UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

APLIKÁCIA METÓD DATA MININGU NA ANALÝZU DÁT V SEKTORE ZDRAVOTNÍCTVA

DIPLOMOVÁ PRÁCA

2017 Bc. Beáta BENKOVÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

APLIKÁCIA METÓD DATA MININGU NA ANALÝZU DÁT V SEKTORE ZDRAVOTNÍCTVA

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Henrieta Tulejová, MSc.

Bratislava 2017 **Bc. Beáta BENKOVÁ**

25171629

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

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

Názov: Aplikácia metód data miningu na analýzu dát v sektore zdravotníctva.
Application of data mining methods to health care data analysis.

Cieľ: Cieľom práce je zvoliť, opísať a aplikovať niektoré z metód data miningu na analýzu zdravotníckych dát.

Vedúci: Mgr. Henrieta Tulejová, MSc.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, DrSc.

Dátum zadania: 21.01.2016

Dátum schválenia: 25.01.2016 prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Touto cestou by som sa chcela poďakovať vedúcej svojej diplomovej práce Mgr. Henriete Tulejovej, MSc. a svojím kolegom zo zdravotnej poisťovne Dôvera za pomoc, ochotu a cenné rady pri písaní práce. Vďaka patrí aj MUDr. Jozefovi Lackovi, CSc., MBA za podnetné medicínske pripomienky a DOC. RNDr. Margaréte Halickej, CSc. za konzultácie a pomoc pri tvorbe modelov.

Abstrakt v štátnom jazyku

BENKOVÁ, Beáta: Aplikácia metód data miningu na analýzu dát v sektore zdravotníctva [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Henrieta Tulejová, MSc., Bratislava, 2017, 61s

Cieľom našej práce je zvoliť, opísať a aplikovať niektoré metódy data miningu na analýzu zdravotníckych dát so špecifikáciou na benchmarking poskytovateľov zdravotnej starostlivosti, konkrétne diabetológov. V prvej kapitole stručne popisujeme ochorenie cukrovky. Následne opisujeme jednotlivé fázy data miningu, ako hierarchického procesu. V tretej kapitole vysvetľujeme základy zvolených štatistických a modelovacích techník, hlavne obálkovej analýzy dát (DEA). V záverečnej kapitole aplikujeme poznatky na dátach z vykázanej zdravotnej starostlivosti a zostavíme celkový benchmarking. V závere výsledky zhodnocujeme a navrhujeme možné vylepšenia.

Kľúčové slová: Data mining, Obálková analýza dát, Benchmarking poskytovateľov zdravotnej starostlivosti

Abstract

BENKOVÁ, Beáta: Application of data mining methods to health care data analysis [Master thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics , Department of Applied Mathematics and Statistics; Supervisor: Mgr. Henrieta Tulejová, MSc., Bratislava, 2017, 61p

The aim of our thesis is to choose, describe and apply some of data mining methods for health care data analysis, focused on health care provider benchmarking, particularly diabetologists. We briefly describe disease of diabetes in the first chapter. Subsequently we profile data mining phases as a hierarchical process. In the third chapter we explain basics of chosen statistical and modelling techniques, mainly Data Envelope Analysis (DEA). In the last chapter we apply findings on reported health care data and form the final benchmarking. In conclusion we review results and suggest possible improvements.

Keywords: Data mining, Data Envelopment Analysis, Healt Care Providers Benchmarking

Obsah

Zoznam obrázkov

Úvod

Dramatický pokrok v oblasti zachytávania, zaznamenávania a skladovania dát nám umožňuje vytváranie rozsiahlych databáz, v ktorých fungujú určité pravidlá a súvislosti. S ich pribúdajúcim množstvom a rozšírením do všetkých oblastí života prichádza rovnako aj motivácia tieto vzťahy identifikovať a skúmať. Za týmto účelom vznikol relatívne nový pojem *data mining*.

Vo všeobecnosti je data mining, alebo hĺbková analýza dát, proces skúmania a analyzovania dát z rôznych perspektív a sumarizácia takto získaných súvislostí do znalostí, ktoré vedú k vylepšeniu stratégií, znižovaniu nákladov, zvyšovaniu ziskov či konkurencieschopnosti. Prvé metódy data miningu sa objavili už v 60. rokoch 20. storočia s rozvojom počítačovej techniky. Postupne si našli uplatnenie v mnohých oboroch, akými sú bankovníctvo, poisťovníctvo, marketing, komunikácie, ale aj medicína a riadenie procesov [5].

Podobný cieľ aj keď iné nástroje má *benchmarking*. Jeho definícia nie je jednotná, no zväčša sa pod pojmom benchmarking (skóring) rozumie súbor činností, pomocou ktorých sa firma snaží porovnávať svoj výkon s inými firmami, prípadne porovnávať subjekty vrámci firmy, vykonávajúce podobnú činnosť. Cieľom benchmarkingu je definovať, kde by mohli nastať zlepšenia a ako ich dosiahnuť [13].

My sme si v našej práci zvolili za cieľ analýzu zdravotníckych dát, pričom práve prienik týchto dvoch procesov bude pre nás kľúčový. V zahraničí je porovnávanie zdravotníckych zariadení stále viac rozšírenejšie a s pribúdajúcimi možnosťami v oblasti spracovania dát sa stáva aj čoraz presnejšie a nápomocnejšie. Počiatky využitia benchmarkingu v zdravotníctve siahajú až do 17. storočia, kedy sa nemocnice porovnávali na základe úmrtnosti. Neskôr sa ho snažili využiť na zníženie nákladov, ako aj zlepšenie výkonnosti. Posledné modifikácie tento koncept dokonca prepojili s potrebou splniť

očakávania pacientov [12].

Týmito prácami sme sa nechali inšpirovať a zamerali sme sa na benchmarking diabetológov, ktorý sme sa pomocou metód data miningu a dát zo zdravotnej poisťovne, pokúsili urobiť čo možno najpresnejší.

V úvodnej kapitole v skratke popisujeme ochorenie cukrovky, jej priebeh a možné komplikácie pre lepšie pochopenie krokov v praktickej časti. Druhá kapitola nám má poskytnúť stručný prehľad procesov, ktoré data mining obsahuje. Ďalej bližšie popisujeme metódy a modelovacie techniky, konkrétne využívané v záverečnej kapitole, ktorá formálne opisuje praktickú časť našej práce.

Kapitola 1

Diabetológia a manažment liečby cukrovky

Diabetológia je medicínsky odbor zaoberajúci sa poruchami látkovej premeny a výživy. Najčastejším ochorením, s ktorým sa diabetológovia dostávajú do kontaktu je ochorenie cukrovky. V tejto kapitole v skratke charakterizujeme toto ochorenie, jeho priebeh, liečbu a motiváciu manažmentu cukrovky z hľadiska zdravotnej poisťovne. V kapitole vychádzame hlavne z [1] a [2].

1.1 Ochorenie cukrovky

Cukrovka, v lekárskej terminológii aj diabetes mellitus, je nevyliečiteľná choroba látkovej premeny, ktorej najvýraznejším prejavom je hyperglykémia, resp. zvýšená hladina cukru v krvi. Jej príčinou je nedostatok hormónu inzulínu, ktorý riadi látkovú premenu a krvný cukor reguluje. Pri jeho zvýšených hodnotách dochádza k poškodeniu ciev, ktoré sa prejavuje rôznymi komplikáciami. Rozlišujeme viacero typov cukrovky:

- **Diabetes mellitus 1. typu** - je typická výrazným alebo úplným nedostatkom inzulínu, pričom je jeho podávanie pacientovi nevyhnutné k životu. Zväčša sa objavuje už v detstve alebo mladom veku.
- **Diabetes mellitus 2. typu** - je civilizačným ochorením spôsobeným kombináciou nedostatočnej tvorby inzulínu a inzulínovej rezistencie. V tele prakticky dochádza k nepravidelnej tvorbe inzulínu, na ktorý tkanivá nereagujú dostatočne

citlivo, čo môže viesť k zvýšenej hladine krvného cukru a prejavom cukrovky. Rizikovými faktormi sú hlavne dedičnosť, nezdravý životný štýl a stres.

- **Iné špecifické formy diabetu** - zahŕňajú prípady cukrovky so známou príčinou, akými môžu byť dedičné poruchy, ako aj cukrovka vyvolaná iným ochorením alebo liekmi.

- **Gestačný diabetes** - je ochorenie matiek počas gravidity, ktoré sa vyskytuje v druhej polovici tehotenstva zhruba pri 3-4% žien. Má nevýrazné príznaky a zvyčajne po tehotenstve zaniká.

Na Slovensku bolo v roku 2015 podľa [4] 345 475 diabetikov v dispenzárnej starostlivosti, čo je 6,37% celkovej populácie. Až 90,6% pritom tvorili diabetici druhého typu a 75% z nich boli osoby vo veku 50 a viac rokov. Ďalej sa pre potreby našej práce zaoberáme len priebehom a liečbou diabetu mellitu 2. typu.

1.2 Priebeh a liečba cukrovky

Cukrovka je chronické ochorenie, ktoré pri zanedbaní postupuje. Liečba cukrovky úzko súvisí s typom cukrovky, ktorým pacient trpí. Vo všeobecnosti však zahŕňa diétny program a zvýšenie fyzickej aktivity. Cieľom liečby je z najväčšej miery odstránenie prejavov ochorenia, zabránenie vzniku chronických a akútnych komplikácií a predĺženie života.

Príčinou chronických komplikácií je hlavne dlhodobo zvýšená hladina cukru, ktorá vedie k poškodeniu malých ciev v citlivých orgánoch. Preto sú najčastejšie postihnuté obličky (diabetická nefropatia), oči a sietnica (diabetická retinopatia) a nervový systém (diabetická neuropatia). Pri väčších komplikáciách dochádza až k ateroskleróze väčších ciev, čo má často za následok rôzne ischemické choroby srdca (srdcový infarkt), cievne mozgové príhody a nedokrvenie dolných končatín, ktoré v krajných prípadoch vedú k amputácii.

Okrem chronických komplikácií, ktoré majú zväčša dlhodobý priebeh, pri cukrovke môže dôjsť aj k akútnym komplikáciám. Tie sa delia podľa hladiny cukru v krvi (glykémie). Hyperglykemické komplikácie vznikajú v prípade neužitia inzulínu alebo dô-

sledkom jeho nedostatočného účinku. K hypoglykémii naopak dochádza v prípade nadmernej dávky lieku alebo zníženia príjmu potravy, prípadne zvýšenej fyzickej aktivite pri užití štandardnej dávky.

Pri liečbe cukrovky 2. typu platí, že v najväčšej miere ju môže ovplyvniť samotný pacient. Dôležitá je takisto prevencia, nakoľko priebeh ochorenia úzko súvisí so štádiom, v ktorom bola cukrovka diagnostikovaná. Pri nameraní zvýšenej hodnoty glukózy v krvi je pacientovi často poradená diéta a pohyb. V prípade možných komplikácií a u pacientov, pri ktorých nepomáhajú režimové opatrenia, nastupuje liečba perorálnymi antidiabetikami (ďalej PAD). Spočiatku je pacientovi predpísaný jeden liek (1 PAD). Ak efekt nie je dostatočný, môžu byť pridané ďalšie lieky v rôznych kombináciach (2 PAD a 3 PAD). Pri nedodržiavaní diéty a nedostatku pohybu však často ani najlepšia kombinácia antidiabetík nemusí znížiť hladinu cukru v krvi a preto často dochádza až k postupu ochorenia do štádia, ktoré si vyžaduje podávanie inzulínu.

Dôležitým ukazovateľom kompenzácie cukrovky je hladina glykovaného hemoglobínu (HbA1c). Ide o vyšetrenie žilovej krvi, ktoré podáva informáciu o priemernej hladine cukru v krvi za posledné 2 až 3 mesiace, a malo by byť pacientom vykonané aspoň 2 až 3 krát do roka. Jeho hladina sa pohybuje na intervale od 2% do 16%, pričom za uspokojivé výsledky sa považuje hladina do 7% [1].

1.3 Program DôveraPomáha diabetikom

Jedným zo základných prvkov starostlivosti o chronicky chorých pacientov s diabetom je edukácia. V západných krajinách Európy (najmä v Holandsku a Nemecku) je bežnou súčasťou tzv. disease management programov. Vplyv edukácií na zlepšenie kvality života pacientov sledovali rôzne štúdie. Zmena nastala u pacientov hlavne v informovanosti a znalosti vlastného ochorenia, ako aj v zlepšení klinických hodnôt, ako sú hladina glukózy v krvi a hodnota glykovaného hemoglobínu, ktoré sú najvýraznejšími špecifikátormi ochorenia. Takéto zmeny môžu viesť k predĺženiu života pacientov, predchádzaniu komplikácií a z hľadiska zdravotnej poisťovne majú význam aj v podobe ušetrenia nákladov za výkony, ktorými sa dá často jednoducho predchádzať. Práve pre takýto program sa rozhodla aj zdravotná poisťovňa Dôvera.

Individuálna edukácia na Slovensku funguje hlavne v ambulanciách, kde ich vykonávajú lekári v spolupráci so zdravotnými sestrami, no značne ich pri tom obmedzujú kapacitné a finančné prostriedky. Túto sa poisťovňa pokúsila posilniť hromadnými edukáciami pacientov vyškolenými edukátormi. Edukácie však nie sú jediným nosným prvkom programu. Okrem nich ponúka pacientom rôzne formy odmien za dodržiavanie liečebného režimu, možnosť objednať sa na vyšetrenie a kvalitnejšiu zdravotnú starostlivosť vďaka vylepšenému informačnému systému.

Do programu DôveraPomáha diabetikom sa zapojilo 21 z 35 oslovených diabetológov z dvoch regiónov Slovenska a za prvý rok sa doň prihlásilo vyše 2500 diabetikov. Cieľom programu je lepšia zdravotná starostlivosť pre pacientov, ktorá je však spojená so zvýšením nákladov pre poisťovňu, pričom v roku 2018 je plánované rozšírenie programu na celé Slovensko. Kvôli rentabilite a kvalite programu, musí tomuto kroku predchádzať dostatočne dobrá analýza a hodnotenie výsledkov z pilotnej fázy. Práve toto bude motiváciou praktickej časti našej práce v kapitole 4.

Kapitola 2

Data mining

Napriek tomu, že pojem *data mining* je relatívne nový, techniky, ktoré využíva poznáme už veľmi dlho. Postupný rozvoj v oblasti zbierania dát však so sebou priniesol potrebu združiť doposiaľ známe metódy z oblasti štatistiky, umelej inteligencie, či strojového učenia a vytvorenie akejsi metodiky práce s veľkým množstvom dát, hlavne na obchodné účely. Postupne bola spísaná a vytvorená metodika CRISP-DM (z angl. cross-industry process for data mining), o ktorej budeme bližšie pojednávať v tejto kapitole.

2.1 Dáta a informácie

Za dáta môžeme považovať akýkoľvek počítačovo zaznamenaný fakt alebo údaj. Vo všeobecnosti môžeme dáta radiť do troch kategórií:

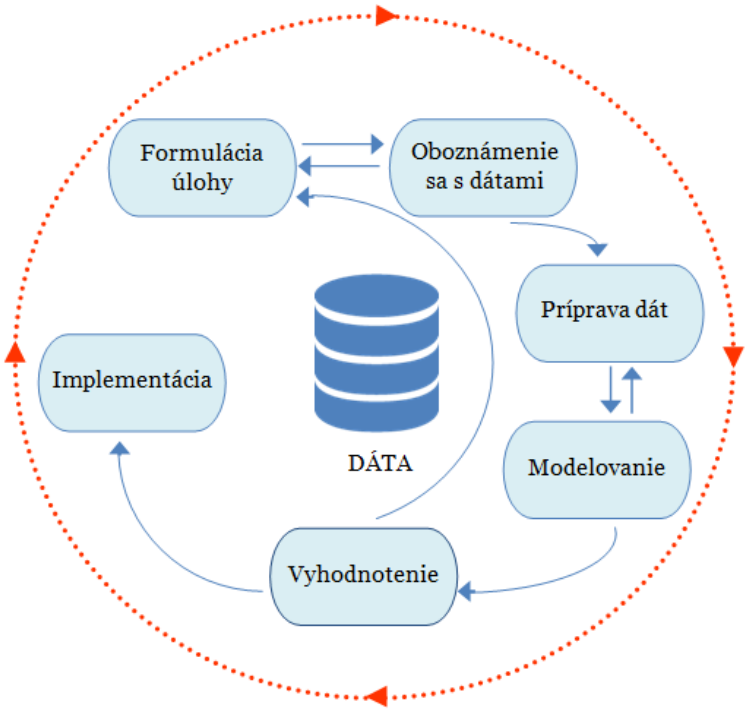
- *aktívne alebo transakčné dáta* - údaje o predaji, nákladoch, transakciách, inventári, mzdách, zúčtovaní a pod.
- *neaktívne dáta* - makroekonomické dáta, predpovede, údaje priemyselného odvetvia a pod.
- *meta dáta* - dáta o dátach samotných, t.j. dizajn databázy, definície pojmov a číselníky

Asociácie a spojitosti v dátach nám môžu poskytnúť informácie, ktoré môžeme pretaviť do znalostí a vzorcov o historickom vývoji, ako aj budúcich trendoch. Obrovský

pokrok v zachytávaní dát nám ponúka vytvárať centralizované databázy, kde môžeme dáta organizovať a neskôr analyzovať pomocou mnohých dataminingových techník.

2.2 Fázy data miningu

Podľa [8] je data mining definovaný ako hierarchický proces, ktorý je založený na štyroch úrovniach: fáza, základná úloha, podúloha a konkrétny príklad. To v praxi znamená, že v každej fáze vieme, čo by malo byť jej výsledkom. K nemu sa dostaneme pomocou bližších zadaní, konkrétnych úloh a ich asociácií. Celkovo vieme data mining rozdeliť na šesť základných fáz (viď Obr. 2.1).



Obr. 2.1: Fázy data miningu

2.2.1 Formulácia úlohy

Úvodná fáza slúži na definovanie cieľov projektu, hlavne z obchodného hľadiska. Dôkladná diskusia môže viesť k formulácií rôznych obmedzení prípadne faktorov, ktoré môžu ovplyvniť výstupy projektu. Výsledkom by mala byť formulácia problému, ktorý

možno vyriešiť nástrojmi data miningu. Mal by obsahovať tieto fázy:

- *Definícia obchodných cieľov* - pochopenie obchodného problému, ktorý chceme primárne riešiť
- *Analýza aktuálnej situácie* - zhodnotenie momentálneho stavu, všetkých súvisiacich informácií a faktorov, ktoré by mohli zohrávať rolu, prípadne rentabilnosť projektu
- *Stanovenie cieľov data miningu* - definícia cieľov projektu v technickej terminológii
- *Vypracovanie projektového plánu* - kroky vykonané v neskorších fázach projektu, vrátane výberu vstupov, výstupov, štatistických nástrojov či modelovacích techník

2.2.2 Oboznámenie sa s dátami

Po zadefinovaní si úlohy projektu a cieľov data miningu je dôležité určiť si, aké požiadavky máme na dáta. Tie je dôležité správne pochopiť a vytvoriť si prvotný náhľad, prípadne identifikovať základné vzorce. Nemali by preto chýbať tieto procesy:

- *Zhromaždenie dát* - vytvorenie datasetu zo zdrojov, ktoré máme k dispozícií
- *Klasifikácia dát* - overenie vlastností dát a vyhodnotenie ich relevantnosti vzhľadom k požiadavkám projektu
- *Skúmanie dát* - vytvorenie dotazov, vizualizácia, zoznamovanie sa s dátami pomocou jednoduchých agregácií a vlastností signifikantných podskupín
- *Overenie kvality dát* - chybovosť dát a ich zastúpenie v datasete

2.2.3 Príprava dát

Táto fáza vo väčšine prípadov zaberá najviac času práce na projekte. Po zoznámení sa s dátami a ich vlastnosťami, nasleduje selekcia potrebných údajov, vyčistenie dát a ďalšie aktivity potrebné ku konštrukcii datasetu v požadovanej forme:

- *Výber dát* - rozhodnutie o tom, ktoré dáta a rovnako aj parametre ideme zahrnúť do modelovania a technické ohraničenia, ako napríklad veľkosť databázy
- *Čistenie dát* - zvýšenie kvality dát na úroveň nutnú pri použití zvolených analytických techník a práca s chýbajúcimi, resp. znehodnotenými údajmi
- *Výstavba finálneho datasetu* - tvorba odvodených premenných alebo transformovanie už existujúcich parametrov
- *Zjednotenie dát* - spájanie a agregácia vytvorených tabuliek do zlúčených záznamov
- *Formátovanie* - modifikácie dát do formátu požadovaného na modelovanie nemeniace ich význam

2.2.4 Modelovanie

Pri modelovaní využívame rôzne modelovacie techniky, často aj niekoľko z nich, pričom sa snažíme optimalizovať hodnoty parametrov a tak získať čo najpresnejší výsledok. Pre potreby našej práce niektoré bližšie opisujeme v kapitole 3.

- *Voľba modelovacích techník* - ak sme už vybrali modelovaciu techniku v úvode, môžeme ju bližšie špecifikovať
- *Tvorba testu* - mechanizmus na testovanie kvality modelu, vytvorený z cvičných dát, ktoré vyberieme z existujúceho datasetu
- *Vytváranie modelu* - aplikácia zvolenej modelovacej techniky na pripravený dataset
- *Zhodnotenie modelu* - sumarizácia výsledkov modelu, hodnotenie jeho kvality a úspešnosti kritéria data miningu

2.2.5 Vyhodnotenie

Akonáhle máme model vytvorený je nutné model vyhodnotiť a posúdiť zvolené kroky. Účelom je zistiť, či sme dostatočne zvážili všetky obchodné aspekty. Po tejto fáze by

sme mali byť schopní rozhodnúť, ako budeme problém pomocou výsledkov získaných data miningom riešiť.

- *Vyhodnotenie výsledkov* - narozdiel od predchádzajúceho hodnotenia, zameraného hlavne na presnosť modelu, v tejto časti posudzujeme, či je splnenie obchodných cieľov dostatočné

- *Preskúmanie procesu* - zváženie existencie nejakých významných faktorov, ktoré sme prehliadli, resp. nezobrali do úvahy alebo možnosť využitia daných atribútov aj pri ďalších analýzach

- *Definovanie ďalších krokov* - na základe predchádzajúcich výsledkov rozhodneme o implementácii, realizácii ďalších iterácií alebo vytvorení novej dataminingovej úlohy

2.2.6 Implementácia

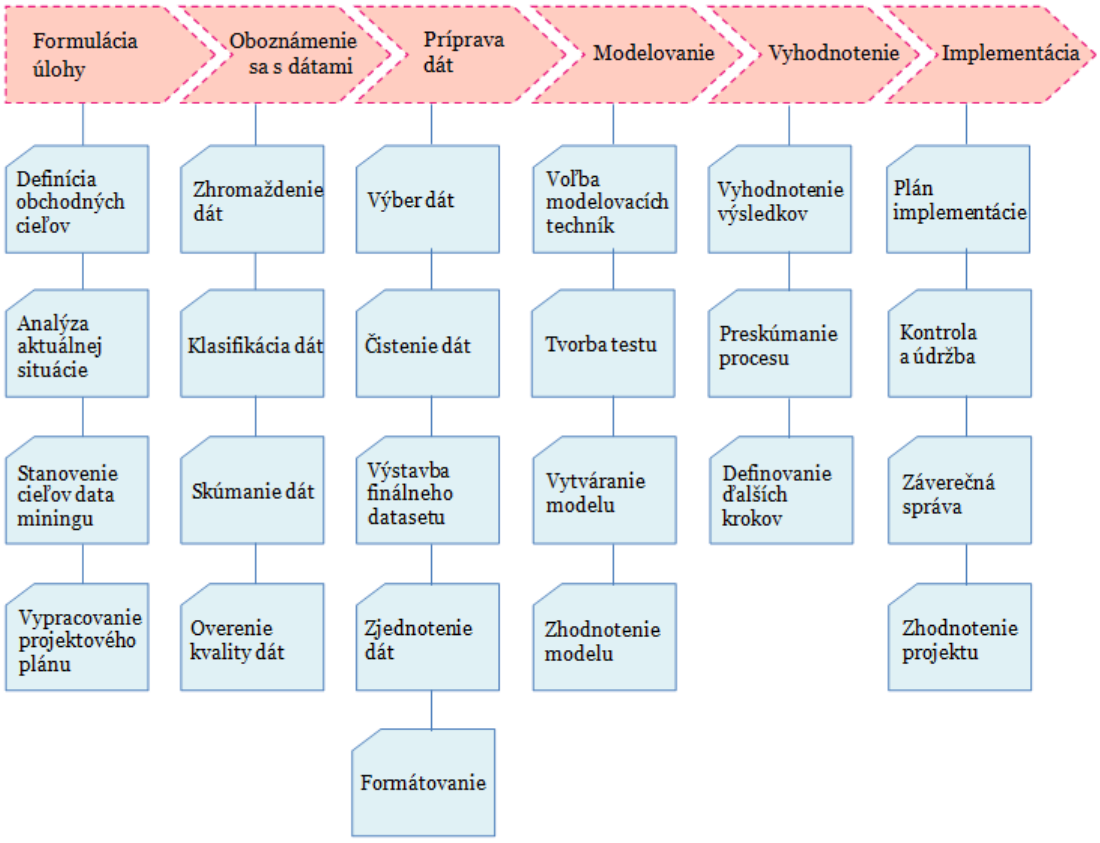
Správnym modelom data mining zväčša nekončí. Keďže analytici a tvorcovia modelov sa obyčajne nedostanú až k tvorbe opatrení, je veľmi dôležité použité postupy a výsledky vysvetliť. Tie musia byť odprezentovateľné, aby ich konečný zákazník vedel využiť. Implementácia je rozhodujúca fáza z hľadiska úspešnosti projektu.

- *Plán implementácie* - definovanie stratégie a z nej vyplývajúcich krokov v obchodnej terminológii

- *Kontrola a údržba* - prevencia nesprávneho využívania modelu a dezinterpretácia výsledkov data miningu používaných v bežnej praxi

- *Záverečná správa* - sumarizácia projektu a jeho prínosov, poznatkov, získaných počas práce a vyhodnotenie výsledkov

- *Zhodnotenie projektu* - zhodnotenie kladov a záporov, ktoré procesy prebehli v poriadku, a ktoré by sa naopak mali zlepšiť



Obr. 2.2: Procesy data miningu

Kapitola 3

Metódy a modelovacie techniky

V tejto kapitole teoreticky popisujeme niektoré štatistické metódy a modelovacie techniky, ktoré v našej práci používame. Prvou z nich sú *DEA modely*, ktoré v praktickej časti využívame na výpočet efektivity, pričom vychádzame hlavne z [9]. Druhou je *analýza zhlukov*, opísaná pomocou [11].

3.1 DEA modely

3.1.1 Základné pojmy

DEA metóda (z angl. *Data Envelopment Analysis*) alebo obálková analýza dát je prostriedok ekonomického manažmentu, hodnotiaci relatívnu efektívnosť vzhľadom na skupinu producentov riadiacich sa tou istou *technológiou*. Technológia T je definovaná m vstupmi $I_1, ...I_m$, ktoré producent spotrebováva a s výstupmi $O_1, ...O_s$, ktoré produkuje. Ich výhodou je, že môžu byť merané v rôznych jednotkách. Keďže sa producenti pri premene vstupov na výstupy rozhodujú samostatne, označujeme ich DMU_j (Decision Making Unit) pre $j = 1, ...n$. Hodnoty vstupov technológie T pre DMU_j označujeme vektorom $x_j \in R_m$ a hodnoty výstupov $y_j \in R_s$. Dvojica (x, y) je pre technológiu T prípustná, ak je z množstva vstupov x možné vyrobiť určité množstvo výstupov y. Množina všetkých takýchto dvojíc sa nazýva *množina prípustných možností* G a je konvexná. *Produkčnou funkciou* f nazývame funkciu, ktorá každej kladnej hodnote vstupu x priradí maximálnu možnú hodnotu výstupu y. Takáto funkcia je rastúca a definuje časť hranice množiny G, ktorú nazývame *hranica efektívnosti*.

- **Efektívnosť**: Pre dvojicu (x, y) platí, že $y \leq f(x)$. Producent je *efektívny* v prípade, že $y = f(x)$ a teda pri danom vstupe produkuje najväčší výstup. Naopak v prípade *neefektívneho* producenta platí $y < f(x)$.

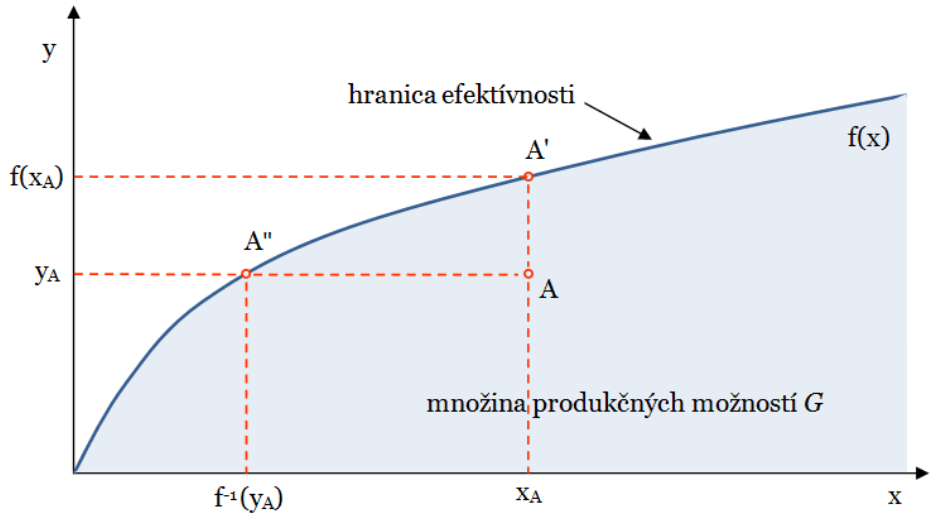
- **Efektivita**: Mieru neefektívnosti môžeme nazvať aj *efektivita* E. Určíme ju pomerom produkcie vzhľadom k efektívnym producentom a teda platí $E \in (0, 1]$, pričom $E = 1$ pre každého efektívneho producenta.

Pri orientovaných modeloch rozlišujeme viac druhov efektivít. Pri výstupne orientovanom modeli sa snažíme maximalizovať hodnotu výstupu pri danej hodnote vstupu. To znamená, že *výstupná efektivita* je definovaná pomerom výstupu útvaru A a maximálnou hodnotou výstupu pri použití danej technológie na vstup x_A:

$$E_{vyst}(A) = \frac{y_A}{f(x_A)}$$

Naopak pri vstupne orientovanom modeli minimalizujeme objem vstupov, za ktoré je možné vyrobiť dané množstvo výstupov. Do podielu teda dávame hodnotu x_A a najmenšiu hodnotu vstupu, za ktoré sa da vyrobiť hodnota výstupu y_A:

$$E_{vst}(A) = \frac{f^{-1}(y_A)}{x_A}$$



Obr. 3.1: Množina produkčných možností. Hranica efektívnosti. Projekcia bodu A.

V oboch prípadoch robíme projekciu bodu A na graf produkčnej funkcie $f(x)$ (Obr. 3.1). Projektovaním najprv v smere $(0, 1)$ a neskôr v smere $(-1, 0)$ získame útvar,

ktorý sa nachádza na hranici efektívnosti a budeme ho preto nazývať *efektívny vzor* A' resp. A''. Viacrozmerný prípad je analógiou jednorozmerného. V skutočnosti však nepoznáme reálny tvar produkčnej funkcie a teda ani hranice efektívnosti. Tu však využívame množinu produkčných možností G vytvorenú z našich DMU. Pomocou nich je možné vytvoriť akúsi obálku, na základe ktorej vieme efektívnu hranicu odhadnúť a dopočítať tak teda aj efektivitu pre jednotlivé DMU. Platí, že čím viac útvarov v obálkovom modeli máme, tým sú naše výsledky presnejšie.

3.1.2 Aproximácie množiny produkčných možností

Podľa [9] sú definované axiomatické vlastnosti produkčných množín. Označme M odhadnutú množinu produkčných možností. Vo všeobecnom prípade predpokladáme **variabilné výnosy z rozsahu**. Pod M_{VRS} (Variable Returns to Scale) rozumieme najmenšiu množinu pre ktorú platí, že obsahuje všetky dvojice (x_j, y_j) pre $j = 1, ...n$, je konvexná a zahŕňa všetky menej efektívne útvary (to znamená, že ak do množiny patrí útvar s určitými vstupmi a výstupmi, bude do nej patriť aj útvar, ktorý vyrába pri väčších vstupoch alebo produkuje menej výstupov):

$$M_{VRS} = \{(x, y) \in \mathbb{R}^{m+s} | \quad \sum_{j=1}^{n} \lambda_j x_j \leq x, \quad \sum_{j=1}^{n} \lambda_j y_j \geq y, \quad \lambda \geq 0, \quad \sum_{j=1}^{n} \lambda_j = 1\}$$

Ak uvažujeme situáciu s **konštantnými výnosmi z rozsahu** musí platiť navyše axióma, ktorá hovorí o tom, že ak bod $(x, y) \in M$, potom aj všetky body ležiace na polpriamke, spájajúcej začiatok súradnicovej sústavy s daným bodom patria do množiny, teda $(cx, cy) \in M, \quad \forall c > 0$. Dostávame tak množinu M_{CRS} (Constant Returns to Scale):

$$M_{CRS} = \{(x, y) \in \mathbb{R}^{m+s} | \quad \sum_{j=1}^{n} \lambda_j x_j \leq x, \quad \sum_{j=1}^{n} \lambda_j y_j \geq y, \quad \lambda \geq 0\}$$

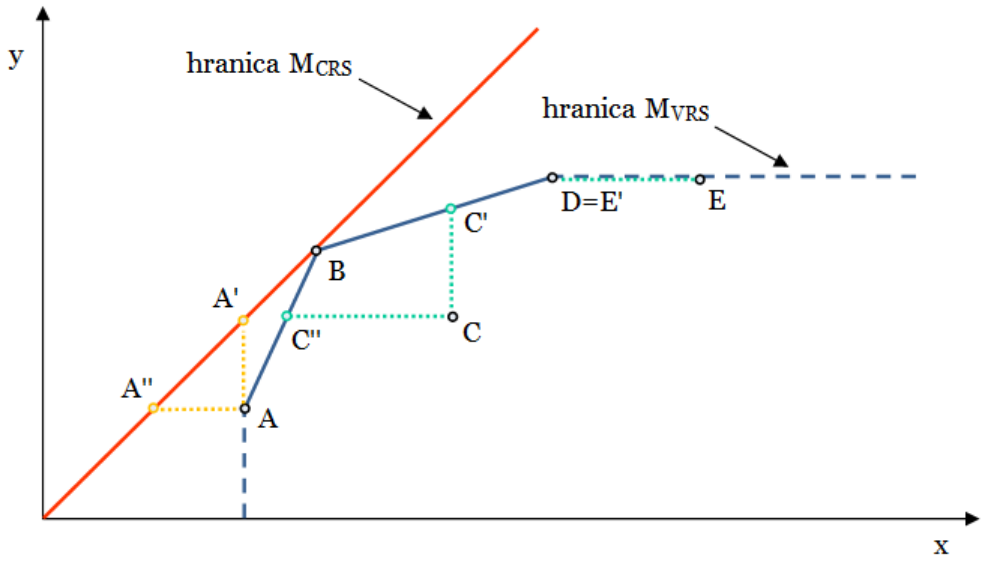
Analogicky potom aproximáciou skutočnej hranice efektívnosti na množine G bude hranica množiny M, totožná s odhadnutou produkčnou funkciou. Takto odhadnutá efektívna hranica ∂M pozostáva z dvoch častí $\partial M = H_E \cup H_P$:

- H_E predstavuje *hranicu efektívnosti* pozostávajúcu z efektívnych bodov ležiacich na hranici:

$$H_E := \{(x, y) \in \partial M \mid \nexists(\bar{x}, \bar{y}) \in M : (\bar{x}, \bar{y}) \neq (x, y), \bar{x} \leq x, \bar{y} \geq y\}$$

- H_P predstavuje *hranicu pseudoefektívnosti* pozostávajúcu z neefektívnych bodov ležiacich na hranici:

$$H_P := \{(x, y) \in \partial M \mid \exists(\bar{x}, \bar{y}) \in M : (\bar{x}, \bar{y}) \neq (x, y), \bar{x} \leq x, \bar{y} \geq y\}$$



Obr. 3.2: Hranice M_{CRS} a M_{VRS}

Vysvetlime si princípy na Obrázku 3.2. Vezmeme si prípad, že máme piatich producentov, ktorí tvoria našu aproximáciu produkčnej množiny. Pri variabilných výnosoch z rozsahu považujeme za **efektívnych** producentov A, B a D. Úsečky, ktoré ich spájajú sú aproximáciou hranice efektívnosti (znázornené plnou modrou čiarou). Bod E je **pseudoefektívny**, pretože sa nachádza hranici pseudoefektívnosti (znázornené prerušovanou modrou čiarou). Jednoducho povedané, existuje producent (v našom prípade D), ktorý je rovnaké množstvo výstupov pri rovnakej technológii T schopný vyprodukovať pri menšom počte vstupov. Útvar C je **neefektívny**. Príslušnou projekciou bodov E do E' a C do C', resp. C do C'' potom môžeme zmerať ich vstupnú resp. výstupnú efektivitu. Z opačného pohľadu sa dá na efektivitu pozerať aj ako na koeficient projekcie na efektívnu hranicu.

Jednoduchšia situácia nastáva pri konštantných výnosoch z rozsahu. Skutočná hranica efektívnosti má v tomto prípade tvar polpriamky vychádzajúcej zo začiatku súradnicovej sústavy. Keďže jej skutočnú smernicu k nepoznáme, vezmeme si za jej aproximáciu polpriamku spájajúcu začiatok sústavy s bodom s najväčšou smernicou (znázornená

červenou čiarou). Efektívnym je v takom prípade producent B a efektivitu ostatných producentov môžeme vypočítať podobne, ako v prvom prípade (na obrázku len pre bod A).

3.1.3 CCR a BBC modely

Teória popísaná v predchádzajúcich častiach tvorí základ pre rôzne DEA modely. Vo všeobecnosti delíme tieto modely na orientované a neorientované. *Orientované modely* sa využívajú pri zameraní na vstupy alebo výstupy. V prípade vstupného modelu normalizujeme vstupy, to znamená, že pri nezmenených hodnotách výstupov sa snažíme hodnoty vstupov minimalizovať. Naopak pri výstupných modeloch normalizujeme výstupy, a teda ich maximalizujeme pri daných hodnotách vstupov. V praxi to znamená, že sa snažíme buď o zníženie nákladov alebo zvýšenie produkcie. Orientáciu rozlišujeme príponou $-I(input)$ pre vstupné modely a $-O(output)$ pre výstupné modely.

CCR model (z iniciálok autorov - Charnes, Cooper a Rhodes, 1978) je historicky prvý využívaný DEA model. Využíva sa pri meraní efektivity producentov, ktorých technológia odpovedá konštantným výnosom z rozsahu pri množine produkčných možností M_{CRS}. Poznáme jeho dve podoby, pod ktorým rozumieme úlohy pre útvar DMU_o v tvare úloh lineárneho programovania:

$(CCR-I)_o$

$$\min_{\theta,\lambda} \quad \theta \tag{3.1}$$

$$s.t. \sum_{j=1}^{n} x_j\lambda_j \leq \theta x_o, \tag{3.2}$$

$$\sum_{j=1}^{n} y_j\lambda_j \geq y_o, \tag{3.3}$$

$$\lambda \geq 0_n. \tag{3.4}$$

$(CCR-O)_o$

$$\max_{\psi,\lambda} \quad \psi \tag{3.5}$$

$$s.t. \sum_{j=1}^{n} x_j\lambda_j \leq x_o, \tag{3.6}$$

$$\sum_{j=1}^{n} y_j\lambda_j \geq \psi y_o, \tag{3.7}$$

$$\lambda \geq 0_n. \tag{3.8}$$

BCC model (z iniciálok autorov Banker, Charnes, Cooper, 1984) je analógiou CCR modelov pre variabilné výnosy z rozsahu s odpovedajúcou množinou M_{VRS}. Práve z podmienky navyše pri definícií množiny produkčných možností M_{VRS} sme získali pridané ohraničenia $\sum_{j=1}^{n} \lambda_j = 1$:

$(BCC-I)_o$ $(BCC-O)_o$

$$\min_{\theta,\lambda} \quad \theta \tag{3.9}$$

$$s.t. \sum_{j=1}^{n} x_j\lambda_j \leq \theta x_o, \tag{3.10}$$

$$\sum_{j=1}^{n} y_j\lambda_j \geq y_o, \tag{3.11}$$

$$\sum_{j=1}^{n} \lambda_j = 1, \lambda \geq 0_n. \tag{3.12}$$

$$\max_{\psi,\lambda} \quad \psi \tag{3.13}$$

$$s.t. \sum_{j=1}^{n} x_j\lambda_j \leq x_o, \tag{3.14}$$

$$\sum_{j=1}^{n} y_j\lambda_j \geq \psi y_o, \tag{3.15}$$

$$\sum_{j=1}^{n} \lambda_j = 1, \lambda \geq 0_n. \tag{3.16}$$

Vysvetlime si teraz postup riešenia, ak uvažujeme model $(CCR-I)$. Zavedením a doplnením nezáporných doplnkových premenných s^x a s^y do nerovností v jednotlivých ohraničeniach ich transformujeme na rovnice. Takéto premenné nazývame rezervy alebo *slacky*. K podmienkam (3.4), (3.8), (3.12) a (3.16) tak ešte pribudnú podmienky $s^x \geq 0_m$ a $s^y \geq 0_s$. Pre každý útvar DMU_o riešime samostatnú úlohu, teda dokopy n úloh a hodnoty účelových funkcií označujeme θ_o^*. Získavame tak riešenie θ^*, λ^*, s^{x*} a s^{y*}. Našou úlohou je maximalizovať súčet rezerv, resp. slackov pre každého producenta pri optimálnej hodnote účelovej funkcie. Takúto úlohu riešime na dve etapy. V prvej etape si pomocou príslušného modelu $CCR-I$ vypočítame θ^*. Tú využívame v druhej etape, kde maximalizujeme súčet slackov nasledovne:

$(CCR-I-S)_o$

$$\min_{\lambda,s^x,s^y} \quad H := e^T s^x + e^T s^y \tag{3.17}$$

$$s.t. \sum_{j=1}^{n} x_j\lambda_j + s^x = \theta^* x_o, \tag{3.18}$$

$$\sum_{j=1}^{n} y_j\lambda_j - s^y = y_o, \tag{3.19}$$

$$\lambda, s^x, s^y \geq 0. \tag{3.20}$$

Dostávame tak zvyšné optimálne hodnoty λ^*, s^{x*} a s^{y*}. Na základe získaných výsledkov môžeme interpretovať riešenie, ktoré sa zhoduje aj s jednoduchými logickými úvahami. Mieru neefektívnosti, resp. efektivitu, definujeme ako optimálnu hodnotu účelovej funkcie a na základe hodnoty slackov vieme rozlíšiť medzi efektivitou a pseudoefektivitou. Efektivitu pre optimálne riešenie úlohy $CCR-I$ definujeme nasledovne:

- Ak platí $(s^{x*}, s^{y*}) = 0$ a $\theta_o^* = 1$, útvar DMU_o je **efektívny**, pričom platí, že **efektivita** $\theta_o^* = 1$.

- Ak platí $(s^{x*}, s^{y*}) = 0$ a $\theta_o^* < 1$, útvar DMU_o je **neefektívny**, pričom platí, že jeho **efektivita** je rovná θ_o^*.

- Ak platí $s^{x*} \neq 0$ alebo $s^{y*} \neq 0$ (to znamená, že niektorý zo slackov je kladný) a $\theta_o^* = 1$, útvar DMU_o je **pseudoefektívny** a teda neefektívny, pričom platí, že **pseudoefektivita** $\theta_o^* = 1$.

- Ak platí $s^{x*} \neq 0$ alebo $s^{y*} \neq 0$ a $\theta_o^* < 1$, útvar DMU_o je **neefektívny**, pričom platí, že jeho **pseudoefektivita** je rovná θ_o^*.

Jednoducho povedané ak $\theta_o^* = 1$, tak daný útvar DMU_o leží na hranici množiny M. Ak by však mohol ešte znížiť hodnotu vstupov pri nezmenených výstupoch (resp. naopak) a teda producent má ešte rezervy, prejaví sa to na kladnosti jedného zo slackov. To je prípad pseudoefektívneho producenta. Podobne to platí pre neefektívne útvary a ich projekciu na hranicu M.
Analogicky by sme postupovali aj pri odvodení ostatných modelov. Pre $BCC - I$ vstupný model by bol postup totožný (v druhej etape by sme modifikovali rovnice (3.18) vypustením θ^* a (3.19) pridaním ψ^*) s rovnakou interpretáciou optimálneho riešenia. Pre výstupné modely $CCR{-}O$ a $BCC{-}O$ je už situácia trochu iná. Kým pri vstupných modeloch optimálna hodnota θ^* predstavovala koeficient projekcie, ktorý projektoval na hranicu efektívnosti skracovaním vstupov, pri výstupných modeloch ho nahradí koeficient ψ^*, ktorý projektuje predlžovaním výstupov. Logicky potom nenadobúda hodnoty patriace intervalu $(0, 1)$, ale práve naopak $\psi^* \geq 1$. Príslušne k tomu sa zmení aj interpretácia efektívnych a neefektívnych útvarov :

- Ak platí $(s^{x*}, s^{y*}) = 0$ a $\psi_o^* = 1$, útvar DMU_o je **efektívny**, pričom platí, že **efektivita** $\psi_o^* = 1$.

- Ak platí $(s^{x*}, s^{y*}) = 0$ a $\psi_o^* > 1$, útvar DMU_o je **neefektívny**, pričom platí, že jeho **efektivita** je rovná $\frac{1}{\psi_o^*}$.

- Ak platí $s^{x*} \neq 0$ alebo $s^{y*} \neq 0$ (to znamená, že niektorý zo slackov je kladný) a $\psi_o^* = 1$, útvar DMU_o je **pseudoefektívny** a teda neefektívny, pričom platí, že **pseudoefektivita** $\psi_o^* = 1$.

- Ak platí $s^{x*} \neq 0$ alebo $s^{y*} \neq 0$ a $\psi_o^* > 1$, útvar DMU_o je **neefektívny**, pričom platí, že jeho **pseudoefektivita** je rovná $\frac{1}{\psi_o^*}$.

Z týchto poznatkov nám vyplýva niekoľko súvislostí. Platí, že vstupné a výstupné modely definujú rovnaké efektívne útvary pri uvažovaní rovnakej produkčnej množiny. Keďže $M_{VRS} \subset M_{CRS}$, efektívne útvary pri konštantných výnosoch z rozsahu sú vždy efektívne aj v prípade variabilných výnosoch z rozsahu, no naopak toto tvrdenie neplatí. Ďalej vieme povedať, že pri $CCR-I$ a $CCR-O$ modeloch pre dané DMU_o je medzi optimálnymi hodnotami θ^* a ψ^* vzťah $E^* = \theta^* = \frac{1}{\psi^*}$ (toto však neplatí v prípade BBC modelov).

3.1.4 Efektívny vzor

Konečným riešením obálkovej analýzy dát nie je len výpočet efektivity v prislúchajúcom modeli. Dôležitým krokom je určenie *efektívneho vzoru* pre producenta. Pre vstupne orientované modely a im príslušné θ^* definujeme efektívne vzory nasledovne:

$$x_o^{\lambda*} := \sum_{j=1}^{n} x_j \lambda_j^* = \theta^* x_o - s^{x*}, \quad y_o^{\lambda*} := \sum_{j=1}^{n} y_j \lambda_j^* = y_o + s^{y*}$$

Podobne tak vieme odvodiť efektívne vzory aj pre výstupné modely:

$$x_o^{\lambda*} := \sum_{j=1}^{n} x_j \lambda_j^* = x_o - s^{x*}, \quad y_o^{\lambda*} := \sum_{j=1}^{n} y_j \lambda_j^* = \psi^* y_o + s^{y*}$$

To že je efektívny vzor skutočne efektívny sme zabezpečili v dvoch etapách. V prvej sme ho posunuli na hranicu množiny produkčných možností a v druhej etape sme overili, či sa nachádza na hranici efektívnosti a pokiaľ nie, tak sme ho pomocou slackov ešte vylepšili. Ako však reálne pomôcť producentovi dostať sa na hranicu efektívnosti? Ako rada slúži tzv. *referenčná množina*, ktorá je definovaná ako:

$$RS_o = \{j \mid \lambda_j > 0\}$$

V podstate predstavuje skupinu útvarov, ktoré slúžia danému producentovi ako efektívne vzory a príslušné koeficienty λ_j tvoria stratégiu, pomocou ktorej sa k nim dopracovať.

3.1.5 Aditívny a SBM model

V prípade, že nemáme preferenciu medzi minimalizáciou vstupov a maximalizáciou výstupov sa využívajú *neorientované modely*. Najznámejším je tzv. **aditívny model**, pod ktorým rozumieme v prípade konštantných výnosov z rozsahu úlohu v tvare:

$(AD - CRS)_o$

$$\min_{\lambda, s^x, s^y} \quad -(e_m^T s^x + e_s^T s^y) \tag{3.21}$$

$$s.t. \sum_{j=1}^{n} x_j \lambda_j + s^x = x_o, \tag{3.22}$$

$$\sum_{j=1}^{n} y_j \lambda_j - s^y = y_o, \tag{3.23}$$

$$\lambda, s^x, s^y \geq 0. \tag{3.24}$$

pričom optimálnu hodnotu účelovej funkcie označíme A_o^*. Pridaním ohraničenia $\sum_{j=1}^{n} \lambda_j = 1$ získavame tvar aditívneho modelu pre variabilné výnosy z rozsahu $(AD - VRS)_o$. Pre efektivitu potom platí, že:

- ak je optimálna hodnota účelovej funkcie $A_o^* = 0$, útvar DMU_o je **efektívny**,
- ak je hodnota $A_o^* \neq 0$, je útvar DMU_o **neefektívny**.

Účelová funkcia úlohy pri aditívnom modeli má vlastnosť monotónnosti a dosahuje hodnoty z intervalu $(-\infty, 0]$. Platí však, že nevyjadruje efektivitu a preto tú musíme pre ďalšie potreby dopočítať na základe analógie s CCR a BCC modelmi. Najväčšou výhodou aditívneho modelu je, že dokáže súčasne zohľadňovať neefektivitu na strane vstupov, ako aj výstupov. To má však za následok, že výsledky modelu sú závislé od jednotiek, v ktorých meriame vstupy a výstupy (tento nedostatok sa často odstraňuje použitím vektorov váh).

Spomenuté nedostatky aditívneho modelu boli motiváciou na sformulovanie **SBM modelu**. Oproti aditívnemu modelu má inú podobu iba jeho účelová funkcia, ktorá je

o čosi zložitejšia, no má zachovanú dôležitú vlastnosť klesajúcosti vzhľadom na rezervy (slacky) s_x a s_y. Odtiaľ pochádza aj jeho názov *SBM - Slack Based Measure*. Pod základným tvarom SBM modelu rozumieme podľa úlohu:

$(Z - SBM)_o$

$$\min_{\lambda, s^x, s^y} \quad \rho := \frac{1 - \frac{1}{m}\sum_{i=1}^{m} \frac{s_i^x}{x_{io}}}{1 + \frac{1}{s}\sum_{r=1}^{s} \frac{s_r^y}{y_{ro}}} \tag{3.25}$$

$$s.t. \sum_{j=1}^{n} x_j \lambda_j + s^x = x_o, \tag{3.26}$$

$$\sum_{j=1}^{n} y_j \lambda_j - s^y = y_o, \tag{3.27}$$

$$\lambda, s^x, s^y \geq 0. \tag{3.28}$$

Účelová funkcia z tejto úlohy už nadobúda hodnoty $\rho \in [0, 1]$, pričom:

$$\rho = 1 \quad \Leftrightarrow \quad s_x = 0, s_y = 0.$$

Opäť platí, že takto naformulovaná úloha zodpovedá konštantným výnosom z rozsahu, a pridaním podmienky $\sum_{j=1}^{n} \lambda_j = 1$ získavame tvar pre variabilné výnosy z rozsahu. Jednotlivé DMU_o sa snažíme sprojektovať v smere $(-s_x, s_y)$. Ak sa bod nachádza na hranici, slacky sú nulové a teda aj hodnota účelovej funkcie $\rho = 1$. Naopak ak sa útvar nenachádza na efektívnej hranici, jeho slacky sú kladné a hodnota $\rho \leq 1$. V SBM modeli teda definujeme *SBM-efektivitu* ako optimálnu hodnotu účelovej funkcie ρ^* pričom platí:

- ak je $\rho^* = 1$ útvar DMU_o je **SBM-efektívny**
- ak je $\rho^* \leq 1$ útvar DMU_o je **SBM-neefektívny**

Takto vypočítaná efektivita je menšia, nanajvýš rovná efektivite získanej z orientovaných modelov. Nedostatky aditívneho modelu sa teda podarilo odstrániť. Problémom základného SBM modelu ale je, že účelová funkcia je nelineárna. Je ju preto ešte potrebné previesť do tvaru úlohy lineárneho programovania, ktorá vyzerá nasledovne:

$$\min_{\Lambda,t,S^x,S^y} \quad \tau := t - \frac{1}{m}\sum_{i=1}^{m}\frac{S_i^x}{x_{io}} \tag{3.29}$$

$$s.t. \quad 1 + \frac{1}{s}\sum_{r=1}^{s}\frac{S_r^y}{y_{ro}} = 1 \tag{3.30}$$

$$\sum_{j=1}^{n} x_j\Lambda_j + S^x = tx_o, \tag{3.31}$$

$$\sum_{j=1}^{n} y_j\Lambda_j - S^y = ty_o, \tag{3.32}$$

$$\Lambda, S^x, S^y \geq 0. \tag{3.33}$$

3.2 Analýza zhlukov

Dôležitým nástrojom na zoskupovanie dát na základe podobnosti (resp. odlišnosti) je **analýza zhlukov** (z angl. *cluster analysis*). Jej základom je rozdelenie n objektov do k zhlukov tak, aby boli podobnosti vrámci jednej skupiny a odlišnosti medzi skupinami čo najväčšie. Súčasne musí platiť, že každý objekt je zaradený presne do jednej skupiny a žiadna skupina nie je byť prázdna. Na takéto zatriedenie existuje viacero spôsobov, ktoré sa vo všeobecnosti dajú rozdeliť do dvoch kategórií:

- **nehierarchické metódy** - tvoria skupinu metód, ktoré rozdelia množinu objektov do neprekrývajúcich sa podmnožín (clustrov), pričom každý objekt patrí práve do jednej podmnožiny a ich počet k je daný
- **hierarchické metódy** - dovoľujú vytváranie tzv. vnorených clustrov, ktoré spolu tvoria hierarchiu objektov, reprezentovanú **dendogramom** (stromom podobnosti)

3.2.1 Metódy K-means a K-medoids

K-means

Táto metóda patrí medzi prototypové techniky. To znamená, že sa snaží nájsť najlepšieho reprezentanta pre daný zhluk. V metóde K-means sa tento reprezentant nazýva **centroid** a najčastejšie je definovaný ako priemer hodnôt vektorov čŕt jednotlivých

objektov v zhluku. Takto získaný bod zvyčajne nekorešponduje so žiadnym objektom. Cieľom metódy je minimalizovať súčet štvorcov vzdialeností objektov od ich centroidov

$$\sum_{i=1}^{k} \sum \rho^2(x_r, c_i)$$

kde

$$c_i = \frac{1}{C_i} \sum_{r \in C_i} x_r$$

je centroid pre zhluk C_i a $\rho(x, y)$ je Euklidovská vzdialenosť

$$\rho(x, y) = \sqrt{\sum_{t=1}^{p} (x_t - y_t)^2}$$

kde x_t a y_t sú t komponenty p rozmerných vektorov čŕt x, y.

Existuje niekoľko algoritmov na získanie optimálnej hodnoty takejto funkcie, pričom najznámejším je tzv. *Lloydov algoritmus*. Pracuje so zvolenými iniciálnymi centroidmi. Následne priradí každý objekt k jeho najbližšiemu centroidu a pre každý zhluk opätovne vypočíta nový centroid s danými objektmi. Takto pokračuje algoritmus kým nepríde do stavu, že centroidy sa nemenia a teda pre žiadny objekt sa nezmení zhluk, do ktorého je priradený. Je zjavné, že rôzne zvolené iniciálne centroidy vedú k rôznym výsledným zhlukom. Preto sa táto metóda obvykle opakuje niekoľko krát s náhodne zvolenými iniciálnymi centroidmi. Jej nevýhodou je, že závisí od jednotiek, v akých meriame jednotlivé parametre vektorov čŕt, nakoľko pracuje so vzdialenosťami euklidovských vektorov.

K-medoids

Metóda K-medoids je veľmi podobná metóde K-means, namiesto centroidov však využíva **medoidy**. Tie predstavujú reprezentantov v podobe konkrétnych objektov, t.j. stredom zhluku je jeden z objektov. Narozdiel od K-means, cieľom tejto metódy je nájdenie takých zhlukov $C_1, .., C_k$, ktoré minimalizujú funkcie vzdialeností $d(r, m_i)$ objektov od ich medoidov m_i:

$$min \quad \sum_{i=1}^{k} \sum_{r \in C_i} d(r, m_i)$$

Opäť existuje viacero spôsobov na výpočet optimálneho riešenia a jedným z nich je algoritmus PAM (z angl. *Partitioning Around Medoids*). Prvým krokom je zvolenie k

objektov za iniciálne medoidy. Všetky ostatné objekty potom priradí k ich najbližším medoidom a vypočíta hodnotu účelovej funkcie. Pre všetky páry objektov (x_s, m_i) (kde x_s je objekt, ktorý nie je medoid) sa snaží zlepšiť hodnotu účelovej funkcie tým, že zvolí x_s za nového reprezentanta. V prípade, že pre daný zhluk nájde lepší medoid, algoritmus zopakuje pridelením všetkých prvkov k ich najbližším medoidom dovtedy, kým sa rozdelenie do zhlukov neprestane meniť. Aj táto metóda sa zvyčajne opakuje s rôznymi počiatočnými medoidmi pre dosiahnutie minimálnej hodnoty účelovej funkcie.

Metóda K-medoids má oproti K-means výhodu v lepšej interpretácií reprezentantov a je menej citlivá na oulayerov.

Kapitola 4

Benchmarking poskytovateľov zdravotnej starostlivosti

V úvode sme definovali benchmarking ako komplexný proces porovnávania rôznych atribútov subjektov, vykonávajúcich rovnakú, resp. podobnú činnosť. Podľa výberu subjektov môžeme klasifikovať *vnútorný* benchmarking, ako skóring útvarov vrámci organizácie a *vonkajší* benchmarking pri porovnávaní s inými podnikmi v rovnakom odvetví [13].

Benchmarking ako proces zahŕňa niekoľko krokov:

- definovanie predmetu benchmarkingu (služba, resp. činnosť)
- identifikácia referenčnej skupiny, ktorú budeme porovnávať
- zbieranie a analyzovanie dát činností a procesov
- porovnanie výkonnosti a definícia cieľov na zlepšenie
- vytvorenie akčného plánu a monitorovanie výsledkov

Cieľom tejto kapitoly je vnútorný benchmarking lekárov zapojených do programu DôveraPomáha diabetikom, pričom predmetom porovnávania sú činnosti vykonávané počas výkonu liečby.

4.1 Dáta a základné pojmy

Základným rámcom dát, z ktorého vychádzame, sú dáta o vykázanej zdravotnej starostlivosti (ďalej VZS). Obsahujú všetku zdravotnú starostlivosť, ktorá bola poistencom vykázaná s informáciami, kto zdravotnú starostlivosť vykonal, prípadne ju odporučil, resp. odoslal pacienta na vyšetrenie, dátumom, kódom diagnózy, cenou a mnohými inými. Fyzické a právnické osoby, ktoré tak môžu urobiť, nazývame poskytovateľmi zdravotnej starostlivosti (ďalej PZS). Na spracovanie dát využívame metódy opísané v kapitole 2 a 3 a na základe výsledkov zostavujeme porovnanie. Podrobnejšie tieto kroky opisujeme v nasledujúcich podkapitolách.

4.2 Definícia účelu benchmarkingu

Diabetológovia zapojení v programe DôveraPomáha diabetikom patria do siete špecializovanej ambulantnej starostlivosti. Ambulantní špecialisti sú štandardne platení za počet výkonov vykazovaných poisťovni, ktorá im následne výkony prepláca. Tie sú však ohraničené tzv. limitmi. Ako benefit poisťovňa (okrem odmeny za prácu v programe) ponúka tzv. bezlimitné prostredie. Zapojenie lekára do programu znamená potenciálne zvýšenie nákladov. Existuje preto prirodzená motivácia hodnotenia diabetológov, ktorá by napomáhala rozhodnutiu, či konkrétneho diabetológa zapojiť alebo nie.

4.3 Parametre hodnotenia

Odpoveď na otázku, ktorý lekár je lepší, a ktorý horší, vôbec nie je jednoduchá. Dá sa na ňu totižto pozerať z viacerých perspektív. Z hľadiska pacienta je hlavným kritériom samozrejme úspešnosť pri liečbe jeho ochorenia. Z pohľadu zdravotnej poisťovne vstupujú do úvahy aj náklady na zdravotnú starostlivosť, s ktorými musí zaobchádzať rozumne. Pri liečbe cukrovky sú však tieto dva parametre výrazne prepojené, nakoľko včasná diagnostika a kontrolovaný priebeh ochorenia indikujú šetrenie nákladov v dlhodobom horizonte. Aj nás preto pri tvorbe benchmarkingu zaujímajú obidva pohľady, pričom väčší dôraz kladieme na kvalitu liečby. Keďže v tejto fáze hodnotíme lekárov,

ktorí sú už zapojení, rozhodli sme sa ako doplnkový pohľad zvoliť doterajšiu aktivitu v programe. Tú sme popísali podľa počtu registrovaných pacientov a pacientov, ktorých odoslali na edukácie vzhľadom na veľkosť kmeňa jeho pacientov (presnejšie vysvetlené neskôr).

Náš benchmarking poskytovateľov sa bude preto skladať z troch častí, ktoré budú do celkového hodnotenia prispievať nasledovne:

- **30%** nákladová efektivita
- **60%** efektivita liečby pacienta
- **10%** aktivita lekára v programe

Za referenčné obdobie sme si vybrali kalendárny rok 2016. Nákladovú efektivitu a efektivitu liečby pacienta sme sa rozhodli merať pomocou obálkovej analýzy dát (DEA), ktorú sme podrobnejšie opísali v podkapitole 3.1. Na každú z nich sme si vytvorili samostatný DEA model s rôznymi vstupmi a výstupmi. Každý z parametrov hodnotenia sme preskúmali a dodefinovali, aby čo najlepšie vystihoval aspekty, ktorý samotný lekár dokáže ovplyvniť.

Definícia diabetika 2. typu

Ešte pred tým, ako sme začali vytvárať modely efektivít bolo nutné zodpovedať niekoľko základných otázok. Tou prvou bola definícia diabetika 2. typu z dát o VZS. Túto informáciu by nám mal poskytnúť kód diagnózy. Diabetes mellitus 1. typu je podľa [16] vykazovaný pod kódom diagnóz E10 a diabetes mellitus 2. typu pod kódom E11. Nie vždy tomu ale tak bolo, takýto spôsob vykazovania diabetu platí od roku 2011 [ICD-10-CM]. Pred týmto obdobím sa pod kódom E10 označoval diabetes mellitus závislý od inzulínu a pod E11 diabetes mellitus nezávislý od inzulínu. Keďže pri cukrovke druhého typu sa závislosť od inzulínu môže časom zmeniť, bolo v dôsledku tejto zmeny vykazovanie diagnóz vo VZS neprehľadné a často chybové. Existuje však niekoľko indikátorov, pomocou ktorých vieme dáta vyčistiť alebo sa minimálne aspoň priblížiť k presnejším výsledkom.

V prvom rade sme sa pozreli na všetkých pacientov, ktorým bola za rok 2016 aspoň raz vykázaná starostlivosť s kódom diagnózy E10 alebo E11. Keďže diabetes 1. typu sa

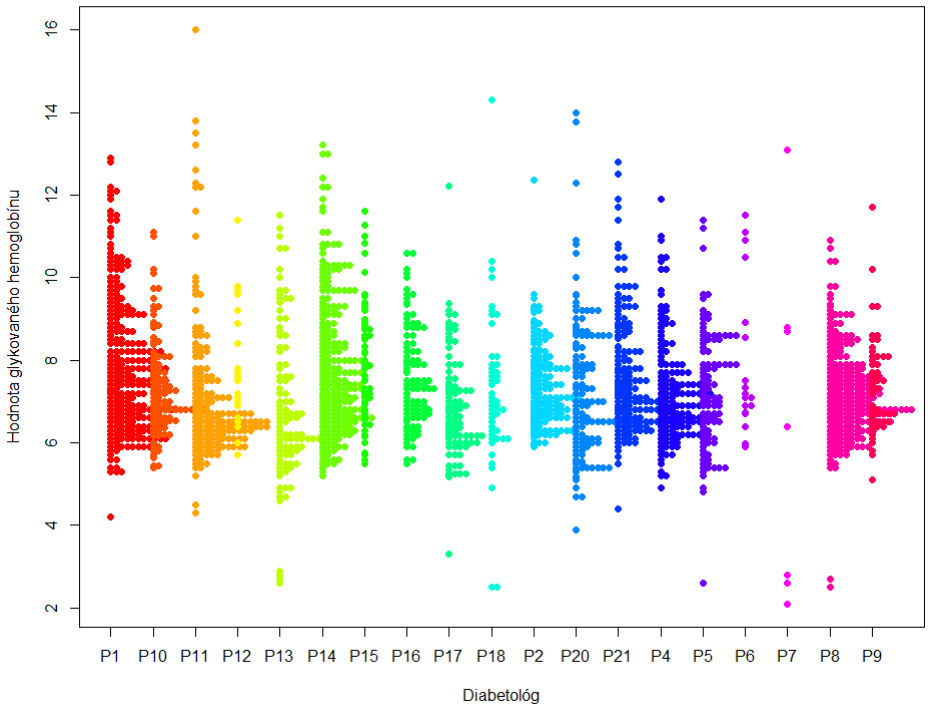
vyznačuje práve nedostatkom, resp. úplným chýbaním inzulínu, je ten v drvivej väčšine prípadov pacientom aj predpísaný. Ak sme teda mali pacienta, ktorému bol predpísaný iba PAD, prípadne PAD v kombinácií s inzulínom, mohli sme ho klasifikovať ako diabetika 2. typu. K týmto sme pridali pacientov, ktorým neboli predpísané žiadne lieky (a teda sú na diéte) a zároveň mali aspoň raz vykázaný kód diagnózy E11 a poskytnutý minimálne jeden výkon v odbornosti diabetológie (050). Komplikovanejší prípad bol v prípade, že mal pacient predpísaný iba inzulín a v jeho vykázanej zdravotnej staroslivosti sa objavili obidva kódy diagnóz. V takomto prípade, sme sa rozhodli pacienta klasifikovať ako diabetika 2. typu, ak bolo diagnóz s kódom E11 aspoň 50%. Túto hranicu sme určili na základe [14] a [3], podľa ktorých je pomer diabetikov 2. typu približne 90% z celkové počtu diabetikov.

Kmeň pacientov

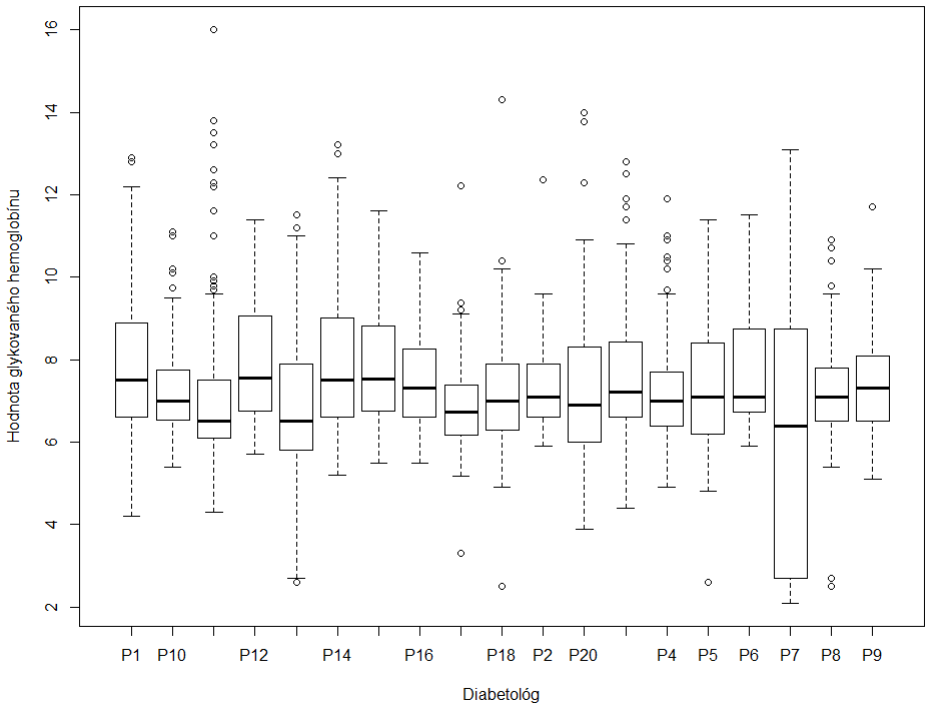
Ďalšou otázkou bolo, akých pacientov možno považovať za pacientov priradených práve danému lekárovi, resp. u ktorých pacientov, s ktorými prišiel lekár do kontaktu zodpovedá za liečbu diabetu 2 typu. Aby sme predišli priradeniu zodpovednosti v prípade, že pacient už navštevuje iného lekára, pridali sme podmienku, že posledná návšteva pacienta musela byť u daného diabetológa. Takto sme vytvorili tzv. kmeň pacientov.

Závažnosť prípadov

Poslednou otázkou bolo, či je vôbec možné daných diabetológov medzi sebou porovnávať, resp. či majú jednotliví lekári vo svojom kmeni rovnako "ťažkých" pacientov. V praxi totiž môže nastať situácia, že vplyvmi, ktoré nemôže lekár ovplyvniť, ako sú napríklad vzdelanie, sociálne prostredie, dostupnosť lekárskej starostlivosti a pod., môže prichádzať do styku s rôznou závažnosťou stavu ochorenia pacientov. Túto hypotézu sme si ale najprv potrebovali overiť. Z klientských dát, ktoré sme získali z vyplnených vstupných formulárov pri registrácií pacienta do programu sme mohli získať obraz o nameraných hodnotách glykovaného hemoglobínu. Podľa [1] vo všeobecnosti platí, že diabetici s hodnotami okolo 7% sú dobre kompenzovaní a pri vyšších hodnotách je riziko komplikácií vyššie. Pre lepšiu názornosť sme si preto zobrazili hodnoty na grafoch (viď Obr. 4.1. a 4.2.). Už z grafov je možno vidieť, že namerané hodnoty a



Obr. 4.1: Scatter plot pre namerané hodnoty glykovaného hemoglobínu diabetológmi



Obr. 4.2: Box ploty nameraných hodnôt glykovaného hemoglobínu diabetológmi

ich zloženie nie je rovnaké u každého lekára. Hypotézu sme si overili pomocou jednofaktorovej analýzy rozptylu (ANOVA). Nulová hypotéza znela, že priemerné hodnoty glykovaného hemoglobínu u pacientov sú u všetkých lekárov rovnaké. Tú sme zamietli ($p - value = 1,7e^{-15}$) a prijali hypotézu, že v zložení pacientov sú u diabetológov signifikantné rozdiely. Treba však brať do úvahy, že neboli splnené predpoklady testu, ktorými sú konštantnosť variácií a normalita dát. Je to spôsobené hlavne chybovosťou pri vypĺňaní vstupných formulárov. Zobrali sme teda tento fakt do úvahy pri tvorbe vstupov a výstupov modelov.

Rovnako treba mať na mysli, že z našich dát máme k dispozícií len údaje poistencov Dôvery z.p., ktorá mala na Slovensku v roku 2016 27,47% podiel poistencov [3] (u lekárov zapojených v programe bol podiel o niečo vyšší).

4.4 Meranie nákladovej efektivity

Prvým modelom, ktorý sme vytvorili, bol DEA model merajúci nákladovú efektivitu jednotlivých diabetológov. Tento model nám mal popísať ako ekonomicky lekár využíva zdroje. Zjednodušene povedané, na jednej strane boli peniaze, ktoré boli lekárovi poskytnuté na výkon práce a na druhej strane zdravotná starostlivosť, ktorú bol schopný za tento balík poskytnúť. Čím viac starostlivosti lekár poskytol za daný obnos, tým lepšie bol ohodnotený.

4.4.1 Vstupy a výstupy nákladového modelu

Vstupy

Základným parametrom modelu bola absolútna hodnota *nákladov*, ktorú lekár vykázal poisťovni. Išlo pritom o náklady na všetky výkony, laboratórne vyšetrenia a predpísané lieky, kde vystupoval ako vyšetrujúci, odosielajúci alebo odporúčajúci lekár nad kmeňom svojich pacientov.

(DMU) PZS	(I) náklady	(O) PAC - diéta	(O) PAC - PAD	(O) PAC - inzulín	(O) # kontaktov	(O) # VV a VL	(O) # liekov
P1	129 297	222	236	95	2 362	20 551	4 922
P2	95 283	32	167	58	1 486	11 532	3 160
P3	79 583	78	206	44	1 107	14 092	2 312
P4	38 985	44	252	44	1 167	8 613	1 951
P5	80 648	83	121	109	2 086	14 267	2 772
P6	354 368	239	789	317	4 807	41 801	12 745
P7	131 191	109	256	101	2 276	17 858	3 276
P8	140 563	122	372	133	2 198	18 213	5 981
P9	46 355	48	114	32	651	6 169	1 760
P10	208 963	202	785	97	4 094	37 539	8 665
P11	74 471	91	209	60	1 642	9 861	2 564
P12	128 305	74	264	96	1 682	19 036	4 281
P13	152 559	126	442	123	2 268	27 917	6 314
P14	179 463	145	424	119	3 661	33 221	7 302
P15	326 141	386	467	169	3 584	43 858	10 650
P16	301 812	193	729	245	6 004	51 309	11 220
P17	222 269	153	407	266	3 666	16 866	9 847
P18	14 596	13	32	7	148	1 642	370
P19	100 911	32	271	52	1 508	13 730	3 808
P20	45 203	13	114	42	1 111	5 014	1 890
P21	41 017	25	73	33	247	4 772	1 642

Obr. 4.3: Hodnoty vstupov a výstupov PZS nákladového DEA modelu

Výstupy

Za výstupy modelu sme si zvolili premenné predstavujúce zdravotnú starostlivosť, ktoré s výškou nákladov priamo úmerne súvisia. Tým prvým bol *počet pacientov* v kmeni. Ako sme už spomínali, pri tomto parametri sme chceli vziať do úvahy, že lekár prichádza do kontaktu s rôzne závažnými prípadmi. Na rozlíšenie lekárov sme preto rozdelili počet pacientov na tri rôzne vstupy a teda počet pacientov v troch rôznych skupinách, podľa liekov, ktoré im boli v daný rok predpísané: *pacienti na diéte* (neberú žiadne lieky), *pacienti na antidiabetikách* (PAD) a *pacienti na inzulíne*. V prípade, že pacient bol na kombinovanej liečbe a teda užíval súčasne antidiabetiká aj inzulín alebo bol na inzulínovej pumpe, bol zaradený do poslednej kategórie.

Ďalšími parametrami boli tie, ktoré náklady priamo indikujú: *počet kontaktov s pacientom*, *počet predpísaných liekov* a *počet výkonov a laboratórnych vyšetrení*. Keďže sa často stáva, že lekár nevykáže každú návštevu pacienta alebo vykáže niektoré úkony o niečo neskôr, zobrali sme do úvahy aj každý predpis lieku, či laboratórne vyšetrenie

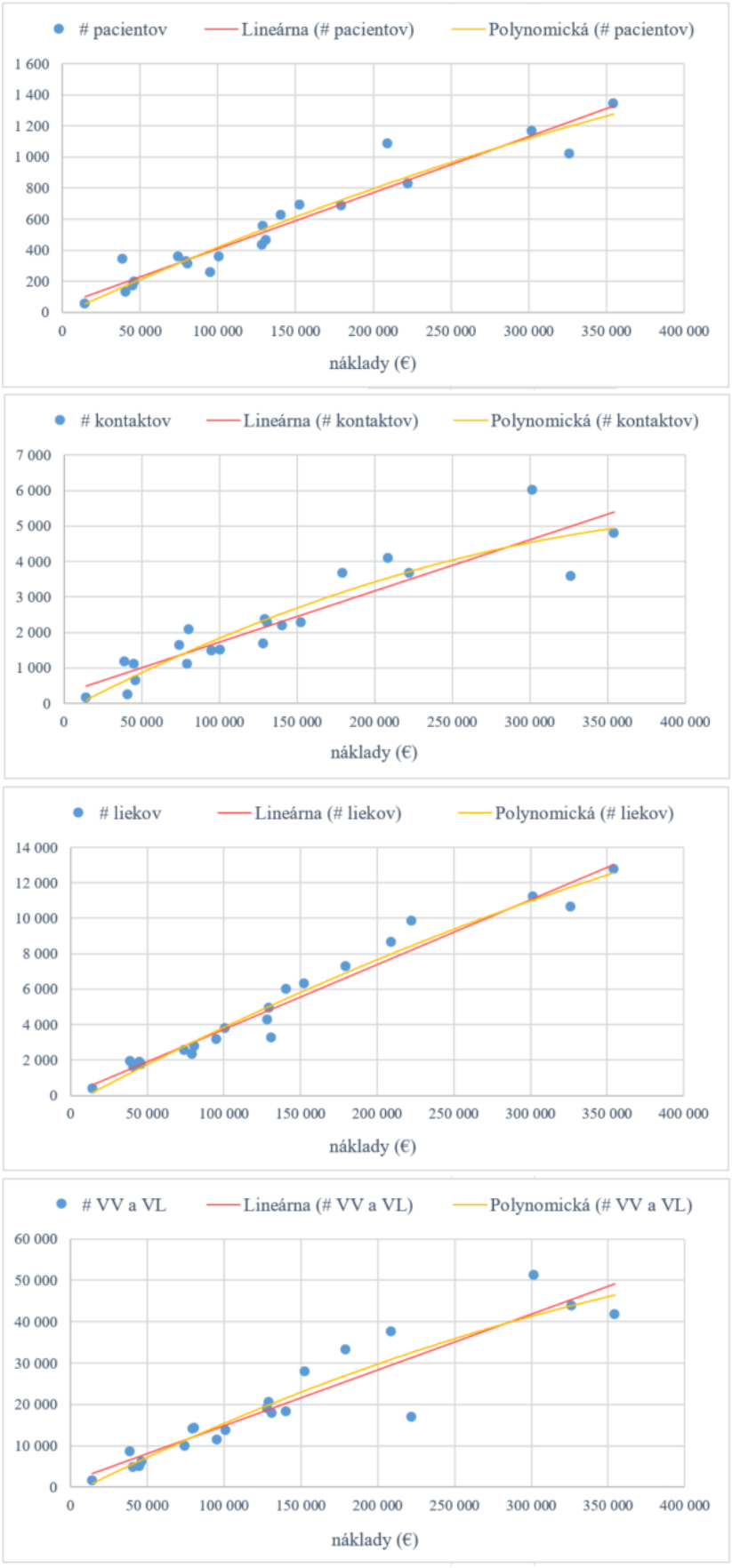
vykázané v unikátny dátum u daného diabetológa. Za kontakt sme pritom považovali každý takýto výkon, ktorý nastal v horizonte aspoň 5 dní od posledného kontaktu. Snažili sme sa tým eliminovať prípady, kedy pacient príde k lekárovi iba na kontrolu alebo nadväzujúce vyšetrenie. Ostatné premenné nám pomôžu v modeli popísať, aké drahé lieky lekár predpisuje a aké drahé zobrazovacie metódy pri svojej práci používa.

Jednotlivých lekárov sme si anonymizovali a z databázy o VZS sme si vytiahli potrebné údaje na vytvorenie tabuľky s hodnotami vstupov a výstupov (Obr. 4.3).

4.4.2 Tvorba nákladového modelu

Po výbere vstupov a výstupov nasledoval výber modelu. V prvom rade sme sa pokúsili o identifikáciu typu výnosov z rozsahu a tým pádom množiny produkčných možností. Nakoľko cena lekárskych výkonov, liekov a inej zdravotnej starostlivosti neklesá s počtom vykázaných výkonov ani predpisov, narozdiel od niektorých iných sektorov, našim apriórnym predpokladom boli konštantné výnosy z rozsahu. Pre názornosť sme si zobrazili závislosť medzi jednotlivými parametrami na grafe s lineárnou a polynomickou (2. stupňa) trendovou čiarou (pre jednoduchosť sme sčítali počet pacientov vo všetkých troch kategóriách). V prípade, že by sa polynomická trendová čiara výrazne líšila od lineárnej a pripomínala by tvar hranice M_{VRS} (viď Obr. 3.2), pracovali by sme s predpokladom variabilných výnosov z rozsahu. Na obrázku však možno vidieť, že tomu tak nie je a preto budeme pracovať s našim prvotným predpokladom.

Následne bolo potrebné zvoliť orientáciu modelu. Ako sme spomenuli v časti 3.1, DEA modely rozdeľujeme na orientované a neorientované. Keďže v našom prípade bolo hlavným cieľom určenie efektivity jednotlivých poskytovateľov, vystačili by sme si s neorientovaným modelom, ktorý dokáže zachytiť neefektivitu na vstupoch aj výstupoch. Ak by sme však chceli pristúpiť ku konkrétnemu riešeniu, vstupne orientovaný model by bol vhodnejší z niekoľkých dôvodov. Tým prvým je, že s nákladmi, ako naším jediným vstupom, dokážu lekári pracovať a do určitej miery aj ovplyvňovať ich výšku. Na druhej strane platí, že našou motiváciou nie je zvyšovanie hodnôt na strane výstupov, akými sú počet predpísaných liekov či počet pacientov. Na výpočet nákladovej efektivity sme teda zvolili $CCR - I$ model, pričom sme použili [10].



Obr. 4.4: Závislosť parametrov nákladového DEA modelu

4.4.3 Výsledky nákladového modelu

PZS	Efektivita	Poradie
P1	1.00	1
P4	1.00	1
P5	1.00	1
P17	0.99	4
P11	0.88	5
P8	0.85	6
P14	0.84	7
P10	0.84	8
P20	0.84	9
P13	0.83	10
P3	0.82	11
P9	0.81	12
P21	0.80	13
P15	0.78	14
P16	0.77	15
P6	0.76	16
P19	0.75	17
P7	0.69	18
P12	0.67	19
P2	0.66	20
P18	0.61	21

Obr. 4.5: Výsledná efektivita a poradie nákladového modelu

Na obrázku 4.5 môžeme vidieť výsledky pre všetkých PZS. Výslednu efektivitu rovnú jednej sme namerali diabetológom P1, P4 a P5. Pozreli sme sa na hodnoty slackov pre dané DMU, či sa v ich prípade nejedná o pseudoefektivitu. Vektory slackov boli nulové a preto ich môžeme vyhlásiť za efektívnych. Ostatní lekári sú podľa modelu neefektívni s efektivitou podľa tabuľky (Obr. 4.5). Výsledky využijeme vo výslednom benchmarkingu s váhou 30%.

4.5 Meranie efektivity liečby

Tvorba parametrov modelu merajúceho efektivitu liečby už bola o niečo zložitejšia, zachovali sme však metodiku pri základných otázkach. To znamená, že kmeň diabetológa bol vytvorený rovnako z pacientov s ochorením cukrovky 2. typu (podľa rovnakej

metodiky) a jeho posledná návšteva bola u daného diabetológa. Zachovali sme aj stratifikáciu pacientov podľa náročnosti. Narozdiel od prvého modelu sme sa pri druhom modeli viac zamerali na samotného pacienta. Kým pri nákladovej efektivite sme sledovali náklady vyprodukované diabetológom pri daných pacientoch, tu sme pozorovali aspekty ochorenia a liečby z pohľadu pacienta, ktorého sme priradili k diabetológovi aj napriek tomu, že neboli priamo indikované daným lekárom, ale s kvalitou liečby súvisia podľa [14].

4.5.1 Vstupy a výstupy liečebného modelu

Vstupy

Pri výbere parametrov sme uvažovali nad liečbou, ako nad procesom, ktorého vstupom je portfólio pacientov, s ktorými prichádza lekár do kontaktu a jeho práca s pacientom, vnímaná ako technológia, má priame a nepriame výstupy. Pod portfóliom pacientom budeme opäť chápať počet *pacientov na diéte*, *pacientov na antidiabetikách* a *pacientov na inzulíne*. Nepriamo svojou prácou lekári ovplyvňujú, koľko komplikácií pri ochorení cukrovky následne majú pacienti (tu treba mať na pamäti, že mieru komplikácií do veľkej miery ovplyvňuje veľa iných faktorov, akými sú napríklad dostupnosť zdravotnej starostlivosti, či postoj pacienta k ochoreniu, tie však pre jednoduchosť modelu nebudú predmetom nášho pozorovania). Komplikácie sú však tzv. nežiadúcim výstupom a budeme s nimi na základe [15] pracovať ako so vstupmi.

Jedným zo vstupov bude *počet komplikácií*, chápaný ako počet jednotlivých diagnóz priradených k výkonom, ktoré boli pacientovi za daný rok poskytnuté. Výkon s danou diagnózou sme pritom brali do úvahy, ak jeho cena presiahla hodnotu dvoch eur a podľa [14] patrí do zoznamu z tabuliek 4.1, 4.2 alebo 4.3. Ak sa u pacienta niektorá z diagnóz opakovala, započítali sme ju len jeden krát. Povedzme, že mal jeden z pacientov napríklad vykázanú diagnózu choroby šošoviek. Našim cieľom bolo zachytiť situáciu, kedy sa táto choroba u neho prejavila a bola natoľko závažná, že k nej priradený výkon bol drahší. Naopak v prípade, že pacient chodí iba na pravidelné kontroly a stav ochorenia ešte nie je vážny, nechceli sme ho brať do úvahy.

Chronické komplikácie	
1) Nefropatia	
E11.2	Diabetes mellitus 2. typu: s obličkovými komplikáciami
N18	Chronická choroba obličiek
N19	Zlyhanie obličiek, bližšie neurčené
Z94.0	Stav po transplantácii obličky
2) Retinopatia	
E11.3	Diabetes mellitus 2. typu: s očnými komplikáciami
H25 - H28	Choroby šošoviek
H30 - H36	Choroby cievovky a sietnice
H43 - H44	Choroby sklovca a očnej gule
H53	Porucha videnia
H54	Slepota a slabozrakosť
3) Neuropatia	
E11.4	Diabetes mellitus 2. typu: s nervovými komplikáciami
G59	Mononeuropatia pri chorobách zatriedených inde
G63	Polyneuropatia pri chorobách zatriedených inde
G64	Iná porucha periférnej nervovej sústavy
R02	Gangréna, nezatriedená inde
Z89	Získané chýbanie končatiny

Tabuľka 4.1: Tabuľka chronických komplikácií súvisiacich s diabetom 2. typu

Akútne komplikácie	
Znaky nekompenzovaného diabetu	
E11.0	Diabetes mellitus 2. typu: s kómou
E11.1	Diabetes mellitus 2. typu: s ketoacidózou
E16	Iná porucha vnútorného vylučovania podžalúdkovej žľazy
E16.0	Hypoglykémia zapríčinená liekmi, bez kómy
E16.1	Iná hypoglykémia
E16.2	Hypoglykémia, bližšie neurčená
E87.0	Hyperosmolalita a hypernatriémia
E87.2	Acidóza

Tabuľka 4.2: Tabuľka akútnych komplikácií súvisiacich s diabetom 2. typu

Komorbiditné ochorenia	
1) Choroby obehovej sústavy	
I20 - I25	Ischemická choroba srdca (angina pectoris, infarkt myokradu a iné)
I50	Zlyhávanie srdca
I60 - I69	Cievne choroby mozgu
I69.3	Následky mozgového infarktu
I70	Ateroskleróza
I70.2	Ateroskleróza končatinových tepien
I73	Iné choroby periférnych ciev
I79	Choroba tepien, tepničiek a vlásočníc pri chorobách zatriedených inde
2) Metabolické poruchy	
E78	Porucha metabolizmu lipoproteínov a iná lipidémia

Tabuľka 4.3: Tabuľka komorbiditných ochorení súvisiacich s diabetom 2. typu

Ďalším vstupom bol *počet výjazdov rýchlej zdravotníckej pomoci* (RZP). Ako sme spomenuli v časti 1.2 u cukrovkárov sa môžu vyskytnúť rôzne akútne komplikácie. K tým podľa [14] môže dôjsť v prípade, že diabetológ nevysvetlí pacientovi dostatočne všetky aspekty užívania liekov, najmä inzulínu. V takých prípadoch následne dochádza často napríklad k hypoglykémii, čo u pacientov indikuje častejšie volanie RZP.

Posledným vstupom bol *počet hospitalizácií* nad kmeňom poistencov daného diabetológa. K tým logicky dochádza pri závažnejších komplikáciách, kedy je nutný pobyt na ložku, prípadne operačný zákrok. Opäť sme však brali do úvahy iba hospitalizácie s vykázanými kódmi diagnóz z tabuliek 4.1, 4.2 alebo 4.3.

(DMU) PZS	(I) PAC - diéta	(I) PAC - PAD	(I) PAC - inzulín	(I) # komplikácií	(I) # výjazdov RZP	(I) # hospitalizácií	(O) # HbA1c
P1	222	236	95	804	40	72	710
P2	32	167	58	358	15	35	305
P3	78	206	44	391	3	19	853
P4	44	252	44	402	10	43	578
P5	83	121	109	699	7	22	492
P6	239	789	317	2 561	87	214	1 344
P7	109	256	101	537	12	93	692
P8	122	372	133	929	28	57	902
P9	48	114	32	203	3	12	185
P10	202	785	97	1 901	45	103	1 339
P11	91	209	60	450	10	29	967
P12	74	264	96	617	10	70	880
P13	126	442	123	885	21	65	1 167
P14	145	424	119	1 104	11	61	1 017
P15	386	467	169	1 672	21	72	2 353
P16	193	729	245	1 807	27	136	2 376
P17	153	407	266	1 012	25	133	1 174
P18	13	32	7	61	1	6	109
P19	32	271	52	551	10	29	803
P20	13	114	42	218	6	29	381
P21	25	73	33	265	9	19	273

Obr. 4.6: Hodnoty vstupov a výstupov PZS liečebného DEA modelu

Výstupy

Premennú, ktorú lekári ovplyvňujú priamo v ordinácií je *počet odmeraných vyšetrení glykovaného hemoglobínu* (HbA1c). Podľa [14] je pravidelné meranie hladiny HbA1c kľúčové z hľadiska nastavenia správnej liečby cukrovky. Lekári využívajú na odmeranie HbA1c viacero spôsobov. Niekedy sa tak deje priamo na ambulancii, alebo môže vzorku poslať do laboratória, kde ju môžu vyšetriť opäť viacerými metódami. My sme zobrali do úvahy všetky známe možnosti. Keďže ide o žiadúci faktor, budeme s ním pracovať

ako s klasickým výstupom.

V tabuľke (Obr. 4.6) vidíme hodnoty vstupov a výstupov pre jednotlivé DMU, ktoré vstúpia do modelu.

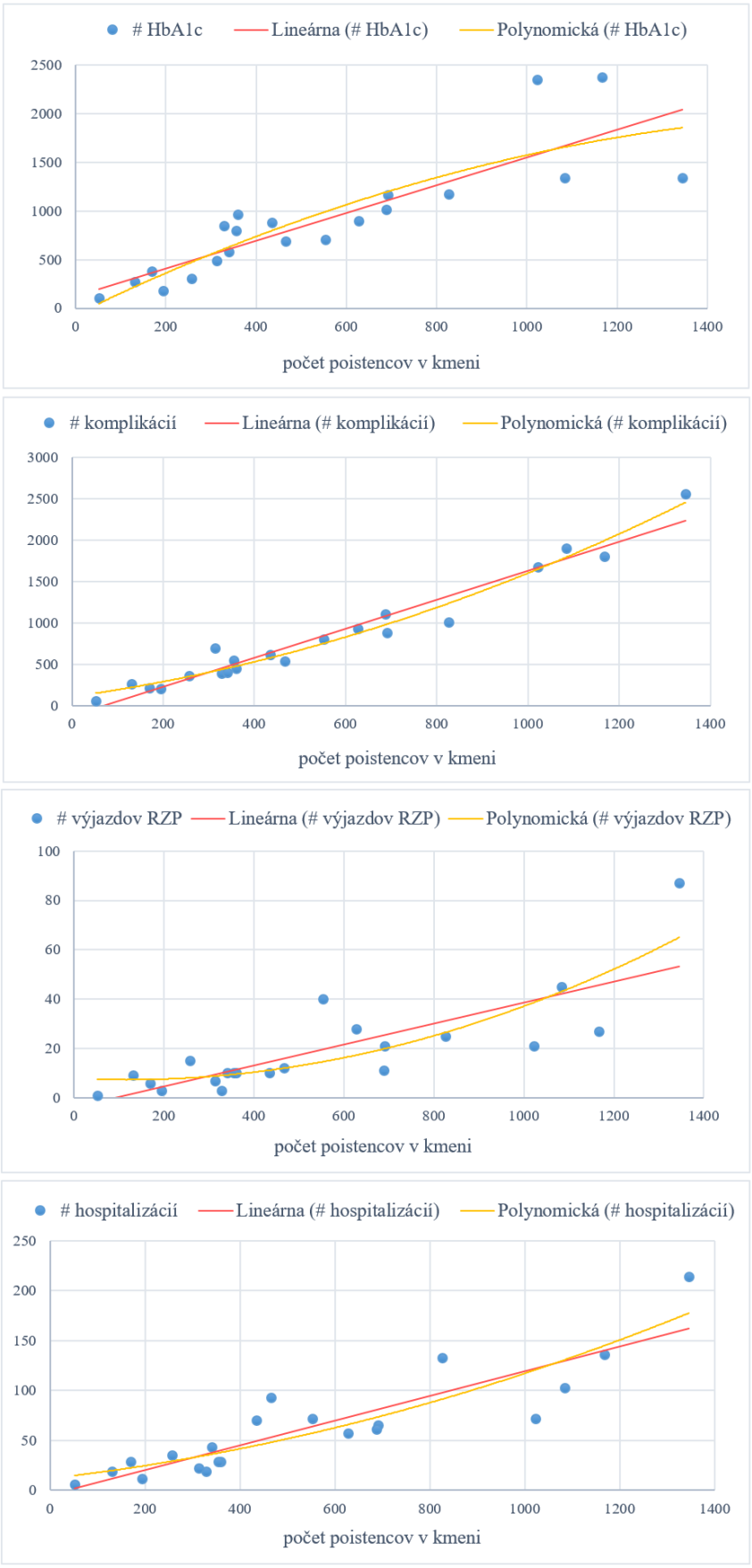
4.5.2 Tvorba liečebného modelu

Pri tvorbe druhého modelu sme postupovali podobne, ako v prvom prípade. Najprv sme sa pokúsili identifikovať typ výnosov z rozsahu. V tomto prípade sa však predpoklad robil ťažšie. Jedným dôvodom bolo, že sme nežiadúce výstupy preniesli na stranu vstupov a druhým, že kvalita liečby pacientov môže s ich rastúcim počtom klesať, napriek tomu, že by tomu tak byť nemalo. Zobrazili sme si teda opäť závislosť medzi výškou nákladov a jednotlivými parametrami. Na grafoch (viď Obr. 4.7) však možno vidieť, že polynomická trendová čiara sa opäť od lineárnej príliš nelíši. Aj preto sme sa opäť rozhodli pre konštantné výnosy z rozsahu.

Kvôli práci s nežiadúcimi výstupmi sme sa rozhodli zvoliť neorientovaný SBM model [15]. Jeho výhodou je, že hodnoty efektivít dostaneme, rovnako ako z prvého modelu, v intervale $[0, 1]$. V prípade, že budeme s výsledkami chcieť neskôr pracovať, pri výpočte efektívnych vzorov dostaneme útvar z hranice efektívnosti, ktorý je najďalej od nášho útvaru, ale nemá menšie výstupy ani väčšie vstupy [9]. Navyše môžeme z účelovej funkcie vynechať parametre počtu pacientov, ktoré diabetológ nemôže ovplyvniť.

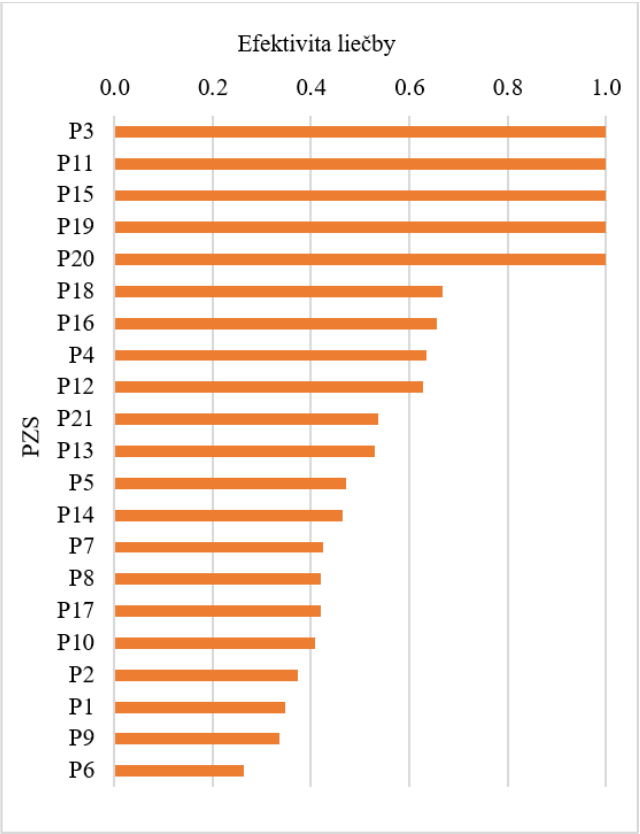
4.5.3 Výsledky liečebného modelu

V druhom modeli sme efektivitu liečby rovnú jednej namerali u piatich diabetológov - P3, P11, P15, P19 a P20. Opäť sme kontrolovali hodnoty slackov, ktoré boli vo všetkých piatich prípadoch nulové, a teda môžeme zamietnuť pseudofektivitu. Vo všetkých piatich prípadoch išlo o iných PZS, ako v prvom modeli. Zostávajúcich 16 lekárov sme označili za neefektívnych s efektivitou podľa tabuľky na obrázku 4.8.



Obr. 4.7: Závislosť parametrov liečebného DEA modelu

PZS	Efektivita	Poradie
P3	1.00	1
P11	1.00	1
P15	1.00	1
P19	1.00	1
P20	1.00	1
P18	0.67	6
P16	0.66	7
P4	0.64	8
P12	0.63	9
P21	0.54	10
P13	0.53	11
P5	0.47	12
P14	0.46	13
P7	0.42	14
P8	0.42	15
P17	0.42	16
P10	0.41	17
P2	0.37	18
P1	0.35	19
P9	0.34	20
P6	0.26	21



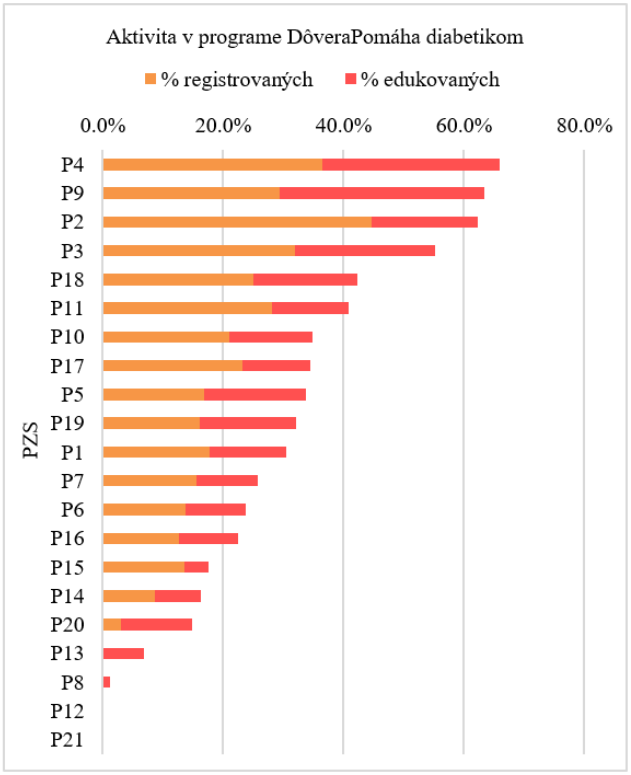
Obr. 4.8: Výsledná efektivita a poradie liečebného modelu

4.6 Aktivita PZS v programe

Poslednú časť benchmarkingu PZS tvorilo vyhodnotenie doterajšej aktivity lekára v programe DôveraPomáha diabetikom. Aby pacienti s ochorením cukrovky 2. typu mohli vstúpiť do programu, je nutná ich registrácia prostredníctvom lekára a jeho následné odporučenie edukácií pacientovi je pre fungovanie programu kľúčové. Tak, ako sme už spomenuli v kapitole 1.3, edukácie napomáhajú k lepšej informovanosti pacientov o svojom ochorení a vyššej adherencii k liečbe. Počet percent zapojených pacientov v oboch kritériách za rok 2016 tvorilo a dopĺňalo celkové hodnotenie vo výške 5% (súčet hodnotenia tak v konečnom dôsledku tvorilo 10 % výsledného skóre). Na obrázku 4.9 môžeme vidieť výsledky. Keďže počet percent bol počítaný z celkového kmeňa, treba dodať, že jeho splnenie dalo viac úsilia lekárom s väčším počtom pacientov v kmeni, narozdiel od lekárov s menším kmeňom (nominálny počet zapojených diabetikov bol vyšší). Aj preto je táto časť doplnková a tvorí výrazne menšiu časť benchmarkingu. Pri

jeho aplikovaní na diabetológov, ktorých by sme ešte len chceli do programu zapojiť by bolo takisto nutné túto časť vynechať.

PZS	% registrovaných	% edukovaných
P4	36.5%	29.4%
P9	29.4%	34.0%
P2	44.7%	17.5%
P3	32.0%	23.2%
P18	25.0%	17.3%
P11	28.1%	12.8%
P10	21.1%	13.7%
P17	23.2%	11.3%
P5	16.9%	16.9%
P19	16.1%	16.1%
P1	17.7%	12.8%
P7	15.7%	10.1%
P6	13.8%	9.9%
P16	12.7%	9.8%
P15	13.5%	4.1%
P14	8.7%	7.6%
P20	3.0%	11.8%
P13	0.0%	6.9%
P8	0.0%	1.3%
P12	0.0%	0.0%
P21	0.0%	0.0%



Obr. 4.9: Plnenie kritéria registrácií edukácií pacientov v programe

4.7 Výsledky benchmarkingu

Výsledkom nášho benchmarkingu bola kombinácia troch kritérií hodnotenia. Do výsledného skóre vstúpili s rôznymi váhami nasledovne:

$$efektivita\ nákladov\ *\ 0.3\ +\ efektivita\ liečby\ *\ 0.6\ +$$
$$+\ \%\ registrovaných\ *\ 0.05\ +\ \%\ edukovaných\ *\ 0.05\ =\ výsledné\ skóre$$

Ak si vezmeme *aktivitu v programe* ako súčet kritéria registrácie a edukácie vzorec bude vyzerať nasledovne:

$$efektivita\ nákladov\ *\ 0.3\ +\ efektivita\ liečby\ *\ 0.6\ +\ aktivita\ v\ programe\ *\ 0.05$$
$$=\ výsledné\ skóre$$

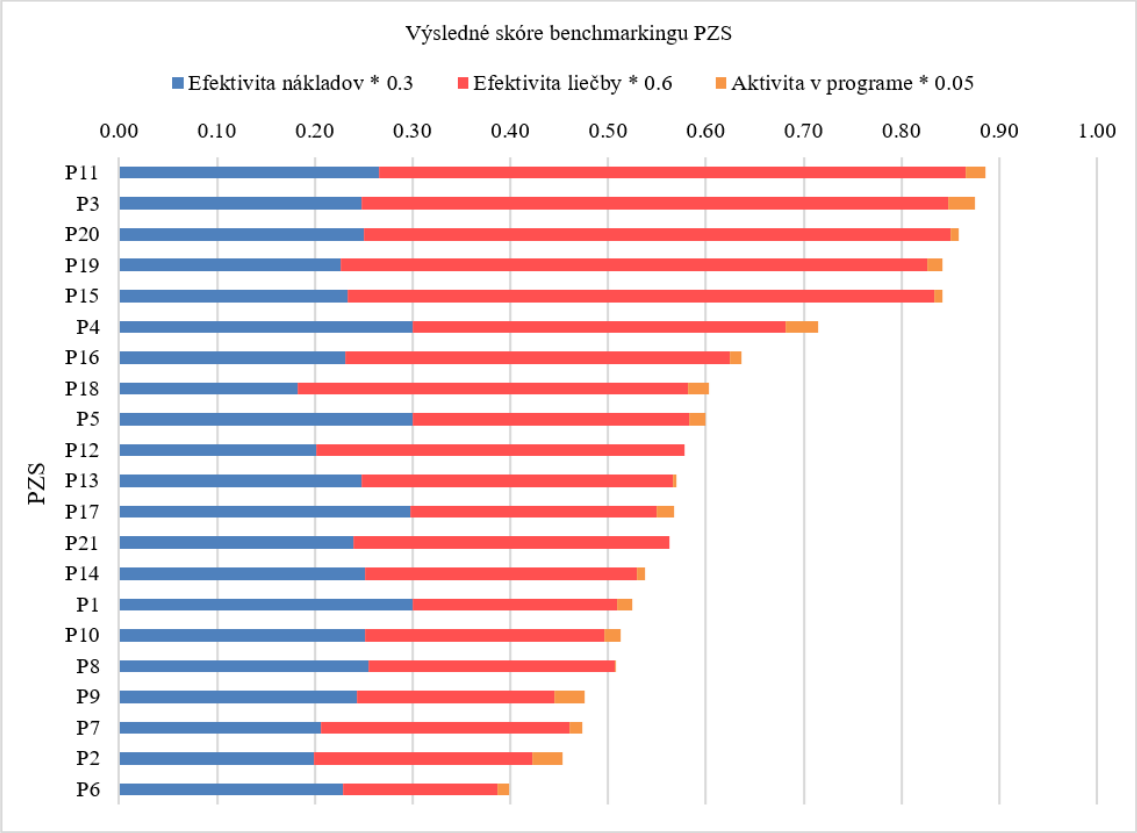
PZS	Efektivita nákladov	Efektivita liečby	Aktivita v programe	Výsledné skóre
P11	0.88	1.00	0.41	0.89
P3	0.82	1.00	0.55	0.88
P20	0.84	1.00	0.15	0.86
P19	0.75	1.00	0.32	0.84
P15	0.78	1.00	0.18	0.84
P4	1.00	0.64	0.66	0.71
P16	0.77	0.66	0.22	0.64
P18	0.61	0.67	0.42	0.60
P5	1.00	0.47	0.34	0.60
P12	0.67	0.63	0.00	0.58
P13	0.83	0.53	0.07	0.57
P17	0.99	0.42	0.35	0.57
P21	0.80	0.54	0.00	0.56
P14	0.84	0.46	0.16	0.54
P1	1.00	0.35	0.31	0.52
P10	0.84	0.41	0.35	0.51
P8	0.85	0.42	0.01	0.51
P9	0.81	0.34	0.63	0.48
P7	0.69	0.42	0.26	0.47
P2	0.66	0.37	0.62	0.45
P6	0.76	0.26	0.24	0.40

Obr. 4.10: Parciálne hodnotenie a výsledné skóre benchmarkingu PZS č.1

Zhrnutie výsledkov z predchádzajúcich častí a získané skóre vidíme na obrázkoch 4.10 a 4.11. Už na prvý pohľad nám z nich vyplýva niekoľko vecí. V prvom rade je vidno, že výsledné poradie PZS približne kopíruje poradie v modeli efektivity liečby. Je to spôsobené hlavne tým, že v najväčšej miere prispieva do hodnotenia, ale aj faktom, že v efektivite nákladov lekárov nie sú výrazné odlišnosti (to je možno vidieť aj na obrázkoch 4.5 a 4.8).

To nás vedie k otázke nakoľko sú výsledky v jednotlivých častiach medzi sebou závislé. Tabuľka z obrázku 4.12 ukazuje, že všetky tri kritériá majú navzájom približne nulové korelácie. To však ešte automaticky neimplikuje nezávislosť.

Keď si však zobrazíme parciálne hodnotenia na grafe podľa lekárov, naše tušenie iba potvrdzuje. Na jednej strane je dobré, že zložky nášho benchmarkingu sú zložené do značnej miery z nezávislých častí, na druhej strane z toho však vyplýva, že efektivita liečby nemusí automaticky viesť k efektivite nákladov a naopak.

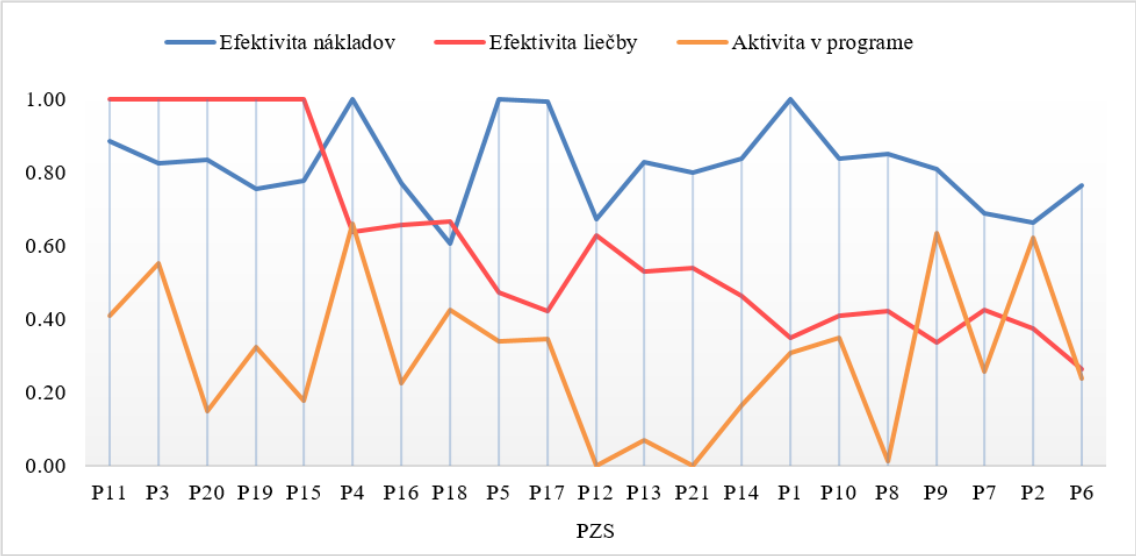


Obr. 4.11: Parciálne hodnotenie a výsledné skóre benchmarkingu PZS č.2

	Efektivita nákladov	Efektivita liečby	Aktivita v programe
Efektivita nákladov	1.000	-0.095	0.132
Efektivita liečby	-0.095	1.000	0.001
Aktivita v programe	0.132	0.001	1.000

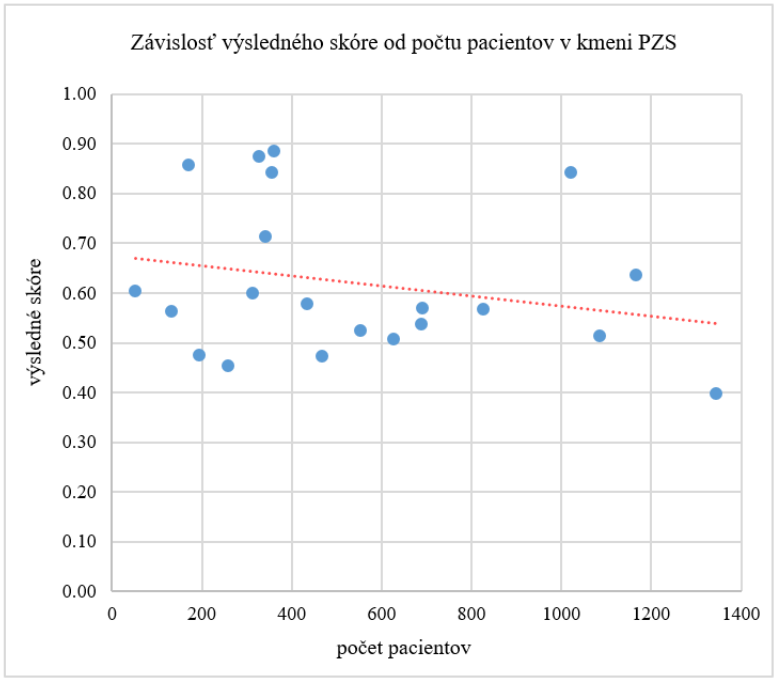
Obr. 4.12: Tabuľka korelácií čiastočných výsledkov

Dá sa teda povedať, ktorí lekári vyšli v hodnotení lepšie? Skúsili sme si vykresliť závislosť výsledného skóre od celkového počtu diabetikov v kmeni (Obr. 4.14). Napriek tomu, že korelácia, medzi týmito dvomi parametrami je rovná -0.239, závislosť v podobe nepriamej úmery nie je jednoznačná. Tu treba poznamenať, že stále pracujeme iba s počtom poistencov zdravotnej poisťovne Dôvera a diabetikov 2.typu. S koľkými inými pacientmi prichádza lekár do kontaktu a ďalšiu náplň jeho práce zohľadniť nevieme. Pri dátach, s ktorými sme pracovali teda môžeme povedať, že výsledok nášho benchmarkingu závisí naozaj od hodnôt v zvolených parametroch, vstupov, výstupov a technológií jednotlivých PZS.



Obr. 4.13: Graf výsledkov benchmarkingu PZS

PZS	PAC	skóre
P11	360	0.89
P3	328	0.88
P20	169	0.86
P19	355	0.84
P15	1022	0.84
P4	340	0.71
P16	1167	0.64
P18	52	0.60
P5	313	0.60
P12	434	0.58
P13	691	0.57
P17	826	0.57
P21	131	0.56
P14	688	0.54
P1	553	0.52
P10	1084	0.51
P8	627	0.51
P9	194	0.48
P7	466	0.47
P2	257	0.45
P6	1345	0.40

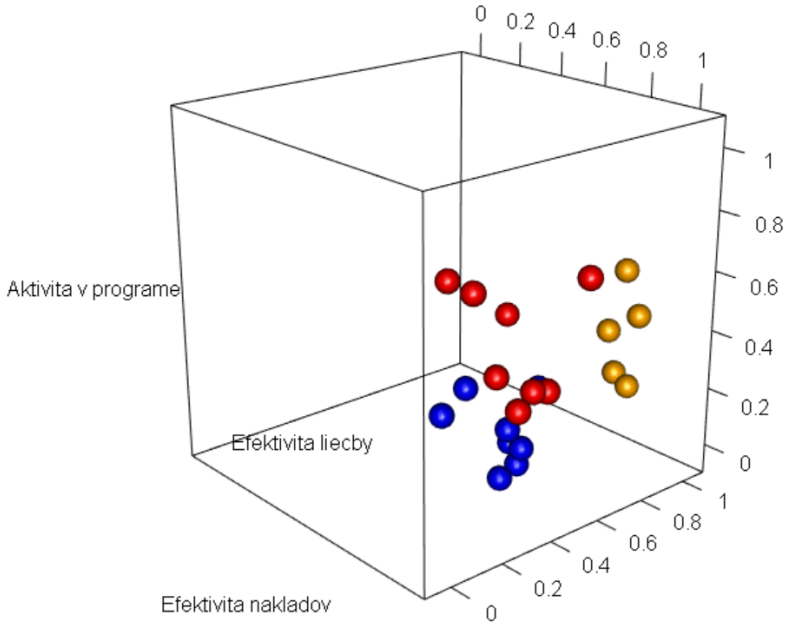


Obr. 4.14: Farebné rozškálovanie PZS podľa počtu paceintov v kmeni a jeho závislosť od výsledného skóre

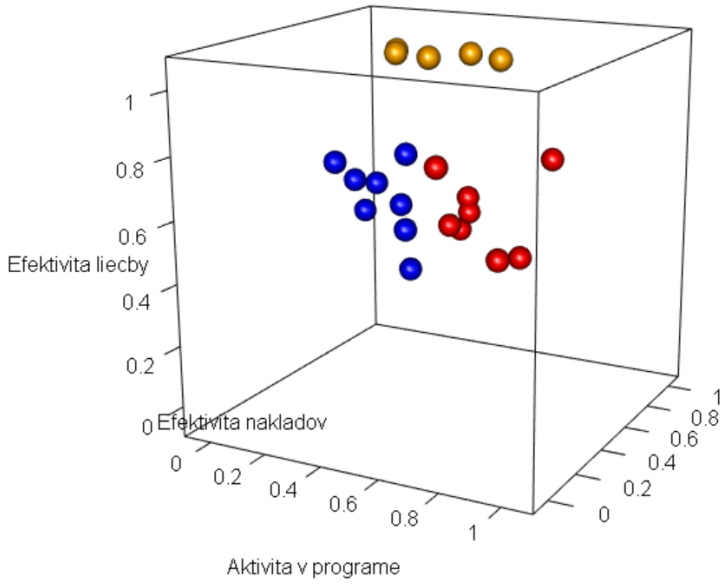
Na záver sme sa rozhodli rozdeliť lekárov do troch zhlukov podľa výsledkov v jednotlivých častiach. Zvolili sme na to metódu K-medoids, opísanú v kapitole 3.2.1 a funkciu *pam* z balíka *cluster* v programe R. Výsledky sme porovnali aj s metódou K-means, no nakoľko sa veľmi nelíšili, ponechali sme prvú metódu pre lepšiu interpretáciu reprezentantov daných zhlukov. Funkcia nám vrátila optimálne rozdelenie do troch zhlukov, ktoré by sa dali popísať nasledovne. Prvý zhluk tvoria lekári, ktorí vyšli v hodnotení najlepšie a sú efektívni z hľadiska liečby. Na obrázkoch 4.16 a 4.17 sú znázornení oranžovou farbou. Ďalšie dva zhluky tvoria lekári, ktorí sú v kvalite liečby neefektívni no líšia sa zvyšnými parametrami. Aktivita v programe nám delí túto skupinu na lepšiu a horšiu polovicu, pričom tá lepšia obsahuje lekárov, ktorí sú nákladovo efektívni. Tí sú vyznačení červenou farbou. Modrá farba predstavuje lekárov, ktorí nevynikajú ani v jednej z častí nášho benchmarkingu.

PZS	Efektivita nákladov	Efektivita liečby	Aktivita v programe	Výsledné skóre	Zhluk
P11	0.88	1.00	0.41	0.89	2
P3	0.82	1.00	0.55	0.88	2
P20	0.84	1.00	0.15	0.86	2
P19	0.75	1.00	0.32	0.84	2
P15	0.78	1.00	0.18	0.84	2
P4	1.00	0.64	0.66	0.71	1
P16	0.77	0.66	0.22	0.64	3
P18	0.61	0.67	0.42	0.60	1
P5	1.00	0.47	0.34	0.60	1
P12	0.67	0.63	0.00	0.58	3
P13	0.83	0.53	0.07	0.57	3
P17	0.99	0.42	0.35	0.57	1
P21	0.80	0.54	0.00	0.56	3
P14	0.84	0.46	0.16	0.54	3
P1	1.00	0.35	0.31	0.52	1
P10	0.84	0.41	0.35	0.51	1
P8	0.85	0.42	0.01	0.51	3
P9	0.81	0.34	0.63	0.48	1
P7	0.69	0.42	0.26	0.47	1
P2	0.66	0.37	0.62	0.45	1
P6	0.76	0.26	0.24	0.40	1

Obr. 4.15: Farebné rozdelenie jednotlivých PZS do zhlukov



Obr. 4.16: 3D graf rozdelenia PZS do zhlukov č.1



Obr. 4.17: 3D graf rozdelenia PZS do zhlukov č.2

Záver

Hlavným výstupom našej práce bol benchmarking poskytovateľov zdravotnej starostlivosti, získaný pomocou zvolených dataminingových metód a modelovacích techník z dát o vykázanej zdravotnej starostlivosti. Našou cieľovou skupinou bola skupina diabetológov zapojených do programu DôveraPomáha diabetikom a ich lekárska činnosť za rok 2016. Tú sme opísali pomocou troch kritérií.

Jedným z nich bola efektivita využívaných nákladov, ktorej hlavnými parametrami boli výška nákladov, počet pacientov (rozdelených v skupinách podľa typu liečby), kontaktov s nimi, počet vykázaných výkonov a laboratórnych vyšetrení a počet predpísaných liekov. Druhým aspektom, a v benchmarkingu aj najdôležitejším, bola efektivita liečby z hľadiska pacienta. Tu sme sa pozreli na kmeň pacientov jednotlivých diabetológov a ich vykázanú zdravotnú starostlivosť. Do modelu vstupoval rovnako počet pacientov, počet komplikácií a hospitalizácií súvisiacich s ochorením diabetu, počet výjazdov rýchlej zdravotníckej pomoci a odmeraných hodnôt glykovaného hemoglobínu. Dodatkovým hodnotením bola miera aktivity v programe, ktorá závisela od percenta pacientov registrovaných a edukovaných v programe.

Výsledky týchto troch typov hodnotenia sa ukázali byť navzájom nezávislé. Lekárov sme na záver rozdelili do troch zhlukov podľa plnenia jednotlivých kritérií. Pri našej práci sme ale narazili na niekoľko možných vylepšení:

- Keďže k dispozícií sme mali iba dáta zo zdravotnej poisťovne Dôvera a mohli sme tým zachytiť len približne štvrtinu až tretinu pacientov daného lekára, bol by model presnejší, keby vychádzal z údajov o všetkých pacientoch.

- Náš benchmarking bol tvorený z 21 PZS a preto sme si mohli dovoliť len sedem parametrov vstupujúcich do jednotlivých DEA modelov. Platí však, že čím viac

vstupov a výstupov by sme mali, tým by bola produkčná množina presnejšia a tým aj výsledky modelu. Napríklad do hodnotenia efektivity liečby by mohli vstúpiť aj počty vyšetrení iných dôležitých klinických parametrov okrem HbA1c.

- Štandardne je v procese data miningu zahrnutá aj validácia modelu, ktorú sme však v našom modeli do veľkej miery vypustili a to hlavne z dôvodu, že nakoľko je takýto spôsob hodnotenia lekárov vo svojej podstate jediný, ktorým zdravotná poisťovňa disponuje, nebolo možné ju s níčím porovnať. Pri ďalšom využití získaného benchmarkingu by ale bolo vhodné výsledky validovať.

- Zhluky, ktoré sme získali v závere našej práce by sa dali porovnať a namodelovať pomocou iných údajov, ktoré o pacientoch máme, napríklad z formulárov pri vstupe do programu. Mohli by sme tak získať lepší prehľad o štruktúre pacientov jednotlivých lekárov a následne porovnať, ako to vplýva na ich výsledky v benchmarkingu

- Pravdepodobne najväčší benefit by mal opakovaný benchmarking PZS v čase. Jednak z hľadiska, ktoré sme navrhli v našej práci, na druhej strane by však bolo zaujímavé sledovať vývoj ochorenia jednotlivých pacientov s cukrovkou. Momentálne máme možnosť sledovať v dátach len krátke časové obdobie na takéto vyhodnotenie, v budúcnosti by ale mohol priniesť nové, zaujímavé pohľady.

Literatúra

[1] Uličiansky V., Schroner Z., Mokáň M.: *ViaDia. Sprievodca diabetika na ceste životom*, Vydavateľstvo P + M Turany, Turany, 2012

[2] Clark, M.: *Diabetes self-management education: A review of published studies*, Prim. Care Diabetes, 2008;2(3):113-120.

[3] ÚDZS: *Oznámenie - oprava výšky podielu poistencov jednotlivých zdravotných poisťovní na celkovom počte poistencov na základe definitívnych počtov poistencov k 1.1.2016 s prepočtom výšky preddavkov na ZZS pre rok 2016*

[4] NCZI: *Činnosť diabetologických ambulancií v SR 2015*, dostupné na internete (24.10.2016): http://www.nczisk.sk/Documents/publikacie/2015/zs1611.pdf

[5] Rud, O. P.: *Data mining Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management*, John Wiley and Sons Inc., New York, 2001

[6] Benková, B.: *Základné metódy datamineringu*, bakalárska práca, FMFI UK, Bratislava 2015

[7] Everitt, B. S.: *The Cambridge Dictionary of Statistics*, Cambridge University Press, New York, 2002

[8] SPSS: *CRISP-DM 1.0*, metodika, SPSS Inc., dostupná na internete (21.11.2016): http://the-modeling-agency.com/crisp-dm.pdf

[9] Halická, M.: *DEA modely (Učebné texty)*, FMFI UK, Bratislava, 2014

[10] Cooper, W.W., Seiford, L.M., Tone, K.: *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References, And DEA-Solver Software*, Springer US, 2007

[11] Harman, R.: *Cluster analysis* (učebné texty), FMFI UK, dostupné na internete (29.4.2017): http://www.iam.fmph.uniba.sk/ospm/Harman/VSAclust.pdf

[12] Ettorchi-Tardy, A., Levif, M., Michel, P.: Benchmarking: A Method for Continuous Quality Improvement in Health (research paper), In: *Healthcare Policy*, Longwoods Publishing, 2012, PMC3359088

[13] Stapenhurst, T.: *The Benchmarking Book: A How-to-Guide to Best Practice for Managers and Practitioners*, Elsevier Ltd., Oxford, 2009, ISBN: 978-0-7506-8905-2

[14] Lacka, J.: *osobná konzultácia*, Dôvera, zdravotná poisťovňa, Bratislava, 2017

[15] Halická, M.: *osobná konzultácia*, Katedra aplikovanej matematiky a štatistiky FMFI UK, Bratislava, 2017

[16] NCZI: *Medzinárodná klasifikácia chorôb MKCH 10*, dostupná na internete (7.2.2017): http://www.nczisk.sk/Standardy-v-zdravotnictve/Pages/MKCH-10-Revizia.aspx