**UNIVERZITA PARDUBICE**
**FAKULTA EKONOMICKO-SPRÁVNÍ**
**ÚSTAV SYSTÉMOVÉHO INŽENÝRSTVÍ A INFORMATIKY**

# ANALÝZA HOSPITALIZAČNÍCH PŘÍPADŮ V OBLASTI ZDRAVOTNICTVÍ

**DIPLOMOVÁ PRÁCE**

**AUTOR PRÁCE:** Eva Sýkorová
**VEDOUCÍ PRÁCE:** doc. Ing. Pavel Petr, Ph.D.

**2007**

**UNIVERSITY OF PARDUBICE**
**FACULTY OF ECONOMICS AND ADMINISTRATION**
**INSTITUTE OF SYSTEM ENGINEERING AND INFORMATICS**

# ANALYSIS OF HOSPITALIZATION´S CASES IN HEALTH CARE

**THESIS**

**AUTHOR:** Eva Sýkorová
**SUPERVISOR:** doc. Ing. Pavel Petr, Ph.D.

**2007**

Prohlašuji:

Tuto práci jsem vypracovala samostatně . Veškeré literární prameny a informace, které jsem v práci využila jsou uvedeny v seznamu použité literatury.

Byla jsem seznámena s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně Univerzity Pardubice.

V Pardubicích dne 25. května 2007

Eva Sýkorová

# Abstrakt

V práci je provedena analýza hospitalizačních případů v oblasti zdravotnictví pomocí data minigových metod, na základně kterých je možné získání základní představy o datech, statistických údajů a predikování jejich budoucího vývoje. Úvod je zaměřen na vymezení základních pojmů z teorie data miningu a popsány jeho základní techniky. V druhé části jsou za použití rozhodovacích stromů, regresní a shlukové analýzy v prostředí Clementine 10.1. analyzovány výsledky.

# Abstract

The aim of the thesis is to analyse hospitalization´s cases in health care using data mining methods which could give us the basic information about records, statistics and to predicate future development. The first part of the thesis describes basic terms and techniques of data mining. The second part presents the process how to realize analysis of hospitalization´s cases. Special software SPSS Clementine was used for generating prediction in decision trees, regression and clustering methods.

Touto cestou bych chtěla poděkovat vedoucímu diplomové práce panu doc. Ing. Pavlu Petrovi, Ph.D. za spolupráci a odbornou konzultaci při vypracování diplomové práce. Dále bych ráda poděkovala Ing. Jaroslavu Tachovskému za poskytnutí potřebných dat a informací k hospitalizačním případům.

# OBSAH

# SEZNAM POUŽITÝCH ZKRATEK

DM - Data Mining

CRISP-DM – Cross-Industry Standard Process for Data Mining

SQL - Structured Query Language

OLAP - On-line Analytical Processing

TDIDT - Top Down Induction of Decision Trees

CHAIM - Chi-square Automatic Interaction Detection

ČR - Česká republika

DRG - Diagnosis Related Group

# ÚVOD

Cílem diplomové práce je analyzovat hospitalizační případy v oblasti zdravotnictví. Najít skryté závislosti mezi daty a využít je pro vytvoření predikčního modelu. Provést statistickou analýzu dat, zhodnotit získané výsledky a na jejich základě následně uvést doporučení pro praxi.

Cílem první kapitoly je vymezit základní pojmy v oblasti data miningu, seznámit se s metodami, které data mining využívá a popsat metodologii CRIPS-DM.

Druhá kapitola podrobně popisuje metodu rozhodovacích stromů. Cílem třetí kapitoly je uvést základní metody shlukové analýzy. Čtvrtá kapitola se zabývá problematikou regresní analýzy.

Pátá kapitola je zaměřena na aplikaci metodologie CRIPS-DM na získaných datech o hospitalizačních případech. Podrobně jsou zde popsány všechny fáze této metodologie. Práce dále obsahuje seznam použitých zkratek, úvod, závěr, datový slovník, seznam použité literatury, seznam obrázků, tabulek a přílohu.

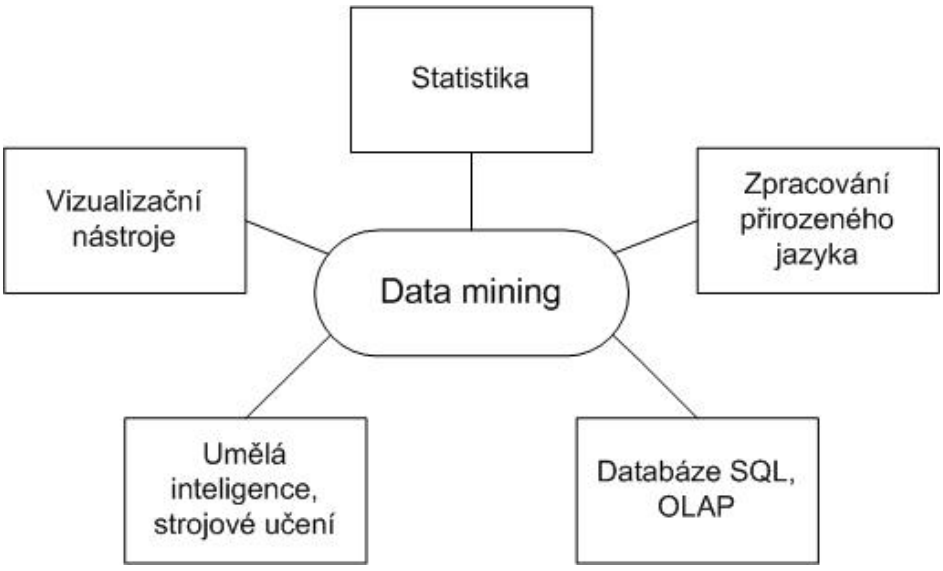
# 1 TEORIE DATA MININGU

## *1.1 DEFINICE DATA MININGU*

Ústředním pojmem této práce je data mining. Český význam slova by se dal přeložit jako: dolování dat, vytěžování dat či dobývání znalostí z databází. Data mining B. S. Everitt, The Cambridge Dictionary of Statistics, pojímá jako: „… netriviální dobývání skrytých, předem neznámých a potencionálně užitečných informací z dat. Při jejich objevování se využívají expertní systémy, grafické a statistické techniky a prezentují se způsobem srozumitelným lidem.“ [1].

## *1.2 ZDROJE DATA MININGU*

Zdroje, které DM využívá, jsou graficky znázorněny na obrázku 1.




**Obrázek 1 - Zdroje DM [1]**

## *1.3 TYPY ÚLOH ŘEŠENÝCH POMOCÍ DATA MININGU*

Data mining je založen na množství matematických a statistických technik. Zde jsou uvedeny pouze některé z nich [2]:

- **Rozhodovací stromy** – prediktivní model, který zobrazuje data v podobě stromu, kde každý uzel určuje kritérium pro následné rozdělení dat do jednotlivých větví.
- **Seskupovací analýza a klasifikace** – technika sloužící k rozdělení dat do skupin s obdobnými charakteristikami. Klasifikace definuje podstatné atributy skupin v podobě klasifikačních kritérií.
- **Regresní analýza** – metoda pro hodnocení závislosti jedné vysvětlované náhodné veličiny (závisle proměnné) $Y$ na jedné nebo několika vysvětlujících veličinách (nezávisle proměnných) $X_1, X_2, \ldots, X_k$.

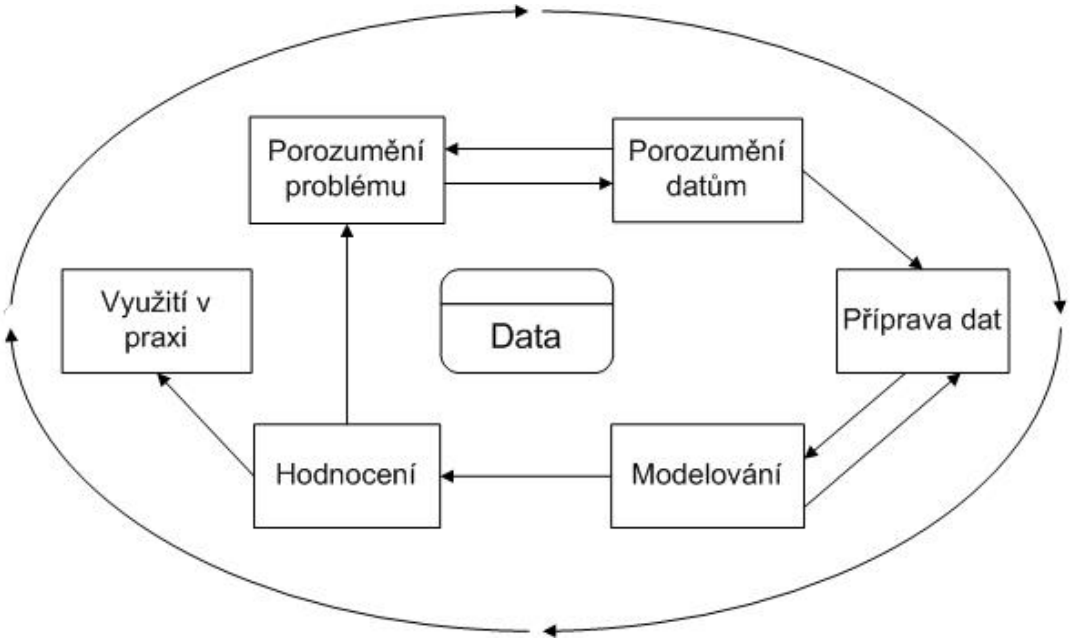
## *1.4 METODOLOGIE PRO DATA MINING*

Protože v současné době existuje spousta firem zabývajících se DM, existuje i několik metodologií. Jednou z nich je metodologie **CRISP-DM**. Cílem této metodiky je umožnění řešení rozsáhlé úlohy dobývání znalostí rychleji, efektivněji, spolehlivěji a s nižšími náklady.

**Fáze CRISP-DM [2]:**

- porozumění problému,
- porozumění datům,
- příprava dat,
- modelování,
- hodnocení,
- využití v praxi.

Jednotlivé fáze a jejich průběh je zobrazen na obrázku 2. Vnější okruh na obrázku symbolizuje cyklickou povahu procesu dobývání znalostí z databází jako takovou.




**Obrázek 2 - Fáze CRISP-DM [2]**

Tato metodologie je zde zmíněna z toho důvodu, že celá tato práce se od ní odvíjí a při její tvorbě bylo postupováno dle jednotlivých etap.

# 2 ROZHODOVACÍ STROMY

Rozhodovací stromy jsou jednou z nejoblíbenějších DM technik. Důvodů proto je několik. Hlavní důvod spočívá v jejich přehlednosti a snadné interpretovatelnosti, která umožňuje uživatelům rychle a lehce vyhodnocovat získané výsledky, identifikovat klíčové položky a vyhledávat zajímavé segmenty případů. Cílem rozhodovacích stromů je identifikovat objekty, popsané různými atributy, do tříd [5].

## *2.1 POUŽITÍ ROZHODOVACÍCH STROMŮ*

Použití rozhodovacích stromů pro klasifikaci odpovídá analogii s klíči k určování rostlin nebo živočichů. Od kořene stromu se na základě odpovědí na otázky (umístěné v nelistových uzlech) postupuje příslušnou větví stále hlouběji, až do listového uzlu, který odpovídá zařazení příkladu do třídy.

Rozhodovací stromy jsou vhodné pro úlohy, kde [5]:

- příklady jsou reprezentovány hodnotami atributů,
- úkolem je klasifikovat příklady do konečného počtu tříd [1],
- hledaný popis konceptů může být tvořen disjunkcemi,
- trénovací data mohou být zatížena šumem,
- trénovací data mohou obsahovat chybějící hodnoty.

## *2.2 ZÁKLADNÍ ALGORITMUS*

Při tvorbě rozhodovacího stromu se postupuje metodou „rozděl a panuj“. Trénovací data se postupně rozdělují na menší a menší podmnožiny (uzly stromu) tak, aby v těchto podmnožinách převládaly příklady jedné třídy. Na počátku tvoří celá trénovací data jednu množinu, na konci máme podmnožiny tvořené příklady z téže třídy. Tento postup bývá často nazýván TDIDT. Postupuje se tedy metodou specializace v prostoru hypotéz (stromů) shora dolů, počínaje stromem s jedním uzlem (kořenem). Cílem je nalézt nějaký strom konzistentní s trénovacími daty, přitom dává přednost menším, jednodušším stromům [4].

**Obecné schéma algoritmu [4]:**

1. Zvol jeden atribut jako kořen dílčího stromu.
2. Rozděl data v tomto uzlu na podmnožiny podle hodnot zvoleného atributu a přidej uzel pro každou podmnožinu.
3. Existuje-li uzel, pro který nepatří všechna data do téže třídy, pro tento uzel opakuj postup od bodu 1, jinak skonči.

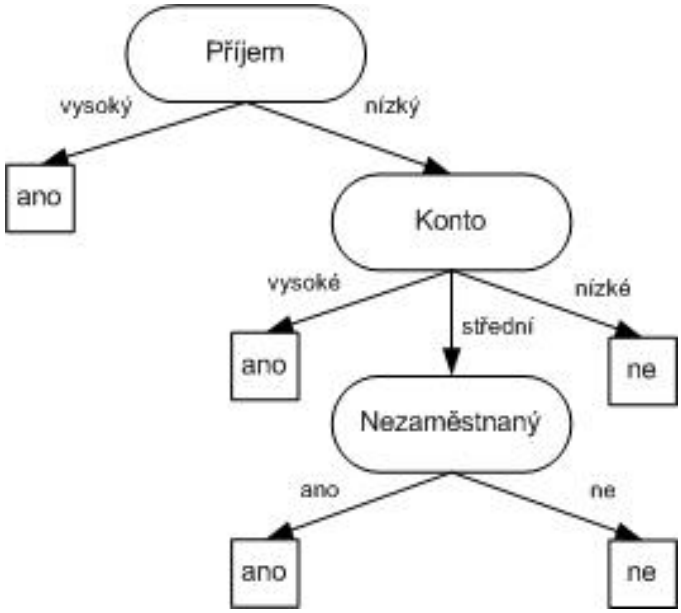
**Omezení obecného algoritmu [4]:**

- Jen pro kategoriální data (počet podmnožin uzlů vytvářený v kroku 2 odpovídá počtu hodnot daného atributu),
- Data nezatížená šumem (růst stromu se podle bodu 3 zastaví v okamžiku, kdy všechny příklady v daném uzlu patří do téže třídy).

[1] V případě regresních stromů můžeme předpovědět i hodnotu numerické veličiny.

## 2.3 SCHÉMA ROZHODOVACÍHO STROMU

Na obrázku 3 je uveden příklad úplného rozhodovacího stromu.




**Obrázek 3 - Úplný rozhodovací strom [4]**

Prvním krokem je vybrání takového atributu, který od sebe nejlépe odliší příklady různých tříd. Vodítkem pro volbu jsou charakteristiky atributu převzaté z teorie informace a pravděpodobnosti: entropie, informační zisk, Gini index.

**Entropie** - slouží pro vyjádření míry neuspořádanosti nějakého systému. Entropie je definovaná jako funkce:

$$H = -\sum_{t=1}^{T} (p_t \log_2 p_t), \tag{1}$$

kde: $H$ vyjadřuje entropii,

$p_t$ je pravděpodobnost výskytu třídy $t$,

$T$ *je* počet tříd.

**Informační zisk** - je míra odvozená z entropie. Zisk(A) se spočítá jako rozdíl entropie pro celá data a pro uvažovaný atribut. Měří redukci entropie způsobenou volbou atributu A:

$$Zisk(\mathbf{A}) = H(\mathbf{C}) - H(\mathbf{A}), \tag{2}$$

kde:

$$H(C) = -\sum_{t=1}^{T} \frac{n_t}{n} \log_2 \frac{n_t}{n}, \tag{3}$$

kde: $n$ je počet hodnot,

$n_t$ je počet hodnot třídy $t$.

Zatímco v případě entropie jsme hledali atribut s minimální hodnotou, v případě informačního zisku hledáme atribut s maximální hodnotou.

**Gini index –** lze použít místo entropie. Je definován následujícím vztahem:

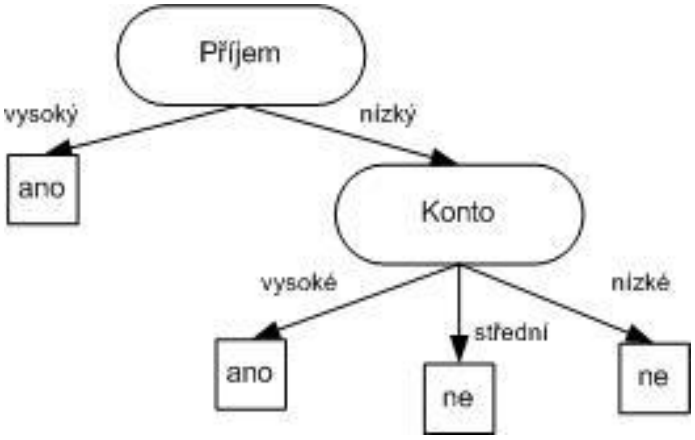
$$Gini = 1 - \sum_{t=1}^{T} (p_t^2), \tag{4}$$

kde: $p_t$ je relativní počet příkladů $t$-té třídy zjišťovaný na nějaké (pod)množině,

$T$ je počet tříd.

## *2.4 PROŘEZÁVÁNÍ STROMŮ*

Nejprve se vytvoří úplný strom. Ve fázi prořezávání se pak pro jednotlivé nelistové uzly posuzuje, do jaké míry úplný strom zhorší náhrada tohoto uzlu (a tedy odpovídajícího podstromu) listem. Náhrada nelistového uzlu listem totiž znamená, že všechny příklady v tomto uzlu, budou zařazeny do téže třídy. Vedlejším efektem této změny je skutečnost, že výsledný strom bývá menší, a tedy srozumitelnější pro interpretaci. Ukázka prořezaného rozhodovacího stromu je na obrázku 4.




**Obrázek 4 - Prořezaný rozhodovací strom [4]**

**Algoritmus prořezávání [4]:**

1. Převeď strom na pravidla.
2. Generalizuj pravidlo odstraněním podmínky za předpokladu, že dojde ke zlepšení odhadované přesnosti.
3. Uspořádej prořezaná pravidla podle odhadované přesnosti; v tomto pořadí budou pravidla použita pro klasifikaci.

## *2.5 REGRESNÍ STROMY*

Zatím jsme předpokládali, že vytváříme stromy pro klasifikaci objektů do tříd. Takovým stromům se obvykle říká klasifikační. Existují ale i stromy regresní, které umožňují odhadovat hodnotu některého numerického atributu. V listových uzlech mají takové stromy místo názvu tříd například konkrétní hodnotu (konstantu), která odpovídá průměrné hodnotě cílového atributu pro příklady v tomto uzlu.

Algoritmus pro tvorbu regresního stromu odpovídá algoritmu TDIDT. Rozdíl je ve způsobu volby atributu pro větvení. Místo entropie se vychází ze směrodatné odchylky hodnot cílového atributu.

$$S_y = -\sum_{v \in Val(A)} \frac{n(A(v))}{n} S_y(A(v)) \tag{5}$$

kde: $S_y$ je větvení,

$n$ je počet měřených znaků,

$A(v)$ je uvažovaný atribut $A$ s hodnotou $v$.

**Algoritmus CHAID**

Algoritmus CHAID se používá jako kritérium pro větvení $\chi^2$. Při větvení se nevytváří tolik větví, kolik má atribut hodnot a hodnoty atributu se postupně seskupují z původního počtu až do dvou skupin. Dále se vybere atribut a jeho kategorizace, která je v daném kroku pro větvení nejlepší.

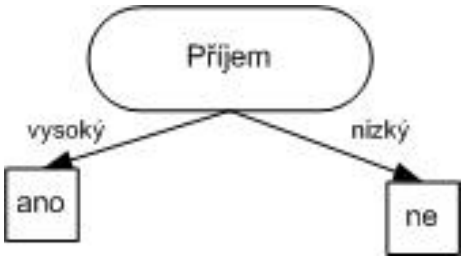
**Algoritmus seskupování hodnot atributu [4]:**

1. Opakuj dokud nevzniknou pouze dvě skupiny hodnot atributu.
    - Zvol dvojici kategorií atributu, které jsou si nejpodobnější z hlediska $\chi^2$ a které mohou být spojeny.

- Považuj novou kategorizaci atributu za možné shlukování v daném kroku.

2. Spočítej pomocí $\chi^2$ testu pravděpodobnost $p$ pro každý z možných způsobů shlukování hodnot.
3. Shlukování s nejnižší pravděpodobností $p$ zvol za nejlepší shlukování hodnot atributu.
4. Zjisti, jestli toto nejlepší shlukování statisticky významně přispěje k odlišení příkladů různých tříd.

Na obrázku 5 je ukázka vytvořeného rozhodovacího pařezu.



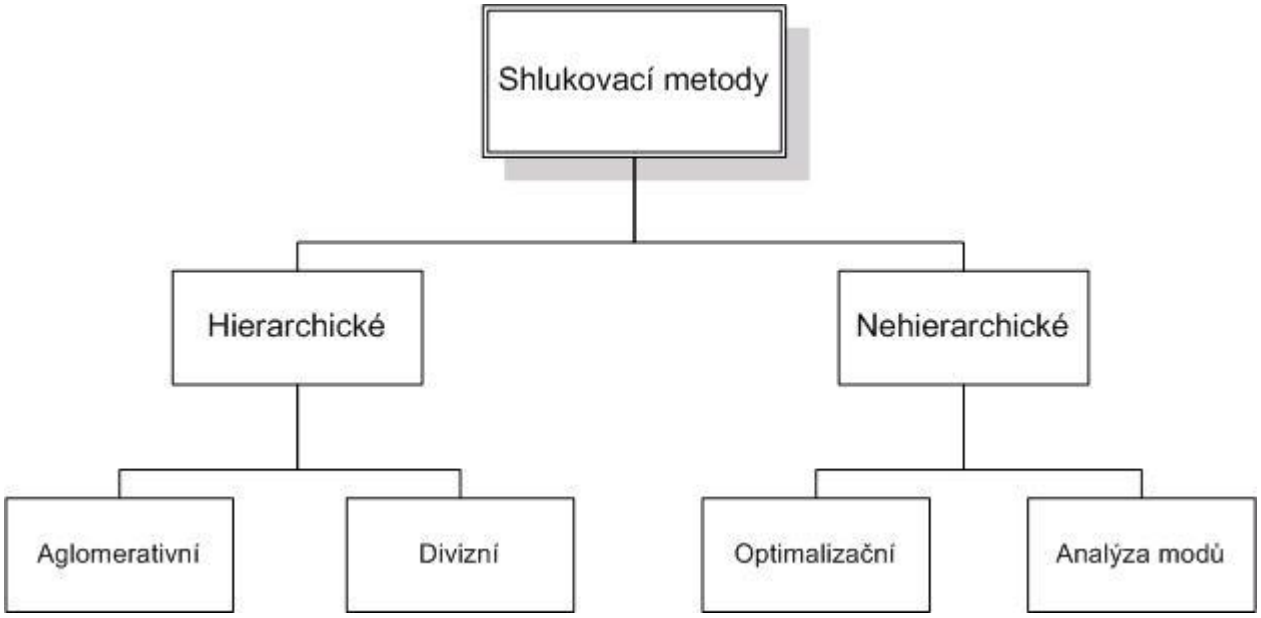

**Obrázek 5 - Rozhodovací pařez [4]**

# 3 SHLUKOVÁ ANALÝZA

Analýza shluků patří mezi metody, které se zabývají vyšetřováním podobnosti vícerozměrných objektů tj. objektů, u nichž je změřeno větší množství znaků a následnou klasifikací přiřazení objektů do shluků [5].

## *3.1 METODY SHLUKOVÉ ANALÝZY*

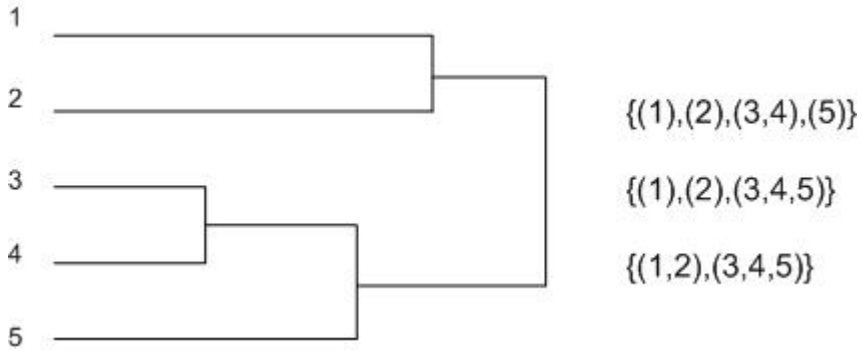
V literatuře se objevuje takové množství shlukovacích metod, že je obtížné nějak rozumně je utřídit. Nejčastější způsob utřídění metod je nikoli podle použitých matematických metod, ale podle systému použité klasifikace [6]. Rozdělení shlukovacích metod je znázorněno na obrázku 6.




**Obrázek 6 - Shlukovací metody [2]**

**Hierarchické shlukovací metody** – tyto shlukovací metody směřují k hierarchické klasifikaci. Z hierarchických metod jsou zde uvedeny dvě základní skupiny, lišící se způsobem shlukování [7].

- Aglomerativní přístup – dva objekty, jejichž vzdálenost je nejmenší, se spojí do prvního shluku a vypočte se nová matice vzdáleností, v níž jsou vynechány objekty z prvního shluku a naopak tento shluk je zařazen jako celek. Celý postup se opakuje tak dlouho, dokud všechny objekty netvoří jeden velký shluk nebo dokud nezůstane určitý, předem zadaný počet shluků. Jednou možností reprezentace tohoto přístupu je dendrogram, který je znázorněn na obrázku 7.




**Obrázek 7 - Dendrogram [zdroj vlastní]**

- Divizní přístup – je obrácený. Vychází se z množiny všech objektů jako jediného shluku a jeho postupným dělením získáme systém shluků, až skončíme ve stádiu jednotlivých objektů [1].

**Nehierarchické shlukovací metody** – hledají nejlepší rozklad množiny objektů iteračním způsobem. Počáteční rozklad zlepšují tak, že hledají rozklad s lepší hodnotou kriteriální funkce.

- Optimalizační – u této metody je počet shluků obvykle předem dán a optimální rozklad se hledá přeřazováním objektů ze shluku do shluku s cílem minimalizovat nebo maximalizovat nějakou charakteristiku rozkladu. Patří sem například k-středové metody.
- Analýza modů – představuje hledání rozkladu do shluků, kde shluky jsou chápány jako místa se zvýšenou koncentrací v *m*-rozměrném prostoru proměnných [7].

**Aglomerativní přístup**

Aglomerativní přístup v rámci hierarchické metody je charakteristický tím, že vycházíme od jednotlivých objektů a jejich postupným seskupováním bodujeme hierarchický systém podmnožin až dospějeme ke konečnému spojení všech objektů do množiny objektů *O*.

Shlukovou analýzu prováníme zpravidla na množině *n* objektů ($O_1$, $O_2$,...,$O_n$), z nichž každý je popsán prostřednictvím *p* ukazatelů ($u_1$, $u_2$,...,$u_p$), které má smysl na dané množině objektů sledovat. Výběr množiny sledovaných ukazatelů rozhoduje o úspěchu závěrů metody, proto je nutné mu věnovat patřičnou pozornost.

Předpokládejme, že na množině objektů $O_i$, $i = 1,2,...,n$ bylo změřeno *p* ukazatelů. Tím získáme vektor $X_i = (x_{i1}, x_{i2},...x_{ip})$. Výsledkem pozorování je matice *X* typu $n \times p$.

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix},$$

kde: $x_{ik}$ vyjadřuje hodnotu *k*-tého ukazatele u *i*-tého objektu.

Tuto matici nazýváme datovou maticí. Jednotlivá měření, čili vektory $X_i$ tvoří řádky matice. Sloupce datové matice *X* vyjadřují hodnoty *k*-tého znaku v množině objektů, $k = 1,2,...,p$. Budeme předpokládat, že byly naměřeny pouze kvantitativní znaky a tedy vektory $X_i$ jsou číselné vektory. Každý z těchto vektorů může pak být zobrazen v *p*-rozměrném Eukleidovském prostoru $E_p$ jako

bod. To znamená, že každý ze sledovaných objektů je zobrazen v Eukleidovském prostoru jako bod a přecházíme ke zkoumání bodové množiny a jejich rozkladů v $E_p$. Úkolem je rozdělení množiny pozorování do disjunktních skupin $S_1, S_2, \ldots, S_m$ $(m \leq n)$, nazývaných shluků.

Základním problémem je kvantitativně vyjádřit podobnost či vzdálenost objektů. V jednotlivých krocích algoritmů posuzujeme podobnost dvou objektů, objektu shluku nebo dvou shluků. U míry podobnosti zpravidla požadujeme, aby nabývala hodnoty 0 pro maximální rozdílnost a hodnoty 1 pro totožnost objektů.

**Mírou nepodobnosti** vektorů $E_p$ je nezáporná reálná funkce $m$, která každé dvojici vektorů $X_i$, $X_j$ z $E_p$ přiřazuje číslo $m_{ij}$, jestliže pro všechny dvojice $X_i$, $X_j$ platí:

- $0 \leq m(X_i, X_j) < 1$ pro $X_i \neq X_j$,
- $m(X_i, X_j) = m(X_i, X_j)$,
- $m(X_i, X_j) = 0$.

Jako míry nepodobnosti jsou typické funkce založené na vzdálenosti objektů. Základní myšlenka je, že čím větší je vzdálenost dvou objektů v $E_p$, tím jsou tyto objekty méně podobné, čili tím mají větší míru nepodobnosti.

Vzdálenost dvou bodů $X_i$, $X_j$ z $E_p$ se definuje jako nezáporná reálná funkce $d(X_i, X_j)$, pro kterou platí:

- $d(X_i, X_j) \geq 0$ pro všechny body $X_i$, $X_j$ z $E_p$,
- $d(X_i, X_j) = 0$ právě tehdy, když $X_i = X_j$,
- $d(X_i, X_j) = d(X_i, X_j)$,
- $d(X_i, X_j) + d(X_j, X_k) \geq d(X_i, X_k)$ pro každou trojici bodů $X_i$, $X_j$, $X_k$ z $E_p$.

K určení vzdáleností objektů se ve shlukové analýze používají různé způsoby výpočtu vzdáleností. Nejběžnější je **Euklidovská vzdálenost**, která se spočítá podle vztahu:

$$d(X_i, X_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \qquad (6)$$

kde: $x_{jk}$ je hodnota $k$-tého pozorování na $i$-tém prvku,

$x_{jk}$ je hodnota $k$-tého pozorování na $j$-tém prvku objektu.

Hodnota Eukleidovské vzdálenosti je závislá na jednotkách proměnných, proto je výhodné provést normovací transformaci proměnných [8].

**Princip hierarchických aglomerativních shlukovacích metod**

Prvním krokem shlukové analýzy je pomocí některé ze vzdáleností vypočítat matici vzdáleností $D_1 = \{d_{ij}\}$, která je symetrická, má na hlavní diagonále nuly a nediagonální prvky vyjadřují vzdálenosti $d(X_i,X_j)$ mezi jednotlivými dvojicemi objektů.

V dalším kroku je nutné se rozhodnout pro některou ze shlukovacích metod. Všechny shlukovací metody pracují na základě stejného algoritmu, liší se však způsobem výpočtu vzdáleností mezi dvěma objekty. Označíme $d_{ij}(S_h,S_k)$ míry vzdálenosti mezi dvěma shluky $S_h$ a $S_k$. Nejbližší dva prvky tvoří první shluk. Dva shluky se vždy spojí v jeden, jestliže je mezi nimi minimální vzdálenost. Po prvním kroku jsme tedy spojili do shluku dva nejbližší objekty a vypočítáme vzdálenost tohoto shluku od zbývajících prvků. Z vypočítaných vzdáleností sestavíme novou matici vzdáleností $D_2$. Opět najdeme nejmenší vzdálenost a tento postup opakujeme. Další nový shluk je pak tvořen buď dvěma nejbližšími prvky, prvkem a shlukem nebo dvěma shluky. Postup se opakuje tak dlouho, dokud není dosaženo požadovaného počtu shluků.

Existuje spousta nejčastěji používaných metod shlukování. Jako příklad je zde uvedena **metoda průměrné vzdálenosti**:

$$d(S_h, S_k) = \frac{1}{n_h n_k} \sum_{x_i \in S_h} \sum d(X_i, X_j) \tag{7}$$

kde: $d(S_h,S_k)$ je míra vzdálenosti mezi shluky $S_h$ a $S_k$,

$d(X_i,X_j)$ je vzdálenost mezi jednotlivými dvojicemi objektů.

**Základní algoritmus nehierarchických optimalizačních shlukovacích metod [7]:**

1. Zadání počátečních typických bodů.
2. Přiřazení každého bodu k nejbližšímu typickému bodu a jeho odpovídajícímu shluku.
3. Výpočet těžiště každého z *k* shluků.
4. Definování nových typických bodů ve vypočtených těžištích.
5. Pokud došlo ke změně v přiřazení bodů shlukům, opakuj od bodu 2.
6. Výpočet kriteriální funkce výsledného rozkladu.

# 4 REGRESNÍ ANALÝZA

**Regresní analýza** je označení statistických metod, pomocí nichž odhadujeme hodnotu jisté náhodné veličiny (závisle proměnné) na základě znalosti jiných veličin (nezávisle proměnných). Hlavním úkolem regresní analýzy je zjištění tvaru stochastické závislosti a parametrů regresní funkce.

Regresní analýza se zabývá závislostí náhodné veličiny $Y$ na nezávisle proměnné $x$, která není náhodná a může být obecně $m$-rozměrná. Náhodná veličina $Y$ má pro danou hodnotu $x = (x_1, x_2, \ldots, x_m)$ a parametry $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$ podmíněnou střední hodnotu $E(Y/x) = g(x, \beta_0, \beta_1, \beta_2, \ldots, \beta_k)$. Funkce $g$ proměnné $x$ se nazývá regresní funkce a parametry $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$ se nazývají regresní koeficienty.

Modely lineární vzhledem k parametrům mají regresní funkci tvaru:

$$g(x, \beta_0, \beta_1, \beta_2, \ldots, \beta_k) = \sum_{i=0}^{k} \beta_i g_i(x), \tag{8}$$

kde: $g_i$ jsou funkce nezávisle proměnných $x = (x_1, x_2, \ldots, x_m)$.

**Regresní funkci** lze definovat následujícím způsobem: nechť $X$ a $Y$ jsou náhodné veličiny. Podmíněnou střední hodnotou $E(Y/x)$, považovanou za funkci proměnné $x$, budeme nazývat regresní funkcí náhodné veličiny $Y$ vzhledem k $X$. Regresní funkce vyjadřuje změny podmíněné střední hodnoty jedné náhodné veličiny při změně druhé náhodné veličiny [8].

## *4.1 LINEÁRNÍ REGRESNÍ MODELY, METODA NEJMENŠÍCH ČTVERCŮ*

Při budování regresních modelů se běžně užívá metody nejmenších čtverců. Tato metoda poskytuje postačující odhady parametrů jenom při splnění předpokladů o datech a o regresním modelu. Pokud tyto předpoklady nejsou splněny, ztrácí výsledky metodou nejmenších čtverců své vlastnosti.

Lineární regresní analýza se používá v těchto případech [8]:

- **Popisu empirických dat** – hledá se vztah, lineární regresní model , který sumarizuje vazby mezi sloupci v datech.
- **Určení parametrů** – běžným cílem regresní analýzy je vyčíslení odhadů neznámých parametrů regresního modelu. Uživatel navrhne regresní model a regresní analýzou se snaží model prokázat. Často tento cíl překrývá i ostatní záměry regresní analýzy.

- **Predikce** – cílem regresní analýzy je často predikce, tj. vyčíslení hodnot závisle proměnných pro zadané kombinace vstupních parametrů. Predikce jsou důležité při plánovaní, monitorování a vyhodnocování procesů.
- **Řízení** – lze využít také k monitoringu a řízení systémů.
- **Výběru důležitých proměnných** – výběr proměnných se provádí s ohledem na nezávisle proměnné, které vysvětlují významný podíl proměnlivosti na závisle proměnné.

**Jednoduchým lineárním modelem** lineární regrese nazýváme takový lineární model, kdy grafem regresní funkce je přímka. Předpokládejme, že $Y_1$, $Y_2$,..., $Y_n$ je $n$-tice náhodných veličin s vlastnostmi $EY_i = \alpha + \beta x_i$, $DY_i = \sigma_2$, $i = 1,2,..., n$, kde $\alpha$, $\beta$, $\sigma_2$ jsou neznáme parametry a $x_1$, $x_2$, …, $x_n$ je $n$-tice známých hodnot.

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \tag{9}$$

kde: $\varepsilon_i$ jsou nezávislé náhodné veličiny, pro které platí $E\varepsilon_i = 0$, $D\varepsilon_i = \sigma^2$, $i = 1,2,...,n$.

Náhodná složka zahrnuje působení náhodných vlivů nebo působení veličin, které nejsou zahrnuty v modelu.

Regresní přímka $y = \alpha + \beta x$ se nazývá regresní přímka, $\beta$ je její směrnice. Úkolem je nyní na základě naměřených dvojic hodnot $(x_1,y_1)$, $(x_2,y_2),...,(x_n,y_n)$ odhadnout neznámé parametry $\alpha$, $\beta$, $\delta^2$ daného modelu. Tyto odhady budeme značit po řadě $a$, $b$, $s^2$.

Bodové odhady $a$, $b$ parametrů $\alpha$, $\beta$ získáme metodou nejmenších čtverců. Princip této metody spočívá v tom, že hledáme takovou funkci $\hat{y} = a + bx$, aby v jistém smyslu co nejvíce přiléhala k bodům $(x_1,y_1)$, $(x_2,y_2),...,(x_n,y_n)$, kde přiléhání měříme součtem rozdílu hodnot $\hat{y}_i$ a $y_i$ [9].

**Předpoklady metody nejmenších čtverců [9]:**

- Regresní parametry $\beta$ mohou nabývat libovolných hodnot.
- Regresní model je lineární v parametrech a platí aditivní model měření.
- Matice nenáhodných, nastavovaných hodnot vysvětlujících proměnných $X$ má hodnost rovnou právě $m$.
- Náhodné chyby $\varepsilon_i$ mají nulovou střední hodnotu $E(\varepsilon_i) = 0$. To musí u korelačních modelů platit vždy. U regresních modelů se může stát, že $E(\varepsilon_i) = K$, $i = 1,...,n$, což znamená, že model neobsahuje absolutní člen. Po jeho zavedení bude $E(\varepsilon_i') = 0$, kde $\varepsilon_i' = y_i - \hat{y}_{P,i} - K$.

- Náhodné chyby $\varepsilon_i$ mají konstantní a konečný rozptyl $E(\varepsilon_i^2) = \delta^2$. Také podmíněný rozptyl $D(y/x) = \delta^2$ je konstantní a jde o homoskedastický případ.
- Náhodné chyby $\varepsilon_i$ jsou vzájemně nekorelované a platí $cov(\varepsilon_i \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0$
- Chyby $\varepsilon_i$ mají normální rozdělení $N(0, \delta^2)$.

Snažíme se najít takové odhady *a*, *b*, aby platilo:

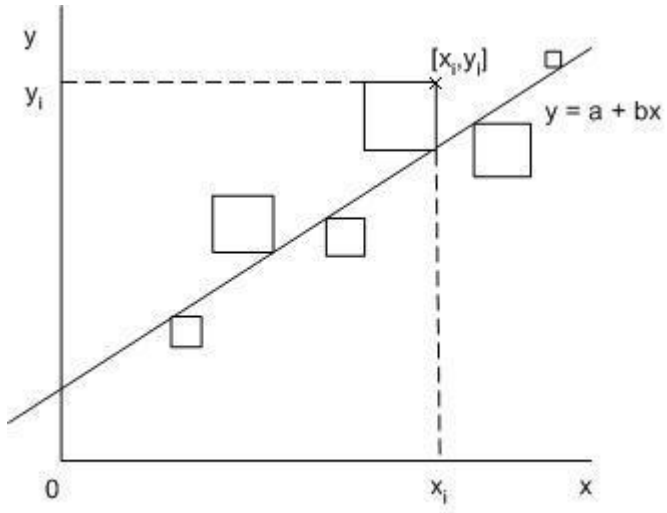
$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow \min \tag{10}$$

Odhady parametrů *a*, *b*:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \tag{11}$$

$$a = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}. \tag{12}$$

Myšlenka metody je grafický znázorněna na následujícím obrázku 8:




**Obrázek 8 - Metoda nejmenších čtverců [8]**

# 5 APLIKACE METODOLOGIE CRISP-DM

Celé zpracování této práce podléhá metodologii CRISP-DM, proto jsou jednotlivé části rozděleny do fází této metodologie.

## *5.1 POROZUMĚNÍ PROBLÉMU*

Údaje o zdravotnické péči představují neuvěřitelný potenciál pro zjištění užitečných informací, které nám poskytují ukazatele kvality lékařské péče, úspěšnosti řešení hospitalizačních případů a veřejného zdraví obecně. A právě tyto informace nám mohou poskytnout údaje o hospitalizačních případech. Najít mezi nimi skryté závislosti a využít je pro vytvoření predikčního modelu. Provést jejich základní statistickou analýzu, zhodnotit získané výsledky a na jejich základě následně uvést doporučení pro praxi.

## *5.2 POROZUMĚNÍ DATŮM*

Tato analýza je zaměřena na administrativní data. Jedná se o data, která jsou v běžné praxi shromažďována povinně pro účely Českého statistického úřadu, pojišťovacích společností a podobně. Pro řešení tohoto problému byl zvolen nástroj pro analýzu dat: Clementine 10.1 od společnosti SPSS.

Data byla získána od společnosti Stapro, s. r. o., která je významným dodavatelem informačních systémů, diagnostických přístrojů, zdravotnické techniky a zároveň i poskytovatelem služeb v oblasti informačních technologií pro zdravotnictví. Po obdržení dat je možné získání základní představy o těchto datech a možnost vytipování zajímavých podmnožin záznamů v databázi. Tato databáze se skládá z informací o asi 370 000 záznamech o hospitalizačních případech ze 14ti nemocnic v České republice, Slovensku, Rakousku a Polsku. Konkrétně 4 nemocnice v České republice a na Slovensku. Dále 3 nemocnice v Rakousku a Polsku.

Jelikož je tato analýza zaměřena na hospitalizační případy, je třeba vymezit základní pojmy týkající se této problematiky:

**Hospitalizační případ**: souvislý pobyt na lůžkovém oddělení v rámci jedné nemocnice.

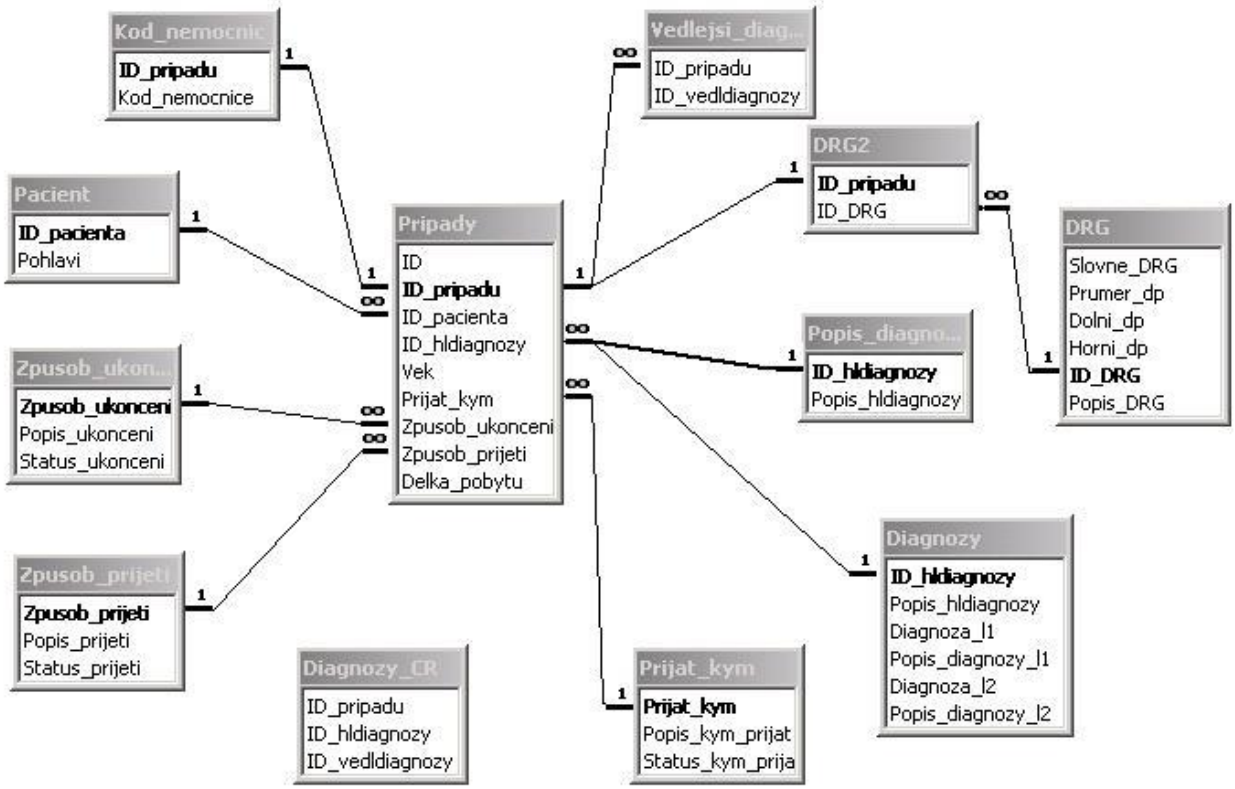
**Hlavní diagnóza:** stav diagnostikovaný na konci období zdravotní péče, primárně odpovědný za potřebu nemocného léčit nebo být vyšetřován. Existuje-li více než jeden takový stav, vybírá se ten, který se považuje nejvíce zodpovědný za největší čerpání prostředků.

**Vedlejší diagnóza**: diagnóza (popřípadě diagnózy), která je zjištěná při vyšetření ve zdravotnickém zařízení a není vyhodnocena jako hlavní diagnóza.

**Délka pobytu**: počet dnů, po kterou je pacient hospitalizován v nemocnici. Vypočítá se jako: (konečný den – počáteční den) – 1. Pokud je výsledek nulový, pobyt je chápán jako jednodenní.

## Relační schéma

Na obrázku 9 je vidět relační schéma získaných dat. Tabulka Diagnózy ČR není v relaci s žádnou další tabulkou, proto analýzy týkající se těchto dat budou dělány odděleně.




**Obrázek 9 - Relační schéma [zdroj vlastní]**

## Ukázka dat

Tabulka Případy

Atributy této tabulky 1 jsou: ID případu, ID pacienta, ID hlavní diagnózy, věk pacienta, kým byl přijat, způsob ukončení, způsob přijetí a délka pobytu.

| | ID_pripadu | ID_pacienta | ID_hldiagnozy | Vek | Prijat_kym | Zpusob_ukonceni | Zpusob_prijeti | Delka_pobytu |
|---|---|---|---|---|---|---|---|---|
| 1 | PL1907726 | 330210bel... | C56 | 71 | PL1 | PL5 | PL3 | 15 |
| 2 | PL1907727 | 790904cihhf | O800 | 24 | PL1 | PL1 | PL1 | 5 |
| 3 | PL1907728 | 380220cfilj | C539 | 66 | PL1 | PL1 | PL3 | 5 |
| 4 | PL1907729 | 860903bllh... | M411 | 17 | PL1 | PL3 | PL2 | 1 |
| 5 | PL1907730 | 740729bln... | O800 | 29 | PL1 | PL1 | PL1 | 7 |
| 6 | PL1907731 | 620510cgjhk | O840 | 42 | PL1 | PL1 | PL1 | 5 |
| 7 | PL1907732 | 510222bje... | N300 | 53 | PL1 | PL2 | PL2 | 6 |
| 8 | PL1907733 | 640624bjk... | N132 | 40 | PL1 | PL2 | PL3 | 4 |
| 9 | PL1907734 | 340101fjegl | N430 | 70 | PL1 | PL2 | PL1 | 18 |
| 10 | PL1907735 | 310212cfg... | N40 | 73 | PL1 | PL2 | PL1 | 1 |

**Tabulka 1- Ukázka tabulky Případy [zdroj vlastní]**

Tabulka Pacient

Atributy této tabulky 2 jsou: ID pacienta a pohlaví. Žena je zde označena písmenem „F“ (Female), muž písmenem „M“ (Male).

| | ID_pacienta | Pohlavi |
|---|---|---|
| 1 | CZ.14cb1d67e25e3332ae2fa0c5f964e044 | M |
| 2 | CZ.14d045bdbaa60bcc42c159d03b5fa458 | M |
| 3 | CZ.14d58dcaefbebb701a4363b44afe88ab | F |
| 4 | 531213bmlgd | M |
| 5 | CZ.14d83f6442986def91e808d6bb474f30 | F |
| 6 | 280109bgjgg | M |
| 7 | 880218bmfmf | M |
| 8 | 320103bfgoj | M |
| 9 | CZ.14e3f19e76aab6c29d527e8c3f59db16 | F |
| 10 | 260617bjkge | M |

**Tabulka 2 - Ukázka tabulky Pacient [zdroj vlastní]**

Tabulka Vedlejší diagnózy

Atributy této tabulky 3 jsou: ID případu a ID vedlejší diagnózy.

| | ID_pripadu | ID_vedl.diagnozy |
|---|---|---|
| 1 | K4X12003184952 | N40 |
| 2 | K4X12003184958 | E039 |
| 3 | K4X12003184958 | I10 |
| 4 | K4X12003184972 | I10 |
| 5 | K4X12003184972 | I48 |
| 6 | K4X12003184973 | S017 |
| 7 | K4X12003184973 | S099 |
| 8 | K4X12003184973 | S501 |
| 9 | K4X12003184973 | S510 |
| 10 | K4X12003184973 | S5250 |

**Tabulka 3 - Ukázka tabulky Vedlejší diagnózy [zdroj vlastní]**

Tabulka Diagnózy

Atributy této tabulky 4 jsou: ID hlavní diagnózy, popis hlavní diagnózy, třída „l1“, do které je hlavní diagnóza zařazena, slovní popis třídy „l1“, třída „l2“, do které je také hlavní diagnóza zařazena a její slovní popis.

| | ID_hldiagnozy | Popis_hldiagnozy | Diagnoza_l1 | Popis_diagnozy_l1 | Diagnoza_l2 | Popis_diagnozy_l2 |
|---|---|---|---|---|---|---|
| 1 | A27 | A27-Leptospirosis | 10 | 01-Infectious | A27 | A27-Leptospirosis |
| 2 | A279 | A279-Leptospiro... | 10 | 01-Infectious | A27 | A27-Leptospirosis |
| 3 | A281 | A281-Cat-scratc... | 10 | 01-Infectious | A28 | A28-Other zoonoti... |
| 4 | A282 | A282-Extraintesti... | 10 | 01-Infectious | A28 | A28-Other zoonoti... |
| 5 | A30 | A30-Leprosy [Ha... | 10 | 01-Infectious | A30 | A30-Leprosy [Han... |
| 6 | A302 | A302-Borderline ... | 10 | 01-Infectious | A30 | A30-Leprosy [Han... |
| 7 | A305 | A305-Lepromato... | 10 | 01-Infectious | A30 | A30-Leprosy [Han... |
| 8 | A308 | A308-Other form... | 10 | 01-Infectious | A30 | A30-Leprosy [Han... |
| 9 | 32 | A32-Listeriosis | 10 | 01-Infectious | A32 | A32-Listeriosis |
| 10 | A320 | A320-Cutaneous... | 10 | 01-Infectious | A32 | A32-Listeriosis |

**Tabulka 4 - Ukázka tabulky Diagnózy [zdroj vlastní]**

Tabulka Popis diagnózy

Atributy této tabulky 5 jsou: ID hlavní diagnózy a popis hlavní diagnózy v českém jazyce.

| | ID_hldiagnozy | Popis_hldiagnozy_CJ |
|---|---|---|
| 1 | Z887 | Alergie na sérum a vakcinu v osobní anamnéze |
| 2 | Z888 | Alergie na j. léky léčiva návykové a biologické látky v osobní anamnéze |
| 3 | Z889 | Alergie na neurčené léky léčiva návykové a biologické látky v osobní anamnéze |
| 4 | Z89 | Získané chybění končetiny |
| 5 | Z890 | Získané chybění prstu(-ů) ruky [včetně palce] jednostranné |
| 6 | Z891 | Získané chybění ruky a zápěstí |
| 7 | Z892 | Získané chybění horní končetiny nad zápěstím |
| 8 | Z893 | Získané chybění obou horních končetin [kterákoliv úroveň] |
| 9 | Z894 | Získané chybění nohy pod kotníkem a kotníku |
| 10 | Z895 | Získané chybění nohy v nebo pod kolenem |

**Tabulka 5 - Ukázka tabulky Popis diagnózy [zdroj vlastní]**

Tabulka Způsob přijetí

Atributy této tabulky 6 jsou: způsob přijetí, popis způsobu přijetí a status způsobu přijetí. Statusy způsobů přijetí jsou zde uvedeny čtyři, a to: „EME“ (Emergency – neodkladné), „OTH“ (Other – jiné), „NOA“ (Not Available – nedosažitelné) a „NOK“ (Not Known – neznámo). Každý stát zde využívá pouze některé z těchto statusů, proto v dalších analýzách budou využívány pouze první dva zmíněné.

| | Zpusob_prijeti | Popis_prijeti | Status_prijeti |
|---|---|---|---|
| 1 | CZ1 | Neodkladné přijetí | EME |
| 2 | CZ2 | Plánované přijetí | OTH |
| 3 | CZ3 | Jiné přijetí | OTH |
| 4 | CZ4 | Léčebný | OTH |
| 5 | CZ5 | Jiný | OTH |
| 6 | CZX | Není přijetí | NOA |
| 7 | CZON | Neznámo | NOK |
| 8 | SK1 | neodkladné prijatie | EME |
| 9 | SK2 | prijatie na objednávku | OTH |
| 10 | SK3 | iný dôvod prijatia | OTH |

**Tabulka 6 - Ukázka tabulky Způsob přijetí [zdroj vlastní]**

Tabulka Způsob ukončení

Atributy této tabulky 7 jsou: kód způsobu ukončení, slovní popis způsobu ukončení a status způsobu ukončení. Statusy způsobu ukončení jsou zde základní tři. A to „DEA“ (Death - zemřel), „DIS“ (Discharge - propuštěn) a „TRA“ (Transfer - převezen). Zbylé dva atributy nejsou příliš významné a jejich zastoupení bylo mizivé. „NOD“ (Not Discharge – nepropuštěn) a „OTH“ (Other - ostatní) tvořily společně pouze 1 %. Proto nebudou v dalších analýzách brány v potaz.

| | Zpusob_ukonceni | Popis_ukonceni | Status_ukonceni |
|---|---|---|---|
| 1 | CZ0 | Pokračuje ústavní péče na stejném lůžku | NOD |
| 2 | CZ1 | Propuštěn do ambulantní péče | DIS |
| 3 | CZ2 | Přeložen do ústavní péče - sociální péče | DIS |
| 4 | CZ3 | Přeložen do ústavní péče - jiná odbornost, stejné ZZ | NOD |
| 5 | CZ4 | Přeložen do ústavní péče - do LDN | TRA |
| 6 | CZ5 | Přeložen do ústavní péče - do jiného ZZ | TRA |
| 7 | CZ6 | Propuštěn do ambulantní péče předčasně | DIS |
| 8 | CZ7 | Zemřel, vystaven poukaz na pitvu | DEA |
| 9 | CZ8 | Zemřel, nevystaven poukaz na pitvu | DEA |
| 10 | CZA | Ostatní | OTH |

**Tabulka 7 - Ukázka tabulky Způsob ukončení [zdroj vlastní]**

Tabulka Přijat kým

Atributy této tabulky 8 jsou: kód kým byl pacient přijat, slovní popis kým byl přijat a status kým byl přijat. Statusy jsou zde podobné jako v předchozích tabulkách. „OTH“ (Other – ostatní), „TRA“ (Transfer – převezen) a „NOA“ (Not Available – nedosažitelné).

| | Prijat_kym | Popis_kym_prijat | Status_kym_prijat |
|---|---|---|---|
| 1 | CZ1 | Praktický lékař | OTH |
| 2 | CZ2 | Jiný ošetřující lékař - ambulantní péče | OTH |
| 3 | CZ3 | Lékař LSPP | OTH |
| 4 | CZ4 | Lékař RZP | OTH |
| 5 | CZ5 | Jiné ZZ | TRA |
| 6 | CZ6 | Jiné oddělení téhož ZZ | NOA |
| 7 | CZ7 | Bez doporučení lékaře | OTH |
| 8 | CZ8 | Sociální pracovník | OTH |
| 9 | CZ9 | Hospitalizace pokračuje | NOA |
| 10 | CZ0 | Narozen v nemocnici | OTH |

**Tabulka 8 - Ukázka tabulky Přijat kým [zdroj vlastní]**

Tabulka DRG

DRG je hospitalizační systém, který zařazuje nemocného léčeného na akutních lůžkách nemocnic podle složitosti jeho onemocnění a ekonomické náročnosti jeho léčby. Toto zařazení umožňuje v určité fázi procesu porovnat získané údaje s obdobnými údaji jiných nemocnic, zároveň je možné provádět na základě tohoto zařazení platbu za jeden případ léčení v akutní lůžkové péči.

Atributy této tabulky 9 jsou: ID DRG, popis DRG, slovní vyjádření DRG, průměrná doba délky pobytu dle DRG, dolní a horní hranice délky pobytu při tomto DRG.

| | ID_DRG | Popis_DRG | Slovne_DRG | Prumer_dp | Dolni_dp | Horni_dp |
|---|---|---|---|---|---|---|
| 1 | IR.D22530 | DRG 22530 | ROZSÁHLÉ POPÁLENINY BEZ KONÍHO TĚPU ... | 4 | 1 | 11 |
| 2 | IR.D22551 | DRG 22551 | POPÁLENINY OMEZENÉHO ROZSAHU POSTIHUJÍCÍ VECHN... | 8 | 3 | 23 |
| 3 | IR.D22553 | DRG 22553 | POPÁLENINY OMEZENÉHO ROZSAHU POSTIHUJÍCÍ VECHN... | 15 | 5 | 45 |
| 4 | IR.D22552 | DRG 22552 | POPÁLENINY OMEZENÉHO ROZSAHU POSTIHUJÍCÍ VECHN... | 8 | 3 | 25 |
| 5 | IR.D22522 | DRG 22522 | MÉNĚ ROZSÁHLÉ POPÁLENINY SKRZ CELOU KŮI. S KONÍM... | 24 | 8 | 71 |
| 6 | IR.D22521 | DRG 22521 | MÉNĚ ROZSÁHLÉ POPÁLENINY SKRZ CELOU KŮI. S KONÍM... | 22 | 7 | 65 |
| 7 | IR.D22523 | DRG 22523 | MÉNĚ ROZSÁHLÉ POPÁLENINY SKRZ CELOU KŮI. S KONÍM... | 40 | 13 | 119 |
| 8 | IR.D22543 | DRG 22543 | POPÁLENINY OMEZENÉHO ROZSAHU POSTIHUJÍCÍ VECHN... | 20 | 7 | 59 |
| 9 | IR.D22541 | DRG 22541 | POPÁLENINY OMEZENÉHO ROZSAHU POSTIHUJÍCÍ VECHN... | 11 | 4 | 34 |
| 10 | IR.D22542 | DRG 22542 | POPÁLENINY OMEZENÉHO ROZSAHU POSTIHUJÍCÍ VECHN... | 12 | 4 | 37 |

**Tabulka 9 - Ukázka tabulky DRG [zdroj vlastní]**

Tabulka DRG2

Atributy této tabulky 10 jsou: ID případu a ID DRG.

| | ID_pripadu | ID_DRG |
|---|---|---|
| 1 | CZ34.26445 | IR.D05422 |
| 2 | CZ34.24708 | IR.D10301 |
| 3 | CZ34.24728 | IR.D06381 |
| 4 | CZ34.24109 | IR.D01431 |
| 5 | CZ34.24131 | IR.D03332 |
| 6 | CZ34.24154 | IR.D05352 |
| 7 | CZ34.24176 | IR.D08371 |
| 8 | CZ34.24479 | IR.D14641 |
| 9 | CZ34.26947 | IR.D18322 |
| 10 | CZ34.26967 | IR.D03351 |

**Tabulka 10 - Ukázka tabulky DRG2 [zdroj vlastní]**

Tabulka Kód nemocnice

Atributy této tabulky 11 jsou: ID případu a příslušný kód nemocnice.

| | ID_pripadu | Kod_nemocnice |
|---|---|---|
| 1 | PL1907726 | PL2 |
| 2 | PL1907727 | PL2 |
| 3 | PL1907728 | PL2 |
| 4 | PL1907729 | PL2 |
| 5 | PL1907730 | PL2 |
| 6 | PL1907731 | PL2 |
| 7 | PL1907732 | PL2 |
| 8 | PL1907733 | PL2 |
| 9 | PL1907734 | PL2 |
| 10 | PL1907735 | PL2 |

**Tabulka 11 - Ukázka tabulky Kód nemocnice [zdroj vlastní]**

Tabulka Diagnózy ČR

Tabulka 12 uvádí případy v rámci České republiky, nicméně záznamů v tomto formátu je jen 27 000, u ostatních tabulek je jich přes 370 000. Proto údaje z této tabulky budou využity pouze samostatně.

Atributy této tabulky 12 jsou: ID případu, ID hlavní diagnózy a ID vedlejší diagnózy.

| | ID_pripadu | ID_hldiagnozy | ID_vedl.diagnozy |
|---|---|---|---|
| 1 | CZ1.54377 | J040 | J00 |
| 2 | CZ1.54380 | I10 | J42 |
| 3 | CZ1.54380 | I10 | I48 |
| 4 | CZ1.54380 | I10 | I500 |
| 5 | CZ1.54380 | I10 | I509 |
| 6 | CZ1.54382 | K263 | K250 |
| 7 | CZ1.54384 | J448 | E119 |
| 8 | CZ1.54384 | J448 | M546 |
| 9 | CZ1.54385 | I200 | R55 |
| 10 | CZ1.54386 | R104 | K590 |

**Tabulka 12 - Ukázka tabulky Diagnózy ČR [zdroj vlastní]**

## *5.3 PŘÍPRAVA DAT*

Kvalita dat se ukázala jako problematická. Důvodem je odlišný způsob shromažďování dat i organizační struktura vybraných zemí. Získané soubory s příponou „txt“ bylo nutné upravit a přizpůsobit dle požadavků programu na zpracování těchto dat a dle požadavků samotné analýzy.

Došlo k odstranění prázdných řádků a z důvodu množství dat, nebylo možné tuto úpravu provádět ručně. Nicméně zde vyvstaly další problémy jako mezery mezi daty a další drobnosti, které ovšem ve výsledném efektu způsobí chybné načtení dat programem. Nakonec se podařilo správně načíst všechna data ve správném formátu, což byl první předpoklad úspěšných analýz. Tato příprava byla z hlediska všech fází časově nejnáročnější.

Všechna vstupní data byla podrobena testu kvality. Úplnosti všech atributů, u číselných proměnných nás zajímalo, zda nějaké hodnoty nejsou odlehlé. Odlehlosti hodnot potvrzeny nebyly a pouze u dvou vstupních souborů byly zjištěny nedostatky. Ze vstupní analýzy bylo patrné, že některé údaje jsou evidentně chybné. Z následujících tabulek 13 a 14 je vidět, že problematický byl věk u tabulky „Případy“, jehož prázdné hodnoty byly z analýz týkajících se věku odstraněny. Dále v tabulce „Diagnózy“ bylo neplatných hodnot poměrně hodně. Nicméně celková kvalita dat byla 99ti %.

Výsledky úplnosti záznamů tabulky Případy

| Field | % Complete | Valid Records |
|---|---|---|
| ID | 100 | 370208 |
| ID_pripadu | 100 | 370208 |
| ID_pacienta | 100 | 370208 |
| ID_hldiagnozy | 100 | 370208 |
| Vek | 97,9 | 362449 |
| Prijat_kym | 100 | 370208 |
| Zpusob_ukonceni | 100 | 370208 |
| Zpusob_prijeti | 100 | 370208 |
| Delka_pobytu | 100 | 370208 |

**Tabulka 13 - Platné hodnoty u tabulky Případy [zdroj vlastní]**

Výsledky úplnosti záznamů tabulky Diagnózy

| Field | % Complete | Valid Records |
|---|---|---|
| ID_hldiagnozy | 100 | 36351 |
| Popis_hldiagnozy | 99,99 | 36348 |
| Diagnoza_l1 | 12,62 | 4586 |
| Popis_diagnozy_l1 | 99,99 | 36349 |
| Diagnoza_l2 | 99,93 | 36326 |
| Popis_diagnozy_l2 | 99,93 | 36327 |

**Tabulka 14 - Platné hodnoty u tabulky Diagnózy [zdroj vlastní]**

## Základní analýza dat

Pro statistickou analýzu byly vybrány tyto ukazatele:

**Četnost (Count) -** počet jednotek daného výběru

**Aritmetický průměr (Mean) -** statistická veličina, která v jistém smyslu vyjadřuje typickou hodnotu popisující soubor mnoha hodnot

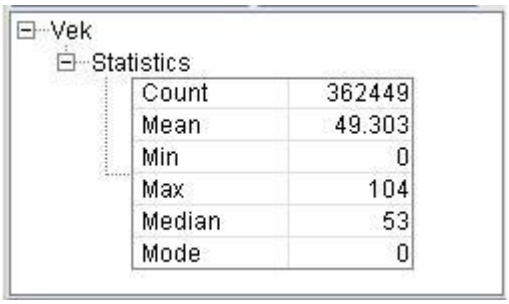
**Minimum (Min) -** minimální hodnota daného výběru

**Maximum (Max) -** maximální hodnota daného výběru

**Medián (Median) -** hodnota, jež dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny

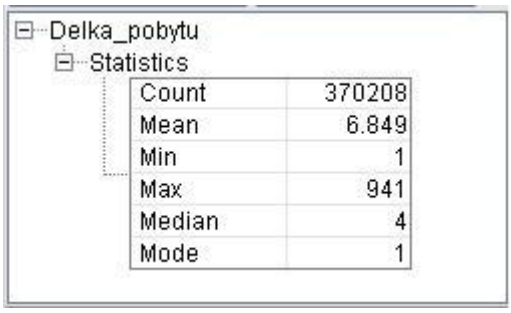
**Modus (Mode) -** hodnota statistického znaku, která má v náhodném výběru největší třídní četnost [4]

Z obrázku 10 je patrné, že nejmladší hospitalizovaný pacient byl ve věku „0“, což znamená rozpětí od narození do jednoho roku. Nejstaršímu bylo 104 let. Nejvyšší třídní četnost (Mode) zde má pacient ve věku „0“, a to ve většině případů z důvodu diagnózy „Z380“ (jediné dítě narozené v nemocnici).




**Obrázek 10 - Statistika atributu Věk [zdroj vlastní]**

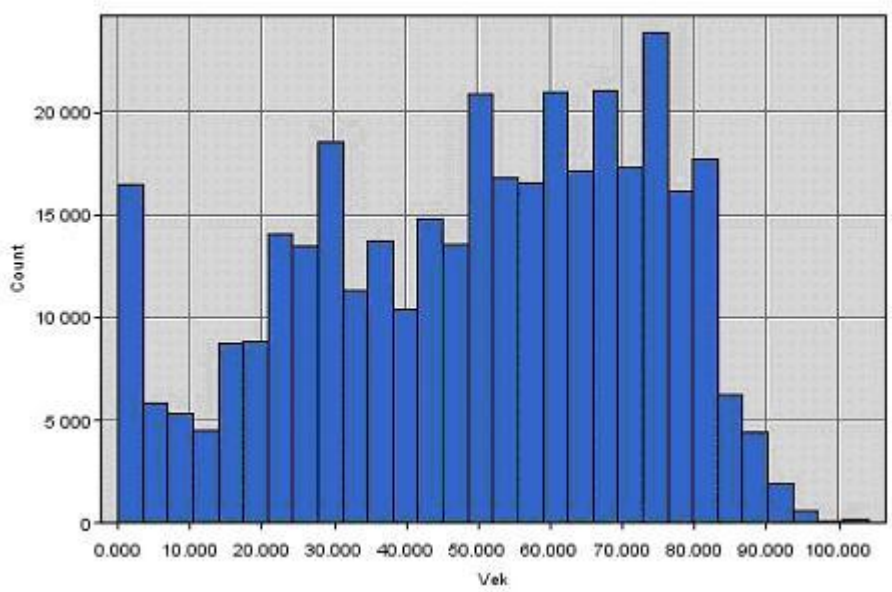
Dále obrázek 11 nám ukazuje, že maximální doba pobytu činila 941 dní, průměrná doba je ale o poznání nižší, necelých 7 dní.




**Obrázek 11 - Statistika atributu Délka pobytu [zdroj vlastní]**

Histogram hospitalizačních případů podle věku

Histogram na obrázku 12 ukazuje věkové rozložení hospitalizovaných pacientů. Vysoký počet u nejmladší věkové skupiny je důsledkem diagnózy „Z380“ – narození jediného dítěte v nemocnici.




**Obrázek 12 - Histogram hospitalizačních případů podle věku [zdroj vlastní]**

Počet hospitalizačních případů dle délky pobytu

Z následující tabulky 15 je vidět, že nejvíce hospitalizačních případů je vyřešeno hned první den. A za týden je propuštěno téměř 72% pacientů.

| | Delka_pobytu | Record_Count | % pripadu |
|---|---|---|---|
| 1 | 1 | 63095 | 17.043 |
| 2 | 2 | 48660 | 13.144 |
| 3 | 3 | 39952 | 10.792 |
| 4 | 4 | 39499 | 10.669 |
| 5 | 5 | 28902 | 7.807 |
| 6 | 6 | 22892 | 6.184 |
| 7 | 7 | 21934 | 5.925 |
| 8 | 8 | 17160 | 4.635 |
| 9 | 9 | 13361 | 3.609 |
| 10 | 10 | 10542 | 2.848 |

**Tabulka 15 - Počet hospitalizačních případů podle délky pobytu [zdroj vlastní]**

Poměr hospitalizačních případů podle pohlaví

Ze všech hospitalizačních případů nepatrně převládají ženy (F – Female) oproti mužům (M – Male), jak je vidět z obrázku 13. S ohledem na skladbu obyvatelstva, není tento rozdíl nijak dramatický.

| Value | Proportion | % | Count |
|---|---|---|---|
| F | | 54,71 | 202532 |
| M | | 45,29 | 167676 |

**Obrázek 13 - Poměr hospitalizačních případů podle pohlaví [zdroj vlastní]**

Poměr hospitalizačních případů podle statusu způsobu přijetí

Ačkoli je u statusu uvedena i hodnota „NOA“, v datech se vůbec nevyskytuje. Obrázek 14 zobrazuje pouze hodnoty, které se v datech skutečně vyskytovaly. Převažuje zde běžné přijetí „OTH“, nejméně je neodkladných přijetí „EME“. Neznámý způsob přijetí „NOK“ tvoří poměrně velkou část. Je to způsobeno tím, že pacient je zde již hospitalizován a data jsou získána v průběhu hospitalizace nebo zde údaj nebyl uveden vůbec.

| Value | Proportion | % | Count |
|---|---|---|---|
| EME | | 17,65 | 64819 |
| NOK | | 37,29 | 136932 |
| OTH | | 45,06 | 165439 |

**Obrázek 14 - Poměr hospitalizačních případů podle statusu přijetí [zdroj vlastní]**

Poměr hospitalizačních případů podle statusu způsobu ukončení

Obrázek 15 zobrazuje poměry dle statusu způsobu ukončení. Jasně zde dominuje propuštění pacienta před zbylými dvěma způsoby ukončení, což naznačuje kvalitní lékařskou péči.

| Value | Proportion | % | Count |
|---|---|---|---|
| DEA | | 2,08 | 7609 |
| DIS | | 94,87 | 346656 |
| TRA | | 3,05 | 11135 |

**Obrázek 15 - Poměr hospitalizačních případů podle statusu způsobu ukončení [zdroj vlastní]**

Celkový počet podle hlavních diagnóz

Z tabulky 16 je patrné, ze mezi nejčastější diagnózu patří „O800“, což je spontánní porod záhlavím. Mezi druhou nejčastější diagnózu patří „Z380“, což je jediné dítě narozené v nemocnici. Diagnóza „H258“ představuje senilní kataraktu neboli šedý zákal.

| | ID_hldiagnozy | Record_Count |
|---|---|---|
| 1 | O800 | 8650 |
| 2 | Z380 | 7880 |
| 3 | H258 | 4550 |
| 4 | I251 | 4090 |
| 5 | Z76 | 3397 |
| 6 | I10 | 3262 |
| 7 | I259 | 2934 |
| 8 | O80 | 2553 |
| 9 | I48 | 2535 |
| 10 | S060 | 2334 |

**Tabulka 16 - Celkový počet dle hlavních diagnóz [zdroj vlastní]**

## *5.4 MODELOVÁNÍ*

### Analýza hospitalizačních případů v rámci ČR

Celkový počet podle hlavních diagnóz

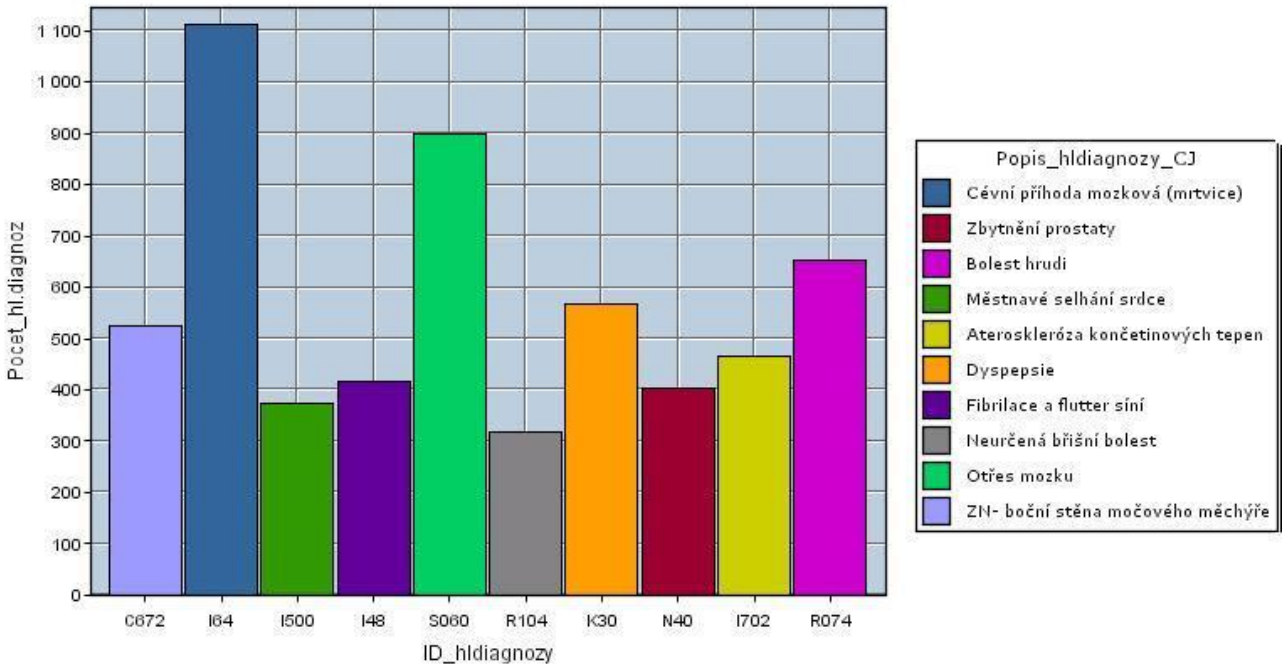
Celkový počet hlavních diagnóz, které se v případech objevují je 1195. Počet vedlejších diagnóz je roven 2898 záznamům. Zde je přehled nejčastějších hlavních diagnóz v rámci České republiky. Využita byla tabulka Diagnózy ČR. Jasně zde dominuje cévní mozková příhoda, což je vidět z tabulky 17. Zaujímá 4 % ze všech ostatních diagnóz.

| | ID_hldiagnozy | Popis_hldiagnozy_CJ | Pocet_hl.diagnoz |
|---|---|---|---|
| 1 | I64 | Cévní příhoda mozková (mrtvice) | 1114 |
| 2 | S060 | Otřes mozku | 898 |
| 3 | R074 | Bolest hrudi | 653 |
| 4 | K30 | Dyspepsie | 568 |
| 5 | C672 | ZN- boční stěna močového měchýře | 525 |
| 6 | I702 | Ateroskleróza končetinových tepen | 464 |
| 7 | I48 | Fibrilace a flutter síní | 418 |
| 8 | N40 | Zbytnění prostaty | 402 |
| 9 | I500 | Městnavé selhání srdce | 374 |
| 10 | R104 | Neurčená břišní bolest | 317 |

**Tabulka 17 - Celkový počet dle hlavních diagnóz ČR [zdroj vlastní]**

Graf celkového počtu dle nejčastějších 10ti hlavních diagnóz

Na obrázku 16 je vidět graf 10ti nejčastějších hlavních diagnóz v rámci ČR.




**Obrázek 16 - Graf dle celkového počtu dle 10ti nejčastějších diagnóz [zdroj vlastní]**

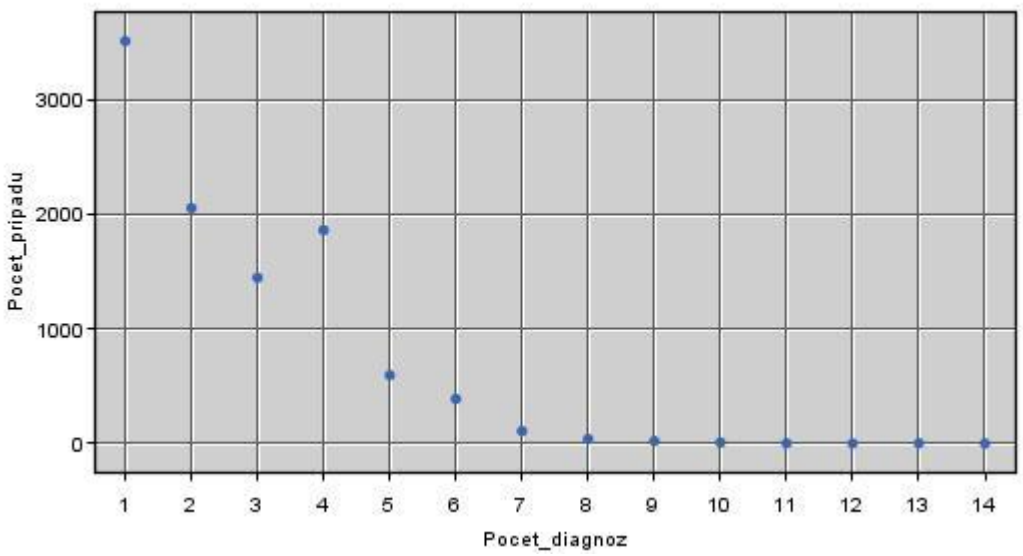
Počet diagnóz v rámci jednoho hospitalizačního případu

Tabulka 18 uvádí, že nejčastěji je pacient léčen na základě jedné diagnózy. Maximální počet diagnóz je roven 14.

| | Pocet_diagnoz | Pocet_pripadu |
|---|---|---|
| 1 | 1 | 3517 |
| 2 | 2 | 2058 |
| 3 | 3 | 1450 |
| 4 | 4 | 1865 |
| 5 | 5 | 597 |
| 6 | 6 | 391 |
| 7 | 7 | 109 |
| 8 | 8 | 40 |
| 9 | 9 | 22 |
| 10 | 10 | 9 |
| 11 | 11 | 2 |
| 12 | 12 | 1 |
| 13 | 13 | 2 |
| 14 | 14 | 1 |

**Tabulka 18 - Počet diagnóz v rámci jednoho hospitalizačního případu [zdroj vlastní]**

Z obrázku 17 je patrné, že s přibývajícím počtem vedlejších diagnóz má počet hospitalizačních případů klesající tendenci. Výjimkou je počet diagnóz s hodnotou „4“.




**Obrázek 17 - Graf počtu diagnóz v rámci jednoho hospitalizačního případu [zdroj vlastní]**

Výskyt vedlejších diagnóz v hlavních diagnózách

Jelikož nás zajímal výskyt vedlejších diagnóz v rámci hlavních diagnóz, bylo vytvořeno makro[2], jehož výsledkem byla matice o hodnotách „1“ a „0“. Kde hodnota „1“ vyjadřuje fakt, že vedlejší diagnóza se vyskytla v rámci hlavní diagnózy alespoň v jednom z hospitalizačních případů. U hodnoty „0“ se nevyskytla v žádném z případů. Jelikož výsledná matice přesahovala možnosti zobrazení v MS Excel, byla vyexportována do formátu „csv“ a následně použita pro analýzy v Clementine. Řádky matice jsou tvořeny hlavními diagnózami a ve sloupcích jsou uvedeny diagnózy vedlejší. Ukázka vytvořené matice je v tabulce 19.

| | field1 | J00 | J42 | I48 | I500 | I509 | K250 | E119 | M546 | R55 | K590 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | J040 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | I10 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| 3 | K263 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 4 | J448 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| 5 | I200 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| 6 | R104 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 |
| 7 | K30 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 8 | E119 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 9 | R509 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 10 | N394 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

**Tabulka 19 - Výsledná matice výskytu vedlejších diagnóz v hlavních [zdroj vlastní]**

[2] Skript vytvořeného makra je uveden v příloze

Vyhodnocením údajů bylo zjištěno, že v 509ti hlavních diagnózách se vyskytovala vedlejší diagnóza „I10“ (hypertenze), což ukazuje tabulka 20. Následuje tabulka 21, kde na druhém místě s 356ti se vyskytuje diagnóza „X599“ (vystavení neurčitým faktorům), která nemá příliš velkou vypovídací hodnotu. Na třetím místě diagnóza „I259“ (ischemická nemoc srdeční) a na čtvrtém diagnóza „E119“ (diabetes). Zobrazeny jsou v tabulkách 22 a 23. Tyto vedlejší diagnózy lze chápat jako rizikové, jelikož se ve velkém počtu vyskytují u hlavních diagnóz.

| | I10 | Record_Count |
|---|---|---|
| 1 | 1 | 509 |
| 2 | 0 | 686 |

**Tabulka 20 - Nejčastější vedlejší diagnóza [zdroj vlastní]**

| | X599 | Record_Count |
|---|---|---|
| 1 | 1 | 356 |
| 2 | 0 | 839 |

**Tabulka 21 - 2. nejčastější vedlejší diagnóza [zdroj vlastní]**

| | I259 | Record_Count |
|---|---|---|
| 1 | 1 | 261 |
| 2 | 0 | 934 |

**Tabulka 22 - 3. nejčastější vedlejší diagnóza [zdroj vlastní]**

| | E119 | Record_Count |
|---|---|---|
| 1 | 1 | 255 |
| 2 | 0 | 940 |

**Tabulka 23 - 4. nejčastější vedlejší diagnóza [zdroj vlastní]**

## Predikce úmrtnosti

Pro predikci úmrtnosti byl použit algoritmus rozhodovacích stromů C5.0. Ostatními algoritmy (C&R Tree, Quest, Chaid) nebylo dosaženo tak dobrých výsledků jako v případě C5.0. Vstupními atributy byly: ID diagnózy, věk, pohlaví, způsob přijetí, způsob propuštění, kým byl pacient přijat a délka pobytu. Výstupem status ukončení „DEA“.

Bylo nutné vybalancování vstupních dat. Jako trénovací data byla použita všechna úmrtí „DEA“, 75% přeložení „TRA“ a 2% propuštění „DIS“. Poměr trénovacích a testovacích dat byl rovnoměrný. Následně byl vygenerovaný model testován na originálních datech.

Z tabulky 24 je patrné, že úspěšnost vytvořeného modelu je téměř 63%. Řádky matice představují skutečný počet případů, ve sloupcích jsou predikované hodnoty. Na hlavní diagonále jsou vidět správně predikované hodnoty.

Results for output field Status_ukonceni

Comparing $C-Status_ukonceni with Status_ukonceni

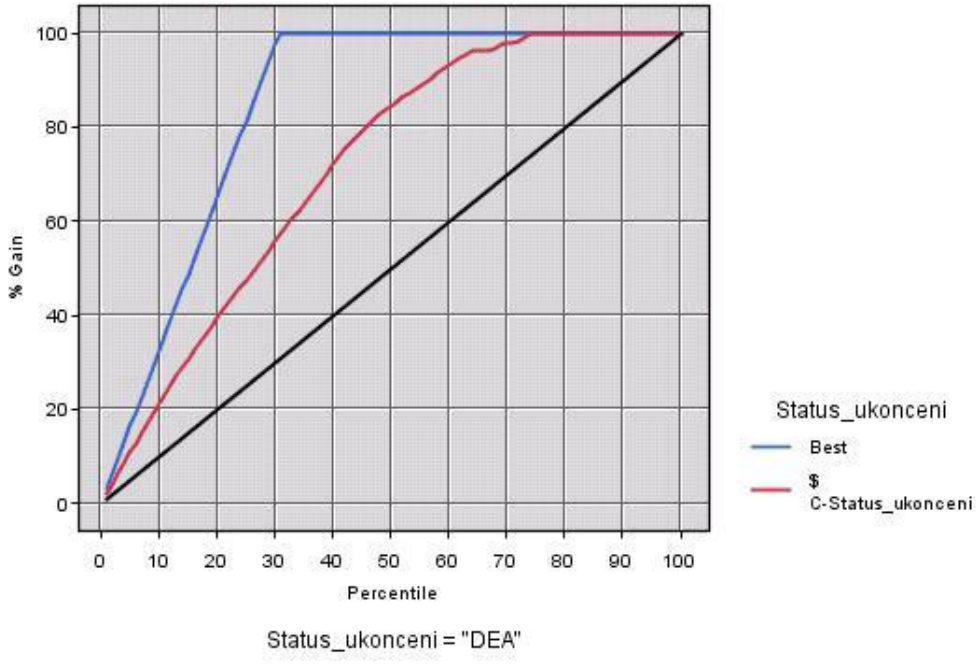
| | | |
|---|---|---|
| Correct | 55 318 | 62,87% |
| Wrong | 32 673 | 37,13% |
| Total | 87 991 | |

Coincidence Matrix for $C-Status_ukonceni (rows show actuals)

| | DEA | DIS | TRA |
|---|---|---|---|
| DEA | 2 834 | 3 355 | 1 420 |
| DIS | 10 982 | 49 647 | 8 618 |
| TRA | 4 280 | 4 018 | 2 837 |

**Tabulka 24 - Analýza predikce úmrtí [zdroj vlastní]**

Na obrázku 18 jsou vidět křivky: náhodného případu (černá barva), nejlepšího možného modelu (modrá barva) a našeho predikovaného případu (červená barva). Je vidět, že v případě 40% trénovacích dat získáme 70% informačního zisku modelu.




**Obrázek 18 - Graf predikce úmrtí [zdroj vlastní]**

## Predikce délky pobytu

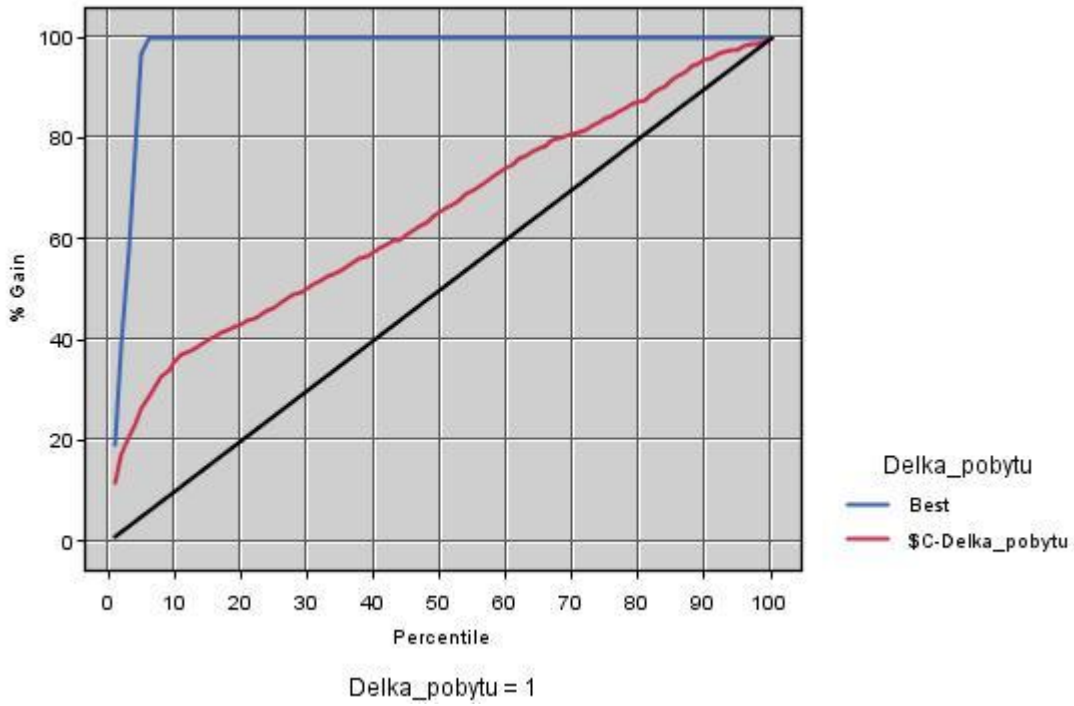
Pro predikci délky pobytu byl opět využit algoritmus rozhodovacích stromů C5.0. Vstupní tabulka byla stejná jako v předchozí predikci úmrtí. Pouze byla omezena délka pobytu na maximální hodnotu 19ti dní a hodnoty byly vybalancovány. Nicméně nebylo dosaženo tak dobrých výsledků jako v předchozím případě.

Výsledný model byl porovnán s absolutním rozdílem skutečných hodnot. Pouze v 15ti % se délka pobytu shodovala se skutečnou. Pokud bychom tolerovali přesnost modelu na 4 dny, dosáhli bychom téměř 57 %, což ukazuje následující tabulka 25.

| | Absolutni_rozdil | Pocet_pripadu | % pripadu |
|---|---|---|---|
| 1 | 0 | 5652 | 15.769 |
| 2 | 1 | 4663 | 13.010 |
| 3 | 2 | 3806 | 10.619 |
| 4 | 3 | 3221 | 8.987 |
| 5 | 4 | 2922 | 8.152 |
| 6 | 5 | 2507 | 6.995 |
| 7 | 6 | 2280 | 6.361 |
| 8 | 7 | 1938 | 5.407 |
| 9 | 8 | 1732 | 4.832 |
| 10 | 9 | 1529 | 4.266 |
| 11 | 10 | 1245 | 3.474 |
| 12 | 11 | 1035 | 2.888 |
| 13 | 12 | 891 | 2.486 |
| 14 | 13 | 717 | 2.000 |
| 15 | 14 | 603 | 1.682 |
| 16 | 15 | 531 | 1.482 |
| 17 | 16 | 404 | 1.127 |
| 18 | 17 | 275 | 0.767 |
| 19 | 18 | 142 | 0.396 |

**Tabulka 25 - Absolutní rozdíl predikované a skutečné délky pobytu [zdroj vlastní]**

Na obrázku 19 je patrný prudký nárůst informačního zisku modelu (červená křivka), který se ovšem na 15ti % trénovacích dat zmírnil a dále stoupá jen pozvolna.




**Obrázek 19 - Graf predikce délky pobytu [zdroj vlastní]**

## Analýza srovnání délky pobytu dle DRG

Průměrná délka pobytu

Agregací dle ID DRG byla zjištěna průměrná doba, po kterou je pacient hospitalizován a porovnávána s tabulkovou hodnotou DRG. Byly zjištěny pomocné hodnoty „0“ a „1“. „0“ v tomto případě představuje průměrnou délku pobytu zjištěnou dle hospitalizačních případů podle DRG, která je rovna nebo kratší než průměrná délka daná dle tabulek DRG. Hodnota „1“ představuje délku pobytu vyšší než doporučený průměrný limit dle DRG. Z obrázku 20 je tedy vidět, že v nadpoloviční většině průměrná doba opravdu tabulkovým hodnotám odpovídá.

| Value | Proportion | % | Count |
|---|---|---|---|
| 0 | | 56,04 | 399 |
| 1 | | 43,96 | 313 |

**Obrázek 20 - Poměr skutečných délek pobytu s hodnotami DRG [zdroj vlastní]**

Celková délka pobytu

Opět porovnání skutečné délky pobytu tentokrát s maximální stanovenou dobou dle DRG. Hodnota „0“ znamená délku pobytu, která splňuje podmínku maximálně stanovené doby dle DRG. Hodnota „1“ nikoli. Z obrázku 21 vyplývá, že v téměř 93% je tato doba dodržena.

| Value | Proportion | % | Count |
|---|---|---|---|
| 0 | | 92,31 | 81975 |
| 1 | | 7,69 | 6827 |

**Obrázek 21 - Poměr délky pobytu s horní hranicí DRG [zdroj vlastní]**

Nejčastější DRG s delší než doporučenou délkou pobytu

Tabulka 26 uvádí nejčastějších 10 diagnóz dle počtu případů, jejichž délka pobytu nekorespondovala s maximální délkou určenou dle DRG. Z velké části se jedná o různé druhy leukémie či rakoviny. Dá se předpokládat, že délka průběhu těchto nemocí se dá jen obtížně předvídat.

| | ID_DRG | Slovne_DRG | Delsi | % pripadu stejne diagnozy |
|---|---|---|---|---|
| 1 | IR.D17312 | LYMFOM A NEAKUTNÍ LEUKÉMIE s CC | 470 | 100.000 |
| 2 | IR.D11301 | ZHOUBNÉ BUJENÍ LEDVIN A MOČOVÝCH CEST A LEDVINOVÉ S... | 401 | 100.000 |
| 3 | IR.D11302 | ZHOUBNÉ BUJENÍ LEDVIN A MOČOVÝCH CEST A LEDVINOVÉ S... | 358 | 100.000 |
| 4 | IR.D17311 | LYMFOM A NEAKUTNÍ LEUKÉMIE bez CC ... | 250 | 100.000 |
| 5 | IR.D15720 | NOVOROZENEC. VÁHA PŘI PORODU > 2499G. SE SYNDROME... | 155 | 100.000 |
| 6 | IR.D17313 | LYMFOM A NEAKUTNÍ LEUKÉMIE s MCC ... | 137 | 100.000 |
| 7 | IR.D00043 | DLOUHODOBÁ MECHANICKÁ VENTILACE S TRACHEOSTOMIÍ s ... | 113 | 100.000 |
| 8 | IR.D11303 | ZHOUBNÉ BUJENÍ LEDVIN A MOČOVÝCH CEST A LEDVINOVÉ S... | 103 | 100.000 |
| 9 | IR.D18303 | SEPTICÉMIE s MCC | 98 | 100.000 |
| 10 | IR.D00042 | DLOUHODOBÁ MECHANICKÁ VENTILACE S TRACHEOSTOMIÍ s ... | 82 | 100.000 |

**Tabulka 26 - Nejčastější diagnózy s delší délkou pobytu než DRG [zdroj vlastní]**

Podrobnější analýzou se ukázalo, že existuje hned 61 diagnóz, které ve všech případech překračují doporučenou maximální délku pobytu dle DRG, což ukazuje tabulka 27. Je řazena sestupně dle % případů stejné diagnózy.

| | % pripadu stejne diagnozy | Pocet pripadu |
|---|---|---|
| 1 | 100.000 | 61 |
| 2 | 92.593 | 1 |
| 3 | 77.778 | 1 |
| 4 | 66.667 | 2 |
| 5 | 50.000 | 6 |
| 6 | 49.565 | 1 |
| 7 | 45.833 | 1 |
| 8 | 37.500 | 1 |
| 9 | 36.364 | 1 |
| 10 | 35.294 | 1 |

**Tabulka 27 - Počet případů dle % případů stejné diagnózy [zdroj vlastní]**

## Analýza počtu hospitalizačních případů podle věkových skupin

Tato analýza měla za úkol rozdělit pacienty do 11 základních skupin dle jejich věku. První kategorie „1“ zahrnuje pacienty s věkem „0“. Tato skupina byla vytvořena záměrně, a to z toho důvodu, že se jedná většinou o různé druhy porodů. Druhá kategorie „2“ zahrnuje pacienty od 1 do 10 let. Kategorie „3“ zahrnuje pacienty od 11ti do 20ti let atd. Do poslední věkové kategorie „11“ patří pacienti starší 90ti let. Rozdělení do věkových kategorií zobrazuje tabulka 28.

| Věková kategorie | Věkové rozpětí |
|---|---|
| 1 | 0 |
| 2 | 1-10 |
| 3 | 11-20 |
| 4 | 21-30 |
| 5 | 31-40 |
| 6 | 41-50 |
| 7 | 51-60 |
| 8 | 61-70 |
| 9 | 71-80 |
| 10 | 81-90 |
| 11 | 91 a více |

**Tabulka 28 - Věkové kategorie [zdroj vlastní]**

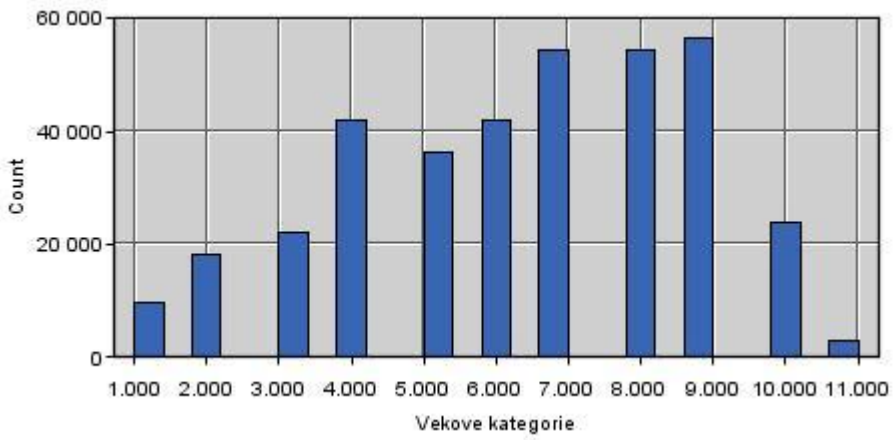
Zastoupení dle věkových skupin:

Tabulka 29 uvádí počet hospitalizačních případů dle věkových kategorií. Nejčastější věkovou skupinou, která je hospitalizována, jsou pacienti ve věku mezi 51. až 80. rokem.

| | Vekove kategorie | Record_Count |
|---|---|---|
| 1 | 1 | 9642 |
| 2 | 2 | 18043 |
| 3 | 3 | 22126 |
| 4 | 4 | 41880 |
| 5 | 5 | 36338 |
| 6 | 6 | 42047 |
| 7 | 7 | 54547 |
| 8 | 8 | 54485 |
| 9 | 9 | 56677 |
| 10 | 10 | 23722 |
| 11 | 11 | 2942 |

**Tabulka 29 - Počet hospitalizačních případů dle věkových kategorií [zdroj vlastní]**

Následující histogram na obrázku 22 uvádí rozložení hospitalizačních případů dle věkových kategorií.




**Obrázek 22 - Histogram dle věkových kategorií [zdroj vlastní]**

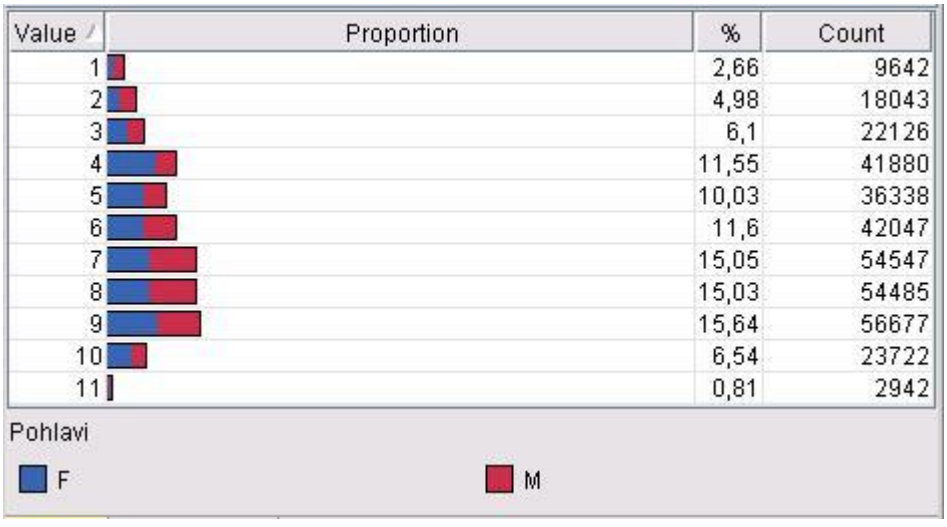
Zastoupení dle pohlaví:

Přehled zastoupení dle věkových kategorií a pohlaví ukazuje tabulka 30. Zajímavé například je, že ve věku 21 – 30 let (věková kategorie 4) jsou téměř 3x více hospitalizované ženy. Podrobnějších zkoumáním bylo zjištěno, že tato situace nastává z důvodu hospitalizace ve spojitosti s porodem. V tomto věku jsou ženy z tohoto důvodu hospitalizovány téměř ze 70ti %.

| | Pohlavi | Vekove kategorie | Pocet pripadu |
|---|---|---|---|
| 1 | F | 11 | 2086 |
| 2 | M | 11 | 856 |
| 3 | M | 10 | 8310 |
| 4 | F | 10 | 15412 |
| 5 | M | 9 | 25695 |
| 6 | F | 9 | 30982 |
| 7 | M | 8 | 28729 |
| 8 | F | 8 | 25756 |
| 9 | M | 7 | 28874 |
| 10 | F | 7 | 25673 |
| 11 | F | 6 | 21967 |
| 12 | M | 6 | 20080 |
| 13 | M | 5 | 13554 |
| 14 | F | 5 | 22784 |
| 15 | M | 4 | 12238 |
| 16 | F | 4 | 29642 |
| 17 | F | 3 | 12092 |
| 18 | M | 3 | 10034 |
| 19 | F | 2 | 7706 |
| 20 | M | 2 | 10337 |
| 21 | F | 1 | 3616 |
| 22 | M | 1 | 6026 |

**Tabulka 30 - Počet hospitalizačních případů dle věkových kategorií a pohlaví [zdroj vlastní]**

Hospitalizovaní pacienti jsou ve věkových skupinách zastoupeny dle pohlaví rovnoměrně, až na zmiňovanou věkovou kategorii 4. Dobře je z obrázku 23 vidět proporcionální zastoupení dle věkových kategorií, kde jasně dominují věkové kategorie 7, 8, 9 s 15ti %.




**Obrázek 23 - Proporcionální zastoupení dle věkové skupiny a pohlaví [zdroj vlastní]**

Nejčastější hlavní diagnóza dle věkové kategorie

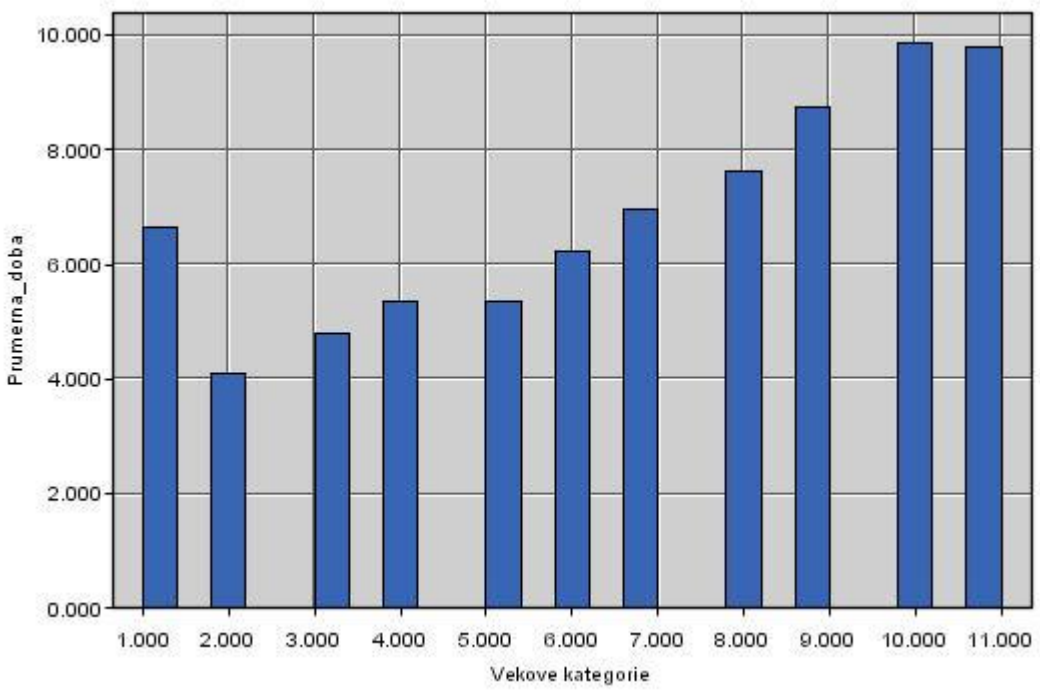
Rozdělením do těchto skupin se ukázalo, že mezi nejčastější diagnózy do 40ti let patří diagnózy spojené s porodem dítěte, ve věku od 51 do 70ti let jsou to srdeční nemoci a v pokročilém věku je to senilní katarakta.neboli šedý zákal. Poslední sloupec tabulky 31 ukazuje, jaké procento ze všech případů v dané věkové kategorii zaujímá nejčastější diagnóza. Například z toho vyplývá, že pacienti (ženy) ve věku od 21 do 30 let v téměř 13ti % ze všech hospitalizačních případů, jsou hospitalizování z důvodu spontánního porodu.

| | Vekove kategorie | ID_hldiagnozy | Popis_hldiagnozy_CJ | % ze vsech pripadu |
|---|---|---|---|---|
| 1 | 1 | Z380 | Jediné dítě narozené v nemocnici | 24.290 |
| 2 | 2 | J352 | Hypertrofie adenoidní tkáně | 10.403 |
| 3 | 3 | O800 | Spontánní porod záhlavím | 2.662 |
| 4 | 4 | O800 | Spontánní porod záhlavím | 12.853 |
| 5 | 5 | O800 | Spontánní porod záhlavím | 7.130 |
| 6 | 6 | M511 | Onemocnění lumbálních plotének s radikulopatií | 1.427 |
| 7 | 7 | I251 | Aterosklerotická nemoc (choroba) srdeční | 1.523 |
| 8 | 8 | I251 | Aterosklerotická nemoc (choroba) srdeční | 2.483 |
| 9 | 9 | H258 | Jiná senilní katarakta | 3.890 |
| 10 | 10 | H258 | Jiná senilní katarakta | 4.376 |
| 11 | 11 | I709 | Generalizovaná a neurčená ateroskleróza | 2.481 |

**Tabulka 31 - Nejčastější hlavní diagnóza dle věkové kategorie [zdroj vlastní]**

Průměrná délka pobytu dle věkové kategorie

Z obrázku 24 je patrné, že s rostoucím věkem dochází ke zvyšování délky pobytu v hospitalizačním zařízení. Nejvyšší průměrnou délku pobytu zde mají lidé ve věku od 81 let. Je zde vidět závislost mezi délkou pobytu a věkem pacienta. Výjimku tvoří kategorie „1“ vlivem diagnóz spojených s narozením dítěte.




**Obrázek 24 - Průměrná délka pobytu dle věkových kategorií [zdroj vlastní]**

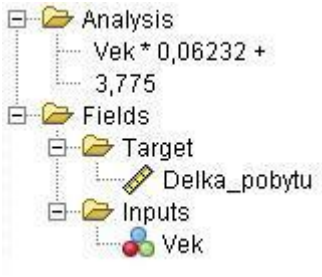
## Regresní analýza délky pobytu a věku

Regresní analýza měla prokázat, zda skutečně existuje závislost mezi délkou pobytu a věkem hospitalizovaného pacienta. Vstupem do lineárního regresního modelu byl věk a délka pobytu pacienta omezená na 10 dní. Data byla stejnoměrně rozdělena na trénovací a testovací. Výsledkem je absolutní rozdíl skutečné délky pobytu a délky pobytu vytvořené regresním modelem, který je zobrazen v tabulce 32. I když porovnáním těchto dvou čísel dochází ke shodě u necelých 10ti %, pokud bychom tolerovali přesnost na 4 dny, dostali bychom se na 80%, což závislost těchto dvou hodnot potvrzuje.

| | Absolutni_rozdil | Pocet_pripadu | % pripadu |
|---|---|---|---|
| 1 | 0 | 31765 | 9.543 |
| 2 | 1 | 65020 | 19.534 |
| 3 | 2 | 61923 | 18.603 |
| 4 | 3 | 60824 | 18.273 |
| 5 | 4 | 47663 | 14.319 |
| 6 | 5 | 23586 | 7.086 |
| 7 | 6 | 8897 | 2.673 |
| 8 | 7 | 7167 | 2.153 |
| 9 | 8 | 6474 | 1.945 |
| 10 | 9 | 5226 | 1.570 |

**Tabulka 32 - Absolutní rozdíl predikované a skutečné délky pobytu [zdroj vlastní]**

Následující obrázek 25 ukazuje výpočet výstupní (predikované) hodnoty délky pobytu na základně vstupního proměnné – věku.




**Obrázek 25 - Tabulka shrnutí regresní analýzy [zdroj vlastní]**

## Analýza závislosti způsobu přijetí a způsobu ukončení

Cílem této analýzy bylo zjistit, zda spolu nesouvisí způsob přijetí (status přijetí) a způsob ukončení (status ukončení). U statusu přijetí byly vybrány pouze hodnoty „OTH“ a „EME“. To nám umožní lepší porovnání mezi běžným a akutním přijetím. U statusu ukončení se analýza zaměřila na základní 3: „TRA“, „DEA“ a „DIS“. Jak už bylo uvedeno výše. Výsledky analýzy jsou v tabulce 33 a 34.

Status přijetí „OTH“

| | Status_prijeti | Status_ukonceni | Pocet_pripadu | % pripadu |
|---|---|---|---|---|
| 1 | OTH | TRA | 6083 | 3.770 |
| 2 | OTH | DEA | 2991 | 1.854 |
| 3 | OTH | DIS | 152287 | 94.377 |

**Tabulka 33 - Status přijetí "OTH" [zdroj vlastní]**

Status přijetí „EME“

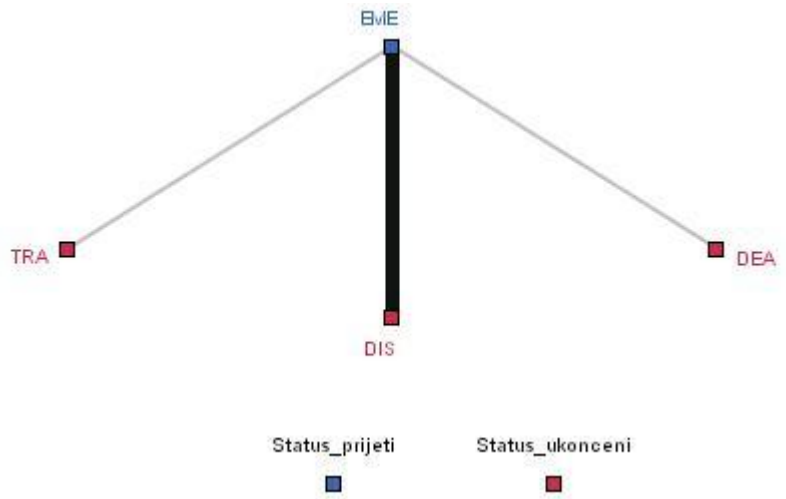
| | Status_prijeti | Status_ukonceni | Pocet_pripadu | % pripadu |
|---|---|---|---|---|
| 1 | EME | TRA | 2427 | 3.780 |
| 2 | EME | DEA | 2541 | 3.958 |
| 3 | EME | DIS | 59231 | 92.262 |

**Tabulka 34 - Status přijetí "EME" [zdroj vlastní]**

Porovnáním obou tabulek je zde nepatrný rozdíl v % podílu případů u statusu „EME“. Jelikož o 2% více pacientů po tomto způsobu přijetí zemřelo. Nicméně jelikož se nejedná o nijak významný

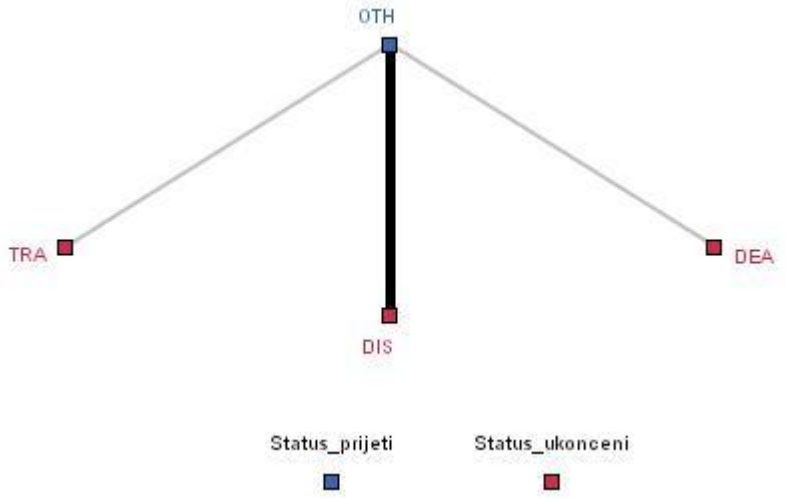
rozdíl, závislost potvrzena nebyla. Naznačují to i následující dva pavučinové grafy na obrázcích 26 a 27, kde je tloušťka čáry, která vyjadřuje míru závislosti, téměř stejná.

Pavučinový graf „EME“



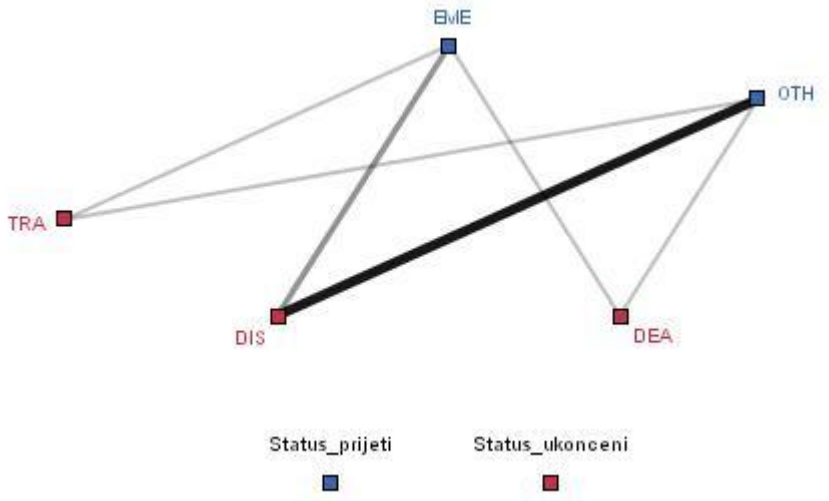

**Obrázek 26 - Pavučinový graf statusu přijetí „EME“ [zdroj vlastní]**

Pavučinový graf „OTH“




**Obrázek 27 - Pavučinový graf statusu přijetí „OTH“ [zdroj vlastní]**

Následující pavučinový graf na obrázku 28 se zaměřil na všechny hospitalizační případy. Jasně ukazuje, že nejvíce případů bylo přijato běžným způsobem („OTH“) a z lékařské péče propuštěno („DIS“). Patrná je i velká vazba mezi akutním přijetí („EME“) a propuštěním („DIS“).




**Obrázek 28 - Pavučinový graf statusu přijetí a ukončení [zdroj vlastní]**

## Analýza způsobu ukončení v závislosti na diagnóze

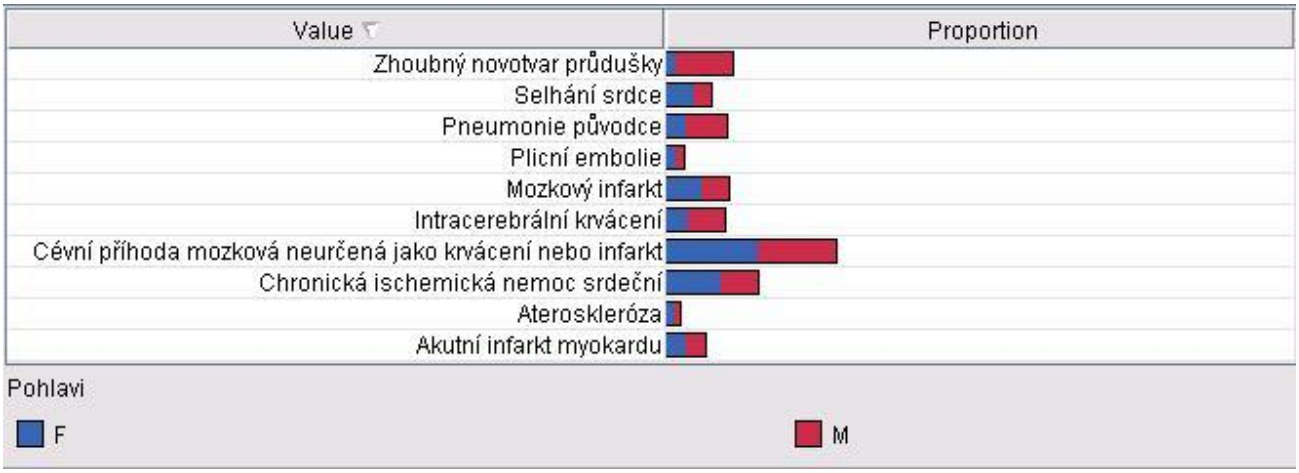
10 nejčastějších diagnóz končící úmrtím ve všech státech

Analýza měla za úkol zjistit nejčastějších 10 diagnóz, které končí úmrtím pacienta. Z následující tabulky 35 je vidět, že ve 389ti případech skončila hospitalizace smrtí z důvodu diagnózy zhoubného novotvaru na průduškách. Druhou nejčastější příčinou úmrtí bylo selhání srdce a třetí pneumonie.

| | ID_hldiagnozy | Popis_hldiagnozy | Vek_prumer | Pocet_pripadu |
|---|---|---|---|---|
| 1 | C34 | Zhoubný novotvar průdušky | 66.165 | 389 |
| 2 | I21 | Akutní infarkt myokardu | 74.957 | 254 |
| 3 | I25 | Chronická ischemická nemoc srdeční | 77.434 | 320 |
| 4 | I26 | Plicní embolie | 72.994 | 159 |
| 5 | I50 | Selhání srdce | 76.108 | 369 |
| 6 | I61 | Intracerebrální krvácení | 70.157 | 217 |
| 7 | I63 | Mozkový infarkt | 75.782 | 275 |
| 8 | I64 | Cévní příhoda mozková neurčená jako krvácení nebo infarkt | 77.071 | 196 |
| 9 | I70 | Ateroskleróza | 80.483 | 172 |
| 10 | J18 | Pneumonie původce | 77.795 | 365 |

**Tabulka 35 - Nejčastější diagnózy končící úmrtím [zdroj vlastní]**

Proporcionální zastoupení můžu a žen u 10ti nejčastějších diagnóz končících smrtí ukazuje obrázek 29. Velký nepoměr je například u diagnózy zhoubného novotvaru na průduškách, kde muži umírají 8x častěji než ženy.




**Obrázek 29 - Zastoupení pohlaví dle nejčastější diagnózy končící úmrtím [zdroj vlastní]**

10 nejčastějších diagnóz končící úmrtím v jednotlivých státech

Následující tabulky 36, 37, 38 a 39 ukazují 10 nejčastějších diagnóz, které končí smrtí pacienta v jednotlivých státech. Porovnáním států bylo zjištěno, že 5 diagnóz se objevuje ve všech vybraných státech. Jedná se o zhoubný nádor na průduškách, selhání srdce, intracerebrální krvácení, pneumonii původce a akutní infarkt myokardu. Všechny tyto diagnózy mají i velmi podobný věkový průměr pacientů. Zřejmě je to způsobeno tím, že všechny státy se nacházejí ve střední Evropě a mají lékařskou péči na podobné úrovni.

| | ID_hldiagnozy | Popis_hldiagnozy | Vek_prumer | Pocet_pripadu_Polsko |
|---|---|---|---|---|
| 1 | I63 | Mozkový infarkt | 75.820 | 89 |
| 2 | I70 | Ateroskleróza | 79.439 | 57 |
| 3 | C34 | Zhoubný novotvar průdušky | 65.509 | 53 |
| 4 | I50 | Selhání srdce | 75.933 | 45 |
| 5 | I61 | Intracerebrální krvácení | 65.487 | 39 |
| 6 | S06 | Nitrolební poranění | 61.974 | 38 |
| 7 | I46 | Srdeční zástava | 64.727 | 33 |
| 8 | J18 | Pneumonie původce | 73.484 | 31 |
| 9 | I21 | Akutní infarkt myokardu | 74.192 | 26 |
| 10 | J44 | Jiná chronická obstruktivní plicní nemoc | 72.875 | 24 |

**Tabulka 36 - 10 nejčastějších diagnóz končících úmrtím – Polsko [zdroj vlastní]**

Rakousko se nepatrně odlišuje od analyzovaných států svým vyšším věkovým průměrem u jednotlivých diagnóz, jak je vidět z tabulky 37.

| | ID_hldiagnozy | Popis_hldiagnozy | Vek_prumer | Pocet_pripadu_Rakousko |
|---|---|---|---|---|
| 1 | I50 | Selhání srdce | 77.750 | 108 |
| 2 | J18 | Pneumonie původce | 78.604 | 91 |
| 3 | I21 | Akutní infarkt myokardu | 77.881 | 67 |
| 4 | C34 | Zhoubný novotvar průdušky | 64.758 | 66 |
| 5 | A41 | Jiná septikémie | 72.424 | 59 |
| 6 | I26 | Plicní embolie | 75.923 | 52 |
| 7 | I25 | Chronická ischemická nemoc srdeční | 77.867 | 45 |
| 8 | I63 | Mozkový infarkt | 77.088 | 34 |
| 9 | I61 | Intracerebrální krvácení | 73.333 | 33 |
| 10 | I64 | Cévní příhoda mozková neurčená jako krvácení nebo infarkt | 78.633 | 30 |

**Tabulka 37 - 10 nejčastějších diagnóz končících úmrtím – Rakousko [zdroj vlastní]**

Zajímavostí následující tabulky 38 je diagnóza fibróza a cirhóza jater, která se u ostatních států vyskytuje až na mnohem nižších pozicích. Například v ČR až na 40. místě. Má i nejnižší průměrný věk (necelých 56 let).

| | ID_hldiagnozy | Popis_hldiagnozy | Vek_prumer | Pocet_pripadu_Slovensko |
|---|---|---|---|---|
| 1 | I25 | Chronická ischemická nemoc srdeční | 77.434 | 122 |
| 2 | I63 | Mozkový infarkt | 73.916 | 95 |
| 3 | I61 | Intracerebrální krvácení | 68.193 | 83 |
| 4 | C34 | Zhoubný novotvar průdušky | 65.952 | 83 |
| 5 | J18 | Pneumonie původce | 75.348 | 69 |
| 6 | I21 | Akutní infarkt myokardu | 71.583 | 60 |
| 7 | I64 | Cévní příhoda mozková neurčená jako krvácení nebo infarkt | 75.830 | 53 |
| 8 | I50 | Selhání srdce | 73.115 | 52 |
| 9 | S06 | Nitrolební poranění | 57.780 | 41 |
| 10 | K74 | Fibróza a cirhóza jater | 55.675 | 40 |

**Tabulka 38 - 10 nejčastějších diagnóz končících úmrtím – Slovensko [zdroj vlastní]**

| | ID_hldiagnozy | Popis_hldiagnozy | Vek_prumer | Pocet_pripadu_CR |
|---|---|---|---|---|
| 1 | C34 | Zhoubný novotvar průdušky | 65.906 | 117 |
| 2 | J18 | Pneumonie původce | 78.948 | 97 |
| 3 | I50 | Selhání srdce | 74.682 | 88 |
| 4 | I25 | Chronická ischemická nemoc srdeční | 78.918 | 49 |
| 5 | I64 | Cévní příhoda mozková neurčená jako krvácení nebo infarkt | 76.089 | 45 |
| 6 | J44 | Jiná chronická obstruktivní plicní nemoc | 68.667 | 42 |
| 7 | I26 | Plicní embolie | 73.951 | 41 |
| 8 | I21 | Akutní infarkt myokardu | 75.634 | 41 |
| 9 | I61 | Intracerebrální krvácení | 72.000 | 33 |
| 10 | J96 | Respirační selhání nezařazené jinde | 67.500 | 32 |

**Tabulka 39 - 10 nejčastějších diagnóz končících úmrtím – ČR [zdroj vlastní]**

## Shluková analýza

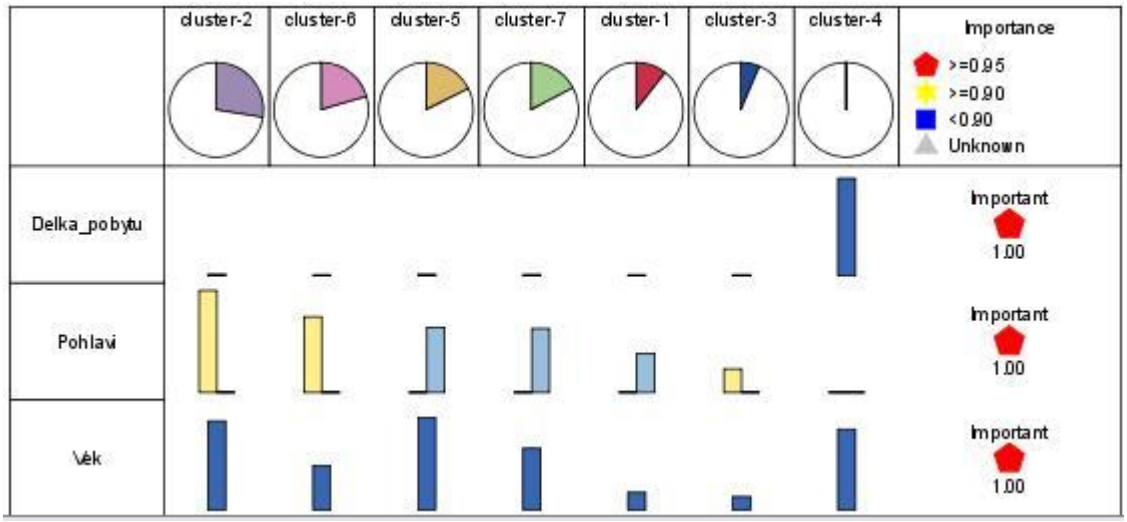
Pro shlukovou analýzu byla použita metoda K-Means, která patří mezi nehierarchické optimalizační metody. Vstupem do modelu byl věk pacienta, délka pobytu a pohlaví. Jako výstup bylo zvoleno 7 shluků (clusterů), do kterých byly vstupní data zařazena.

U vytvořených shluků byl pro větší přehlednost vytvořen sloupec průměrný věk a průměrná délka pobytu. Z tabulky 40 je vidět, že průměrný věk se u všech shluků liší, průměrná délka pobytu má také nepatrně zvyšující se tendenci v závislosti na věku. Výjimku tvoří shluk 4, který má průměrnou délku pobytu 740 dní a je tvořen pouze 3 hospitalizačními případy. Všechny tyto případy měly společnou diagnózu: zhoubný novotvar na průduškách.

| | $KM-K-Means | Pocet_pripadu | Vek_Mean | Delka_pobytu_Mean |
|---|---|---|---|---|
| 1 | cluster-5 | 63589 | 72.100 | 8.121 |
| 2 | cluster-4 | 3 | 59.333 | 740.000 |
| 3 | cluster-2 | 99909 | 69.439 | 8.273 |
| 4 | cluster-7 | 62506 | 48.162 | 6.693 |
| 5 | cluster-6 | 74393 | 34.648 | 5.439 |
| 6 | cluster-3 | 23414 | 10.162 | 4.762 |
| 7 | cluster-1 | 38635 | 13.490 | 5.172 |

**Tabulka 40 - Vytvořené shluky [zdroj vlastní]**

Na obrázku 30 je grafické znázornění vytvořených shluků, které nám uvádí koláčové grafy počtu případů ve shlucích. Dále pak zastoupení vstupních atributů – délky pobytu, pohlaví a věku.




**Obrázek 30 - Obrázek rozdělení do shluků [zdroj vlastní]**

Každý shluk zahrnuje pouze jeden druh pohlaví. Ženy jsou zastoupeny ve shlucích 2, 3, 6. Muži ve zbylých shlucích. Pohlaví tedy mělo zásadní vliv na rozdělení do shluků.

## *5.5 HODNOCENÍ*

### Hodnocení výsledků

Analýza hospitalizačních případů v rámci ČR

První analýzou zabývající se hospitalizačními případy v České republice jsme získali základní představu o 10ti nejčastějších diagnózách, se kterými jsou pacienti hospitalizovány. Zároveň je z ní zřejmé, že nejčastěji jsou pacienti hospitalizováni na základě pouze jedné diagnózy. Dále byl zkoumán výskyt vedlejších diagnóz v rámci hlavní diagnózy a bylo zjištěno, že například vysoký krevní tlak a cukrovka patří mezi rizikové faktory u hospitalizovaných pacientů.

Predikce úmrtnosti

Tato analýza měla na základě vstupních parametrů, které by mohly ovlivňovat způsob ukončení pobytu pacienta, predikovat úmrtnost. Byly využity všechny atributy hospitalizačních případů, jelikož při vyřazení některého z atributů se správnost predikce snižovala. To poukazuje na to, že čím více údajů od pacientů získáme, tím lepší možnost předpovídat průběh jednotlivých nemocí budeme mít. Podařilo se dosáhnout téměř 63% úspěšnosti v predikci při zachování poměru u statusů ukončení („DEA“, „DIS“, „TRA“). Tento výsledek má i přes individualitu každého člověka dobrou vypovídací hodnotu.

Predikce délky pobytu

Tato analýza, pomocí rozhodovacích stromů, se zaměřila na co nejpřesnější stanovení délky pobytu pacienta. Vstupními parametry byly stejné hodnoty jako v predikci úmrtnosti. Výsledný model nevykazoval tak dobré výsledky, jako model předchozí. Absolutní přesnost dosáhla 15ti %. Pokud bychom se spokojili s tolerancí 4 dní, přesnost by dosáhla 57%.

Srovnání délky pobytu dle DRG

Touto analýzou bylo zjištěno, že 92% všech případů je ukončené s délkou trvání, která nepřekračuje limit stanovený DRG. Nicméně existuje 61 diagnóz, které ve všech svých případech tuto stanovenou délku pobytu překračují. U těchto diagnóz by bylo dobré zvážení změnění hodnoty u DRG.

Analýza počtu hospitalizačních případů podle věkových skupin

Věkové skupiny měli zpřehlednit, zda zde existuje závislost mezi věkem pacienta a délkou jeho pobytu v nemocnici. Tento předpoklad se potvrdil. Starší pacienti jsou hospitalizováni déle než mladší. Zvolení do skupin po 10ti letech se neukázalo jako příliš optimální, jelikož se nejčastější diagnózy u věkových kategorií opakovaly. Mohlo by proto dojít ke snížení počtu skupin z 11 na 7.

Regresní analýza délky pobytu a věku

Touto analýzou měla být opět potvrzena závislost mezi délkou pobytu a věkem pacienta. Ta se opět potvrdila, jelikož s tolerancí 4 dnů, získáme 80 % správnost výsledků.

Analýza závislosti způsobu přijetí a způsobu ukončení

Tato závislost se nepotvrdila, jelikož proporcionální podíl statusů ukončení je u obou způsobů přijetí téměř totožný. Nejčastěji jsou pacienti přijímáni běžným způsobem a léčba je ukončena pacientovým propuštěním.

Analýza způsobu ukončení v závislosti na diagnóze

První část této analýzy se zaměřila na zjištění 10ti nejčastějších diagnóz, které končí smrtí pacienta. Dále zjistit průměrný věk a rozdělení pacientů podle pohlaví. Podařilo se zjistit, že muži 8x častěji umírají ve spojitosti s diagnózou zhoubného nádoru na průduškách než ženy. V druhé části je porovnání 4 států a jejich nejčastějších 10 diagnóz. Tady nebyly nalezeny žádné výraznější rozdíly.

Shluková analýza

U této analýzy nás zajímalo, jakým způsobem budou shluky vytvořeny. Zvoleno jich bylo 7, jelikož dle ostatních analýz se zdál tento počet optimální. Nejvíce shluky ovlivnilo pohlaví a věk pacienta. Délka pobytu hrála velkou roli u shluku 4. Opět se zde potvrdila závislost mezi věkem pacienta a délkou pobytu.

## Posouzení procesu

Nejnáročnější na celém procesu byla fáze přípravy dat. Kdy data bylo nutné přizpůsobit dle požadavků programu a samotným analýzám. Tato fáze představovala zhruba 70 % z celkového procesu. Bylo nutné odstranit prázdné řádky a z důvodu množství dat, nebylo možné tuto úpravu provádět ručně. Proto byl vytvořen skript pomocí PHP, který celou práci usnadnil. Dalším problémem byly mezery mezi daty a další drobnosti, které ve výsledném efektu způsobily chybné

načtení dat programem. Nakonec se podařilo všechna data načíst ve správném formátu, což byl první předpoklad úspěšných analýz.

### Vymezení dalších kroků

Dalším krokem by byla hlubší analýza s možností využití jiných metod predikce. Na tuto analýzu by bylo samozřejmě potřeba více prostoru a času. Ale jelikož tato data představují neuvěřitelný potenciál, byla by škoda jich nevyužít.

## *5.6 DOPORUČENÍ PRO PRAXI*

Zásadním problémem je samotný sběr informací o hospitalizovaných pacientech. Tato databáze se skládala ze 4 různých států, což může způsobit odlišnou metodiku sběru a klasifikaci údajů. Doporučením by mělo být zavedení jednotného systému ve všech státech a získání co nejpřesnějších informací o pacientech. Toto zavedení by mohlo přispět ke zvýšení přesnosti predikovaných hodnot. Dále se zaměřit na evidenci o zdravotním stavu rodičů, na základě které by bylo možné lépe predikovat proces hospitalizace potomků pacientů.

Dalším problémem je stanovení maximální délky pobytu dle DRG pro určité diagnózy. Například u neakutní leukémie či zhoubného bujení ledvin a močových cest tato hodnota nebyla dodržena v žádném ze sledovaných případů. Proto by bylo dobré zvážení prodloužení této doby. Toto doporučení se týká zejména pojišťoven, které jsou kompetentní k určování těchto limitů.

# ZÁVĚR

Cílem diplomové práce je provést statistickou analýzu dat, nalézt závislosti mezi daty za účelem vytvoření predikčních modelů a formulovat doporučení pro praxi. Za tímto účelem jsou definované následující kapitoly:

1. Teorie data miningu
2. Rozhodovací stromy
3. Shluková analýza
4. Regresní analýza
5. Aplikace metodologie CRISP-DM

První kapitola je zaměřena na definici pojmů spojených s DM. Jsou zde uvedeny příklady úloh, které jsou vhodné pro DM a v práci byly využity. Dále uvedena metodologie CRISP-DM, která se stala klíčovou při vypracovávání této práce.

V rámci splnění druhé dílčí kapitoly je zde uvedena definice rozhodovacích stromů, popsány základní algoritmy, které se při tvorbě rozhodovacích stromů uplatňují. Dále uvedeny různé metody zvolení vhodného atributu pro následné dělení.

Ve třetí kapitole jsou popsány různé shlukovací metody a popsány jejich základní algoritmy, které je možné při tvorbě shlukové analýzy využít.

Čtvrtá kapitola se zabývá problematikou regresní analýzy. Popsáním modelu jednoduché lineární regrese a vysvětlení principu metody nejmenších čtverců.

K naplnění posledního dílčího cíle je v páté kapitole podrobně rozepsáno zpracovávání dat pomocí metodologie CRISP-DM. Byla učiněna základní statistická analýza hospitalizačních případů. Nalezena závislost mezi věkem pacienta a délkou pobytu, která byla následně využita pro vytvoření predikčních modelů délky pobytu a úmrtnosti pacientů. Využito bylo i shlukové a regresní analýzy. V závěru kapitoly je vše shrnuto a uvedena doporučení pro praxi.

Na základě splnění všech dílčích cílů, je možné konstatovat, že cíl diplomové práce (provést statistickou analýzu dat, nalézt skryté závislosti v datech a doporučení pro praxi) byl splněn. Hlavní přínos práce je orientovaný na statistickou analýzu dat a na návrhy modelů s cílem predikce délky pobytu a případného úmrtí pacienta. Modely dosahují přijatelné testovací chyby, která umožňuje

použít tento vybraný model jako podpůrný či doplňkový nástroj v lékařství. Další návrhy modelů vhodných pro predikci by mohly být provedeny například pomocí neuronových sítí. Tím by se otevřel prostor pro hlubší poznání a zpracování dané problematiky.

# DATOVÝ SLOVNÍK

| Název sloupce | Popis sloupce | Datový typ | Rozsah |
|---|---|---|---|
| ID_pripadu | ID případu | řetězec | |
| ID_pacienta | ID pacienta | řetězec | |
| ID_hldiagnozy | ID hlavní diagnózy | řetězec | |
| Popis_hldiagnozy | Slovní popis (anglicky) | textová hodnota | |
| Popis_hldiagnozy_CJ | Slovní popis (česky) | textová hodnota | |
| ID_vedl.diagnozy | ID vedlejší diagnózy | řetězec | |
| Prijat_kym | Kód lékařského přijetí | řetězec | |
| Popis_kym_prijat | Popis lékařského přijetí | textová hodnota | |
| Status_kym_prijat | Status lékařského přijetí | textová hodnota | |
| Zpusob_ukonceni | Kód způsobu ukončení | řetězec | |
| Popis_ukonceni | Popis způsobu ukončení | textová hodnota | |
| Status_ukonceni | Status způsobu ukončení | textová hodnota | |
| Způsob_prijeti | Kód způsobu přijetí | řetězec | |
| Popis_prijeti | Popis způsobu přijetí | textová hodnota | |
| Status_prijeti | Status způsobu přijetí | textová hodnota | |
| Delka_pobytu | Počet dnů v nemocnici | číselná hodnota | (1-941) |
| Vek | Stáří pacienta | číselná hodnota | (0-104) |
| Pohlavi | Druh pohlaví | textová hodnota | (M, F) |
| Diagnoza_l1 | Třída diagnózy L1 | číselná hodnota | |
| Popis_diagnozy_l1 | Popis třídy diagnózy L1 | textová hodnota | |
| Diagnoza_l2 | Třída diagnózy L2 | řetězec | |

| | | | |
|---|---|---|---|
| Popis_diagnozy_l2 | Popis třídy diagnózy L2 | textová hodnota | |
| ID_DRG | ID DRG | řetězec | |
| Popis_DRG | Popis DRG | řetězec | |
| Slovne_DRG | Slovní popis DRG | textová hodnota | |
| Průměr_dp | Průměrná doba dle DRG | číselná hodnota | (0-68) |
| Horni_dp | Maximální doba dle DRG | číselná hodnota | (0-205) |
| Dolni_dp | Minimální doba dle DRG | číselná hodnota | (0-23) |
| Kod_nemocnice | Kód nemocnice | řetězec | |

# SEZNAM POUŽITÉ LITERATURY


[1] POŠÍK, P.: Co je data mining?, Část 1. [on-line], [cit. 2007-05-20], dostupné z <http://cyber.felk.cvut.cz/gerstner/teaching/zbd/DataMining1-hout.pdf>.

[2] Petr, P.: Data Mining Díl 1., Pardubice: Univerzita Pardubice, 2006, 144 s., ISBN: 80-7194-886-1.

[3] MELOUN, M.: Analýza shluků [on-line], [cit. 2007-05-20], dostupné z <http://meloun.upce.cz/kapitoly/4gmetody.pdf>.

[4] Berka, P.: Dobývání znalostí z databází, Academia, 2003, 366 s., ISBN: 80-200-1062-9.

[5] MelouN, M., Militký J.: Kompendium statistického zpracování dat, Academia, Praha, 2006, 974 s., ISBN: 80-200-1396-2.

[6] ŽÁk, l.: Jeden ze způsobů zpracování nepřesných dat [on-line], [cit. 2007-05-20], dostupné z <http://www.volny.cz/elzet/Libor/F_Shluk.pdf>.

[7] LUKASOVÁ, A., ŠARMANOVÁ, J.: Metody shlukové analýzy, SNTL, Praha, 1985, 210 s.

[8] Kubanová, J.: Statistické metody pro ekonomickou a technickou praxi, Static, Bratislava, 2003, 187 s., ISBN: 80-85659-31-X.

[9] HEBÁK, P.: Regrese, I. část, Vysoká škola ekonomická v Praze, Praha, 1998, 138 s., ISBN: 80-7079-909-9.

# SEZNAM OBRÁZKŮ

# SEZNAM TABULEK

# PŘÍLOHA

Skript vloženého makra

```
Public Sub do_it()

' r ... radek v tabulce na aktivnim listu
' dh,dv ... pocet popisu diagnoz v poli diah/diav
' h,v ... cislo prvku v poli diag odpovidajicimu popisu hlavni/vedlejsi diagnozy
' sh,sv ... popis hlavni/vedlejsi diagnozy
' diah/diav ... jednorozmerne pole popisu diagnoz
' incl ... dvojrozmerne pole zavislosti
' max ... maximum diagnoz ve vysledku

    Const max = 2400

    Dim r, dh, dv, h, v, i As Integer
    Dim sh, sv, s As String
    Dim diah(1 To max) As String
    Dim diav(1 To max) As String
    Dim incl(1 To max, 1 To max) As Boolean
   ' nulovani pole (pro jistotu)
    For h = 1 To max
       For v = 1 To max
          incl(h, v) = False
       Next v
    Next h

    dh = 0
    dv = 0
    r = 1
    Do
' nacteni diagnoz ze sloupcu 2 a 3
       sh = ActiveSheet.Cells(r, 2).Value
       If sh = "" Then Exit Do
       sv = ActiveSheet.Cells(r, 3).Value
       If sv = "" Then Exit Do
' vyhledani h. diagnozy v poli diag
       h = 0
       For i = 1 To dh
          If diah(i) = sh Then
             h = i
             Exit For
          End If
       Next i
' vyhledani v. diagnozy v poli diag
       v = 0
```

```
        For i = 1 To dv
           If diav(i) = sv Then
              v = i
              Exit For
           End If
        Next i
' doplneni h. diagnozy do pole diag
        If h = 0 Then
           If dh = max Then Exit Do
           dh = dh + 1
           diah(dh) = sh
           h = dh
        End If
' doplneni v. diagnozy do pole diag
        If v = 0 Then
           If dv = max Then Exit Do
           dv = dv + 1
           diav(dv) = sv
           v = dv
        End If
' zaevidovani vztahu do pole incl
        incl(h, v) = True
        r = r + 1
    Loop
' otevreni souboru
    Set fs = CreateObject("Scripting.FileSystemObject")
    Set a = fs.CreateTextFile("c:\export.csv", True)
' 1. radek = hlavicka CSV souboru
    s = ""
    For i = 1 To dv
       s = s + ";" + diav(i)
    Next i
    a.WriteLine (s)
' export ostatnich radku
    For h = 1 To dh
       s = diah(h)
       For v = 1 To dv
          If incl(h, v) Then
             s = s + ";1"
          Else
             s = s + ";0"
          End If
       Next v
       a.WriteLine (s)
    Next h
'uzavreni souboru
    a.Close
End Sub
```

# ÚDAJE PRO KNIHOVNICKOU DATABÁZI

| | |
|---|---|
| Název práce | Analýza hospitalizačních případů v oblasti zdravotnictví |
| Autor práce | Eva Sýkorová |
| Obor | Informatika ve veřejné správě |
| Rok obhajoby | 2007 |
| Vedoucí práce | doc. Ing. Pavel Petr, Ph.D. |
| Anotace | V práci je provedena analýza hospitalizačních případů v oblasti zdravotnictví pomocí data minigových metod, na základně kterých je možné získání základní představy o datech, statistických údajů a predikování jejich budoucího vývoje. Úvod je zaměřen na vymezení základních pojmů z teorie data miningu a popsány jeho základní techniky. V druhé části jsou za použití rozhodovacích stromů, regresní a shlukové analýzy v prostředí Clementine 10.1. analyzovány výsledky. |
| Klíčová slova | data mining, shluková analýza, rozhodovací stromy, regresní analýza, hospitalizační případ |