proběhlo tak, aby každý pacient měl pro každý týden účasti v programu ITAREPS záznam. Takto upravená data lze poté zobrazovat a zpracovávat v čase.

3.3 Trendy

Kromě statistických údajů je zajímavý i vývoj samotných hodnot SMS v čase. Práce se potýká s těžkou úlohou díky variabilitě dat a vysokému počtu nulových

SMS. Pro lepší představu, jak moc se jednotlivé průběhy mohou lišit, slouží obrázek níže.



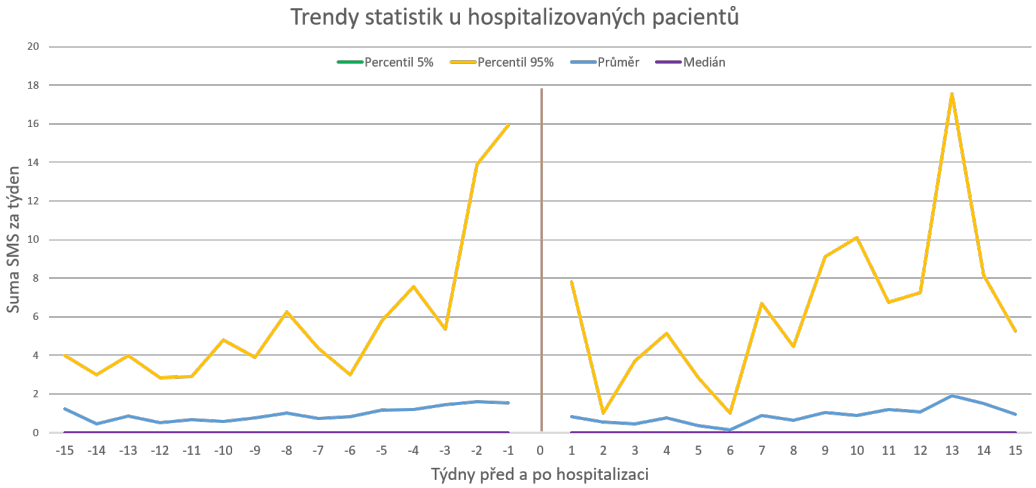
Obrázek 3.10. Ukázka vybraných průběhů. *Průběhy 4 velmi rozdílných pacientů. V čase 0 je znázorněn týden počátku hospitalizace, období samotné hospitalizace je vynecháno. Záporné časy značí týdny před hospitalizací, kladné časy pak počet týdnů po hospitalizaci.*

Obrázek 3.10 ukazuje hned několik zajímavých průběhů. Pacient číslo 1 posílá SMS poměrně poctivě. Před hospitalizací je pozorovatelný útlum v celkové sumě SMS. Po hospitalizaci pacient 15 týdnů neposílá vůbec nic. Podobné chování vykazuje i pacient číslo 3. Ovšem před samotnou hospitalizací je zaznamenáno mírné zvýšení celkové sumy. Pacient číslo 4 po celou dobu posílá nulové hodnoty, poté pošle několikrát v celkové sumě vysokou hodnotu. Pacient číslo 2 představuje kolísavý průběh, ale před hospitalizací je patrné postupné narůstání celkové sumy.

Data po doplnění obsahují záznamy pro každý týden. Nyní je možné jednotlivé pacienty vůči sobě porovnávat v čase, jak lze vidět na dvojici grafů 3.10 a 3.11. V čase 0 je znázorněn týden počátku hospitalizace, přičemž samotné období hospitalizace je zde vynecháno. Záporné časy značí počet týdnů před hospitalizací, kladné časy pak počet týdnů po hospitalizaci. Graf zobrazuje průměrnou sumu pro všechny poslané SMS pro všechny pacienty v daném týdnu, 95 % a 5 % percentil a medián. Hodnota mediánu je rovna 0 protože mnoho (alespoň polovina) pacientů posílá nulové hodnoty, tedy zřejmě nepociťují žádné zhoršení nebo je jejich hodnocení ovlivněno negativními příznaky onemocnění naznačenými v úvodu (neaktivní přístup a podobně).

Z doplněných dat byl vypočten průměrný průběh před a po hospitalizaci, viz obrázek 3.11. Průměrný průběh, sestavený z dat všech pacientů, má očekávané tendence. Na obrázku se jedná o modrou křivku. Průměrná SMS přes všechny pacienty pro daný týden dosahuje velmi malých hodnot, a to i v případě, že se blíží

hospitalizace (maximální hodnota průměrné SMS je rovna 2). V klidovém období jsou hodnoty nižší a před hospitalizací nastává mírné zvýšení, po léčbě naopak mírný pokles. Poté je opět pozorovatelné navýšení hodnot posílaných SMS. Velkou variabilitu posílaných SMS ukazuje 95 % percentil a mediánové hodnoty se drží na hodnotě nula díky velkému počtu nulových SMS.



Obrázek 3.11. Trendy statistik u hospitalizovaných pacientů. *V čase 0 je znázorněn týden počátku hospitalizace. Záporné časy značí počet týdnů před hospitalizací, kladné časy pak počet týdnů po hospitalizaci. Graf zobrazuje průměrnou sumu pro všechny poslané SMS pro všechny pacienty v daném týdnu, 95 % a 5 % percentil a medián*

Z obrázku 3.11 je patrné, že pacienti před nadcházejícím relapsem vykazují zhoršení příznaků. Po hospitalizaci, zřejmě vlivem léků a terapie, pacienti mají průměrně nižší hodnoty příznaků. Jakmile účinky terapie odezní, hodnoty příznaků opět narůstají.

3.4 Rozdělení na období

Na základě pozorování z předchozí kapitoly lze průběh hospitalizovaného pacienta rozdělit na několik typických období. Jednotlivá období lze vyjádřit následujícím intervalovým grafem.

Obrázek 3.12. Intervaly typických období hospitalizovaných pacientů

Jednotlivé intervaly značí klidové období před hospitalizací <-inf; -11>, přechodové období před hospitalizací <-10; -7>, kritické období <-6; -1>, období hospitalizace <0>, přechodové období po hospitalizaci <1; 6> a klidové období po hospitalizaci <7; inf>.

Pro týdny před kritickým obdobím existuje možnost, že k příznakům již dochází a data jsou výrazně zkreslená. Týdny po hospitalizaci bývají pacienti utlumení a nějakou dobu trvá, než se dostanou zpět do svého běžného režimu. Jelikož je pro další zpracování potřeba znát typické průběhy jednotlivých období, týdny <-10; -7> a <1; 6> jsou odstraněny, dále budou označeny jako přechodové.

3.5 Porovnání období

SMS odpovědi nenabývají normálního rozdělení, protože je v datech velké množšství nulových SMS. Jedná se o dotazníkové odpovědi, kde rozsah celkové sumy může nabývat hodnot nejméně 0 a nejvíce 40 v případě, že pošle na všech 10 otázek maximální možnou hodnotu. K porovnání posloužila průměrná hodnota sumy všech otázek, tentokrát pro různá období. Porovnání histogramů je provedeno pouze pro pacienty, jejichž rodina posílá SMS. Velmi mnoho pacientů v kritickém období začíná posílat nulové, či velmi malé ohodnocení.

Lze předpokládat, že klidové části budou mít podobné rozložení, jelikož byly odstraněny týdny přechodových období, které by mohly klidové průběhy zkreslit. Protože není jasné, jakého rozložení jednotlivá období nabývají, pro jejich porovnání vůči sobě byl použit dvouvýběrový Kolmogorov-Smirnov test (dále KS test) s hladinou významnosti $\alpha = 5\,\%$. Pro mnohočetná porovnání, je potřeba použít Bonferroniho korekci, kdy pro získání korektní p-hodnoty, je hladina významnosti α dělena počtem porovnání m. V tomto případě je počet zároveň provedených porovnání roven třem. $P - hodnota_b$ po korekci se rovná rovna 0,0167. Nulová hypotéza H0 pro KS test předpokládá, že data dvou výběrů patří do stejného rozložení. Zamítnutí nulové hypotézy H0 (nabývá hodnoty 1) nastává, kdy její $p - hodnota_{H0}$ je menší nebo rovna $p - hodnotě_b$.

PAT	Klidové období před hospitalizací	Kritické období
Klidová část před	-	-
Kritické období	$H = 1, p = 5,8575 \cdot 10^{-7}$	-
Klidové část po	$H = 0, p = 0,7821$	$H = 1, p = 8,4167 \cdot 10^{-5}$

Tabulka 3.2. Porovnání rozložení posílaných SMS pro tři období pro hospitalizované pacienty

Klidové části v porovnání s kritickým obdobím jsou daleko za hranicí signifikance a vyvrací hypotézu H0 o podobnosti jejich rozdělení. Naopak pro klidové části před a po hospitalizaci KS test potvrdil hypotézu H0 a pochází ze stejného rozdělení.

RODINA	Klidové období před hospitalizací	Kritické období
Klidová část před	-	-
Kritické období	$H = 1, p = 4,4558 \cdot 10^{-7}$	-
Klidové část po	$H = 1, p = 2,0034 \cdot 10^{-5}$	$H = 1, p = 8,8845 \cdot 10^{-7}$

Tabulka 3.3. Porovnání rozložení posílaných SMS pro tři období pro rodinné příslušníky hospitalizovaných pacientů

U rodinných příslušníků u hospitalizovaných pacientů jsou všechny hypotézy H0 zamítnuty a nepotvrzují podobnosti mezi jednotlivými obdobími.

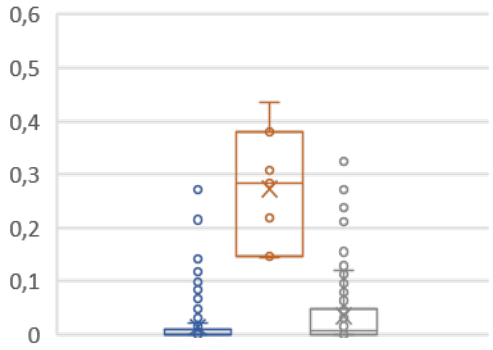
Na základě předchozích KS testů (3.2) byla sloučena období před a po hospitalizaci. Pro sloučená klidová období hospitalozovaných pacientů a klidová období nehospitalizovaných byly provedeny KS a Mann-Whitney U testy s hladinou významnosti α 5 %. Bonferroniho korekci v tomto případě není potřeba provádět, protože se jedná o dva různé testy.

$Test, alfa = 5\,\%$	Mann-Whitney U test	Kolmogorov-Smirnov test
Klidové období hospitalizovaného vs. klidové období nehospitalizovaného pacienta	$H = 1, P = 2,7298 \cdot 10^{-48}$	$H = 1, P = 3,4629 \cdot 10^{-44}$

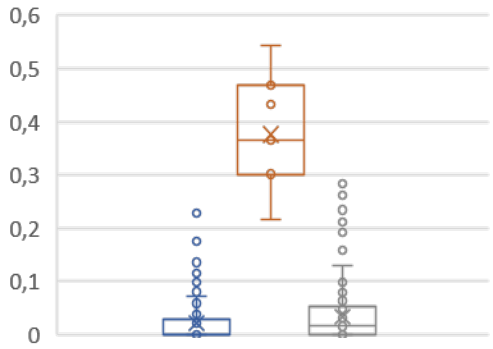
Tabulka 3.4. Porovnání klidových sloučených období hospitalizovaného pacienta a nehospitalizovaného pacienta

Mann-Whitney U test zamítá hypotézu H0, že sloučená klidová období u hospitalizovaného pacienta a klidová období nehostapitalizovaného pacienta pochází z rozdělení se stejnými mediány. Kolmogorov-Smirnov test zároveň vyvrátil, že pochází ze stejného rozdělení. Podle výsledků v tabulce 3.4 se klidová období hospitalizovaného a nehospitalizovaného pacienta významně liší.

Další metodou, jak prozkoumat povahu jednotlivých období a ověřit si výsledky Mann-Whitney U testu, jsou tzv. krabicové grafy se zobrazenými mediány.



Obrázek 3.13. Krabicové grafy pro hospitalizované pacienty



Obrázek 3.14. Krabicové grafy pro rodinné příslušníky hospitalizovaných pacientů

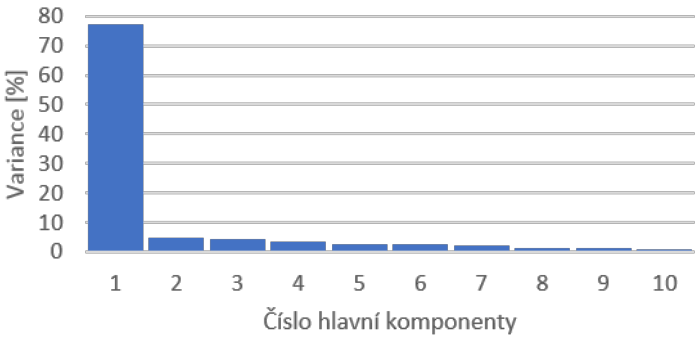
Testy pro sloučená klidová období a klidové období ukazují, že hospitalizovaný pacient během nerizikového období posílá SMS jiných hodnot než pacient, který hospitalizovaný nebyl. Na základě testů, histogramů, krabicových grafů byla klidová období sloučena a dále bylo s daty pracováno jako s celkovým klidovým obdobím. Naopak u rodinných příslušníků nevychází žádné podobnosti mezi jednotlivými obdobími, a proto se dále zaměříme na skupinu hospitalizovaných a nehospitalizovaných pacientů.

3.6 Vnitřní struktura dotazníku

V první části se tato kapitola zaměřuje na možné závislostí otázek EWSQ dotazníku, viz 1.2. Pro analýzu těchto závislostí je použita metoda Principal Component Analysis (PCA). Další část se věnuje otázce, zda na základě podobností odpovědí lze rozdělit pacienty do skupin. K tomu budou použity shlukovací metody K-means a hierarchické shlukování.

3.6.1 Závislost otázek mezi sebou

Zvolená metoda PCA ukazuje, jak moc jsou jednotlivé otázky v dotazníku na sobě závislé. Vstupem pro výpočet PCA byla matice 10 × počet odpovědí od každého pacienta pro danou otázku 1 až 10 (cca 300 000 záznamů). Pro vyhodnocení komponent a jejich variance byly odstraněny z dat NaN hodnoty a výpočtu byla předložena matice 10 × cca 233 000 záznamů. Pokud by každá komponenta přispívala rovnoměrně, variance pro každou komponentu by musela být 10 %.



Obrázek 3.15. Hlavní komponenty dotazníkových otázek pro kritické období hospitalizovaného pacienta. *První hlavní komponenta popisuje téměř 80 % variance všech dat.*

Pouze jedna komponenta přesahuje hranici 10 %, celkově pak zastává 78 % variance všech dat. Všechny otstaní komponenty jsou zastoupeny podstatně méne, a tak tabulka 3.5 ukazuje jednotlivé zastoupení otázek pouze pro hlavní komponentu první.

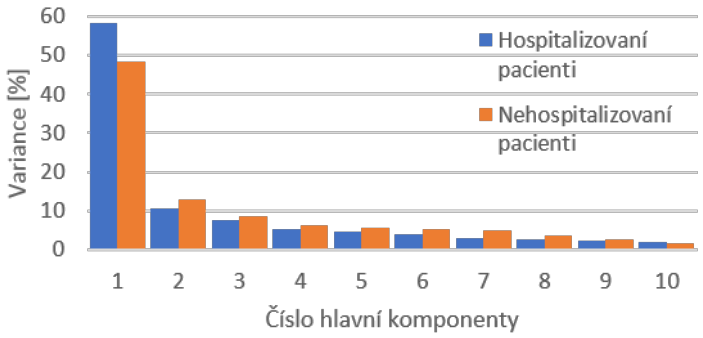
Popis otázky	Komponenta 1
Zhoršení spánku	0,260824
Zhoršení chuti k jídlu	0,324723
Zhoršení soustředění se	0,303701
Pociťování obav	0,35655
Zvýšený neklid	0,321623
Neporozumění dějům kolem	0,331083
Ztráta zájmu a energie	0,340977
Zhoršení řešení každodenních problémů	0,319094
Slyšení neexistujících hlasů	0,283614
Výskyt jiného varovného příznaku	0,309165

Tabulka 3.5. Loadingy pro hospitalizované kritické období

Z tabulky 3.5 je patrné, že jednotivé zastoupení dotazníkových otázek je rovnoměrné, a tedy jendotlivé otázky v kritickém období mají přibližně stejnou váhu.

Popis otázky	Komponenta 1		Komponenta 2		Komponenta 3	
	Hosp	Nehosp	Hosp	Nehosp	Hosp	Nehosp
Zhoršení spánku	0,3511	0,3205	-0,5888	0,4999	0,6896	0,6865
Zhoršení chuti k jídlu	0,1968	0,1754	0,0085	-0,004	-0,1051	-0,1118
Zhoršení soustředění se	0,248	0,3411	0,0089	-0,2407	-0,1074	-0,2078
Pociťování obav	0,3816	0,3618	-0,0555	-0,1683	-0,1319	-0,2181
Zvýšený neklid	0,2928	0,3689	0,0844	-0,178	-0,3077	-0,145
Neporozumění dějům kolem	0,3995	0,2367	-0,0248	0,0011	-0,1234	-0,2219
Ztráta zájmu a energie	0,3378	0,3848	0,1995	-0,2741	-0,0513	0,3263
Zhoršení řešení každodenních problémů	0,2576	0,3529	-0,026	-0,2946	-0,2527	0,2634
Slyšení neexistujících hlasů	0,2767	0,2392	0,7568	0,5533	0,5223	-0,3169
Výskyt jiného varovného příznaku	0,358	0,3114	-0,1699	0,4048	-0,1874	-0,2805

Tabulka 3.6. Loadingy pro klidová období hospitalizovaných a nehospitalizovaných pacientů



Obrázek 3.16. Hlavní komponenty dotazníkových otázek pro klidová období hospitalizovaných a nehospitalizovaných pacientů. *První hlavní komponenta pro obě skupiny popisuje více jako 5 % variance všech dat.*

Situace pro klidové období obou skupin pacientů je ale jiná, první dvě komponenty přesahují hranici 10 %. Nejsilnější komponenta pro hospitalizované pacienty vysvětluje necelých 60 % variance dat a pro nehospitalizované téměř 50. Třetí hlavní komponenta je těsně na hranici významnosti, proto se jí tabulka 3.6 bude též věnovat.

První komponenta pro obě skupiny obsahuje téměř rovnoměrné zastoupení všech otázek, s výjimkou otázky číslo 2. Druhá komponenta je sycena nejvíce otázkou číslo 1 a 9. U nehospitalizovaných má silné zastoupení i otázka číslo 10, skoro žádné otázky 2 a 6, naopak u hospitalizovaných otázky 2, 3, 5, 6, 8 nejsou prakticky zastoupeny. Třetí komponenta je sycena hlavně otázkou 1, u hospitalizovaných se pak silně projevuje i otázka 9, naopak otázka číslo 7 se neprojevuje skoro vůbec. U nehospitalizovaných jsou ostatní otázky zastoupeny podobně a

kromě otázky číslo 1 není žádná významnější. Otázka číslo 1 se objevuje pro obě skupiny ve druhé a třetí komponentě jako významně sycená. Otázka číslo 9 je pro hospitalizované zastoupena silně též pro 2 a 3 hlavní komponentu.

První hlavní komponenta jasně ukazuje silnou závislost jednotlivých otázek mezi sebou a jejich rovnoměrné zastoupení. V dalších komponenntách se vyskytují i další otázky, které jsou výrazněji zastoupeny, ale samotné komponenty vysvětlují málo variance dat.

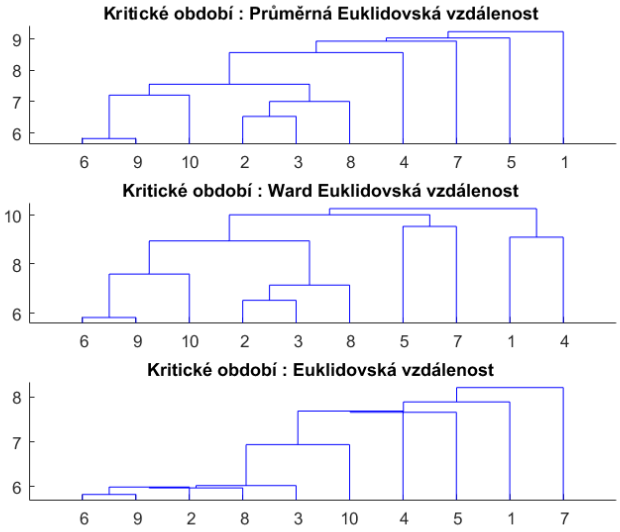
3.6.2 Podobnost otázek mezi sebou

Na základě odpovědí pacientů nachází shlukovací metody podobnosti a rozdělují pacienty do shluků.

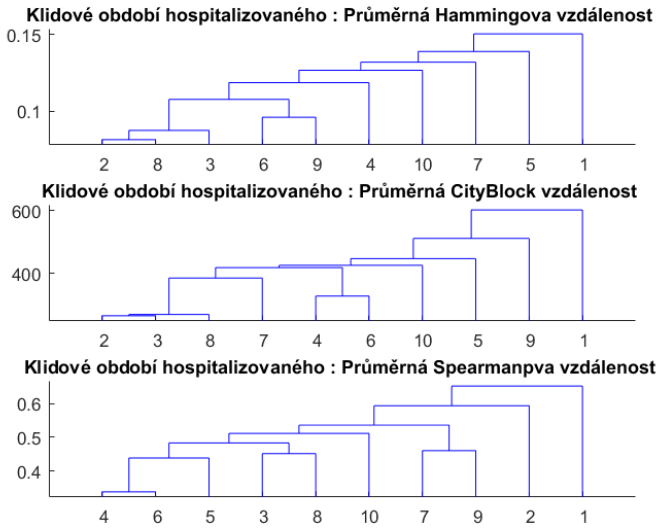
A) Hierarchické shlukování

- Euklidovská metrika, pouze jiné metody výpočtů vzdáleností

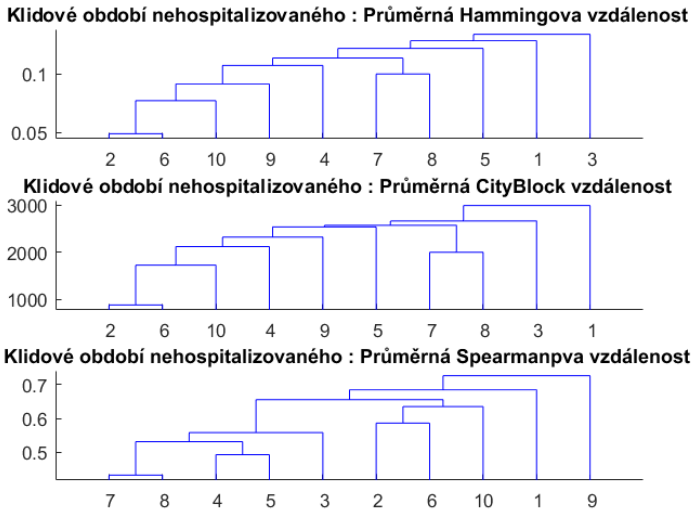
Euklidovská metrika patří mezi nejběžnější a pro ni budou vyzkoušeny různé metody výpočtů.



Obrázek 3.17. Pacienti hospitalizovaní, kritické období



Obrázek 3.18. Pacienti hospitalizovaní, klidové období

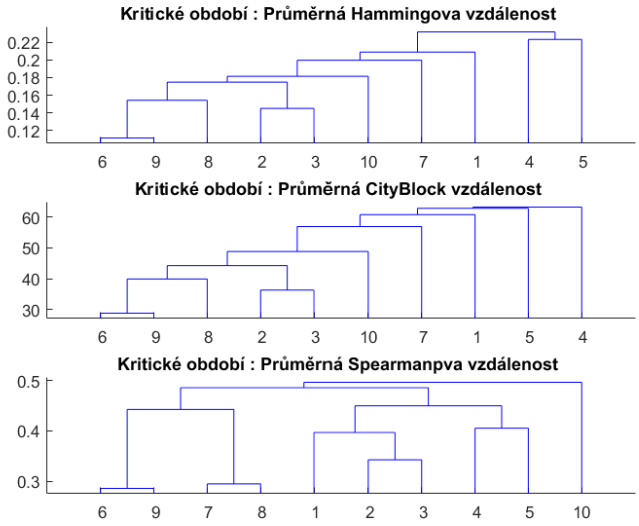


Obrázek 3.19. Pacienti nehospitalizovaní, klidové období

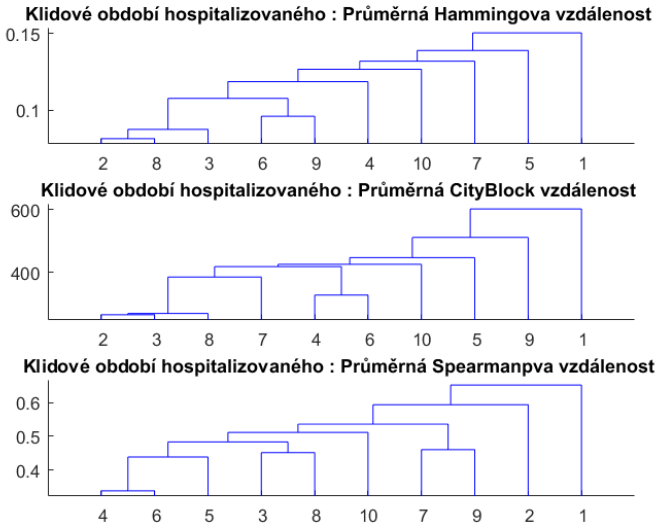
Na obrázku nahoře je patrné, že se vytváří rozdílné klastry pro období kritické a klidové (zde již sloučené). U pacientů pro klidové období vzniká jeden výraznější klastr otázek 2 a 8. Pro kritické období lze pozorovat shluky 7, 8 a 4 a 6. Vzhledem k celkovým vzdálenostem v grafech, tyto nejvýznamnější shluky nejsou až tak příliš výrazné, ale mohou býti zajímavé.

- Průměrovácí metoda výpočtu vzdáleností, jiné metriky

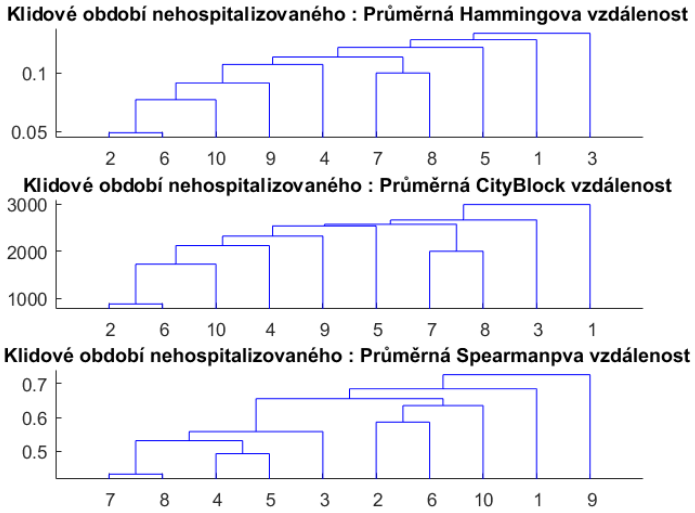
Průměrovací metoda výpočtu vzdáleností mezi shluky je používána nejčastěji. Nicméně v tomto případě byla použita spíše pro zajímavost, jelikož vstupní data jsou dotazníkového typu a obsahují mnoho nul.



Obrázek 3.20. Pacienti hospitalizovaní, kritické období



Obrázek 3.21. Pacienti hospitalizovaní, klidové období

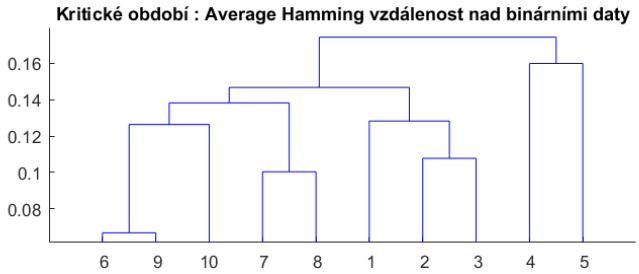


Obrázek 3.22. Pacienti nehospitalizovaní, klidové období

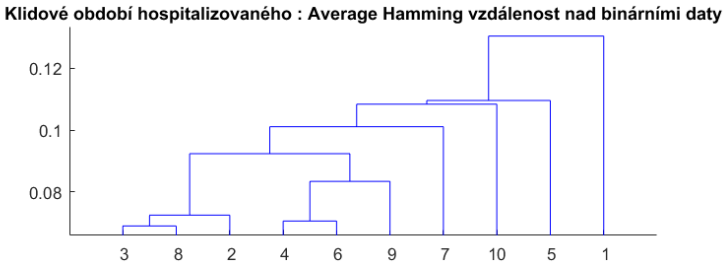
U pacientů se pro kritické období objevuje klastr 4, 6 jako nevýznamnější. Zdá se, že pro kritické období jsou shluky výraznější, a tak se projeví i za použití různých metrik.

- Průměrovácí metoda výpočtu vzdáleností, jiné metriky nad binárními daty

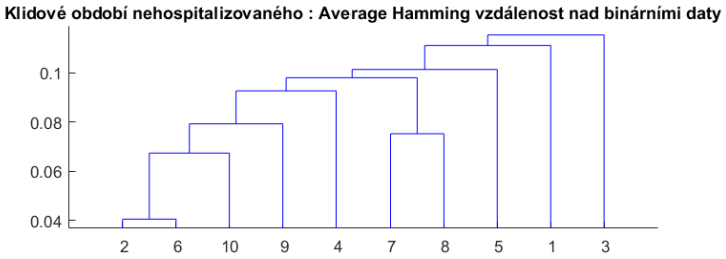
Vstupní data obsahují velký poměr nulových odpovědí, proto byla vytvořena binární data, která lze popsat pseudokódem ($(hodnota\ odpovědi > 0) = 1$). Tedy všechny nenulové odpovědi nabývají hodnoty 1 a nulové odpovědi zůstávají nulovými. Výsledná data obsahují pouze příznaky, zda byla otázka kladně hodnocena (zhoršení stavu), ale již nezohledňuje váhy jednotlivých odpovědí.



Obrázek 3.23. Dendrogram nad binárními daty pro odpovědi pacientů, kritické období



Obrázek 3.24. Dendrogram nad binárními daty pro odpovědi pacientů, kritické období u hospitalizovaných pacientů

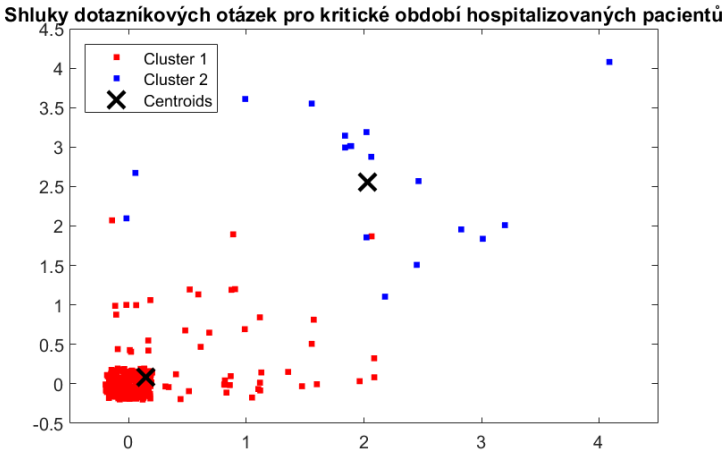


Obrázek 3.25. Dendrogram nad binárními daty pro odpovědi pacientů, kritické období u nehospitalizovaných pacientů

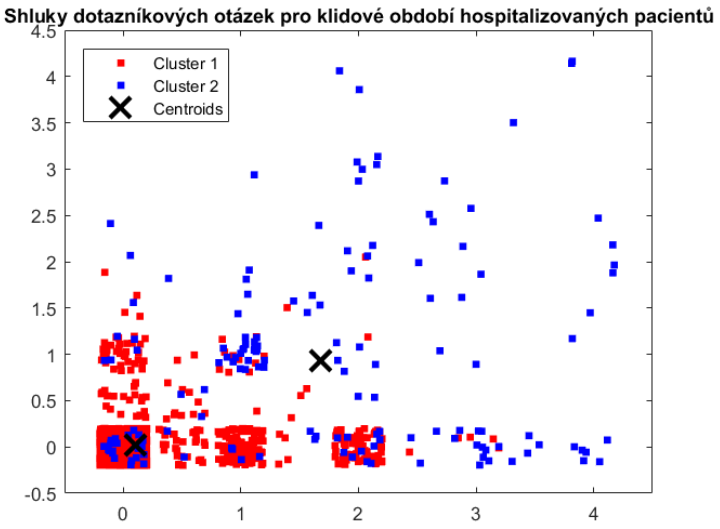
Vzhledem k vzdálenostem jednotlivých klastrů a nesourodosti otázek, které spadají do klastrů, lze říci, že mezi otázkami nejsou žádné výrazné klastry. Nejspíše je to způsobeno tím, že vzniká podobnost mezi jednotlivými otázkami i na základě toho, že pacienti posílají hodně nulových hodnot pro jednotlivé otázky. Z obrázku je ale patrné, že nalezené skupiny podobných otázek se ve většině neshodují s předchozími postupy.

B) Nehierarchické K-means

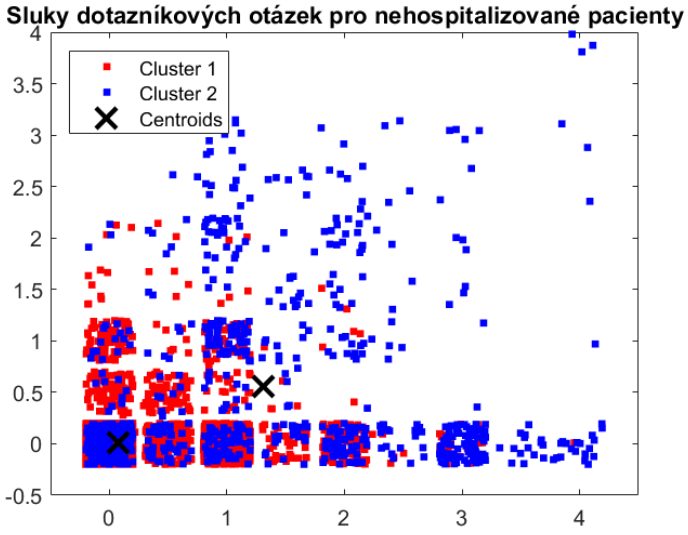
Ani metoda K-means nenachází klastry, kvůli velkému množství nulových odpovědí většina otázek spadá do jednoho klastru i za použití různých metrik.



Obrázek 3.26. Shluky dotazníkových otázek pro kritické období hospitalizovaných pacientů



Obrázek 3.27. Shluky dotazníkových otázek pro klidové období hospitalizovaných pacientů



Obrázek 3.28. Shluky dotazníkových otázek pro nehospitalizované pacienty

Obrázky výše ukazují, že vždy jeden shluk odděluje pacienty s mnoha nulovými odpověďmi, první shluk se vždy umístil v nulových hodnotách. U hospitalizovaných pacientů v kritickém období lze pozorovat, že jednotlivé shluky se téměř nepřekrývají, zatímco v klidovém se už objevují překryvy obou shluků. Střed shluku 2 klesl, zřejmě proto, že pacienti posílají více nulových SMS, což je v klidovém období očekávané. Pro nehospitalizované pacienty je vidět velký překryv obou shluků a střed shluku 2 je ještě níže než pro hospitalizované pacienty v klidovém období. To je též v souladu s tím, že celkový stav pacientů se tolik nezhoršuje.

Kapitola 4
Analýza pacientů dle symptomů

Každý pacient by měl být při vstupu do programu ohodnocen na základě symptomů, které vykazuje. Symptomy diagnostikuje lékař a jsou uloženy ve vstupních datech. Tyto symptomy pomáhají doktorům určit, kterým typem schizofrenie pacient trpí. Kapitola se zaměří na souvislost symptomů mezi sebou, na souvislost s otázkami (například posílání velkého množství nul či žádných hodnot může znamenat útlum, nezájem). Díky tomu lze posuzovat pacienty a shlukovat je i podle jednotlivých symptomů, které u nich byly pozorovány.

Podle typu symptomů lze předznamenávat konkrétní chování pacientů, například jak jsou pacienti aktivní v programu ITAREPS (posílají/neposílají SMS). Závažnost symptomu pak může poukazovat například na vyšší riziko hospitalizací a spolu s tím lze očekávat vyšší průměrnou sumu SMS.

Symptomy jsou uloženy jako vektor s následujícími souřadnicemi:

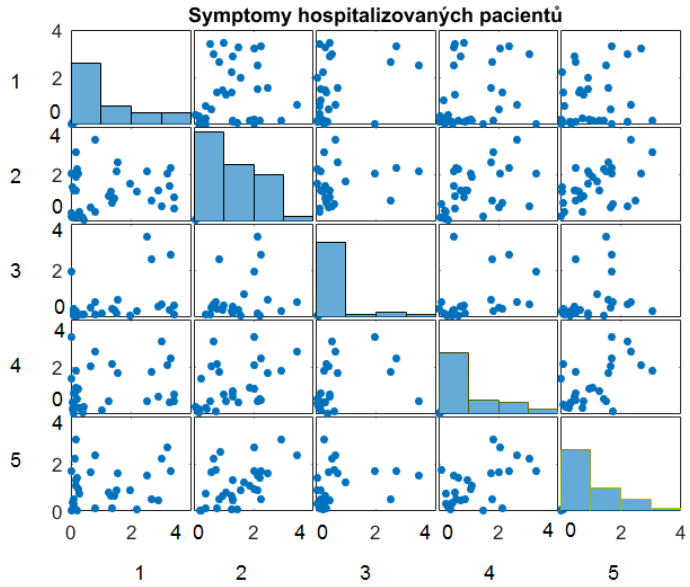
1. Pozitivní symptomy (bludy, halucinace).
2. Negativní symptomy (oploštěná emotivita, pasivita, chudost řeči, sociální stažení, absence nonverbální komunikace (gesta, mimika)).
3. Dezorganizační symptomy (dezorganizovaná řeč, dezorganizované chování).
4. Afektivní symptomy (přítomnost deprese a/nebo mánie).
5. Kognitivní symptomy (narušení kognitivních funkcí, které se promítá do snížení profesní a sociální kompetence pacienta).

Souřadnice vektoru nabývají hodnot 0-4. Jejich popis v tabulce 4.1

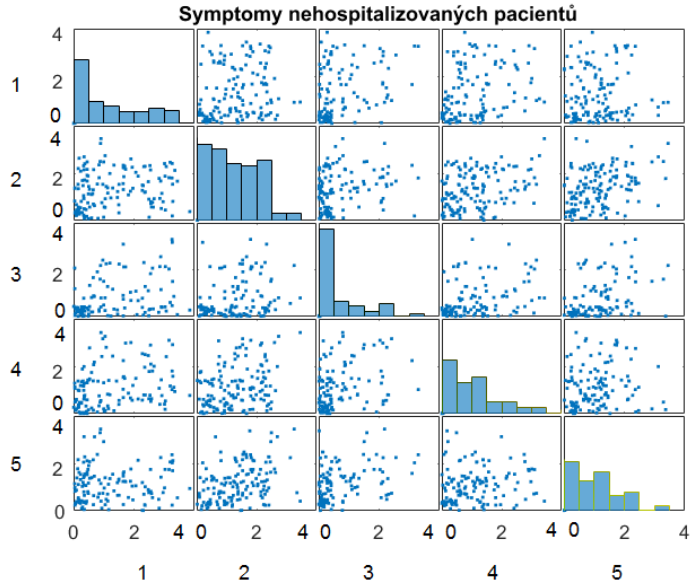
Škála	Význam
0	Nepřítomny
1	Slabě vyjádřeny
2	Středně vyjádřeny
3	Výrazně vyjádřeny
4	Zcela dominují

Tabulka 4.1. Škála symptomů

Krátký pohled na symptomy a jejich škálu umožňuje obrázek 4.1, kde je porovnána osa x, představena číslem symptomu a osa y představena závažností symptomu. Histogram ukazuje rozložení hodnot závažností pro daný symptom.



Obrázek 4.1. Porovnání symptonů u hospitalizovaných vůči sobě. *Obrázek znázorňuje porovnání jednotlivých symptomů vůči závažnosti.*



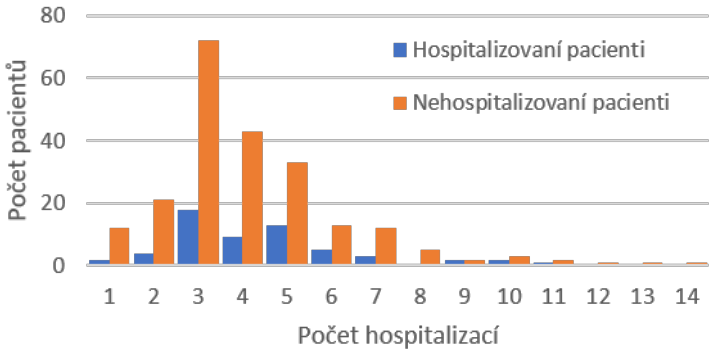
Obrázek 4.2. Porovnání symptonů u hospitalizovaných vůči sobě. *Obrázek znázorňuje porovnání jednotlivých symptomů vůči závažnosti. Na diagonální ose je zobran histogram hodnot pro konkrétní symptom.*

Obrázky ukazují, že mezi jednotlivými symptomy neexistuje lineární závislost, a proto se kapitola zaměří na další metody, jak symptomy prozkoumat a na jejich základě rozdělit pacienty.

4.1 Základní přehledy

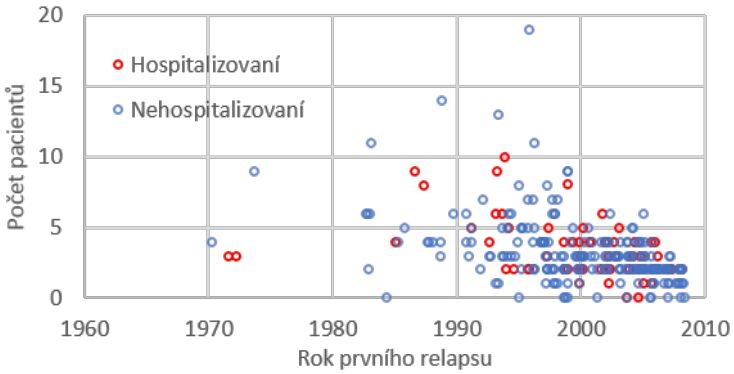
Vzhledem k tomu, že pacient trpí nemocí, nejspíše se u něj projevují nějaké symptomy zmíněné výše. Pakliže pacient nemá symptomy vůbec v systému zaznamenány, pravděpodobně je ošetřující lékař neuvedl do zprávy. Mezi pacienty existuje poměrně velká skupina, která má všechny symptomy nulové. To znamená, že u daného pacienta ještě symptomy nepropukly, nebo mohly být chybně zpracovány neexistující záznamy. Skupina čítá celkem 72 pacientů - 22 z 59 hospitalizovaných během programu ITAREPS a 50 z 221 nehospitalizovaných. Některé přehledy se budou zabývat pacienty jak s vyjádřenými, tak s nevyjádřenými symptomy.

Zajímavý může být i počet hospitalizací ještě před vstupem do samotného programu ITAREPS. Pro doktory je také užitečná informace, kdy byl pacient vůbec poprvé hospitalizován, protože často teprve poté je stanovena diagnóza schizofrenie.



Obrázek 4.3. Počet pacientů a počet hospitalizací před vstupem do programu ITAREPS . *Histogram zde není normovaný a uvádí skutečné, ne relativní, počty pacientů.*

Obrázek níže ukazuje skutečné počty pacientů a počet hospitalizací. Většina pacientů prošla 0 až 8 hospitalizacemi. Vyskytují se též pacienti s až 19 hospitalizacemi. Nejčastěji hospitalizovaný pacient měl první kolaps v roce 1996 a poslal 100 % všech SMS a 64 % z nich obsahovalo nuly, tedy jeho stav se nezhoršoval. Pacient, který měl pouze 2 hospitalizace do ITAREPS měl první kolaps v roce 2006, posílal 90 % SMS a 46 % z toho obsahovalo nuly. Dle dalších obrázků ale nelze říci, že čím déle u pacienta pozorujeme symptomy, tím vícekrát byl hospitalizován. Mohl mít včasnou detekci od lékaře a včasné nasazení léků, nebo trpí mírnějším typem schizofrenie.



Obrázek 4.4. Počet hospitalizací před ITAREPS od prvního výskytu symptomu. *Hospitalizovaní pacienti (nahoře) se setkali s prvními symptomy průměrně v roce 1991, nehospitalizovaní pacienti (dole) až v roce 1999.*

U nehospitalizovaných pacientů je korelační koeficient mezi rokem prvního výskytu a počtem hospitalizací -0,4380. U hospitalizovaných korelační s hodntou -0,3816 koeficient také ukazuje na střední až podstatnou závislost. Je celkem očekávané, že čím starší pacient je, tím je pravděpodobnější, že prošel více relapsy než pacient mladší.

U nehospitalizovaných pacientů během programu ITAREPS se první symptomy vyskytly průměrně později, a to kolem roku 1999, zatímco u hospitalizovaných to bylo kolem roku 1991. Průměrný věk narození pacienta se v obou skupinách liší pouze o půl roku a pohybuje se přibližně kolem roku 1975. Hospitalizovaným pacientům v průběhu ITAREPS bylo průměrně 16 let, když nastaly první symptomy, zatímco nehospitalizovaným bylo 24. Schizofrenie se nejčastěji projeví mezi 15. a 35. rokem života, alespoň polovina všech nemocných je detekována již před 25. rokem [12].

Zajímavým faktem je, že rozdělení pohlaví je u nehospitalizovaných pacientů rovnoměrně vyváženo. U hospitalizovaných je poměr pohlaví 1:2 (16 žen, 33 mužů).

Skupina	Průměrná závažnost symptomu (nenulové symptomy)	Průměrný počet hospitalizací před ITAREPS	Průměrná SMS
Hospitalizovaní	5,22	3,49	1,4604
Nehospitalizovaní	4,79	3,20	0,9249

Tabulka 4.2. Základní přehledy o průměrných hodnotách na jednoho pacienta dané skupiny

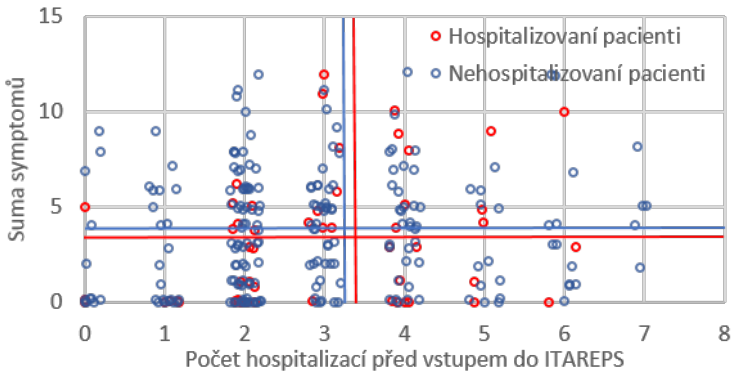
Lze očekávat, že u pacientů, kteří byli hospitalizovaní během ITAREPS, se některé syndromy projevovaly výrazněji než u pacientů, u kterých hospitalizace nenastala. Podobným číslům napovídá i průměrná hodnota odpovědí v zasílaných SMS. Hospitalizovaní pacienti posílají v průměru vyšší hodnoty. Průměrný počet hospitalizací před vstupem do programu je pro obě skupiny podobný.

4.2 Vztah mezi závažností symptomů a počtem hospitalizací

Čím závažnější symptom je, tím nabývá vyšší hodnoty na škále od 0 do 4. Kapitola 4.2 se zaměřuje na celkovou sumu symptomů a počet hospitalizací před a během programu ITAREPS, například zda byli pacienti s vysokou celkovou sumou hospitalizováni častěji a naopak pacienti s nižší celkovou sumou symptomů hospitalizováni méně často.

4.2.1 Počet hospitalizací před programem ITAREPS

Jak je uvedeno v přehledové tabulce 4.2 průměrná suma symptomů pro hospitalizované je 5,22, pro nehospitalizované 4,79. Průměrný počet hospitalizací před ITAREPS pro hospitalizované během programu ITAREPS je 3,49, pro nehospitalizované během programu 3,2. Průměrné hodnoty jednotlivých skupin se od sebe příliš neliší.

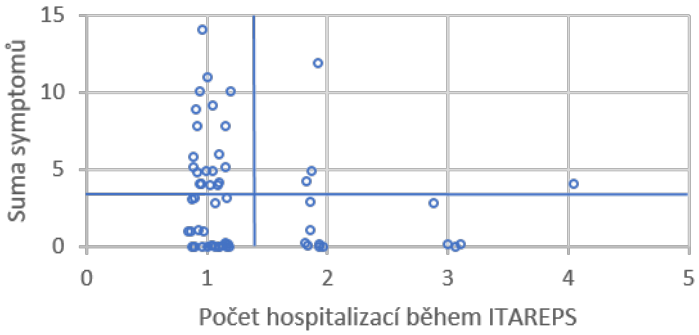


Obrázek 4.5. Závislost sumy symptomů na počtu hospitalizací. *Vodorovná čára znázorňuje průměr sumy symptomů a svislá průměrný počet hospitalizací před vstupem do ITAREPS.*

V datech v grafu 4.5 není na první pohled patrná závislost mezi symptomy a počty hospitalizací před ITAREPS. To potvrzují korelační koeficienty, pro hospitalizovaného pacienta 0,1915 a pro nehospitalizovaného 0,0479. V prvním případě se jedná pouze o nízkou až střední závislost, avšak ve druhém případě závislost skoro žádná neexistuje.

4.2.2 Počet hospitalizací během programu ITAREPS

Počet hospitalizací během ITAREPS se týká pouze pacientů, kteří byli hospitalizování po nastoupení do programu ITAREPS.



Obrázek 4.6. Závislost sumy symptomů na počtu hospitalizací během ITAREPS. *Vodorovná čára znázorňuje průměr sumy symptomů a svislá průměrný počet hospitalizací během programu ITAREPS.*

Obrázek opět ukazuje, že mezi celkovou sumou symptomů a počtem hospitalizací během ITAREPS závislost neexistuje. To potvrzuje i hodnota korelačního koeficientu -0,1652. To odpovídá korelaci nízké až střední, tedy téměř zanedbatelné.

4.3 Vztah mezi posílanými SMS a symptomy

Mezi celkovými sumami a počtem hospitalizací není významná závislost. Proto se tato část zaměřuje na jednotlivé závažnosti jednotlivých symptomů. Jak již bylo zmíněno v kapitole 1, lze očekávat, že některé symptomy, například negativní, mohou ovlivnit pacientovu aktivitu.

Symptom/závažnost symptomu	0	1	2	3	4
Pozitivní					
Hospitalizovaní	0,8514	0,1683	0,0973	0,3378	0
Nehospitalizovaní	0,3903	0,2070	0,1314	0,1926	0,0036
Negativní					
Hospitalizovaní	0,5929	0,5852	0,1653	0,1114	0
Nehospitalizovaní	0,3004	0,3728	0,2131	0,0386	0
Dezorganizační					
Hospitalizovaní	566858	0,0752	0,0180	0,0053	0,0109
Nehospitalizovaní	0,7193	0,0823	0,1099	0,0134	0
Afektní Hospitalizovaní					
Nehospitalizovaní	0,7049	0,5303	0,1528	0,0547	0,0119
	0,2828	0,4343	0,1794	0,0282	0,0002
Kognitivní					
Hospitalizovaní	0,9353	0,3037	0,1249	0,0908	0
Nehospitalizovaní	0,3405	0,4809	0,0668	0,0367	0

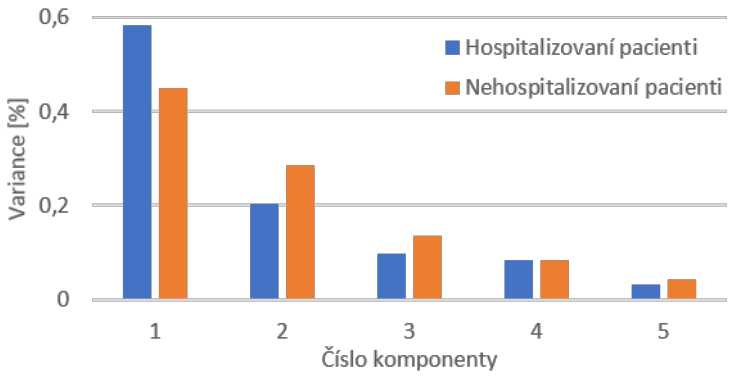
Tabulka 4.3. Průměrná SMS na pacienta ve skupině pro symptom a závažnost symptomu pro všechny pacienty, hospitalizované a nehospitalizované. *Průměrná SMS pro všechny pro 0,2703, pro hospitalizované 0,2909, nehospitalizované 0,1850.*

Pro symptomy, které nejsou vůbec vyjádřeny, tedy doktor je ohodnotil závažností 0 – nepřítomny, posílají hospitalizovaní pacienti průměrně vyšší hodnoty oproti pacientům, kteří hospitalizovaní nebyli. Z tabulky je zřetelná klesající tendence průměrných hodnot s rosoutcí závažností. Symptomy závažnosti 3. a 4. stupně prakticky nejsou zastoupeny.

4.4 Závislost symptomů

Stejně jako v kapitole 3.6.1 byla zkoumána závislost posílaných SMS pomocí metody PCA, zde je metoda aplikována ke zjištění závislosti symptomů.

Dle výsledků PCA na obrázku 4.7 je patrné, že v případě zahrnutí pacientů s nulovým vektorem symptomů je rozdíl mezi první hlavní a dalšími komponentami mnohem výraznější, než u pacientů s alespoň jedním vyjádřeným symptomem.

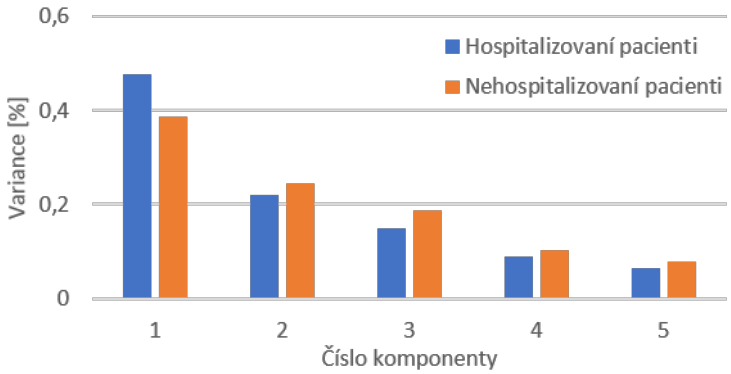


Obrázek 4.7. Analýza hlavních komponent symptomů pro všechny pacienty. *První hlavní komponenta u hospitalizovaných pacientů vysvětluje necelých 60 % a nehospitalizovaných necelých 50 % variance. Druhá komponenta pro obě skupiny popisuje 20 % variance všech dat.*

Obě skupiny pacientů jsou vyjádřeny významně prvními dvěmi hlavními komponentami, u nehospitalizovaných je navíc ještě výraznější i komponenta třetí. Jednotlivé zastoupení symptomů v komponentách popisuje tabulka

Symptom	Komponenta 1		Komponenta 2		Komponenta 3	
	Hosp	Nehosp	Hosp	Nehosp	Hosp	Nehosp
Pozitivní	0,4922	0,6153	0,8106	-0,5848	0,0647	-0,4885
Negativní	0,4339	0,4165	-0,2266	0,5269	-0,4087	-0,2178
Dezorganizační	0,389	0,3274	0,0782	0,0805	-0,3574	0,009
Afektní	0,4839	0,4907	-0,2825	-0,1494	0,8017	0,8413
Kognitivní	0,4286	0,3159	-0,4533	0,5929	-0,2413	-0,0774

Tabulka 4.4. První tři hlavní komponenty pro všechny pacienty



Obrázek 4.8. Analýza hlavních komponent symptomů pro pacienty s vyjádřenými symptomy. *První hlavní komponenta u hospitalizovaných pacientů vysvětluje necelých 45 % a nehospitalizovaných necelých 40 % variance. Druhá komponenta pro obě skupinu hospitalizovaných popisuje 30 %, u nehospitalizovaných 25 % variance všech dat.*

Pro pacienty s pouze vyjádřenými symptomy, tedy alespoň jeden symptom je ozančen závažností 1 je variance první komponenty nižší. Pro druhou a třetí komponentu ale platí opak. Ty popisují mnohem více variance, než je tomu u pacientů i s nevyjádřenými symptomy. Jednotlivé zastoupení symptomů v komponentách popisuje tabulka 4.4 a 4.5

První hlavní komponenta u hospitalizovaných pacientů má zastoupeny všechny symptomy téměř rovnoměrně. Velmi podobně je tomu i u nehospitalizovaných. Poukazuje to na celkovou závažnost všech symptomů, nicméně u nehospitalizovaných jsou pozitivní symptomy zastoupeny silněji, než všechny ostatní.

Druhá hlavní komponenta pro obě skupiny vychází obdobně. U nehospitalizovaných pacientů je patrné, že je sycena velmi významně hned 3 symptomy – pozitivní, negativní a kognitivní. Pro hospitalizované to jsou pouze pozitivní a kognitivní. Pro obě skupiny je zastoupení dezorganizačních symptomů zanedbatelné.

Třetí hlavní komponenta je zastoupena ve velké převaze afektivním symptomem.

Symptom	Komponenta 1		Komponenta 2		Komponenta 3	
	Hosp	Nehosp	Hosp	Nehosp	Hosp	Nehosp
Pozitivní	0,5004	0,7802	0,7799	-0,3181	0,0680	-0,4883
Negativní	0,2800	0,1319	-0,2310	0,5831	-0,3973	-0,2207
Dezorganizační	0,5596	0,3590	0,0089	0,3276	-0,3756	0,0042
Afektní	0,4720	0,4917	-0,3244	-0,0321	0,7983	0,8403
Kognitivní	0,3676	0,0564	0,4827	0,6712	-0,2433	-0,0819

Tabulka 4.5. První tři komponenty pouze pro pacienty s vyjádřenými symptomy.

Pro pacienty s pouze vyjádřenými symptomy nejsou jednotlivé komponenty vnitřně tak sourodě zastoupeny jako pro všechny pacinety i s nulovými symptomy. Hlavní komponenta taktéž nemá rovnoměrné zastoupení jednotlivých symptomů. Pro hospitalizované první komponenta je nejvíce sycena pozitivními a dezorganizačními symptomy, druhá kompoentna zejména pozitivními a třetí komponentna afektními symptomy. U hospitalizovaných první komponenta je nejvíce zastoupena pozitivními symptomy, druhá hlavně kognitivními a negativními. Třetí komponenta se shoduje a silné zastoupení je také afektiními symptomy.

4.5 Vzájemná podobnost symptomů

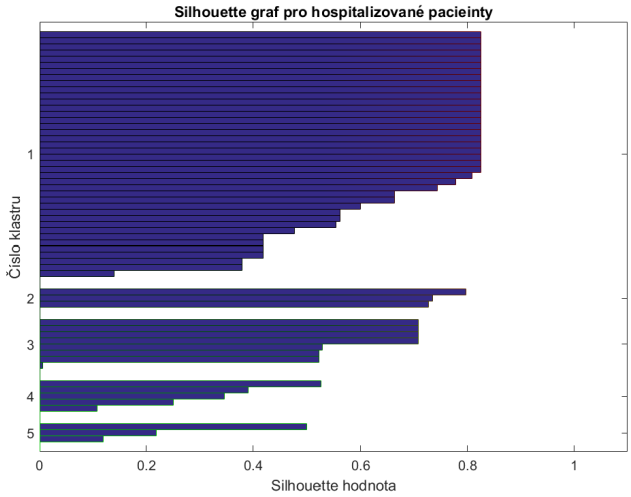
K nalezení podobných skupin pacientů na základě jejich symptomů byl použit algoritmus k-means. Nejprve je uveden přehled všech pacientů v tabulce, kde optimální počet shluků je dle silhouette koeficientů 2. Poté je algoritmus aplikován na skupinu hospitalizovaných a nehospitalizovaných pacientů. Graficky je zastoupení pacientů v jednotlivých shlucích znázorněno v silhoutte grafu 4.9 a 4.10.

	1. Shluk	2. Shluk
Celkem počet pacientů	99	181
Počet hospitalizovaných pacientů	19	41
Průměrný vektor symptomů	[2,0000 1,4949 0,8788 1,6162 1,101]	[0,1768 0,6243 0,0608 0,3149 0,5138]
Průměrná SMS	474452	0,8834
% poslaných SMS	76,7619	78,9213
% nulových SMS	67,0164	70,4709

Tabulka 4.6. Vyhodnocení 2 shluků nad všemi pacienty. *Výsledky ukazují, že průměrné hodnoty symptomů se v jednotlivých shlucích významně liší.*

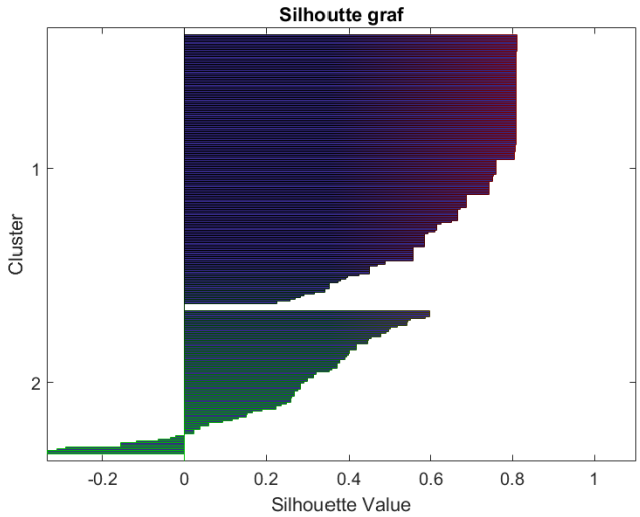
První shluk je zhruba 2x méně početnější než shluk druhý. Tento poměr platí i pro počet hospitalizovaných v jednotlivých shlucích. Přestože je poměr nehospita-

lizovaných a hospitalizovaných v obou shlucích podobný, průměrné symptomy se podstatně liší. Pro minoritní shluk jsou symptomy průměrně vyšší než pro majoritní shluk, zřejmě proto, že pohltil většinu pacientů s nulovými symptomy. Tomu také napovídá i hodnota průměrně SMS, kde pacienti s nevyjádřenými symptomy mohou být stabilnější. Pacienti v obou shlucích jsou podobně aktivní a i % nulových SMS se téměř shoduje.



Obrázek 4.9. Silhoutte graf pro 5 shluků dle symptomů hospitalizovaných pacientů. *Průměrné hodnoty koeficientů pro každý shluk 0,9203 0,1958 0,6699 0,3996 0,4337, celková průměrná pak 0,5239.*

V dalších bězích se průměrná nejvyšší hodnota pohybovala okolo 0,6 pro právě 5 shluků pro hospitalizované pacienty během ITAREPS. Pro nehospitalizované pacienty a jejich symptomy se jako optimální počet shluků ukázal být počet dva. Nejvyšší průměrná hodnota pro tyto shluky byla mírně pod 0,5. Hospitalizovaní pacienti vstupují do programu s více různorodými symptomy.



Obrázek 4.10. Silhoutte graf pro 2 shluky dle symptomů nehospitalizovaných pacientů. *Průměrné hodnoty koeficientů pro každý 0,6860, 0,2606 celková průměrná pak 0,4733.*

Pro zpracování přehledů o jednotlivých shlucích byly odstraněni pacienti s odlehlými pozorování pro průměrnou SMS, celkem se jedná o 4 pacienty.

Legenda:	Shluk 1	Shluk 2	Shluk 3	Shluk 4	Shluk 5
Počet pacientů,					
Průměrná suma SMS,					
počet hospitalizací před ITAREPS,					
% poslaných SMS,					
% nulových SMS					
Průměrně symptom					
Hospitalizovaní	28	13	7	4	6
	1,0619	1,0174	0,855	0,947	0,7583
	3,3	3,4	3,9	2,4	3,3
	79,4 %	68,8 %	61,5 %	68,9 %	81,1 %
	64,6 %	61,3 %	54,4 %	64,3 %	77,0 %
	1,0571	0,9846	0,7714	0,9200	0,9667
Nehospitalizovaní	142	76	-	-	-
	39814	0,9666			
	3,4	3,2			
	78,7 %	78,4 %			
	68,2 %	69,6 %			
	59,86 %	76,88 %			

Tabulka 4.7. Přehledové informace o shlucích

- Hospitalizovaní

Pacienti v shluku číslo 5 jsou nejvíc aktivní v posílání SMS, ale díky nejvyššímu procentu poslaných nulových SMS mají nejmenší průměrnou celkovou SMS. Pacienti ve shluku číslo 4 mají nejnižší počet hospitalizací. Shluk číslo 3 je složen z pacientů, kteří posílají nejméně nulových SMS, ale zároveň i nejméně aktivní v posílání SMS a jejich průměrná celková SMS je nízká. Zajímavé ale je, že mají nejvyšší počet hospitalizací. Pacienti z nejpočetnějších shluk§ 1 a 2 mají podobné charakteristiky, ale procento poslaných SMS se výrazně liší.

- Nehospitalizovaní

Pacienti v obou shlucích mají podobné charakteristiky, průměrná SMS se mírně liší. Zajímavým faktem ale je, že shluk číslo 2 je 2x početnější než shluk číslo 1. Pacienti v početnějším shluku posílají průměrně vyšší SMS.

Kapitola 5
Současný klasifikátor

Pro návrh nového klasifikátoru je třeba více prozkoumat stávající systém. Systém se dívá na sumu aktuálně poslané SMS pacienta daný týden a na sumu otázek 4, 6, 9. Celkovou sumu SMS porovnává s prahem hodnoty 8 a sumu otázek s prahem hodnoty 3. Jestliže suma překročí jeden z prahů, jedná se o primární alert. Dále se dívá na celou předchozí SMS a její sumu. Jestliže suma současné a minulé SMS překročí práh velikosti 12, pak se také jedná o alert, sekundární. Nastane-li alert primární, či sekundární, je vysláno upozornění o možném zhoršení pacienta. Pacientova SMS, je systémem zařazena do třídy podezřelých pro hospitalizaci, tudíž pacient je pro daný okamžik také považován jako podezřelý pro hospitalizaci. Závažnost onemocnění je diagnostikována lékařem.

V datech dostupných pro tuto práci vyvstalo 1244 alertů pro celkovou sumu s prahem 8, pro sumu sloupců s prahem 3 jich vyvstalo 830. Alertů pro sumu dvou po sobě jdoucích SMS s prahem 12 se vyskytlo pouze 298. Celkem nastalo 1631 primárních a sekundárních upozornění.

5.1 Vyhodnocení systému

Struktura dat obsahuje datové záznamy s datem, kdy nastal alert. Tato práce se obecně zabývá tím, jak vypadají data pro hospitalizované a nehospitalizované pacienty, je tedy žádoucí nalézt alerty, které skutečně předcházely hospitalizacím. Struktura dat ale také obsahuje datové záznamy s datem podání medikace. Nelze spoléhat na to, že všechny aplikované medikace byly zpětně zaznamenány. Je ale známo, že je-li podána včas dávka medikace, lze relapsu předejít, proto se vyhodnocení zaměří i na podanou medikaci.

Vyhodnocení pro hospitalizované a nehospitalizované pacienty je spočteno zvlášť, zejména kvůli vyhodnocení úspěšnosti alertů před hospitalizací. Úspěšnost systému udávají statistické metriky popsané v použitých metodách.

Úspěšnost stávajícího systému pro hospitalizované pacienty je hodnocena dle toho, kdy alerty nastávají v časové ose vzhledem k hospitalizaci, viz kapitola 3, a kdy byla podána medikace po vyvstání alertu.

Nastal-li alert v kritickém období před hospitalizací, či byla-li podána po alertu medikace do 1 týdne, je alert označen za kladně pozitivní (TP). Alerty kdy nebyla podána následně medikace a nenastala hospitalizace během 6 týdnů jsou falešně pozitivní (FP).

Nenastal-li alert v kritickém období, jedná se opominutí a falešně negativní (FN) upozornění, jelikož alert nastat měl. V případě, že alert nenastal v klidovém období a nebyla podána medikace, pak se jedná o chtěné chování systému a kladně negativní upozornění (TN).

V přechodových obdobích se jedná o alert, který nelze s jistotou zařadit do výše zmíněných tříd, viz kapitola 3. Přechodové období před hospitalizací je zařazeno do třídy přechod 1 a po hospitalizaci přechod 2.

Pro nehospitalizované pacienty je vyhodnocení snazší. Celé jejich období je chápáno jako klidové a vyhodnocuje se pouze, zda byla podána medikace po alertu do 1 týdne. Pro určení přehlédnutých možných hospitalizací slouží právě data o podané medikaci. Nastal-li alert a byla-li následně podána medikace, pak se jedná o kladně pozitivní (TP) třídu. V opačném případě, pokud medikace podána nebyla, alert spadá do třídy falešně pozitivních (FP). Pokud pacientovi byla podána medikace, ale nepředcházel jí alert, jde o přehlédnuté riziko, které spadá do třídy falešně negativních (FN). Ostatní týdny, kdy alert nenastal a medikace nebyla pacientovi podána, spadají do třídy kladně negativních (TN).

	Všichni	Hospitalizovaní	Nehospitalizovaní
True positive	99	53	46
False positive	1532	406	1126
True negative	27774	6240	21534
False negative	502	434	68

Tabulka 5.1. Vyhodnocení stávajícího systému

Dle tabulky výše lze vidět, že přehlédnuté alerty u hospitalizovaných v poměru k alertům, jež nenastaly a neměly nastat, se jedná o velmi malé množství. Ale kdyby je systém detekoval a vyvstal alert, počet chtěných alertů by se významně zvýšil. Skupiny přechodů nabývají nulových počtů. Lze tedy usuzovat, že zvolené intervaly jsou správné. Přechodové období před hospitalizací by mohlo vyvolávat alerty, jelikož pacienti mohou mít různě dlouhá kritická období. Naopak v přechodovém období po hospitalizaci neočekáváme, vzhledem k účinkům léčby, zhoršení a tedy ani alerty.

Jak lze vidět z tabulek, systém trpí malou sensitivitou. Je žádoucí, aby pacient byl spíše brán jako podezřelý, že u něj dojde k relapsu a dostavil se k doktorovi,

	Všichni	Hospitalizovaní	Nehospitalizovaní
sensitivita	0,1647	0,1088	0,4035
specificita	0,9477	0,9389	0,9503
accuracy	0,9319	0,88224	0,94757
F1	0,0887	0,11205	0,0715
Youden J index	0,1124	0,0477	0,3538

Tabulka 5.2. Vyhodnocení úspěšnostních metrik současného systému

tedy jedná se o možný falešný poplach, než aby bylo nebezpečí přehlédnuto a nebyla tak podána včas medikace.

5.2 Možnosti vylepšení stávajícího systému

Na základě získaných metrik lze nastavovat jednotlivé globální prahy tak, aby se zlepšila zejména sensitivita - zlepšení detekce vůči přehlédnutým alertům. Lze zohlednit také to, že upozornění pro 2 po sobě jdoucí SMS nastalo téměř o třetinu méně než suma sloupců a téměř o čtvrtinu méně než celková suma. Je možné, že práh je příliš vysoký, nebo se takovéto případy vyskytují jen málo díky častým neposlaným SMS - pacineti často vynechávají, viz kapitola 3.

Současný klasifikátor se zaměřuje na konkrétní otázky a též jim nastavuje práh, i zde je prostor pro optimalizaci, pokusit se najít nejvhodnější kombinaci otázek.

5.3 Vyhodnocení vylepšení

Výsledky v tabulkách jasně ukazují, že nastaví-li se pouze jeden z prahů (práh pro sumu dou jdoucích SMS za sebou, práh pro sumu nad celou SMS, práh pro vybrané otázky) a ostatní zůstanou na původních hodnotách, nejlepší hodnoty vychází pro nejmenší možný práh, a to práh velikosti 1. Přesto je sensitivita pro všechny pacienty nízká, v nejlepším možném případě dosahuje 49 %. Jedná se o změněný práh pro dvě po sobě jdoucí SMS, z čehož lze usuzovat, že případ nastává často. Ze shlukové analýzy pro všechny pacienty víme, že shluk obsahující 70 % všech hospitalizovaných pacientů ukázal průměrnou SMS 0,8834. Ze shlukové analýzy pouze pro hospitalizované pacienty víme, že průměrná nejnižší SMS je 0,7583 a nejvyšší 1,0619.

Suma dvou sobě jdoucích SMS s prahem 1 je tedy tak nízká, že to nelze považovat za alertní hodnotu. Za alertní hodnotu by se mohl považovat práh pro sumu 2 za sebou jdoucích SMS s hodnotou 2 a 3, kde je sensitivita 43,7 %, resp. 41,2 %.

	1	2	3	4	5	12
sensitivita	0,4904	0,4371	0,412	0,3593	0,3197	0,1647
specificita	0,7093	0,763	0,7909	0,8321	0,8527	0,9477
accuracy	0,7046	0,7562	0,7828	0,8222	0,8417	0,932
F1	0,066	0,0705	0,0744	0,0782	0,0769	0,089
Youden J index	0,1997	0,2002	0,202884	0,1915	0,1724	0,1124

Tabulka 5.3. Práh pro sumu 2 po sobě jdoucích SMS

	1	2	3	4	5	6	7	8
sensitivita	0,4264	0,3614	0,3249	0,2516	0,2144	0,1881	0,1697	0,1647
specificita	0,7719	0,8313	0,8796	0,9105	0,9299	0,9421	0,9463	0,9477
accuracy	0,7645	0,8216	0,8682	0,897	0,9153	0,9269	0,9307	0,932
F1	0,0713	0,0774	0,0923	0,0905	0,0933	0,0943	0,0896	0,0887
Youden J index	0,1983	0,1927	0,2045	0,1621	0,1443	0,1303	0,116	0,1124

Tabulka 5.4. Práh pro sumu celé SMS

	1	2	3
sensitivita	0,237864	0,182119	0,164725
specificita	0,90859	0,938148	0,947724
accuracy	0,894864	0,922933	0,931989
F1	0,084751	0,086854	0,08871
Youden J index	0,146454	0,120267	0,112449

Tabulka 5.5. Práh pro otázky 4, 6 a 9

U prahu pro sumu jedné SMS lze považovat za alertní hodnotu 2, kde je sensitivita 36,1 %. Tedy podobný výsledek jako pro práh s hodnotou 4 pro sumu dvou po sobě jdoucích SMS.

I nejmenší práh 1 pro vybrané otázky ukazuje velmi malou sensitivitu, pouhých 23 %. Nemá velký význam nastavovat pouze práh pro tyto vybrané otázky.

Tabulka níže představuje nejlepší výsledky pro kombinace všech třech prahů [suma SMS, suma otázek, suma 2 SMS].

Hodnota prahu	[8,3,1]	[8,2,1]	[6,3,1]	[7,3,1]	[7,2,1]	[6,2,1]
sensitivita	0,4941	0,4859	0,4844	0,4837	0,4821	0,4821
specificita	0,7093	0,7107	0,7129	0,7108	0,7118	0,7134
accuracy	0,7046	0,7059	0,7081	0,7060	0,7070	0,7085
F1	0,0660	0,0655	0,0659	0,0653	0,0650	0,0655
Youden J index	0,2034	0,1966	0,1973	0,1945	0,1939	0,1955

Tabulka 5.6. Prahy pro kombinace všech tří prahů

	[4, 2, 1]	[10, 8, 2]	[4, 6, 9]
sensitivita	0,4941	0,4934	0,4904
specificita	0,7093	0,7088	0,7093
accuracy	0,7046	0,7042	0,7046
F1	0,0660	0,0666	0,066
Youden J index	0,2034	0,2022	0,1997

Tabulka 5.7. Výsledky metrik pro nejlepší kombinace tří otázek při prahu velikosti 3

Původní prahy [8, 3, 12] odpovídají nejhorší možné sensitivitě ze všech kombinací, pouhých 16,47 %. Z výsledků je patrné, že vyšších sensitivit je dosaženo hlavně díky sníženému prahu pro sumu dvou po sobě jdoucích SMS.

Pro vyhodnocení nejlepší kombinace tří otázek byly nastaveny prahy s nejlepšími výsledky z tabulky 5.6. Nejlépe vyšly kombinace otázek [4,2,1] a [10,8,2]. V analýze hierarchickým shlukováním se podobné skupiny otázek nevyskytují. Je nutné ale podotknout, že všechny kombinace tří otázek měly velmi podobné výsledky. Lze tak soudit, že žádné tři kombinace nemají výraznější váhu pro klasifikaci.

Kapitola 6
Návrh a implementace nového klasifikátoru

Při návrhu nového systému se vychází ze získaných znalostí během explorativní analýzy. Vzhledem k tomu, že původní systém je poměrně jednoduchý a využívá pouze nastavených prahů, společných pro všechny pacienty, jeho výsledky nejsou optimální. Cílem je tedy vyzkoušet jinou metodu, která je sofistikovanější a nenastavuje pouze pevný práh. Jelikož chceme dosáhnout klasifikace do dvou tříd (podezřelý/nepodezřelý), vystačíme si s lineárním klasifikátorem.

6.1 Apriorní rozřazení pacientů do shluků

Při vstupu do programu ITAREPS lékaři uvádějí o pacientovi informace jako je pohlaví, datum narození a počet zaznamenaných relapsů. Dle předchozího zkoumání toto nejsou vhodná data, která by mohla sama o sobě významně sloužit pro klasifikátor. Shlukování pacientů dle symptomů ukazuje zajímavé výsledky, a tak se nabízí jej použít jako primární pohled na pacienty.

Celkový pohled na pacienty a jejich shluky slouží jako dobré vodítko, a proto budou pacienti rozděleni do 2 shluků, čímž bude dosaženo individualizace díky podobnosti pacientů dle jejich symptomů. Každého nového pacienta zařadíme do nejvhodnějšího shluku dle vzdálenosti od středu vybudovaných shluků.

6.2 Klasifikace na podezřelé a nepodezřelé

První přístup oštítkování pacienta spočívá v tom, zda byl hospitalizován (tedy 100 % rizikový), či nikoliv (třídy -1, 1), jsou zohledněny i medikace podané včas do jednoho týdne od vyvstání alertu.

Každý shluk z apriorního rozřazení je naučen pomocí lineárního klasifikátoru SVM pomocí těchto štítků a poté predikuje každého nového pacienta, zda je rizikový, či nikoliv.

Jako vstupní data pro klasifikátor byly použity všechny SMS, předchozí SMS, byla-li přítomna a přidružené příznaky. Jeden řádek příznaků obsahoval 2 SMS ($2 \times$ odpověď pacienta na 10 otázek daného pacienta.

Jak například taková sada příznaků mohla vypadat pro nehospitalizovaného a hospitalizovaného pacienta:

[2 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0]

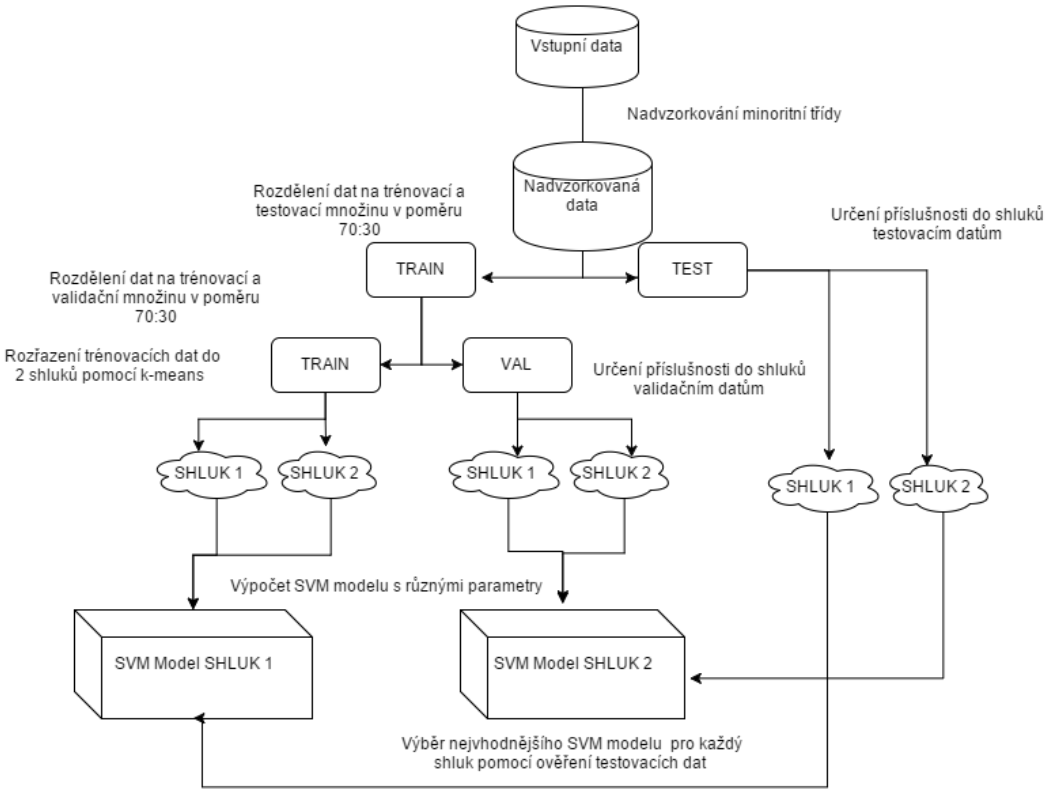
[0 0 0 1 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0]

Neplatí, že čím více příznaků, tím lépe se klasifikátor natrénuje, a proto budeme postupně zkoušet kombinace, jejichž výsledky budou uvedeny v kapitole 6.4.

Přiřazení daného row-setu do příslušné třídy (podezřelý/nepodezřelý z hospitalizace) zajišťují štítky podezřelý/nepodezřelý (pro SVM klasifikátor to jsou štítky 1/-1) popsané výše.

6.3 Průběh experimentu

Celý proces učení probíhá na trénovací množině, nad kterou se vyhodnocuje trénovací chyba. Pro vyhodnocení nastavení nejlepších parametrů klasifikátoru je použita validační množina, nad kterou se iterují různé hodnoty pro daný kernel, či penalizační konstantu. K finálnímu vyhodnocení slouží testovací množina dat za použití nejlepšího možného modelu nalezeného nad množinou validační. Data jsou rozdělena na trénovací/testovací množinu v poměru 70/30. Při vytváření skupin je dbáno na to, aby bylo vždy bráno stejné procento hospitalizovaných a nehospitalizovaných. Taktéž je žádoucí, aby vstupní data byla kvalitní, a tedy aby nebyla neúplná. Jelikož víme, že někteří pacienti nemají určené symptomy vůbec, nebo že neposílali některé týdny SMS, je třeba tyto rowsety dle potřeby odstranit. Trénovací data jsou dále dělena na skutečně trénovací a validační v poměru 70/30. Pomocí testovací sady dat je pouze ověřen natrénovaný SVM model.



Obrázek 6.1. Diagram průběhu experimentu trénování a validování SVM modelů.

Jedná se o časově náročný postup, jelikož se jedná o vyhodnocení trénování pro oba shluků, tak i celých dat. Výsledky z trénování SVM modelu nad celými daty slouží jako porovnání k individualizovanému trénování nad shluky. Výsledky budou reportovány v podobě metrik sensitivity, specificity, přesnosti, F1, Youden J indexu v přílohách. Níže budou zmíněny a popsány nejlepší výsledky.

6.3.1 Nadvzorkování, podvzorkování

Vstupní data obsahují nevyvážená data. Poztiviní třída je zastoupena vůči negativní třídě v poměru 1:50. První experimenty testovaly učení SVM modelu nad podvzorkovanými daty, ale ukázalo se, že výsledky nebyly tak příznivé jako pro nadvzorkovanou sadu dat.

6.3.2 Volba penalizace

Pro každou sadu příznaků (1 SMS, 2 SMS) se náhodně vygenerují nadvzorkovaná data a 5 $\times$ se křížově validují Monte Carlo metodou, kdy náhodně se vybere trénovací a validační množina v poměru 70:30.

6.3.3 Volba příznakového prostoru

Zaměříme se na to, zda si klasifikátor vystačí se stejnými daty, jako má k dispozici stávající systém. Určitě je pro nás zajímavé, zda si stávající klasifikátor poradí (a jak) s méně informacemi, případně stejnými a zda lépe. Modelu bude nejprve předložena sada dat obsahující pouze první SMS samostatně, poté dvě po sobě jdoucí SMS. Jestliže neexistují dvě po sobě jdoucí SMS, pak je záznam brán jako neúplný a je pro učení vyřazen.

6.4 Výsledky

Výsledky pro rozdělená data do shluků vychází lépe, než nad celými daty. To je pozitivní zpráva, jelikož to potvrzuje, že má smysl aplikovat metody, které napomohou systému být více individualizovaným. Celkový proces trénování, validace a testování je velmi náročný, a tak po předchozích experimentech byl zvolen rozsah pro penalizační konstantu od 1 do 100 logaritmicky.

Tabulka níže ukazuje vybrané vhodné výsledky pro učení nad daty obsahující 1 SMS.

Shluk 1, C = 10	TRAIN	VALID	TEST
sensitivita	0,4306	0,4181	0,437
specificita	0,9041	0,8828	0,9045
accuracy	0,6743	0,6542	0,6768
F1	0,5620	0,5433	0,5684
Youden J index	0,33477	0,301	0,3416
Shluk 2, C = 100	TRAIN	VALID	TEST
sensitivita	0,4599	0,4687	0,4599
specificita	0,8988	0,8971	0,8988
accuracy	0,68575	0,688	0,6857
F1	0,5869	0,5945	0,5899
Youden J index	0,3587	0,3658	0,3587
Úplná data, C = 215	TRAIN	VALID	TEST
sensitivita	0,3956	0,2915	0,2989
specificita	0,9040	0,8931	0,8928
accuracy	0,6618	0,5978	0,5977
F1	0,5271	0,4157	0,4247
Youden J index	0,2996	0,1846	0,1917

Tabulka 6.1. Výsledky trénování a validace nad daty obsahující 1 SMS.

Specificita je podobná u všech dat, vychází velmi vysoká okolo 90 %, to značí o správné detekci vzorků spadajících do negativní třídy. Sensitivita je ale o poznání

lepší u dat rozdělených do shluků. Pro shluk druhý je sensitivita až 90 %. Výsledky pro data ve shlucích jsou dobré, oproti úplným datům, kde je sensitivita okolo 25 % a penalizační konstanta nejvyšší.

Výsledky pro data obsahující 2 za sebou jdoucí SMS pro oba shluky jsou podobné, sensitivita 68 % a specificita 87 %. Ačkoliv je výsledek pro druhý shluk o něco horší než pro data obsahující pouze 1 SMS, pro první shluk je významně lepší. Metrika specificity pro úplná data ukazuje selhání v detekci negativních vzorků, specificita naopak ukazuje výbornou detekci pozitivní třídy.

Shluk 1, C = 10	TRAIN	VALID	TEST
sensitivita	0,305	0,2989	0,3088
specificita	0,8688	0,858	0,8612
accuracy	0,5749	0,5687	0,5693
F1	0,4279	0,4177	0,4311
Youden J index	0,1738	0,1569	0,17
Shluk 2, C = 10	TRAIN	VALID	TEST
sensitivita	0,2365	0,185	0,2239
specificita	0,9671	0,9709	0,9726
accuracy	0,6948	0,6843	0,6986
F1	0,3662	0,2994	0,3523
Youden J index	0,2036	0,1559	0,1966
Úplná data, C = 10	TRAIN	VALID	TEST
sensitivita	0,225	0,2188	0,222
specificita	0,9173	0,9106	0,9144
accuracy	0,5736	0,57	0,5675
F1	0,3438	0,3338	0,3396
Youden J index	0,1423	0,1294	0,1364

Tabulka 6.2. Výsledky trénování a validace nad daty obsahující 2 SMS.

Výsledky potvrzují, že individualizovaný klasifikátor nad daty rozdělených do shluků, vybraných dle předchozí analýzy, funguje v případě volby správných parametrů a vstupních dat výrazněji lépe než původní systém i po nalezení optimálních prahů.

6.5 Možná vylepšení

Práce používá standardní implementaci MatLab SVM a výsledky byly lepší v porovnání s knihovnou LibSVM [13] za nastavení stejných podmínek. Knihovna LibSVM patří k jedné z nejpoužívanějších knihoven a je to jistě jen otázkou správné kombinace parametrů pro SVM klasifikaci. Cílem této práce ale není najít nejlepší možný klasifikátor, ale vhodnou alternativu, která dá základ pro další zkoumání a vytváření více individualizovaného systému.

Systému by mohlo pomoci zvětšení počtu pozitivních prvků, a tak o něco snížit nevybalancovaná data. Toho lze docílit například zvážením doby, kdy již před podáním medikace pacient může vykazovat zhoršení stavu. Je nepravděpodobné, že by se změna udála během 1 týdne, kdy vyvstal alert, ale nejspíše již pár týdnů předem.

Také metody, jak pracovat s nevybalancovanými daty mohou být mnohem sofistikovanější, například SMOTE a ADASYN [14–15]. Obecně trénování nad nevyváženými daty je v posledních letech velmi zkoumané a živé téma, jelikož takových případů v reálném světě existuje mnoho.

Kapitola 7
Závěr

Práce se věnuje analýze dat dostupných z programu ITAREPS. Celkový počet pacientů čítá 280 pacientů, z čehož 59 podstoupilo hospitalizaci během ITAREPS. Dvěma hospitalizacemi si prošlo 15 pacientů a třemi pouze 5, intervaly mezi jednotlivými hospitalizacemi se zkrátily a délka samotné hospitalizace se prodloužila. Během zkoumání bylo zjištěno, že zhruba 20 % pacientů a 64 % rodinných příslušníků posílá alespoň 70 % očekávaných SMS. Jedná se o chybějící údaje, které bylo třeba nahradit Not A Number hodnotou. Po doplnění chybějících týdnů bylo možné zkoumat odpovědi účastníků v čase a pozorovat tak typické průběhy. Ze všech poslaných SMS je 70 % obsahující nulové odpovědi na otázky. To není překvapující, jelikož pacienti jsou většinu času v klidu a také nehospitalizovaní tvoří majoritní skupinu. I přes velmi variabilní data u hospitalizovaných pacientů během ITAREPS byly objeveny trendy, které odpovídají očekávaní. V klidovém období pacienti posílají průměrně nižší SMS, v kritickém období (6 týdnů před hospitalizací) posílají zvýšené hodnoty, které opět klesají po odeznění účinků medikace podávané během hospitalizace. Z vyhodnocování jsou vyřazeny přechodová období, která by mohla zkreslovat průběhy. Jedná se o přechodové období před kritickým obdobím, kdy je těžké určit přesný týden, kdy se pacient začal zhoršovat, a o přechodové období po hospitalizaci, kdy pacient bývá utlumený po léčbě. Klidová období před a po hospitalizaci byla na základě statistik sloučena. Klidová období hospitalizovaného a nehospitalizovaného pacienta nepochází ze stejného rozdělení a ani z rozdělení se stejnými mediány. Protože chceme zjistit co nejvíce informací o rozdílnosti o pacientech, další zkoumání se týkalo hlavně skupiny hospitalizovaných a nehospitalizovaných během ITAREPS. Rodinní příslušníci nebyli dále zkoumáni z důvodu toho, že jejich účast není povinná, a proto mnoho pacientů rodinného příslušníka nemělo. Také statistiky ukazují, že se neshodují hodnocení rodinných příslušníků během klidových období hospitalizovaných pacientů.

Metoda PCA ukázala, že jednotlivé otázky jsou hodně závislé. První hlavní komponenta pro kritické období a klidová období obou skupin je sycena rovnoměrně všemi otázkami. U klidových období je zajímavá druhá a třetí komponenta. Druhá hlavní komponenta je pro klidová období obou skupin sycena zejména otázkou č. 1

(zhoršení spánku) a č. 9 (slyšení nepřítomných hlasů). Třetí komponenta je také zastoupena otázkou č. 1. Pro hospitalizované navíc ještě č. 9 a nehospitalizované č. 7 (ztráta energie a zájmu). Dále byla zkoumána podobnost jednotlivých otázek za pomocí hierarchického shlukování různých vzdáleností a metrik. Některé shluky se pro různé kombinace výpočtů opakovaly. Díky malým vzdálenostem mezi shluky nelze ale říci, že existují významnější shluky. Nehierarchické shlukování ukazuje, že pacienti posílající nulové SMS spadají do společného shluku. Střed pro druhý shluk nabývá největších hodnot pro kritické období hospitalizovaného a naopak nejmenší pro klidové období nehospitalizovaného.

Další část se zabývala symptomy, které zadává offline do systému doktor při vstupní prohlídce do programu ITAREPS. Data ale ukazují, že někteří pacienti nemají symptomy zaznamenány. Další zajímavou skupinu tvoří pacienti, kteří mají symptomy nulové. To znamená, že jsou buď nevyjádřené a symptomy u nich nepropukly, nebo jsou špatně zpracovány a zaznamenány. Protože se jedná o početnou skupinu, 72 pacientů, tedy 1/4 z celkového počtu. Práce dále graficky ukazuje, kolik pacientů prodělalo určitý počet hospitalizací do zapojení do programu ITAREPS. Korelační koeficient ukazuje na střední až mírně významnou korelaci pro počet hospitalizací a rokem prvního výskytu. To je v celku logické, jelikož lze usuzovat, že čím dříve byl zaznamenán první relaps, tím více relapsů a hospitalizací celkem pacient utrpěl. Není totiž vůbec snadné podchytit příznaky včas. Výpočty nepotvrdily významné závislosti mezi počtem hospitalizací před a během ITAREPS a celkovou sumou symptomů. Mezi posílanými SMS a jednotlivými stupni závažností symptomů je těžké vypozorovat závislosti. Je ale patrné, že pro symptomy 4. stupně pacienti posílají velmi nízké hodnoty. Nízké hodnoty SMS také náleží příznakům 3. stupně, až na příznak pozitivní, který má v celkovém pohledu nízkou průměrnou SMS, ale oproti ostatním symptomům o řád vyšší. Pro zkoumání závislosti symptomů byla opět použita metoda PCA. Odhaluje, že pro pacienty s pouze vyjádřenými příznaky je první hlavní komponenta zastoupena nesourodě, zatímco pro pacienty včetně nevyjádřených symptomů je zastoupení všech symptomů poměrně rovnoměrné.

Důležitou částí této práce je objevení podobností mezi pacienty na základě jejich symptomů. Pacienti ze shluku s průměrně vyššími hodnotami symptomů posílají průměrně vyšší SMS, než pacienti s průměrně nižšími hodnotami symptomů.

Druhá část práce se věnovala současnému systému, jeho vyhodnocení a návrhu nového, individualizovaného, systému pomocí strojového učení a výsledků z explorativní analýzy. Vyhodnocení pomocí statistických metrik ukázalo, že celková sensitivita systému je příliš malá. Nízká sensitivita je nežádoucí a to z toho dů-

vodu, že to znamená přehlédnutá rizika. Ke zlepšení sensitivity pomohla změna jednotlivých prahů. Nicméně nejlepší výsledek přinesly původní prahy pro celkovou sumu SMS a sumu tří původních otázek, změnil se pouze práh pro dvě SMS jdoucí za sebou z 12 na 1. Konkrétní kombinace otázek nemá příliš velkou váhu, jelikož všechny kombinace vychází dle statistik velmi podobně.

Statistické metriky klasifikace nad shlukovanými daty pomocí SVM modelu přináší skvělé výsledky, až 89 % sensitivitu a 90 % specificitu pro jeden ze shluků se vstupními daty obsahující 1 SMS. Výsledky pro 2 po sobě jdoucí SMS nejsou tak vysoké, ale stabilní pro oba dva shluky, a to sensitivita 70 % a specificita necelých 90 %. Výsledky jak pro 1 SMS, tak pro 2 SMS nad úplnými daty, tedy nerozdělených do shluků, jsou výsledky podstatně horší, sensitivita je vysoká, ale specificita nízká. Ačkoliv explorativní analýza poukazuje na několik problémů (chybějící záznamy, mnoho nulových SMS), ale SVM algoritmus si s nimi dokázal poradit. Výhodou je, že se nemusí nastavovat jednotlivé prahy pro konkrétní SMS. Avšak doba trénování a validací penalizační konstanty byla poměrně dlouhá. Jelikož nebylo cílem hledat optimální nastavení algoritmu, práce se tím zabývala spíše okrajově. Jednotlivé parametry SVM trénování mohou být předmětem dalšího zkoumání. Stejně tak využití jiných knihoven. Například SVMLight byla použita pro trénování nad nevyváženým data setem bez jakékoliv úpravy dat a přináší dobré výsledky [16]. Možnou variantou jsou i jiné algoritmy, například neuronové sítě, které si mohou poradit lépe s neúplnými a nevyváženými daty.

Jednoduchou možností optimalizace pro učení SVM modelem je úprava vstupních dat. V práci bylo implementováno podvzorkování a nadvzorkování. Existují ale i více sofistikované algoritmy vycházející z rozložení a heuristiky dat. Jelikož byla pozitivní třída 50x menší než třída negativní, stojí za uvážení, zda lze třídu pozitivní zvětšit. Bylo by možné diskutovat s doktory, v jakých případech pacienti dostávají medikaci. Například zda už před podáním medikace pacienti vykazují zhoršení několik týdnů předem. Pak by se dala třída pozitivních vzorků rozšířit o několik týdnů pro každou medikaci a zpřesnit tak detekci klasifikátoru. Pro přesnější detekci by i pomohlo, kdyby doktoři zaznamenávali, zda se jednalo pouze o úpravu medikace, či skutečně podání dávky pro předejití dalšímu zhoršování stavu pacienta. Momentálně vstupní data tento záznam neobsahují, je nutné brát na zřetel, že to může zanášet do systému nepřesnosti.

Práce nastínila některé problémy, které vznikají již při samotném sběru dat (nepochopení systému hodnocení zhoršování/zlepšování stavu, mnoho neposlaných SMS, chybějící symptomy, dodatečná informace o důvodu podání medikace). Vzhledem k dobrým výsledkům kontrolované studie [2], kde se dbalo na častou

kontrolu pacienta, vyplňování a posílaní SMS, se může jednat o problémy zanášející nepřesnosti bránící přesnějšímu vyhodnocení. Experimentální vyhodnocení současného systému pomocí změny prahů ukazuje, že lze dosáhnout lepších výsledků i v současném systému. Nejlepší výsledky dosahují 49 % sensitivity, což je oproti současným 17 % velké zlepšení. Strojové učení v kombinaci s individualizací pomocí shlukovací metody k-means dle symptomů potvrzuje, že má smysl se dále zabývat individualizací, zpřesňováním vstupních dat a řešení otázek vznikajících s úlohami strojového učení -konkrétní parametry, nevyvážené zastoupené pozitivní a negativní třídy apod.

Literatura

[1] et al. Španiel, F. ITAREPS: Information Technology Aided Relapse Prevention Programme in Schizophrenia. *Schizophrenia Research*. 2008, 98 (1-3),

[2] H. et al. Komatsu. Effectiveness of Information Technology Aided Relapse Prevention Programme in Schizophrenia excluding the effect of user adherence: a randomized controlled trial.. *Schizophr Res.*. 2013, 150 (1),

[3] A. Barbato. *Schizophrenia and public health*. 1998. `http://www.who.int/mental_health/media/en/55.pdf?ua=1`.

[4] American Psychiatric Association. *DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS 4TH EDITION*. 2000.

[5] Institut biostatistikya analýz. *PSYCHIATRRICKÁ PROPEUDETIKA*. `http://telemedicina.med.muni.cz/psychiatricka-propedeutika/index.php?pg=text`.

[6] American Psychiatric Association. *DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS FIFTH EDITION*. 2014.

[7] M. et al. Herz. A program for relapse prevention in schizophrenia: a controlled study.. *Arch Gen Psychiatry*. 2000, 57 (3),

[8] Chih-Jen Lin Chih-Chung Chang. *LIBSVM – A Library for Support Vector Machines*. 2016. `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

[9] G. James. *An Introduction to Statistical Learning*. Springer, 2013. ISBN 978-1-4614-7138-7.

[10] Michael K. Skinner M. Muksitul Haque a Lawrence B. Holder. Imbalanced Class Learning in Epigenetics. *Journal of Computational Biology*. 2014, 21 (17),

[11] Louise C Showe Malik Yousef, Segun Jung a Michael K. Showe. Learning from positive examples when the negative class is undetermined- microRNA gene identification. *Algorithms Mol Biology*. 2008, 3 (2),

[12] J. Rachoch. *PSYCHIATRIE, První vydání*. Galén, 2001. ISBN 80-7262-140-8.

[13] MATHWORKS. *Support Vector Machines for Binary Classification*. 2016. https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html.

[14] V. Chawla. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002,

[15] H Haibo. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Neural Networks*. 2008,

[16] J. Brank. Training text classifiers with SVM on very few positive examples. *Microsoft Research*. 2003,

Příloha A

Výsledky trénování nad daty obsahující 1 SMS

Trénovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,1931	0,1921	0,1918	0,4438	0,1975
Specificita	0,9399	0,941	0,9386	0,8995	0,9399
Accuracy	0,5666	0,5651	0,5657	0,6628	0,5691
F1	0,3081	0,3073	0,3068	0,5735	0,314
Youden	0,1329	0,133	0,1304	0,3433	0,1374
Validační					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,1844	0,1893	0,1969	0,433	0,1827
Specificita	0,9358	0,9323	0,9356	0,9051	0,9373
Accuracy	0,5621	0,5683	0,5685	0,6587	0,5615
F1	0,2968	0,3013	0,312	0,5652	0,2935
Youden	0,1202	0,1216	0,1325	0,3381	0,1199
Testovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,1925	0,1916	0,1942	0,4197	0,1942
Specificita	0,9402	0,9393	0,9391	0,9005	0,9397
Accuracy	0,5666	0,5658	0,567	0,6564	0,5673
F1	0,3074	0,306	0,3094	0,5495	0,3096
Youden	0,1326	0,1309	0,1333	0,3202	0,1339

Obrázek A.1. Výsledky trénování nad daty obsahující 1 SMS pro shluk 1

Trénovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,4106	0,4461	0,4368	0,194	0,4371
Specificita	0,9058	0,9033	0,9039	0,937	0,9026
Accuracy	0,6643	0,6789	0,6743	0,5685	0,6742
F1	0,544	0,5769	0,5686	0,3078	0,5683
Youden	0,3163	0,3494	0,3406	0,131	0,3396
Validační					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,406	0,9078	0,6596	0,5412	0,3138
Specificita	0,4386	0,9035	0,6735	0,5669	0,3352
Accuracy	0,4285	0,9066	0,6732	0,561	0,3344
F1	0,197	0,9362	0,5673	0,3121	0,1317
Youden	-0,1554	0,8113	0,333	0,1081	-0,3511
Testovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,3981	0,908	0,6643	0,5313	0,3061
Specificita	0,4329	0,8991	0,676	0,5608	0,3313
Accuracy	0,4221	0,908	0,6758	0,5544	0,33
F1	0,1924	0,9386	0,5705	0,306	0,1293
Youden	-0,169	0,807	0,3403	0,0921	-0,3626

Obrázek A.2. Výsledky trénování nad daty obsahující 1 SMS pro shluk 2

Trénovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,2439	0,2451	0,2438	0,2464	0,2662
Specificita	0,9215	0,9197	0,9197	0,9211	0,9102
Accuracy	0,5853	0,5836	0,5823	0,5851	0,5895
F1	0,3688	0,371	0,3678	0,3717	0,3929
Youden	0,1654	0,1648	0,1635	0,1676	0,1764
Validační					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,24	0,2416	0,2451	0,2448	0,2628
Specificita	0,9205	0,9178	0,9215	0,9222	0,9089
Accuracy	0,5838	0,587	0,5853	0,5862	0,5912
F1	0,3647	0,3645	0,3683	0,3703	0,3896
Youden	0,1604	0,1595	0,1666	0,1669	0,1717
Testovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,2339	0,2359	0,2342	0,2356	0,2556
Specificita	0,9232	0,9219	0,923	0,9246	0,9129
Accuracy	0,5838	0,586	0,5838	0,5853	0,5897
F1	0,3562	0,3594	0,3566	0,3589	0,3808
Youden	0,1571	0,1578	0,1572	0,1603	0,1685

Obrázek A.3. Výsledky trénování nad daty obsahující 1 SMS pro úplná data

Příloha B

Výsledky trénování nad daty obsahující 2 SMS

Trénovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,9558	0,9763	0,9288	0,6804	0,9772
Specificita	0,1314	0,0941	0,1882	0,8706	0,1204
Accuracy	0,5906	0,5889	0,5548	0,6104	0,5905
F1	0,7278	0,7267	0,6748	0,7408	0,727
Youden	0,0872	0,0704	0,1169	0,5511	0,0977
Validační					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,9503	0,9757	0,9261	0,6886	0,9722
Specificita	0,1326	0,1035	0,1926	0,872	0,1165
Accuracy	0,5801	0,594	0,556	0,5989	0,5872
F1	0,7124	0,7244	0,6704	0,7321	0,7215
Youden	0,0829	0,0792	0,1187	0,5606	0,0887
Testovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,9561	0,9755	0,9265	0,7022	0,9772
Specificita	0,1292	0,0974	0,1857	0,8847	0,1171
Accuracy	0,5714	0,5755	0,5432	0,5859	0,5773
F1	0,7047	0,7133	0,6611	0,7202	0,7153
Youden	0,0853	0,0728	0,1122	0,5869	0,0943

Obrázek B.4. Výsledky trénování nad daty obsahující 2 SMS pro shluk 1

Trénovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,0525	0,0575	0,6793	0,5356	0,0218
Specificita	0,9901	0,9899	0,8641	0,847	0,9883
Accuracy	0,6757	0,745	0,763	0,5932	0,5863
F1	0,0971	0,1061	0,7582	0,6453	0,042
Youden	0,0426	0,0474	0,5434	0,3825	0,0102
Validační					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,0505	0,0505	0,6835	0,5288	0,0141
Specificita	0,988	0,989	0,8768	0,8491	0,9844
Accuracy	0,6605	0,7452	0,765	0,5906	0,5676
F1	0,0945	0,0929	0,7635	0,6428	0,0274
Youden	0,0385	0,0395	0,5602	0,3778	-0,0015
Testovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,0458	0,0458	0,7077	0,5509	0,0166
Specificita	0,988	0,9861	0,8764	0,85	0,9857
Accuracy	0,6882	0,7435	0,7891	0,5949	0,586
F1	0,0849	0,0848	0,7764	0,6543	0,032
Youden	0,0339	0,0319	0,5841	0,4009	0,0023

Obrázek B.5. Výsledky trénování nad daty obsahující 2 SMS pro shluk 2

Trénovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,9011	0,9023	0,9021	0,8997	0,8997
Specificita	0,196	0,1912	0,1953	0,1935	0,1992
Accuracy	0,5528	0,5484	0,5516	0,5488	0,5505
F1	0,6701	0,6677	0,6698	0,6673	0,667
Youden	0,097	0,0935	0,0974	0,0932	0,0989
Validační					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,8989	0,9043	0,9017	0,9035	0,8973
Specificita	0,1929	0,1959	0,1969	0,1991	0,2028
Accuracy	0,5417	0,5486	0,5493	0,5519	0,5502
F1	0,6618	0,6664	0,6658	0,6686	0,6665
Youden	0,0917	0,1002	0,0986	0,1027	0,1
Testovací					
Cost	**1**	**3,1623**	**10**	**31,6228**	**100**
Sensitivita	0,898	0,9013	0,8985	0,8991	0,8976
Specificita	0,2027	0,2009	0,2018	0,2014	0,2088
Accuracy	0,5414	0,5419	0,541	0,5426	0,5406
F1	0,6562	0,657	0,6559	0,656	0,6577
Youden	0,1007	0,1022	0,1003	0,1005	0,1064

Obrázek B.6. Výsledky trénování nad daty obsahující 2 SMS pro úplná data