ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

Fakulta elektrotechnická

Katedra kybernetiky

Analýza dat z porodnického modulu nemocničního informačního systému

Hospital information system Obstetrics-module data analysis

Diplomová práce

Studijní program: Biomedicínské inženýrství a informatika

Studijní obor: Biomedicínské inženýrství

Vedoucí práce: Ing. Václav Chudáček, Ph.D.

Bc. Lenka Nejedlá

Praha 2014

České vysoké učení technické v Praze
Fakulta elektrotechnická

Katedra kybernetiky

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: Bc. Lenka N e j e d l á

Studijní program: Biomedicínské inženýrství a informatika (magisterský)

Obor: Biomedicínské inženýrství

Název tématu: Analýza dat z porodnického modulu nemocničního informačního systému

Pokyny pro vypracování:

1. Seznamte se s problematikou dolování dat v medicínském prostředí.
2. Seznamte se s daty dostupnými z porodnického modulu NIS FN Brno.
3. Na základě explorativní analýzy formulujte na podmnožině dat několik zajímavých hypotéz – zajímavých klinicky, technicky či finančně. Spolupracujte s klinickým expertem.
4. Ověřte validitu navržených hypotéz.
5. V závěru práce zformulujte možné směry pokračování Vaší práce, popište technické problémy, se kterými jste se musela potýkat, navrhněte jejich řešení.

Seznam odborné literatury:

[1] Mařík, V.; Štěpánková, O.; Lažanský, J. a kol.: Umělá inteligence 4. Praha: Academia, 2003.
[2] Orozova-Bekkevold, I., et al.: Maternal vaccination and preterm birth: using data mining as a screening tool, J Pharmacy World & Science, 2007.

Vedoucí diplomové práce: Ing. Václav Chudáček, Ph.D.

Platnost zadání: do konce letního semestru 2014/2015

L.S.

doc. Dr. Ing. Jan Kybic
vedoucí katedry

prof. Ing. Pavel Ripka, CSc.
děkan

V Praze dne 10. 1. 2014

Prohlášení

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

V Praze dne …… ……..……

Podpis autora práce

Poděkování

Děkuji Ing. Václavu Chudáčkovi, Ph.D. za odborné vedení mé diplomové práce, za cenné rady a připomínky. Poděkování patří také mé rodině a přátelům za podporu a trpělivost po celou dobu mého studia.

ABSTRAKT

Tato práce se zaměřuje na dolování dat, především na statistické metody dolování dat v lékařství. V teoretické části práce jsou vymezeny základní pojmy z oblasti dolování dat a popsány jednotlivé metodiky a techniky. Praktickou část tvoří analýza datového souboru z porodnického modulu nemocničního informačního systému Fakultní nemocnice Brno. Pro práci s daty byly zvoleny počítačové programy pgAdmin III, Matlab, RStudio a Microsoft Excel 2010. Při analýze dat je postupováno podle metodiky CRISP-DM. Užity byly statistické neparametrické testy: Wilcoxonův dvouvýběrový rank sum test, Kruskal-Walisův test, Wilcoxonův párový test, Spearmanův test nezávislosti, test dobré shody, logistická regrese.

KLÍČOVÁ SLOVA: dobývání znalostí z databází, data mining, porodnictví, statistické metody

ABSTRACT

This thesis deals with data mining in medicine. The theoretical part is an overview of common methods that are used in data mining, especially statistical methods applied in medicine. The practical part is an analysis of the obstetrics database from Faculty Hospital Brno. Software - pgAdmin III, Matlab, RStudio and Microsoft Excel 2010 were used to help with this problem. Data analysis is followed by the methodology CRISP-DM. For data analysis were used statistical nonparametric tests: Wilcoxon two-sample rank sum test, Kruskal - Walis test, Wilcoxon signed-rank test, Spearman's test, Pearson's chi-squared test, logistic regression.

KEY WORDS: knowledge discovery in databases, data mining, obstetrics, statistical methods

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Historické počátky analýzy medicínských dat řadíme do 19. století. Šlo pouze o lokální výzkumy, které neměly velkou vypovídající hodnotu. K rozvoji statistického zkoumání v lékařství došlo s využitím nových laboratorních metod, měřících postupů a uchováváním záznamů o pacientech. Za pomoci jednoduchých statistických metod, jež prováděli sami lékaři díky svým záznamům a zkušenostem, byly vytěženy základní závislosti v pozorovaných datech. V dnešní době se uchovává velké množství dat, téměř v každé nemocnici se setkáme s nemocničním informačním systémem (NIS), kde jsou uchovávány záznamy o pacientech a jejich vyšetřeních. NIS umožňuje uložená data exportovat pro další zpracování. Nesou však všechna data plnohodnotnou informaci? S touto otázkou se začíná zviditelňovat fakt, že nejde primárně o shromažďování informací, ale jedná se hlavně o jejich interpretaci a praktické využití. Data mining je v současné době jedním z nejpoužívanějších nástrojů pro analýzu dat. Nalezl uplatnění v mnoha oborech a tak ani medicína není výjimkou. Z nemocničních databází dostáváme tedy statisticky významná data, která reprezentují určitý populační výběr, což jsou ideální podmínky pro data mining, ovšem data získaná přímo z NIS, jsou v syrové podobě a zatížena množstvím chyb. Proto je nutné pečlivé předzpracování s ohledem na použitý typ úlohy. [7, 21]

Hlavní náplní, jak již samotný název diplomové práce napovídá, je provést analýzu datového souboru z porodnického modulu nemocničního informačního systému Fakultní nemocnice Brno. Pro analýzu jsme si vybrali statistické metody. Cílem je tedy zjistit, která data nesou důležitou informaci pro specifické problémy spojené s porodem, např. jaké příznaky ovlivňují nízké pH novorozence, provedení císařského řezu nebo délku porodních dob. Dalším úkolem je popsat technické problémy, s nimiž se při předzpracování i analýze dat potýkáme a navrhnout jejich řešení.

V úvodní části práce jsou vymezeny základní pojmy z oblasti data miningu a popsány jednotlivé metodiky a techniky, které data mining využívá. Dále pak práce shrnuje informace o statistických metodách data minigu v medicíně. Poslední kapitola teoretické

části se zabývá obdobnými pracemi jiných autorů, ukazuje na problémy, se kterými se při analýze dat v lékařství setkáváme, jaká témata se zkoumají a jakými metodami se řeší. Praktická část se zabývá analýzou dat za pomoci statistických dataminingových metod (Wilcoxonův dvouvýběrový rank sum test, Kruskal – Walisův test, Wilcoxonův párový test, Spearmanův test nezávislosti, test dobré shody, logistická regrese), a je zde postupováno podle metodiky CRISP-DM.

Teoretická část

1 Dobývání dat (data mining)

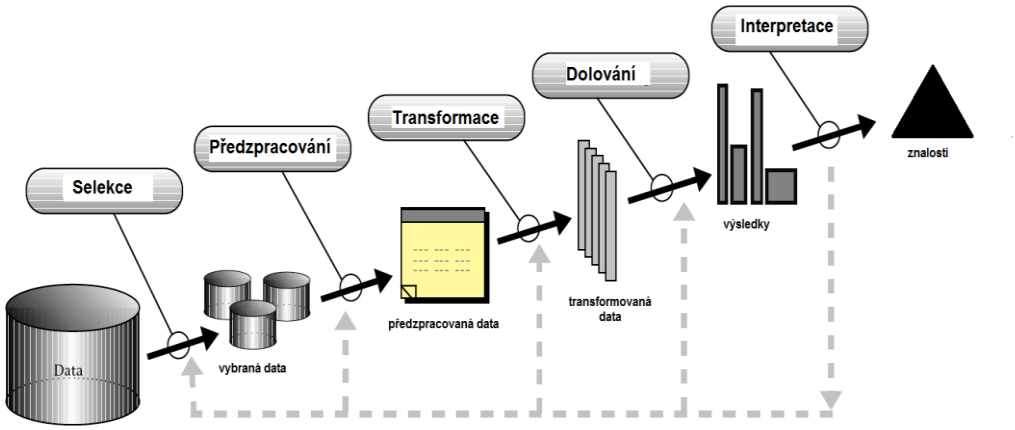
Termín data minig (DM), který je nejčastěji překládán jako dobývání (či dolování) dat (znalostí), podle uznávaného výzkumníka Usamy Fayyada je definován jako netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat.

Historické počátky aktivit, jež dnes označujeme jako data mining, jsou datovány do 60. let 20. století a souvisí s rozvojem počítačové techniky. Získané postupy sloužily pouze pro výzkumné účely a jejich zavádění do praxe bylo velmi ojedinělé. To se změnilo v sedmdesátých a osmdesátých letech díky narůstající rychlosti a paměti počítačů. Nějvětší rozvoj data miningu ale nastal až v 90. letech minulého století a to díky rozmachu umělé inteligence (přesněji strojového učení). Právě v této době se začíná používat termín dobývání znalostí z databází (Knowledge Discovery in Databases, dále jen KDD).

Většina autorů se shoduje na tom, že jde o postup, při němž se ze surových dat, která jsou nejčastěji k dispozici ve formě databáze či relačních tabulek datového skladu, získávají pomocí statistických a logických metod znalosti, které mohou být využity ke strategickému rozhodování. Dobývání znalostí z databází vzniklo tedy propojením poznatků ze tří oblastí: databází (slouží pro uchování velkého množství dat a hledání informací v nich), statisitky (umožňují analýzu dat a hledání souvislostí v nich) a strojového učení (oblast umělé inteligence zabývající se problematikou vytvoření programů schopných učit se ze zkušeností). KDD se používá nejen ve vědeckém výzkumu, ale i ve většině sfér běžného život. Nárůst aplikací v oblasti data miningu se projevil i na softwarovém trhu, existuje již poměrně široká nabídka specializovaných softwarů pro tento účel. Mezi komerční aplikace patří například SAS Enterprise Miner a STATISTICA Data Miner, mezi známé nekomerční softwary patří Weka a Orange. [3, 4, 7, 11]

V současné době tedy chápeme termín data mining jako jednu fázi širšího procesu dobývání znalostí z databází. KDD je chápáno jako interaktivní a iterativní proces tvořený kroky:

- selekce - vybrání dat z databáze, jež jsou relevantní pro řešenou úlohu
- předzpracování - odstranění šumu, odstranění odlehlých hodnot, doplnění hodnot aj.
- transformace - převod dat do podoby vhodné pro analýzu, často početně náročné operace
- data mining (dolování z dat) - aplikování metod umělé inteligence a získání vzorů v datech (data patterns), častý je iterativní průběh a kombinování více typů analytických metod
- interpretace - vizualizace nebo jiná prezentace znalostí v podobě snadno pochopitelné pro uživatele. [3, 11]



Obrázek 1 : Data mining[1]

[1] Upraveno z : ALTHAUS, Kevin et al. Anwendungsmöglichkeiten von Text Mining im Web Content Mining. In: WinfWiki [online]. [cit. 2014-05-13]. Dostupné z:http://winfwiki.wi-fom.de/index.php/Anwendungsm%C3%B6glichkeiten_von_Text_Mining_im_Web_Content_Mining

1.1 Metodiky KDD

Aby mohlo být dobývání dat z databází co nejefektivnější a přehledné, začaly vznikat různé metodologie, jejichž cílem je poskytnout uživatelům jednotný rámec. Metodologie je tedy standardizovaný návod, který po jednotlivých krocích popisuje, jak během celého procesu KDD postupovat.

Tři nejznámější metodiky jsou Metodika 5A, Metodika SEMMA a Metodika CRISP-DM.

1.1.1 Metodika 5A

Kroky metodiky jsou:

- *Assess* (posouzení) – stanovení kontextu-cílů, strategií a procesů
- *Access* (získávání) – shromáždění a příprava potřebných dat
- *Analyze* (analyzování) – provedení datových analýz, používá se více metod a porovnávají se jejich výsledky a efektivita
- *Act* (provedení) – přeměna znalostí na akční znalosti, rozhodnutí
- *Automate* (automatizace) – převedení výsledků analýz do praxe a následné užívání

1.1.2 Metodika SEMMA

Kroky metodiky jsou:

- *Sample* (vzorek) – výběr vhodných dat
- *Explore* (poznávání) – průzkum a redukce dat
- *Modify* (úprava) – datové transformace, seskupování hodnot atributů,
- *Model* (modelování) – analýza dat
- *Assess* (posouzení) – porovnávání modelů a jejich interpretace srozumitelná pro uživatele [4, 16]

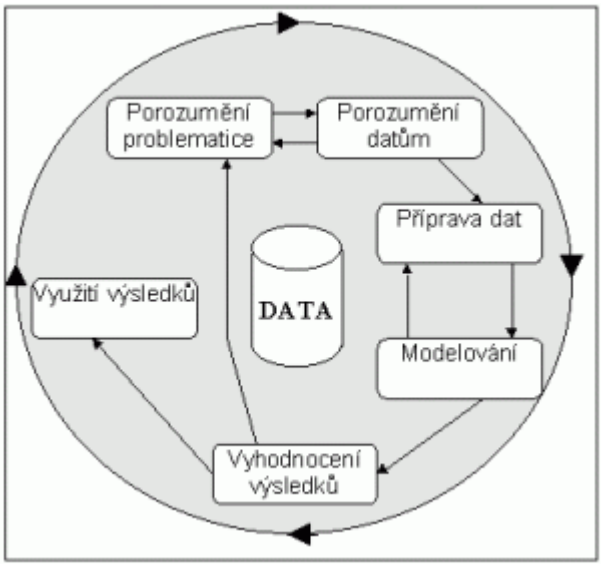
1.1.3 Metodika CRISP-DM

Tato metodika je nejrozšířenější, vznikla v roce 1996 v rámci projektu Evropské komise. Cílem metodiky je navrhnout univerzální postup při řešení projektů, použitelný v různých komerčních softwarových aplikacích, a návrh řešení problémů, které mohou během projektu nastat. CRISP-DM (CRoss-Industry Standard Proces for Data Mining) rozděluje proces do 6 kroků, výsledky jednotlivých etap se navzájem ovlivňují a na základě těchto výsledků je často potřeba vracet se k předchozím fázím.

Kroky metodiky jsou:

- *Business understanding* (Porozumění problému) – Pro tuto počáteční fázi je vyžadováno pochopení cílů úlohy a požadavků na její řešení. Posuzují se zde rizika a přínos projektu. V této fázi tedy dochází ke stanovení předběžného plánu projektu a k analýze přínosů.
- *Data understanding* (Porozumění datům) – Tato fáze je charakterizována sběrem data a následným seznámením s nimi, například pomocí popisné statistiky (četnosti hodnot atributů, průměry, minima, maxima atd.).
- *Data preparation* (Příprava dat) - Fáze přípravy dat zahrnuje veškeré činnosti potřebné k vytvoření konečného datového souboru. Provádí se zde selekce dat, čištění dat (odstraňování odlehlých hodnot), transformace dat, odvozování dat atd. Tato fáze je nejpracnější a často vyžaduje opakování.
- *Modeling* (Modelování) - Úkolem této fáze je výběr vhodných algoritmů pro analýzu dat. Doporučuje se vyzkoušet více různých metod s různým nastavením parametrů a výsledky pak porovnat a zkombinovat. Některé techniky mají specifické požadavky na podobu dat, obvykle je tedy potřeba vrátit se zpět k fázi přípravy dat.
- *Evaluation* (Vyhodnocení výsledků) - V této fázi je již vytvořen kvalitní model. Před konečným využitím modelu je nutné přezkoumat, zda skutečně dosáhneme daných cílů. Na konci této fáze by mělo být rozhodnuto o využití výsledků data miningu.

- *Deployment* (Využití výsledků) - Získané znalosti je třeba interpretovat v dostatečně srozumitelné podobě, aby je uživatel mohl efektivně využít. Výstupem celého procesu může být jak prosté sepsání závěrečné zprávy, tak složitější zavedení systému pro automatickou klasifikaci a predikci nových případů. [3, 4, 15, 16]



Obrázek 2 : Metodika CRIPS – DM[2]

1.2 Kategorie úloh dolování dat

Predikce a deskripce jsou dva základní cíle data miningu v praxi.

- Predikce: předvídání budoucí hodnoty atributu na základě nalezených vzorů v datech. Typickým úkolem je najít určitou hodnotu atributu díky znalosti jiných atributů.

[2] Převzato z: BERKA, Petr. Aplikace systémů dobývání znalostí pro analýzu medicínských dat. In: *EuroMISE* [online]. 2001 [cit. 2014-05-11]. Dostupné z: http://euromise.vse.cz/kdd/index.php?page=procesk dd

- Deskripce: popis nalezených vzorů a vztahů v datech, které mohou ovlivnit rozhodování.

Predikce a deskripce je nejčastěji dosaženo pomocí klasifikace, regrese, shlukování, sumarizace, modelování závislostí a detekce změn a odchylek.

Klasifikace (Classification)

Podstatou klasifikace je rozdělit data do jednotlivých tříd pomocí modelu, který byl vytvořen na tréninkové množině dat (každý objekt je možné zařadit do jedné z předem daných tříd). Nejjednodušším typem klasifikace je binární, jenž má jen dvě možné hodnoty. Výsledkem klasifikace jsou diskrétní, kategorické hodnoty (př.: podle kombinace atributů teplota (28), množství srážek (0) určujeme výsledek roční období -> léto).

Regrese (Regression)

Předpovídá číselnou hodnotu, je to řada dříve zjištěných hodnot, která se používá pro predikci následujících hodnot. Regresní modely jsou testovány určením rozdílu mezi předpovídanou a očekávanou hodnotou. Výsledkem regrese je tedy reálné číslo (př.: teplota se měří po celou dobu - v dešti, při zataženém nebo slunečném počasí a v každém období. V případě kombinace "déšť", "slunečné počasí" a "léto", budeme pomocí regrese očekávat teploty 26,5 stupňů Celsia).

Shlukování (Clustering)

Je podobné klasifikaci s tím rozdílem, že nevyužívá cílových hodnot. To znamená, že hledá přirozené skupiny dat, pro které platí, že podobnost dat mezi shluky je velmi malá a uvnitř shluků velmi vysoká. Využívá se například při prozkoumávání dat nebo detekci anomálií, kdy se některá data nemusí hodit do žádné skupiny dat.

Modelování závislostí (Dependency Modeling)

Hledání modelu, který popisuje významné závislosti mezi proměnnými. Model má dvě úrovně:

1. Strukturální úroveň udává, které proměnné jsou na sobě závislé,

2. Kvantitativní úroveň popisuje síly závislostí za použití číselné stupnice.

Sumarizace (Summarization) a Detekce změn a odchylek (Change and Deviation Detection)

Sumarizace obsahuje metody pro nalezení uceleného popisu podmnožiny dat. Cílem detekce změn a odchylek je nalézt takové případy, které jsou neobvyklé ve zdánlivě homogenních datech. [11, 29]

1.3 Techniky data miningu

Nejznámějšími metodami užívanými při řešení zmíněných kategorií úloh jsou statistické metody, rozhodovací stromy, asociační pravidla, neuronové sítě.

1.1.1 Statistické metody

Statistika nabízí celou řadu teoreticky dobře prozkoumaných a léty praxe ověřených metod pro analýzu dat. Pro oblast dobývání znalostí z databází mají význam:

- kontingenční tabulky – pro zjišťování vztahu mezi dvěma kategoriálními veličinami,
- regresní analýza – pomocí ní odhadujeme hodnotu jisté náhodné veličiny (takzvané závisle proměnné, nazývané též cílová proměnná, regresand anebo vysvětlovaná proměnná) na základě znalosti jiných veličin (nezávisle proměnných, regresorů, kovariát anebo vysvětlujících proměnných)
    - závisle proměnná je skalár nebo vektor z nějakého lineárního prostoru =>lineární regrese
    - závisle proměnná je diskrétní => diskriminační analýza – pro odlišení příkladů (pozorování) patřících do různých tříd
    - závisle proměnná je binární (nabývá pouze dvou hodnot) => logistická regrese
- shluková analýza – pro nalezení skupin (shluků) navzájem si podobných příkladů.

- Bayesovská klasifikace - využívá důsledků Bayesovy věty pro klasifikaci záznamu přiřazením třídy z množiny možných tříd. [3, 4]

1.3.1 Rozhodovací stromy

Cílem rozhodovacích stromů je identifikovat objekty, popsané různými atributy, do tříd. Velkou výhodou rozhodovacích stromů je jejich přehlednost a snadná interpretovatelnost znalostí. Rozhodovací strom se skládá z uzlů stromu, ty představují body, v nichž se strom na základě hodnoty některého z atributů větví. Na konci rozhodovacího stromu jsou tzv. listy stromu, podmnožiny, které reprezentují jednotlivé třídy cílového atributu. Rozhodovací strom zařazujeme do metod učení s učitelem to znamená, že se nejprve musí vytvořit z množiny daných objektů, které jiný algoritmus (učitel) zařadit do tříd (třída se obvykle označuje jako závislý atribut a zapisuje se do tabulky do posledního sloupce). Nejčastěji se při tvorbě rozhodovacích stromů postupuje metodou rozděl a panuj (divide and conquer). Trénovací data se rozdělují na menší podmnožiny tak, aby v těchto podmnožinách převládaly příklady jedné třídy až do té doby, dokud nejsou všechny příklady z trénovací množiny pokryté v jednotlivých podmnožinách stromu. Tento postup bývá nazýván top down induction of decision trees (TDIDT). [3, 11, 21]

1.3.2 Asociační pravidla

Pojem asociační pravidla velmi zpopularizoval Agrawal počátkem 90. pomocí analýzy nákupního košíku, při níž je zjišťováno jaké druhy zboží si současně zákazníci kupují (např. supermarketech - pivo a párek). Jedná se tedy o hledání všech zajímavých vzájemných vazeb mezi jednotlivými atributy, přičemž žádný atribut není vyčleněn jako cíl klasifikace. Asociační pravidlo má tvar ANT => SUC, kde levá část pravidla se nazývá předpoklad (antecedent), pravá potom závěr (sukcedent). Základní charakteristikou pravidel jsou dvě odvozené veličiny a to podpora (support) – rozumíme tím počet objektů, které splňují předpoklad i závěr a spolehlivost (confidence) – ta je podmíněná pravděpodobností závěru, pokud platí předpoklad. V medicíně se asociační pravidla užívají pro identifikaci nových závislostí v datech při dlouhodobějším pozorování a v expertních

systémech. Velkou nevýhodou této metody je velká výpočetní náročnost celého procesu, protože se při hledání asociačních pravidel vytváří všechny kombinace vstupních atributů. [3, 21]

1.3.3 Neuronové sítě

Umělé neuronové sítě vycházejí z analogie s lidským mozkem. Podobně jako mozek jsou tvořeny množstvím navzájem propojených neuronů. Neuron je chápán jako element, který přijímá podněty od jiných neuronů, jenž jsou k němu připojeny „na vstupu“. Pokud souhrnný účinek těchto vstupních podnětů překročí určitý práh, dojde k aktivaci neuronu a začne svým výstupem působit na další neurony. První modely neuronů a neuronových sítí se zkoumaly v rámci umělé inteligence již v 50. letech. Neuronové sítě se dají využít jak pro učení s učitelem – vícevrstevné dopředné sítě a Hopfieldovy zpětnovazební sítě, tak i pro učení bez učitele – Kohonenovy samoorganizující se mapy nebo metoda SVM (Support Vector Machine). Neuronové sítě jsou uspořádané ve vrstvách, první vrstva je vrstva vstupních neuronů, pak následuje několik skrytých vrstev a poslední je vrstva výstupní. Všechny neurony mezi sousedními vrstvami jsou propojeny vahami. [3, 21]

2 Statistické metody používané v medicíně

2.1 Jednovýběrový Kolmogorovův-Smirnovův test

Pomocí KS testu ověřujeme, zda náhodná proměnná má předpokládané (teoretické) rozdělení, nejčastěji se jím ověřuje normalita dat. Nulová hypotéza H_0 předpokládá, že testovaný výběr odpovídá normálnímu rozdělení. Data jsou rozdělena do k tříd, do stejného počtu tříd je rozděleno i předpokládané normální rozdělení. Nad každou třídou testovaného i teoretického výběru se spočítají četnosti n_{1i}, n_{2i}. Hodnotícím kritériem je pak $D = \frac{1}{n} \max|N_{1i} - N_{2i}|$, kde n je celkový počet prvků výběru a $|N_{1i} - N_{2i}|$ je absolutní hodnota rozdílu kumulativních četností výběru a testovaného rozdělení. Hodnotící kritérium se porovnává s tabelovanou kritickou hodnotou pro danou hladinu významnosti α.

2.2 Dvouvýběrový t-test

Dvouvýběrový (nepárový) t-test, slouží k porovnání střední hodnoty jednoho souboru se střední hodnotou druhého souboru. V lékařském výzkumu se obvykle zajímáme o rozdíl mezi populačním průměrem sledované veličiny v ošetřované skupině (treatment mean) a populačním průměrem této veličiny v kontrolní skupině (control mean). Sledujme tedy rozdíl mezi dvěma výběrovými průměry.

2.3 Wilcoxonův dvouvýběrový rank sum test (neparametrický pořadový test Mann – Whitney)

Tento test slouží k porovnávání mediánů dvou různých výběrových souborů, které nemají normální rozdělení pravděpodobnosti. Nulová hypotéza H_0 zní: Mediány obou výběrů jsou shodné. Alternativní hypotéza H_A: Mediány obou výběrů se liší. Nejprve je nutné seřadit hodnoty všech pozorování do neklesající posloupnosti a určit jejich pořadí. Poté se vypočítají testovací statistiky $U_1 = S_1 - \frac{n_1(n_1+1)}{2}$, $U_2 = S_2 - \frac{n_2(n_2+1)}{2}$, kde n_1, n_2 jsou rozsahy výběrů a S_1, S_2 jsou součty pořadí

jednotlivých výběrů. Nulovou hypotézu zamítáme, pokud je menší z hodnot U_1 a U_2 < než tabelovaná kritická hodnota $U(n_1, n_2, \alpha)$.

2.4 Kruskal – Wallisův test

Kruskal – Walisův test je neparametrický test, který je obdobou jednoduchého třídění analýzy rozptylu (ANOVA pro jeden faktor). Je rozšířením Wilcoxonova dvouvýběrového testu pro k výběrů, přičemž k ≥ 3. Nulová hypotéza H_0 zní: Mediány všech výběrů jsou shodné. Alternativní hypotéza H_A: Mediány všech výběrů se liší. Opět je nejprve nutné seřadit hodnoty všech pozorování do neklesající posloupnosti a určit jejich pořadí. Poté se vypočítá pro každý výběrový soubor suma pořadí T_k a určí se celkový rozsah výběru. $N = n_1 + n_2 + \dots + n_k$, kde n_i označuje počet hodnot každého souboru. Nakonec se vypočítá testovací statistika pomocí následujícího vztahu $Q = \frac{12}{N^2+N} \sum_{n_i}^{k} \left(\frac{T_i^2}{n_i}\right) - 3(N+1)$. Za předpokladu, že $n_i \rightarrow \infty$ a za platnosti H_0 má Kruskal-Walisův test asymptoticky X^2 rozdělení o (k – 1) stupních volnosti. Nulovou hypotézu nezamítáme, pokud je testovací veličina ≥ než tabelovaná kritická hodnota $X^2_{k-1}(\alpha)$. Zamítneme-li H_0 pak ještě určujeme, které dvojice výběrů se od sebe statisticky významně liší. Postupuje se následovně: označíme $t_i = \frac{T_i}{n_i}\ pro\ i = 1, \dots, k$, potom můžeme říci, že se od sebe distribuční funkce i-tého a j-tého výběru statisticky významně liší, jestliže platí: $|t_i - t_j| > \sqrt{\frac{1}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right) N(N+1) X^2_{k-1}(\alpha)}$.

2.5 Wilcoxonův párový test

Je to neparametrický pořadový test založený na porovnávání párových hodnot jednoho výběrového souboru. Obvykle to bývá měření před a po nějakém zásahu. V našem případě ho používáme k porovnávání aktuálního a předchozího těhotenství. Nulová hypotéza H_0: Medián rozdílů je nulový. Alternativní hypotéza H_A: Medián rozdílů je různý od nuly. Pro testování je nejprve nutné vypočítat rozdíly mezi párovými hodnotami, přičemž nulové rozdíly do dalšího výpočtu nezařazujeme

(n- počet párů s nenulovým rozdílem). Poté určíme pořadí rozdílů v absolutních hodnotách a nakonec vypočteme součet pořadí kladných rozdílů S_+ a součet pořadí záporných rozdílů S_-. Nulovou hypotézu zamítáme, když je menší z S_+ a S_- < tabelovaná kritická hodnota $S_{(n,\alpha)}$.

2.6 Spearmanův test nezávislosti

Neparametrický pořadový test, jenž zjišťuje, zda jsou sledované veličiny (X,Y), které nemají normální rozdělení dat, na sobě závislé. Nulová hypotéza H_0: Veličiny jsou nezávislé. Alternativní hypotéza H_A: Sledované veličiny jsou na sobě závislé. Spearmanův korelační koeficient je $r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$, kde d_i jsou rozdíly mezi pořadím hodnot X_i a Y_i a n je počet korelačních dvojic. Nulovou hypotézu zamítáme, pokud absolutní hodnota $|r_s|$ > tabelovaná kritická hodnota $r_{s\,(n,\alpha)}$. Korelační koeficient nabývá hodnot pouze od -1 do 1. Znaménko minus ukazuje na opačné pořadí sledovaných veličin. Čím více se korelační koeficient blíží nule, tím méně jsou veličiny na sobě závislé.

2.7 Test dobré shody, test nezávislosti a homogenity v kontingenční tabulce

Test dobré shody je metoda matematické statistiky, která umožňuje ověřit, zda má náhodná veličina určité předem dané rozdělení pravděpodobnosti. Test se mimo jiné často používá pro ověřování hypotéz v kontingenční tabulce.

Kontingenční tabulka se užívá k přehledné vizualizaci vzájemného vztahu dvou statistických znaků. Kategorie jednoho znaku určují řádky (r) a kategorie druhého znaku sloupce (s). Klasický test nezávislosti nebo homogenity je založen na testu dobré shody, tedy porovnání očekávaných četností a skutečných četností a slouží ke zjištění, zda mezi dvěma znaky existuje prokazatelný výrazný vztah.

Testovací statistika se počítá následovně $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$, kde

- n_{ij} je četnost v řádku i a sloupci j (počet pokusů, při nichž má první znak hodnotu odpovídající řádku i a druhý znak hodnotu odpovídající sloupci j),
- m_{ij} očekávaná četnost v řádku i a sloupci j.

Očekávanou četnost vypočteme podle vztahu $m_{ij} = \frac{R_i S_j}{N}$, kde

- R_i je součet všech četností v řádku i (počet pokusů, při nichž má první znak hodnotu odpovídající řádku i bez ohledu na druhý znak),
- S_j je součet četností ve sloupci j (počet pokusů, při nichž má druhý znak hodnotu také odpovídající sloupci j bez ohledu na první znak),
- N je součet četností v celé tabulce (počet všech pokusů).

Pokud hypotéza nezávislosti (resp. homogenity) platí, má testová statistika přibližně rozdělení chí kvadrát o (r-1) (s-1) stupních volnosti. Hodnota testové statistiky se tedy porovná s kritickou hodnotou (kvantilem) příslušné hladiny významnosti. [25, 33]

2.8 Logistická regrese

Logistická regrese je metodou matematické statistiky, která slouží k modelování vztahů mezi vysvětlujícími (nezávislými) proměnnými a vysvětlovanou (závislou) proměnnou, která má binomické rozdělení (nabývá pouze dvou hodnot, např. 0 - jev nenastal, 1 - jev nastal). Nezávislé proměnné označujeme vektorem x = (x_1, x_2, ... , x_p), mohou být jak spojité, tak kategorizované. V logistické regresi odhadujeme pravděpodobnost výskytu jevu P(Y=1).

Šance (angl. odds), že nastal jev Y = 1, je definována jako podíl pravděpodobnosti, že Y = 1 a pravděpodobnosti, že Y ≠ 1, tedy: $šance(Y = 1) = \frac{P(Y=1)}{1-P(Y=1)}$. Šance vyjadřuje, kolikrát je vyšší pravděpodobnost, že Y nabude hodnoty 1, než pravděpodobnost, že nabude hodnoty 0. Hodnoty šance leží v intervalu (0; ∞), proto se provádí logitová transformace, tzv. zavedení přirozeného logaritmu, čímž dostaneme interval (- ∞; ∞). $logit\left(P(Y = 1)\right) = \ln\left(šance(Y = 1)\right) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$

Pravděpodobnost jevu Y=1 pak vyjádříme jako: $P(Y=1)=\frac{1}{1+e^{-x'\beta}}$, $\beta = (\beta 0, \beta 1, \ldots, \beta m)$ – koeficienty modelu, $x = (x_1, x_2, \ldots, x_p)$ – nezávislé proměnné.

Jelikož pro různé realizace x náhodného vektoru X nabývá pravděpodobnost různých hodnot, je proto tato pravděpodobnost podmíněná: $P(Y=1|X=x)=\frac{1}{1+e^{-x'\beta}}$

Pro odhady koeficientů modelu $\beta = (\beta 0, \beta 1, \ldots, \beta m)$ se používá metoda maximální věrohodnosti (maximum-likelihood).

Poměr šancí (angl. odds ratio, OR) určuje velikost změny šance jevu (kolikrát se zvětší šance Y =1), při jednotkové změně nezávisle proměnné, jestliže zbývajících n - 1 veličin je fixovaných. $OR(X_i)=e^{\beta i}$ [25, 32]

3 Související práce

V následující kapitole budou popsány prameny a literatura související s DM v porodnictví, konkrétně zde budou popsány problémy se zpracováním dat, se kterými se autoři podobných prací zabývají. Dále bude zmíněno, jaká témata jsou nejvíce zkoumána a jaké DM statistické metody se nevíce používají.

3.1 Problémy se zpracováním medicínských dat

Problémů se zpracováním medicínských dat je mnoho. Na rozdíl od data miningu spojeného se zpracováním technických či ekonomických dat (kde se tento obor prosazuje nejvíce), vstupuje do problematiky v medicíně složka složitého subjektivního vyhodnocování výsledku. Pro analytiky jsou často lékařská data nesrozumitelná, a proto by měly být všechny kroky při zpracování dat konzultovány s expertem – lékařem. Tak se může zabránit ztrátě cenných informací při samotném zpracování.

3.1.1 Nestrukturovaná data

Medicínská data jsou uložena v databázových systémech, jež jsou součástí nemocničních informačních systémů. Dobře vyvinutý NIS poskytuje dobrý způsob pro ukládání dat, které se mohou stát dostupné prostřednictvím internetu nebo intranetu. Z většiny NIS lze exportovat data ve formě tabulek, což je výhodné, protože většina data miningových metod je navržena pro práci s tabulkami. Řádky tabulky reprezentují jednotlivé pacienty. Sloupce tabulky znázorňují atributy, které představují hodnoty různých naměřených údajů a výstupy z rozmanitých vyšetření. Některé atributy (např. poznámka, adresa) obsahují více informací. Pokud mají být tyto informace automaticky zpracovávány, musí být nejdříve převedeny do srozumitelnější podoby, ve které jsou jednotlivá fakta přímo přístupná.

Článek [18] uvádí, že 50% klinických údajů popisujících stav pacienta během léčby je uloženo v nestrukturovaných textech. Analýzou volného textu za účelem získání informací se zabývá taxt maning. [4, 7, 21]

3.1.2 Velikost databáze

Jedním z problémů data miningu medicínských dat souvisí se zpracováním databáze jako celku. Celé tabulky bývají velmi rozsáhlé, obsahují až desítky tisíc pozorování (počet pacientů) a desítky až stovky atributů. Pro jednodušší zpracování se proto provádí selekce nebo extrakce příznaků. Selekce – výběr pouze některých příznaků, které mají pro danou úlohu nějaký význam (u úloh klasifikace se často využívá výběru příznaků podle informačního zisku). Extrakce – odvození nových příznaků z původních (výška, váha → BMI). [21]

3.1.3 Nesourodost dat

Medicínská data jsou velmi nesourodá. Informace u jednotlivých pacientů se liší, protože ne všichni pacienti trpí stejným onemocněním stejně tak nepodstupují stejné vyšetření a laboratorní testy. [31]

Další nesourodost se projevuje v oblasti používaných standardů pro skladování medicínských dat. Tyto standardy jsou často specifické pro dané oddělení a pro srovnání s jinými hodnotami je potřeba data transformovat. Transformovat data musíme také podle charakteru řešeného problému a použité metody, některá data v databázi totiž mohou být spojitá (hodnota krevního tlaku, teplota), jiná kategoriální (pohlaví - muž/žena, kouření – ano/ne). Transformaci spojitých dat provádíme např. pomocí prahování podle jednoho nebo více parametrů.

V databázích se také často objevují chybějící hodnoty. Postupů, jak pracovat s nulovými hodnotami, je několik, například ve zdroji [26] používají tři různé metody, nahrazení průměrem, odstranění atributů, které obsahují 90 % chybějících údajů a odstranění instancí, jež obsahují 6 nebo více chybějících hodnot a zbývající chybějící hodnoty doplnili mediánem.

Další chyby vznikají při zápisu dat lékařem, například používají různé názvy (synonyma) pro popis stejné nemoci. Nebo překlepem mohou vznikat odlehlé hodnoty. Ve zdroji [10] se píše, že odlehlé hodnoty svádějí k tomu, aby byly odstraněny

ze souboru dat. Je rozšířená domněnka, že mají špatný vliv na vypočtené statistiky. Např. že falešně zvyšují hodnotu směrodatné odchylky jako míry rozptylu dat nebo že mohou způsobit vychýlení (bias) počítaného průměru. Existuje však zlaté pravidlo, které říká, že ze sady dat se nikdy nemá vyloučit nějaká hodnota pouze ze statistických důvodů. V článku [26] byly odlehlé hodnoty zjišťovány pomocí euklidovy vzdálenosti a algoritmu k- nejbližší soused.

Reprezentace hodnot ve formě dvourozměrné tabulky přináší také nevýhodu a to nemožnost sledovat časový vývoj hodnot určitého atributu Časový vývoj hodnot je zpravidla k dispozici, ale jeho začlenění do klasické dvourozměrné tabulky je takřka nemožné. Většinou se používaní pouze dva údaje, které odpovídají hodnotám daných atributů při příjmu a při propuštění pacienta. Jakákoliv práce s takovou informací vyžaduje dobrou znalost problému. [4, 7, 21]

3.1.4 Etické problémy

Další problémy souvisejí s etickými a společenskými otázkami. Před samotnou prací s daty je nutné údaje anonymizovat, v medicíně se tím rozumí zbavit je údajů, podle kterých by mohli být pacienti identifikováni (jméno, příjmení, rodné číslo). A dále je nutné přijmout opatření, aby se data nedostala k rukám třetí osoby. [21]

3.2 Témata podobných prací a využité statistické metody

Předčasný porod

V oblasti porodnictví se autoři dataminingových úloh věnují nejčastěji tématu předčasných porodů. Přičemž předčasný porod je definován jako narození dítěte před 37. týdnem těhotenství (těhotenství kratší než 259 dnů) a je hlavním příčinným faktorem neonatální mortality a morbidity. Zdroj [20] uvádí četnost předčasných porodů v rozmezí 5 až 10 procent, udává také, že až 20 % předčasných porodů jsou porody

indukované a to z důvodu závažných těhotenských patologií (např. preeklampsie[3], intrauterinní růstová retardace plodu). Jako prediktory předčasného porodu označuje cervikální inkompetenci, předčasný odtok plodové vody, infekce, stres, těžká fyzická práce, sociální faktory, onemocnění matky a jako jeden z nejvýznamnějších rizikových faktorů předčasného porodu uvádí předčasný porod v anamnéze.

Problematice předčasného porodu se také věnuje článek [6], který také zkoumá rizikové faktory spojené s předčasným porodem (věk matky, pohlaví dítěte, výšku a hmotnost matky, její návyky – kouření, alkohol, atd.), ale především se věnuje závislosti mezi předčasným porodem a mateřskými volnočasovými pohybovými aktivitami před a během těhotenství. Studii provádí z 1714 dotazníků, k testování využívá **test dobré shody** pro jednorozměrnou analýzu a pro více rozměrnou analýzu **logistickou regresi**. Výsledky ukazují, že mateřské volnočasové pohybové aktivity před nebo během těhotenství, nemají vliv na předčasný porod. Článek [28] se také zabývá jednorozměrnou analýzou prediktorů předčasného porodu, kromě testu dobré shody, používá ještě test **ANOVA.** Tím odpadá nutnost převádět intervalové prediktory jako je např. věk matky, hmotnost, atd. na data kategoriální. Avšak tento test může být použit pouze pro data s normálním rozdělením.

Test dobré shody byl též použit v článku [2], kde jím porovnávali soci-demografické charakteristiky mezi staršími a mladšími matkami (kouření, rodinný stav, vzdělání, atd). Zjistili, že pro starší matky je více pravděpodobné, že jsou bílé, vdané, mají za sebou více porodů, obézní a mají vyšší úroveň vzdělání.

Gestační diabetes

V článku [13] testem dobré shody nebyly zjištěny rozdíly v počtu těhotenství (primiparita x multiparita) a rizikovým faktorem pro **diabetes** (obesita, diabetes prvního

[3] Těhotenská toxikóza

stupně, macrosomia u předchozího dítěte, porucha glukózové tolerance, věk<25, glykosurie) mezi skupinou, která trpí gestačním diabetem a kontrolní skupinou.

Císařský řez, anestezie

Dalším častým tématem data miningu v porodnictví jsou císařské řezy. Císařský řez (lat. sectio caesarea) je porod chirurgickou cestou, během které je novorozenec vybaven z děložní dutiny otevřenou břišní stěnou. Obecně císařský řez dělíme na plánovaný a neplánovaný, i u neplánovaných indikací je nutný souhlas rodičky. V zahraničí je dokonce možné provést císařský řez na přání, např. v USA takto rodí 25% žen, aby se vyhnuly porodním bolestem. České zdravotnictví tuto možnost nenabízí, důvodem pro tento způsob porodu je pouze zdravotní komplikace např. nepoměr velikosti hlavičky plodu a pánve, překážky v porodních cestách, některé poruchy placenty a pupečníku, některá celková onemocnění matky, akutní tíseň plodu, nepravidelná uložení plodu, umírající a mrtvá matka. Porod císařským řezem je pro ženu šestkrát rizikovější než porod přirozený, nejčastějšími komplikacemi jsou krvácení, embolie, poškození močového měchýře nebo tenkého střeva, infekce a kýla v jizvě. Data miningu, týkajicího se císařského řezu, se věnuje zdroj [28], jenž na základě 5 atributů predikuje pomocí rozhodovacího stromu, kdy je nutno použít chirurgickou cestu porodu.

Statistické metody pro téma císařského řezu jsou použity ve zdroji [5] a to v souvislosti s použitím anestezie při tomto výkonu. Je zde uvedeno, že v roce 2011 bylo v Česku registrováno 107570 porodů, z toho 24 % jich bylo ukončeno císařským řezem. Celková anesteziologie byla podána u 47 % císařských řezů, v 53 % pak byla využita anestezie regionální (neuroaxiální). Z regionálních technik převažovala anestezie spinální (76 %) před anestezií epidurální (24 %). V článku je zastáván názor, že neuroaxiální metody mají mít v porodnictví přednost před celkovou anestezií. Tradičním zdůvodněním je vyšší riziko celkové anestezie pro adaptaci novorozence i bezpečnost matky. Vlivem anestezie na zdraví novorozence se zabývá zdroj [9], který opět používá test dobré shody a navíc ještě **t – testem** porovnává průměrnou porodní váhu novorozenců od matek, které během těhotenství prodělali operaci s anestezií a matek které pod anestezií nebyly.

Ve zdroji [30] autoři logistickou regresí prokázali, že epidurální analgezie protahuje porod, zvyšuje potřebu užití oxytocinu k augmentaci porodu a zvyšuje pravděpodobnost instrumentálního porodu. Dále článek také předkládá souvislost mezi denní dobou a porodem, konkrétně augmentací porodu (podání oxytocinu k urychlení a zesílení porodní činnosti), nástřihem hráze či instrumentálním porodem. Pro všechna měření byly zaznamenány vyšší hodnoty od 10 hodin dopoledne po 10 hodin večer ve srovnání s časem od 2 hodin ráno po 8 hodin ráno. Přesněji řečeno instrumentální porod je o 43 %, epiziotomie o 10 %, augmentace porodu o 86 % vyšší během dne/časných večerních hodin ve srovnání s nočním časem/časnými ranními hodinami.

Vliv kouření

Dalším zajímavým tématem v porodnictví je vliv kouření. Vlivem kouření ve spojení s věkem matky na porození mrtvého plodu se zabývají autoři článku [2], využívají Kaplan-Meierovu analýzu přežití – odhad intervalu mezi dvěma časovými událostmi. Ve zdroji [23] zjistili pomocí **Spearmanova korelačního koeficientu**, že kouření u matek během těhotenství se snižuje se rostoucím dosaženým vzděláním otce i matky, naopak konzumace alkoholu stoupá. Pomocí Spearmanova testu v článku [14] zjistili negativní korelaci mezi BMI před těhotenstvím a fyzickou aktivitou.

Vliv antibiotik

Velmi často se také zkoumá vliv antibiotik. Ve zdroji [8] se využívá McNemarův test pro posouzení, zda po léčbě penicilinem v průběhu těhotenství dochází k nějakým patologiím u novorozenců. Ukázalo se, že při užívání penicilinu je větší riziko rozštěpu rtu/patra a rektální/anální atrézie /stenózy.

Mimoděložní těhotenství

Jedním z častých témat je rovněž mimoděložní těhotenství. V článku [27] je **Wilcoxon rank sum test** požíván pro zjištění, že u žen, které již jednou prodělali mimoděložní těhotenství je větší pravděpodobnost ruptury. Tento test je použit i ve zdroji [19] pro porovnání délky a šířky děložního hrdla, věku a BMI u čtyř skupin žen

(netěhotné nulipary, netěhotné primipary/pluripary, těhotné nulipary, těhotné primipary/pluripary). Ukázalo se, že délka děložního hrdla je statisticky větší u těhotných žen než u netěhotných. Dále bylo zjištěno, že netěhotné nulipary jsou mladší, mají nižší BMI a menší délku i šířku děložního hrdla než netěhotné primipary/pluri pary. Dále byl v tomto článku proveden **Kruskal – Wallisův test,** stejně jako Wilcoxonův test slouží pro porovnání délky a šířky děložního hrdla, věku, BMI u čtyř skupin žen, avšak porovnává všechny skupiny najednou a neukazuje na žádné podstatné rozdíly mezi skupinami.

.

Praktická část

4 Realizace procesu dobývání znalostí z databází

Pro práci s daty byly využity počítačové programy pgAdmin III, Matlab, RStudio a Microsoft Excel 2010. V rámci této práce je postupováno podle metodiky CRISP-DM, která byla již popsána. První 4 kroky metodiky se nachází v této kapitole, pátý krok tvoří samostatnou kapitolu a poslední krok je zahrnut v závěru.

4.1 Porozumění problému

Cílem této práce je statisticky ověřit několik hypotéz na datech z porodnického oddělení Fakultní nemocnice Brno.

Statistickou analýzou výběrových dat získaných sledováním měřené náhodné veličiny jsme schopni rozhodnout o platnosti určitého obecného tvrzení (statistické hypotézy) na úrovni celé populace. Statistickou hypotézou rozumíme jakékoliv tvrzení, které se může týkat neznámých parametrů, daných funkcí parametrů, ale také tvaru rozdělení a dalších vlastností základního souboru.

Při testování statistických hypotéz vždy porovnáváme dvě hypotézy. První z nich, nulová hypotéza H_0 je hypotéza, kterou testujeme. Druhou hypotézou je alternativní hypotéza, kterou obvykle značíme H_A nebo H_1. Alternativní hypotéza přesně vymezuje, do jaké situace se dostáváme, když nulová hypotéza neplatí.

Druhým krokem při testování statistických hypotéz je určení hladiny významnosti testu α, což je pravděpodobnost, že se zamítne nulová hypotéza, ačkoliv platí. Pro medicínská data se využívá hladina 0,05 (příp. 0,01) a tím dostaneme 95% (99%) jistotu správného rozhodnutí.

Poté, co zformulujeme nulovou hypotézu a určíme hladinu významnosti, spočteme pravděpodobnost, s jakou bychom mohli obdržet pozorovaná data nebo data stejně, či ještě více odporující nulové hypotéze, za předpokladu, že je nulová hypotéza pravdivá. Tato pravděpodobnost se nazývá p-hodnota (p-value, p-level). P-hodnotu porovnáváme

s hladinou významnosti α, pokud $p \geq \alpha$, nezamítáme H_0. Čím menší je p - hodnota, tím méně důvěryhodná je nulová hypotéza. [33]

4.2 Porozumění datům

Úkolem této fáze je sběr dat a seznámení se s charakterem dat samotných. Sběr dat probíhal v rámci porodnického modulu NIS ve FN Brno od roku 2003 do roku 2014. Pro testování jsme měli k dispozici tabulku s názvem jpt_all2, která obsahovala 56835 záznamů o 191 příznacích. Atributy byly jak spojité veličiny (hmotnost novorozence, výška novorozence, věk matky, atd.) tak kategoriální (dichotomické – pohlaví novorozence, způsob porodu, aj.; vícekategoriální – rodinný stav, státní příslušnost, aj.).

Tabulka jpt_all2 shrnuje informace o novorozenci (výška, váha, pH, pohlaví, Apgar skóre), o matce (věk, rodinný stav, státní příslušnost, adresa, zaměstnání), o otci (datum narození), o porodu (datum, čas, místo, způsob, doba trvání jednotlivých porodních dob, poloha plodu při porodu, medikace, komplikace – pupečník kolem krku, preeklampsie, dystokie ramének, atd.), o placentě a plodové vodě (hmotnost, porucha placenty, infarkt placenty, zabarvení plodové vody, datum a čas odtoku), o předchozím těhotenství (datum, pohlaví, hmotnost a výška novorozence, komplikace)

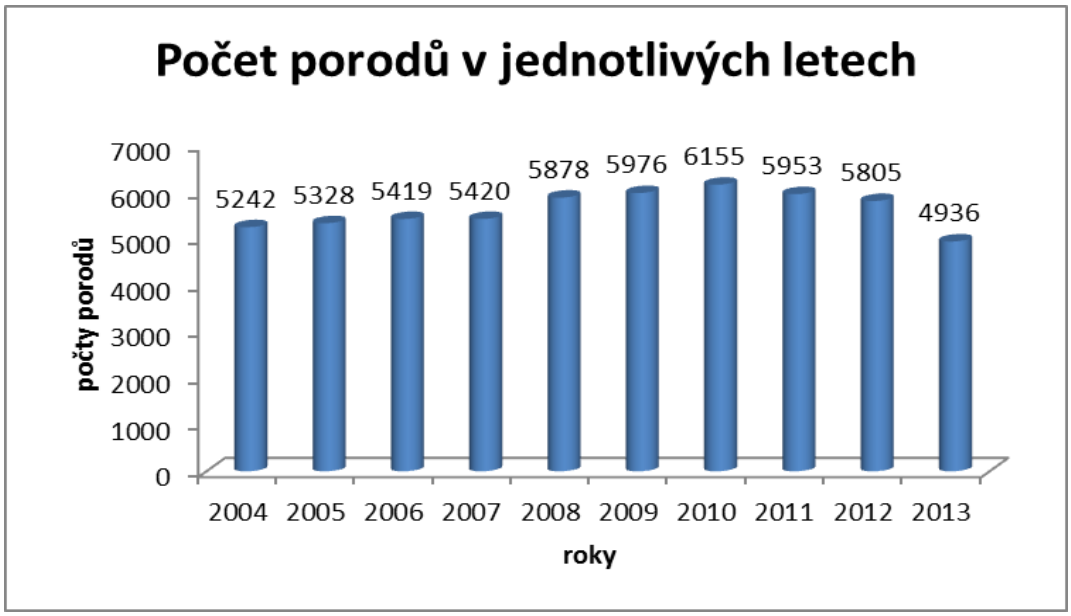
K databázi jsme přistupovali pomocí programu pgAdmin, v němž jsme SQL příkazy vybírali data pro jednotlivé testy. Data byla vyexportována do souboru csv a dále zpracovávána v programech Matlab, Excel, a RStudio.

Před samotným testováním hypotéz jsme provedli zobrazení některých deskriptivních charakteristik pomocí tabulek grafů v programu Excel.

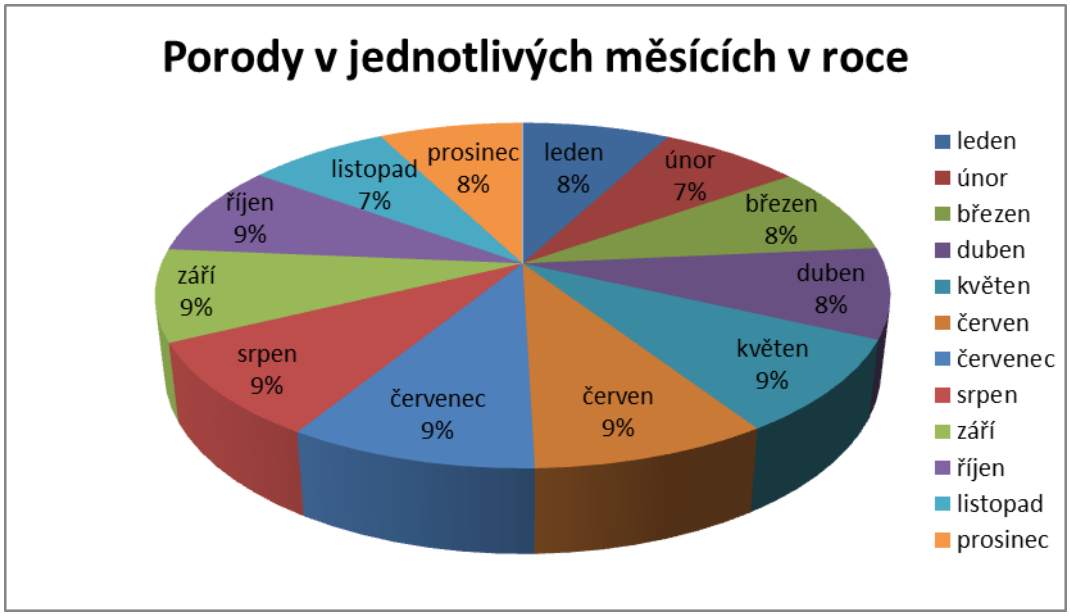
Tabulka 1 : popisná statistika porodů I

počet porodů v jednotlivých letech			počet porodů v jednotlivých měsících v roce			počet porodů v jednotlivých dnech v týdnu		
rok	četnost	relat. čet.	měsíc	četnost	relat. čet.	den	četnost	relat. čet.
2003	722	1 %	leden	4417	8 %	pondělí	8141	14 %
2004	5242	9 %	únor	4216	7 %	úterý	8978	16 %
2005	5328	9 %	březen	4727	8 %	středa	8872	16 %
2006	5419	10 %	duben	4793	8 %	čtvrtek	9015	16 %
2007	5420	10 %	květen	5012	9 %	pátek	9055	16 %
2008	5878	10 %	červen	5002	9 %	sobota	6701	12 %
2009	5976	11 %	červenec	5250	9 %	neděle	6072	11 %
2010	6155	11 %	srpen	5107	9 %			
2011	5953	10 %	září	4888	9 %			
2012	5805	10 %	říjen	4885	9 %			
2013	4936	9 %	listopad	4221	7 %			
			prosinec	4316	8 %			
celkově	56834	100 %	celkem	56834	100 %	celkem	56834	100 %

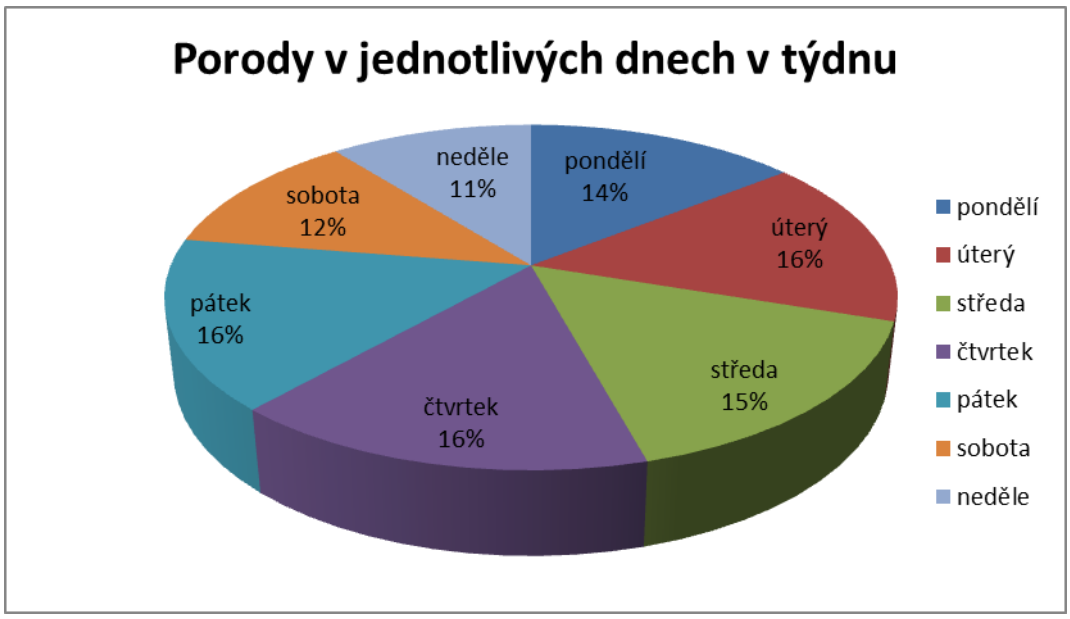
Z popisné statistiky porodů (Tabulka 1) jsme zjistili, že nejvíce porodů bylo v roce 2009 a 2010. Sběr dat byl zahájen na podzim roku 2003, tudíž je v tomto roce registrován menší počet porodů a proto tento rok do další statistiky a grafů nezařazujeme. Vývoj počtu porodů (Obrázek 3) v jednotlivých letech má rostoucí charakter až do roku 2010, poté má klesající trend. Počty porodů v jednotlivých měsících jsou téměř vyrovnané, nejvíce dětí se rodí od května do října (Obrázek 4). Popisná statistika porodů v jednotlivých dnech v týdnu ukázala, že nejméně dětí se rodí o víkendu (Obrázek 5).



Obrázek 3: Graf - počet porodů v jednotlivých letech

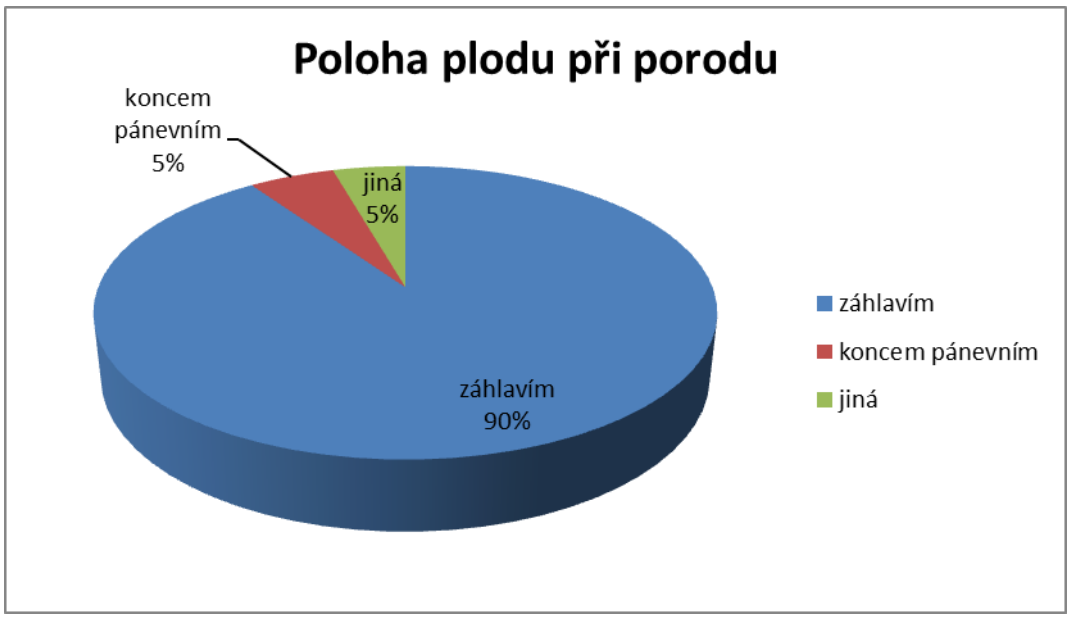


Obrázek 4: Graf - porody v jednotlivých měsících v roce



Obrázek 5: Graf - porody v jednotlivých dnech v týdnu

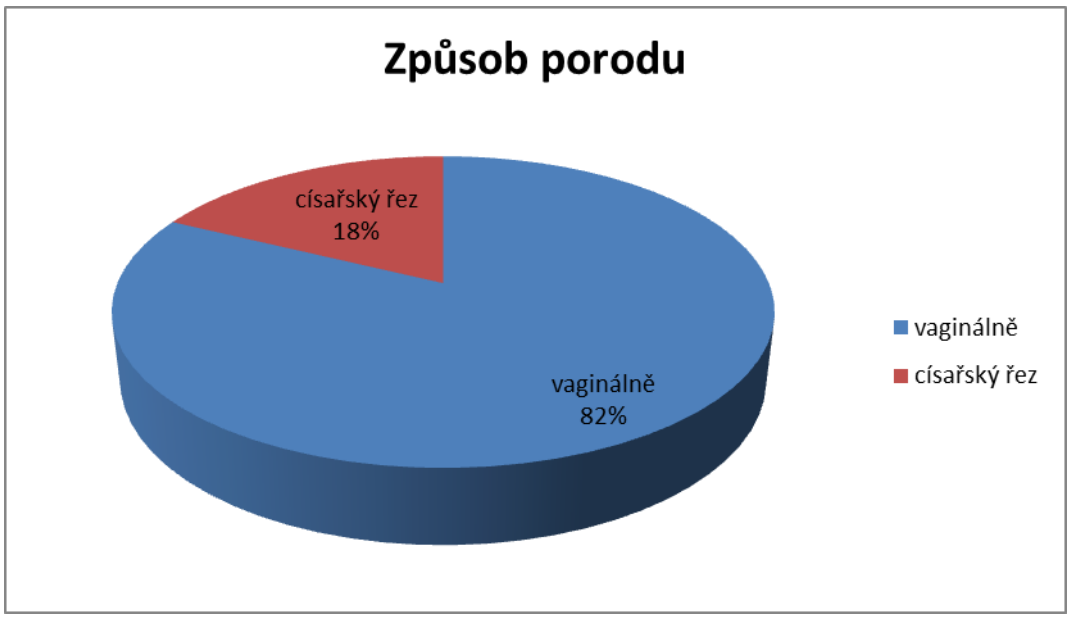
Z popisné statistiky porodů (Tabulka 2) také vyplývá, že císařským řezem se narodilo 18 % dětí (Obrázek 7), nejčastější poloha při porodu byla záhlavím (Obrázek 6) a předčasných porodů bylo zaznamenáno 7 % (Obrázek 9). Dále jsme vykreslili grafy pro vývoj počtu předčasných porodů a císařských řezů od roku 2004 do roku 2013. Do grafů jsme vynášeli četnosti relativní vzhledem k celkovému počtu porodů v souboru. U císařských řezů (Obrázek 8) pozorujeme do roku 2010 rostoucí trend, po roce 2010 naopak počty císařských řezů mírně klesají. Vývoj předčasných porodů (Obrázek 10) nevykazuje žádný trend, četnosti porodů před 34. týdnem se pohybují okolo 3 % a porody mezi 34. a 37. týdnem okolo 4 - 5 %.



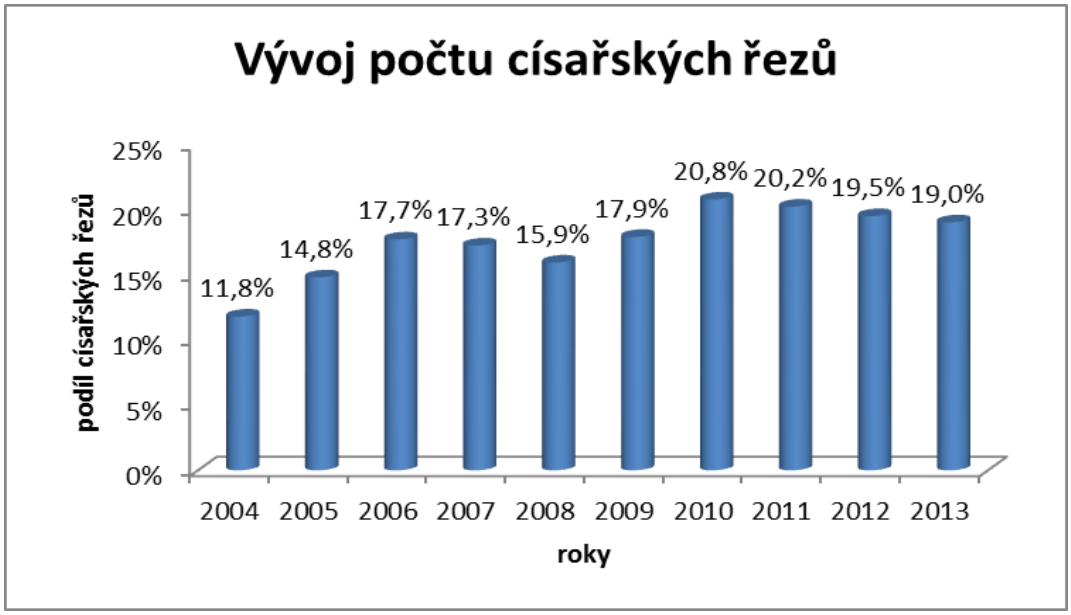
Obrázek 6: Graf - poloha plodu při porodu

Tabulka 2 : Popisná statistika porodů II

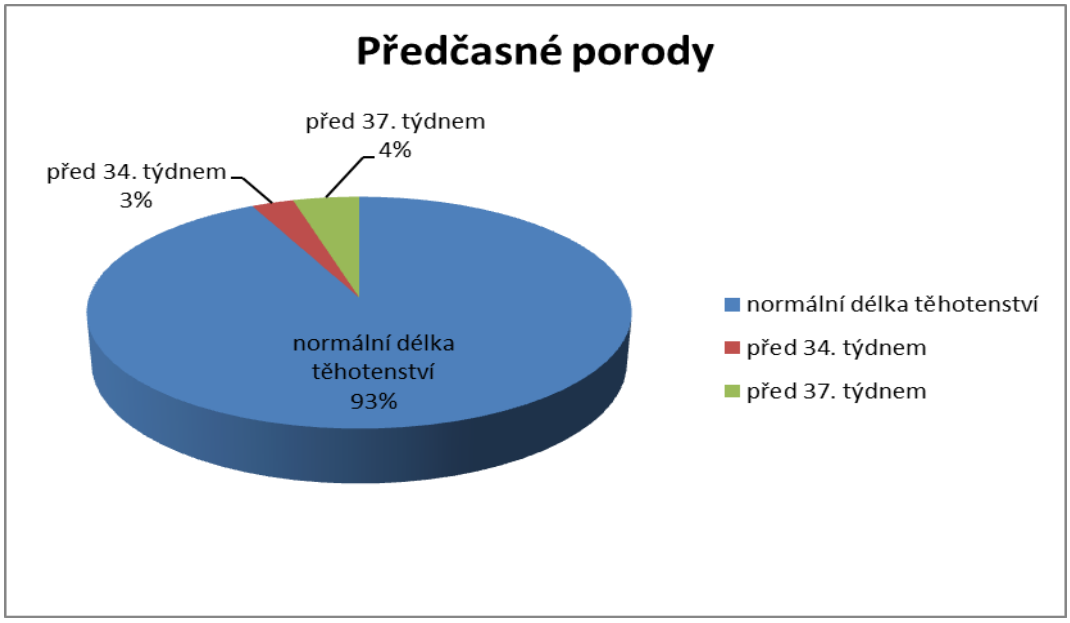
atributy		četnost	relat. čet.
způsob porodu	vaginálně	46721	82 %
	císařský řez	10096	18 %
	celkem	56817	100 %
poloha plodu při porodu	záhlavím	50622	90 %
	koncem pánevním	2946	5 %
	jiná	2497	4 %
	celkem	56065	100 %
předčasné porody	před 34. týdnem	1612	3 %
	před 37. týdnem	2524	4 %
	po 37. týdnu	52588	93 %
	celkem	56724	100 %



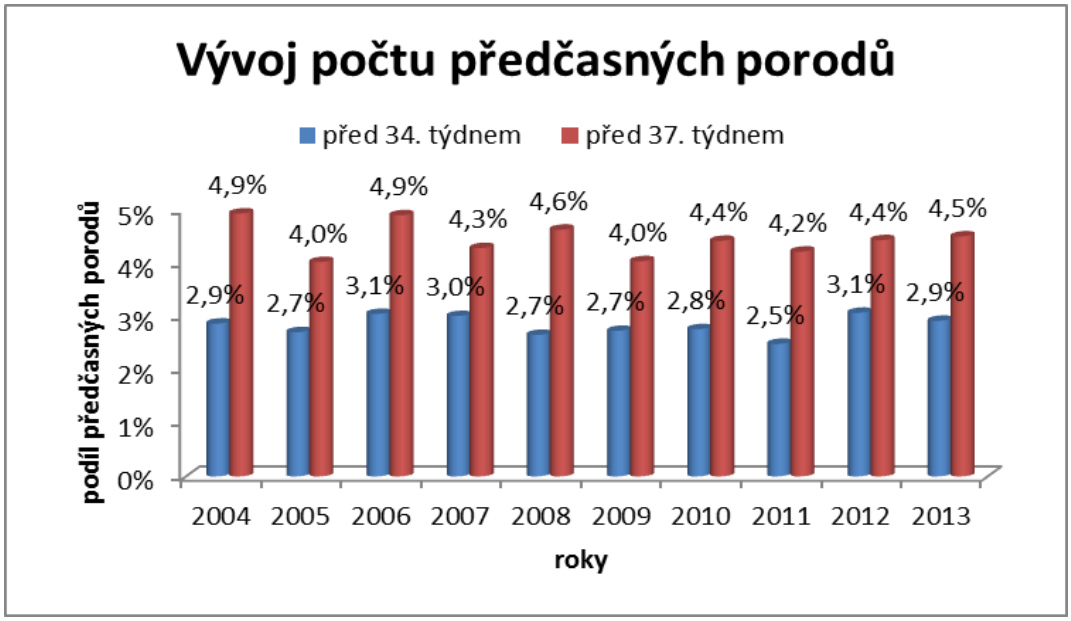
Obrázek 7: Graf - způsob porodu



Obrázek 8: Graf – vývoj počtu císařských řezů, vyjádřen v procentech z celkového počtu porodů



Obrázek 9: Graf - předčasné porody



Obrázek 10: Graf – vývoj předčasných porodů, vyjádřen v procentech z celkového počtu porodů

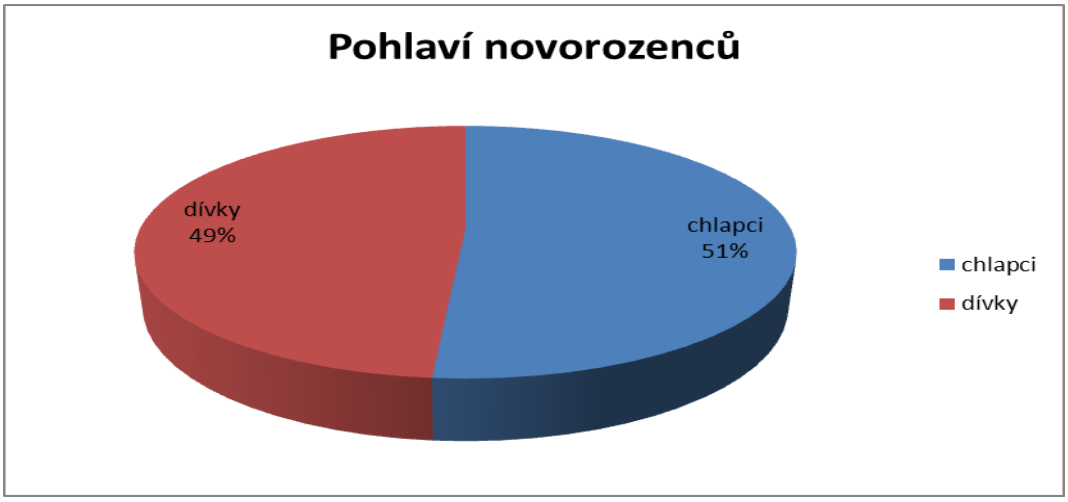
Při popisné statistice novorozenců (Tabulka 3, Tabulka 4) jsme zjistili, že se narodilo více chlapců než děvčat (Obrázek 11), dále jsme se zaměřili na jejich vitalitu, hmotnost, výšku a pH, kde jsme vypočítali medián, minimální a maximální odchylku.

Tabulka 3 : Popisná statistika novorozenců I

atributy		četnost	relativní četnost
pohlaví	chlapci	29154	51 %
	dívky	27641	49 %
	celkem	56795	100 %
vitalita	živé	56748	100 %
	mrtvé	9	0 %
	celkem	56757	100 %

Tabulka 4: Popisná statistika novorozenců II

atributy	četnost	medián	minimum	maximum
hmotnost (g)	56816	3350	300	6060
výška (cm)	56410	50	5	59
H	24738	7,27	6,50	7,61



Obrázek 11: Graf - pohlaví novorozenců

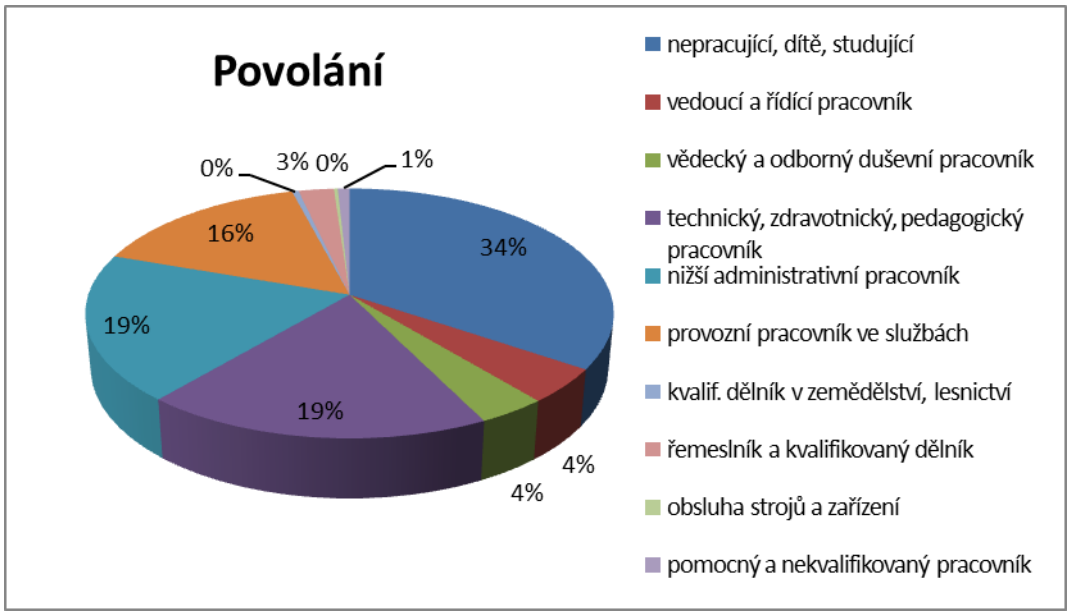
Při popisné statistice matek jsme se zaměřili na jejich věk (medián = 30, min = 13, max = 52), rodinný stav a státní příslušnost (Tabulka 5) a také na zaměstnání (Tabulka 6, Obrázek 12). Matky byly převážně češky (95%), vdané (69 %) a nejčastěji nezaměstnané (34 %).

Tabulka 5 : Popisná statistika matek I

rodinný stav	četnost	relat.čet.	státní příslušnost	četnost	relat.čet.
svobodná	14612	26 %	CZ	54209	95 %
vdaná	39415	69 %	SK	1092	2 %
rozvedená	2658	5 %	UA	423	1 %
ovdovělá	102	0 %	VN	420	1 %
družka	20	0 %	ostatní východ	614	1 %
nezjištěno	27	0 %	ostatní západ	76	0 %
celkem	56834	100 %	celkem	56834	100 %

Tabulka 6 : Popisná statistika matek II

zaměstnání	počet	procenta	zaměstnání	počet	procenta
nepracující, dítě, studující	19589	34 %	provozní pracovník ve službách	8859	16 %
vedoucí a řídící pracovník	2557	4 %	kvalif. dělník v zemědělství, lesnictví	258	0 %
vědecký a odborný duševní pracovník	2085	4 %	řemeslník a kvalifikovaný dělník	1424	3%
technický, zdravotnický, pedagogický pracovník	10578	19 %	obsluha strojů a zařízení	156	0 %
nižší administrativní pracovník	10858	19 %	pomocný a nekvalifikovaný pracovník	471	1 %



Obrázek 12: Graf - povolání matek

4.3 Příprava dat

V této fázi je úkolem upravit data tak, aby je bylo možno dále zpracovávat pomocí statistických testů. Příprava zahrnuje selekci příznaků, čistění dat a převod typů dat.

Při selekci atributů vybíráme pouze ty, které zahrneme do testování. Po selekci nám tedy z původních 191 příznaků zbylo pouze 88, odstranili jsme příznaky, které obsahovali identifikační čísla, souhrné poznámky, data co se opakovali a citlivé údaje (telefon, jméno kontaktní osoby atd.).

Čistění dat zahrnuje práci s chybějícími a odlehlými hodnotami. Vzhledem k velkému počtu dat jsme chybějící hodnoty nenahrazovali průměrem ani mediánem a rozhodli jsme se je z datového souboru odstranit. Odlehlé hodnoty jsme v souboru ponechali v souladu se zdrojem [10], který uvádí pravidlo, jenž říká, že ze sady dat se nikdy nemá vyloučit nějaká hodnota pouze ze statistických důvodů.

S ohledem na zvolené statistické metody bylo nutné, pro další práci některé atributy upravit na dichotomické (rozumíme data, která nabývají hodnot z dvouprvkové množiny (0/1; true/false)). Také časové údaje, uváděné v tabulce, ve formátu hh:mm:ss jsme převedli na čísla v minutách. [15, 21]

4.4 Modelování

Ve fázi modelování dochází k výběru DM technik. Tato práce se zaměřuje na statistické metody. Teoretické pozadí použitých metod bylo vysvětleno v kapitole 3. K testování hypotéz jsou využívány následující testy:

4.4.1 Jednovýběrový Kolmogorovův-Smirnovův test

Normalitu dat pomocí tohoto testu jsme ověřovali v programu Matlab funkcí h = kstest(x), která vrací výsledek h = 1, pokud nulovou hypotézu zamítáme na 5% hladině významnosti (data nepocházejí z normálního rozdělení) nebo h = 0, pokud nulovou hypotézu nezamítáme.

4.4.2 Wilcoxonův dvouvýběrový rank sum test

Výpočet tohoto testu jsme prováděli rovněž v programu Matlab, funkcí [P,H] = ranksum(x,y). P udává p-hodnotu testu a H nabývá hodnot 0, když nulovou hypotézu nemůžeme zamítnout (mediány obou výběrů se rovnají) nebo 1, když nulovou hypotézu zamítáme na hladině významnosti 5 %.

4.4.3 Kruskal – Wallisův test

Kruskal – Walisův test jsme počítali v programu Matlab pomocí funkce P = kruskalwallis(X), jejímž výsledkem je tabulka ANOVA obsahující stupně volnosti, hodnotu $X^2_{k-1}(\alpha)$, p-hodnotu testu a další hodnoty. Výsledkem této funkce je také krabicový graf zachycující mediány (červeně) a horní a dolní kvartily jednotlivých výběrů. Vousy rozšiřují graf o extrémní datové body, které ještě nejsou považovány za odlehlé hodnoty. Odlehlé hodnoty (outliers) jsou vykresleny na grafu červenými křížky.

4.4.4 Wilcoxonův párový test

Tento test je v programu Matlab dán funkcí [P,H] = signrank(x,y), která dává stejné hodnoty jako ranksum test, tedy p-hodnotu a H = 0, když nemůžeme zamítnout nulovou hypotézu na 5% hladině významnosti.

4.4.5 Spearmanův test nezávislosti

Spearmanův test nezávislosti byl počítán také v programu Matlab pomocí funkce [RHO,PVAL] = corr(x,y,'type','Spearman'), kde RHO udává hodnotu Spearmanova korelačního koeficientu a PVAL p-hodnotu, která jestliže je menší než 0,05 pak je výsledek statisticky významný od nuly.

4.4.6 Test dobré shody

Tento test jsme počítali v programu Excel 2010 pomocí funkce CHISQ.TEST(aktuální, očekávané). Jejímž výsledkem je p-hodnota testu.

4.4.7 Logistická regrese

Logistickou regresi jsme počítali v programu RStudio. Použili jsme funkci m = glm(formula, family = binomial(link="logit")). Z výsledných koeficientů jsme vypočítali poměr šancí exp(coef(m)), a intervaly spolehlivosti confint(m), tyto údaje jsme pak využili do funkce pro sestrojení grafu forest plot v programu Matlab.

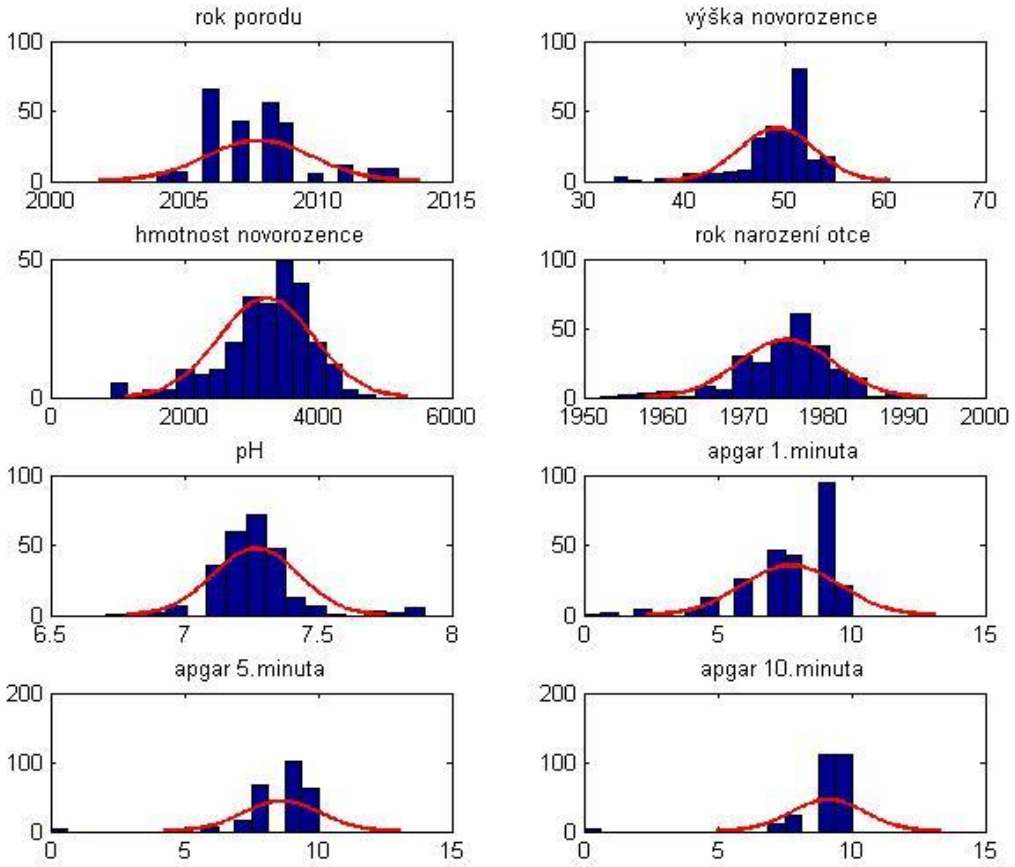
Poměr šancí je v grafu vyznačen čtverečkem, který je protnut horizontální čarou představující 95% interval spolehlivosti. Do grafu se také vynáší svislá linie protínající osu v bodě, který představuje nulový efekt, v případě poměru šancí je to bod 1. Pokud se některé intervaly spolehlivosti protínají s touto svislou linií, ukazuje to, že při dané hladině významnosti je vliv rizika nulový.

5 Vyhodnocení výsledků

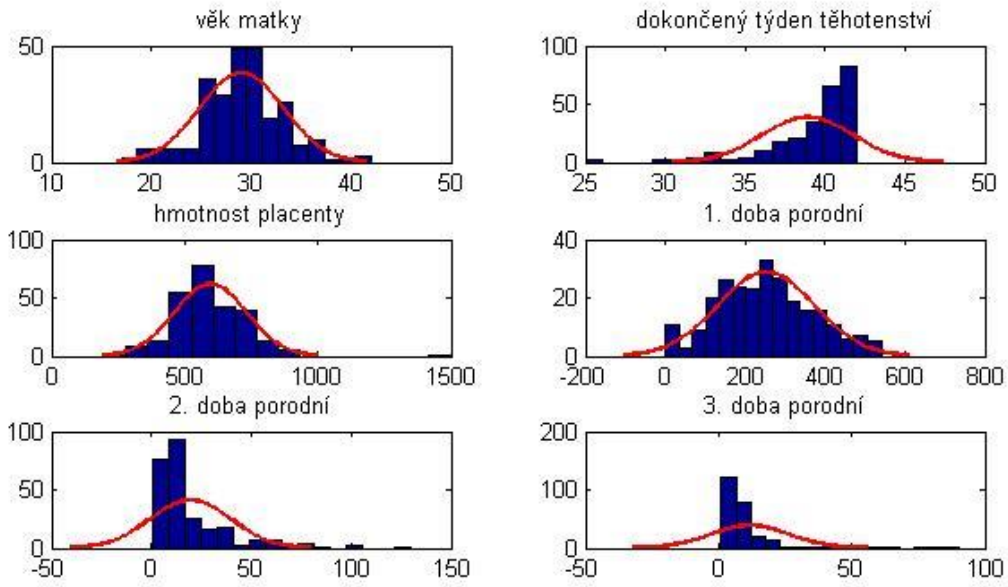
Všechny statistické testy byly počítány na hladině významnosti 0,05.

5.1 Výsledky Kolmogorovova-Smirnovova test

Prvním testem, který jsme počítali, byl jednovýběrový kolmogorovův-Smirnovovův test pro ověření normality dat. Protože většina příznaků v tabulce jpt_all2 jsou kvalitativní (dichotomická a nominální) data, počítali jsme tento test pouze u 14 kvantitativních příznaků. Pro názornost jsme sestrojili histogramy (Obrázek 13, Obrázek 14) proložené Gausovou křivkou normálnho rozdělení.



Obrázek 13: Histogramy jednotlivých příznaků proložené křivkou normálního rozložení I



Obrázek 14: Histogramy jednotlivých příznaků proložené křivkou normálního rozložení II

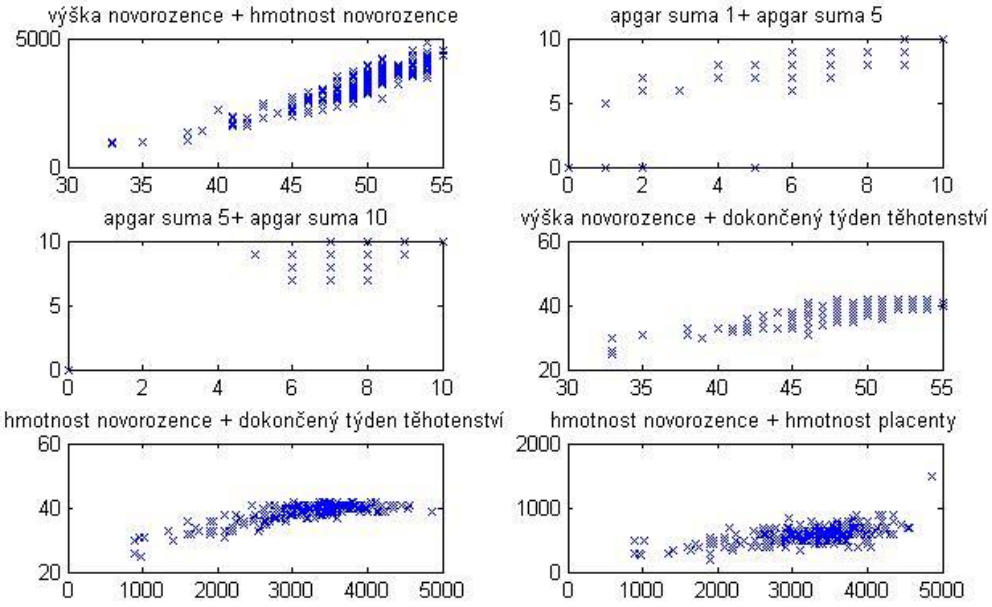
U všech testovaných příznaků nám vyšlo, že nepocházejí z normálního rozdělení. Proto jsme pro další testování volili neparametrické testy, které nevyžadují normální rozdělení. Pro testování kvantitativních dat dvou závislých výběrů jsme použili Wilcoxonův párový test, pro dva nezávislé výběry Wilcoxonův rank sum test, pro tři nezávislé výběry Kruskal-Walis test. Spearmanovým testem nezávislosti jsme hodnotili korelaci mezi kvantitativními znaky. Pro data kvalitativní jsme využili testu dobré shody. Logistickou regresí jsme testovali jak data kvalitativní tak i kvantitativní.

5.2 Výsledky Spearmanova testu nezávislosti

Spearmanovým testem jsme zjišťovali, zda existuje závislost mezi jednotlivými kvantitativními příznaky. Testovali jsme každý příznak s každým. Jako významné jsme vyhodnotili výsledky, které měli absolutní hodnotu Spearmenova koeficientu větší než 0,5, tyto výsledky jsou uvedeny v Tabulka 7. Opět jsou data pro názornost vykreslena v Obrázek 15.

Tabulka 7: Významné závislosti mezi příznaky: Spearmanův test

korelované atributy		Spearmanův koeficient
hmotnost novorozence	výška novorozence	0,88
apgar suma 1	apgar suma 5	0,83
apgar suma 5	apgar suma 10	0,77
délka těhotenství	výška novorozence	0,60
délka těhotenství	hmotnost novorozence	0,59
hmotnost novorozence	hmotnost placenty	0,53



Obrázek 15: Korelovaná data

5.3 Výsledky Wilcoxonova rank sum testu

Tímto testem jsme porovnávali dva nezávislé soubory. Zjišťovali, rozdíly mezi novorozenci s dystokií ramének a bez ní, dále rozdíly mezi pohlavími novorozenců a nakonec jsme zkoumali rozdíly mezi jednotlivými porodními dobami.

Tabulka 8 shrnuje výsledky rank sum testu pro dystokii ramének plodu, kde jsme zjišťovali, zda má věk matky a délka těhotenství vliv na dystokii ramének a také jsme testovali, jestli dystokie ovlivňuje pH novorozence. Nulové hypotézy zněly:

- Není rozdíl mezi věkem matky u porodů s dystokií a bez dystokie.
- Není rozdíl mezi délkou těhotenství u porodů s dystokií a bez dystokie.
- Není rozdíl mezi pH novorozence u porodů s dystokií a bez dystokie.

Porodů s dystokií ramének bylo 66, bez dystokie 24584. Nezamítáme pouze hypotézu, jež říká, že není rozdíl mezi věkem matky u porodů s dystokií a bez dystokie.

Tabulka 8: Dystokie ramének plodu: Wilcoxon rank sum test - u příznaků, jejichž p-hodnota je <0,05, jsou rozdíly signifikantní.

příznaky	medián		p-hodnota
	bez dystokie, n = 24584	s dystokií, n= 66	
věk matky	30	30,5	0,213
dokončený týden těhotenství	40	40	< 0,05
pH novorozence	7,28	7,24	< 0,05

Tabulka 9 obsahuje výsledky rank sum testu pro pohlaví, kde jsme zkoumali 6 hypotéz:

- Není rozdíl mezi výškou chlapců a dívek.
- Není rozdíl mezi hmotností chlapců a dívek.
- Není rozdíl mezi pH u chlapců a dívek.
- Není rozdíl mezi věkem matky u chlapců a u dívek
- Není rozdíl mezi délkou těhotenství u matek chlapců a u matek dívek.
- Není rozdíl mezi hmotností placenty u chlapců a dívek.

Pro testování jsme měli k dispozici 11908 záznamů mužského pohlaví a 10505 ženského pohlaví. Pouze hypotézu, která tvrdí, že není rozdíl v délce těhotenství, nemůžeme zamítnout.

Tabulka 9: Pohlaví: Wilcoxon rank sum test - u příznaků, jejichž p-hodnota je <0,05, jsou rozdíly signifikantní.

atributy	medián		p-hodnota
	chlapci, n = 11905	dívky, n= 10501	
porodní výška [cm]	50	49	< 0,05
porodní hmotnost [g]	3400	3260	< 0,05
pH novorozence	7,28	7,29	< 0,05
věk matky	30	30	< 0,05
dokončený týden těhotenství	40	40	0,058
hmotnost placenty [g]	600	600	< 0,05

Dále jsme zkoumali, zda na délky porodních dob má vliv medikace, analegetika, spazmolytika, antibiotika, epidural, oxytocin, preeklampsie, diabetes, indukce. Nulové hypotézy zněly:

- Není rozdíl mezi délkami porodních dob u porodů s medikací a bez medikace.
- Není rozdíl mezi délkami porodních dob u porodů analgetiky a bez nich.
- Není rozdíl mezi délkami porodních dob u porodů se spazmolytiky a bez nich.
- Není rozdíl mezi délkami porodních dob u porodů s antibiotiky a bez nich.
- Není rozdíl mezi délkami porodních dob u porodů s epiduralem a bez něj.
- Není rozdíl mezi délkami porodních dob u porodů s oxytocinem a bez něj.
- Není rozdíl mezi délkami porodních dob u porodů s preeklampsií a bez ní.
- Není rozdíl mezi délkami porodních dob u porodů s diabetem a bez něj.
- Není rozdíl mezi délkami porodních dob u indukovaných porodů a přirozených porodů.

Výsledky shrnuje Tabulka 10, ve které vidíme, že téměř všechny zmíněné faktory ovlivňují porodní doby. Nulové hypotézy nemůžeme zamítnout pouze u vlivu analgetik a indukovaného porodu na 3. porodní dobu.

Tabulka 10: Porodní doby: Wilcoxon rank sum test - u příznaků, jejichž p-hodnota je <0,05, jsou rozdíly signifikantní.

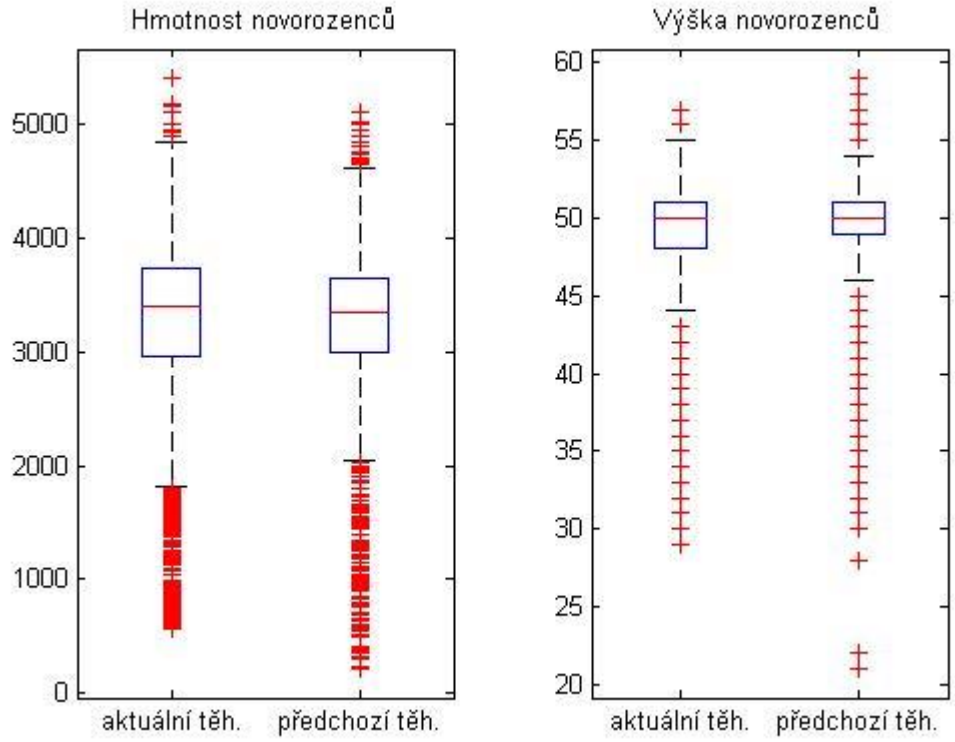
atributy	četnost	1. doba porodní		2. doba porodní		3. doba porodní	
		medián [min]	p -hodnota	medián [min]	p -hodnota	medián [min]	p -hodnota
s medikací	48151	190	< 0,05	5	< 0,05	10	< 0,05
bez medikace	1311	230		10		5	
s analgetiky	6502	305	< 0,05	10	< 0,05	10	0,298
bez analgetik	42960	215		10		5	
se spazmolytiky	16120	280	< 0,05	10	< 0,05	10	< 0,05
bez spazmolytik	33342	205		5		5	
s antibiotiky	8042	255	< 0,05	10	< 0,05	5	< 0,05
bez antibiotik	41420	225		10		10	
s epiduralem	10643	300	< 0,05	10	< 0,05	5	< 0,05
bez epiduralu	38819	210		5		5	
s oxytocinem	15955	235	< 0,05	10	< 0,05	10	< 0,05
bez oxytocinu	38566	220		5		5	
s preeklampsií	18	99,5	< 0,05	5	< 0,05	2	< 0,05
bez preeklampsie	49389	230		10		10	
s diabetem	175	145	< 0,05	5	< 0,05	3	< 0,05
bez diabetu	49408	230		10		10	
indukovaný porod	13647	200	< 0,05	10	< 0,05	5	0,799
neindukovaný porod	35815	240		10		5	

5.4 Výsledky Wilcoxonova párového testu

Párovým testem jsme porovnávali, zda exisutuje rozdíl mezi aktuálním a předchozím těhotenstvím v hmotnosti a výšce novorozence, test byl prováděn na 6866 záznamech. Před testováním jsme si data ještě vykreslili pomocí krabicových grafů (Obrázek 16). Poté jsme testovali tyto nulové hypotézy:

- Není rozdíl mezi hmotností novorozence u aktuálního a předchozího porodu.
- Není rozdíl mezi výškou novorozence u aktuálního a předchozího porodu.

Výsledky shrnuje Tabulka 11. Obě nulové hypotézy zamítáme.



Obrázek 16: Krabicové grafy k Wilcoxonovu párovému testu – statisticky významný rozdíl je jak u výšky tak i u váhy novorozenců

Tabulka 11: Předchozí a aktuální těhotenství: Wilcoxonův parový test- u příznaků, jejichž p-hodnota je <0,05, jsou rozdíly signifikantní.

atributy	medián (25% kvartil, 75% kvartil)		p-hodnota
	předchozí těhotenství	aktuální těhotenství	
hmotnost novorozence [g]	3340 (2960, 3730)	3400 (3000, 3650)	< 0,05
výška novorozence [cm]	50 (48, 51)	50 (49, 51)	< 0,05

5.5 Výsledky Kruskal – Walis testu

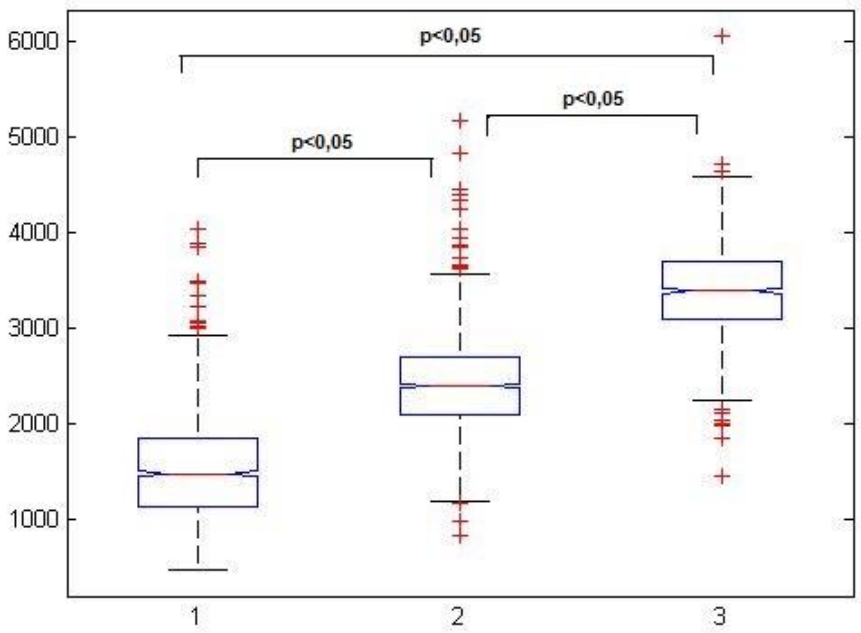
Pomocí testu Kruskal – Walis jsme zjišťovali, zda má délka těhotenství vliv na hmotnost, výšku a pH novorozence, hmotnost placenty a zda věk matky ovlivňuje délku těhotenství. Délku těhotenství jsme rozdělili do třech kategorií – předčasný porod před 34. týdnem těhotenství (1273 záznamů), předčasný porod před 37. týdnem těhotenství (1546 záznamů) a porod po 37. týdnu těhotenství (21034 záznamů). Testovali jsme tyto nulové hypotézy:

- Není rozdíl mezi hmotností novorozenců u předčasných a normálních porodů.
- Není rozdíl mezi výškou novorozenců u předčasných a normálních porodů.
- Není rozdíl mezi věkem matky u předčasných a normálních porodů.
- Není rozdíl mezi pH novorozenců u předčasných a normálních porodů.
- Není rozdíl mezi hmotností placenty u předčasných a normálních porodů.

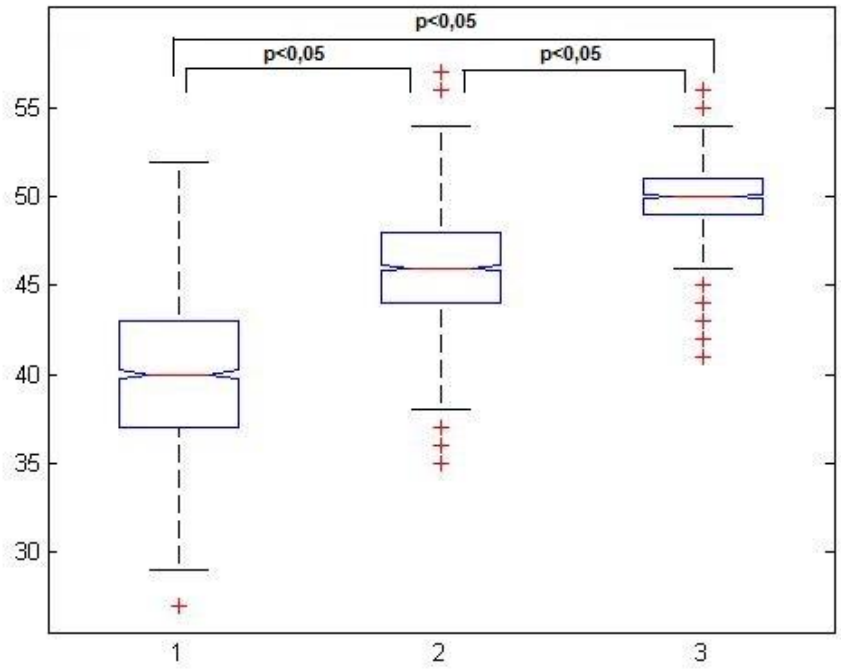
Výsledky shrnuje Tabulka 12. Nezamítáme pouze hypotézu o hmotnosti placenty. Pro ostatní hypotézy jsme dále zjišťovali, které z tří výběrů jsou statisticky odlišné, tyto výsledky obsahuje Tabulka 12. Statisticky významná odlišnost není mezi věkem matky a pH novorozence u obou předčasných porodů. Jednotlivá data byla pro znázornění vykreslena do krabicových grafů (Obrázek 17, Obrázek 18, Obrázek 19, Obrázek 20, Obrázek 21).

Tabulka 12: Předčasné porody: Kruskal - Walis test – u příznaků, jejichž p-hodnota je <0,05, jsou rozdíly signifikantní.

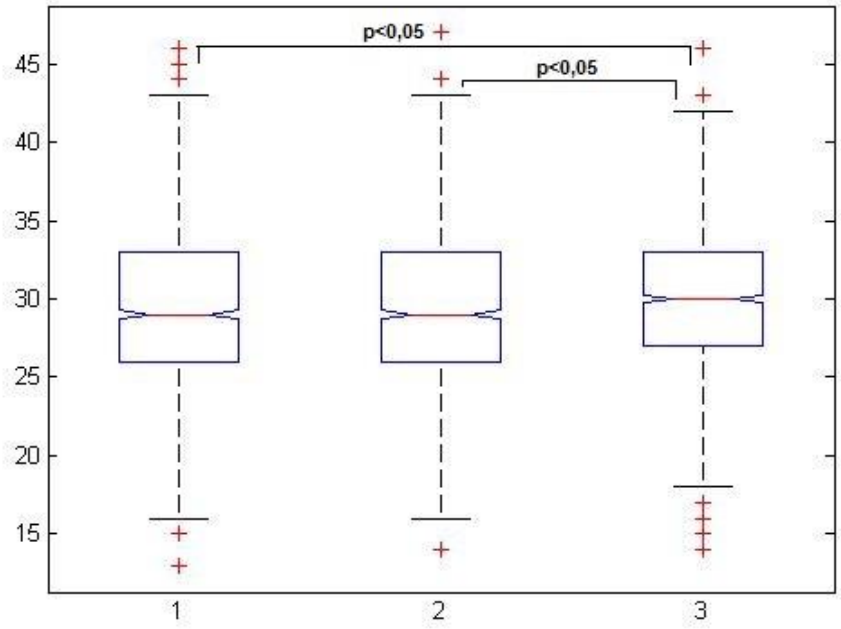
atributy	medián			p-hodnota
	před 34. týdnem	před 37. týdnem	po 37. týdnu	
hmotnost novorozence [g]	1700	2650	3420	< 0,05
výška novorozence [cm]	42	47	50	< 0,05
věk matky	30	29	30	< 0,05
pH novorozence	7,3	7,29	7,28	< 0,05
hmotnost placenty [g]	400	500	600	0,109



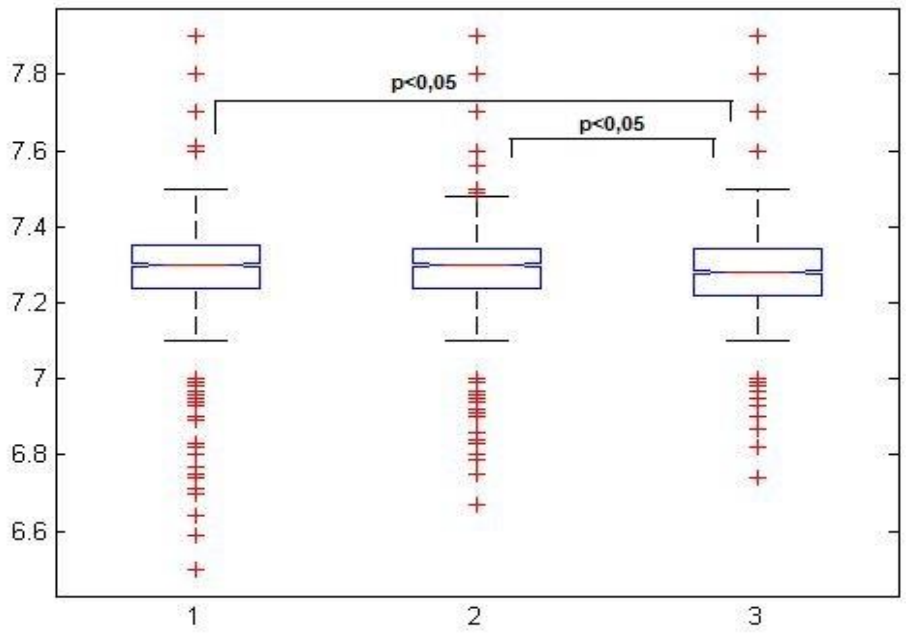
Obrázek 17: Boxplot hmotnost novorozence: Kruskal – Walis - statisticky významný rozdíl je mezi všemi třemi výběry



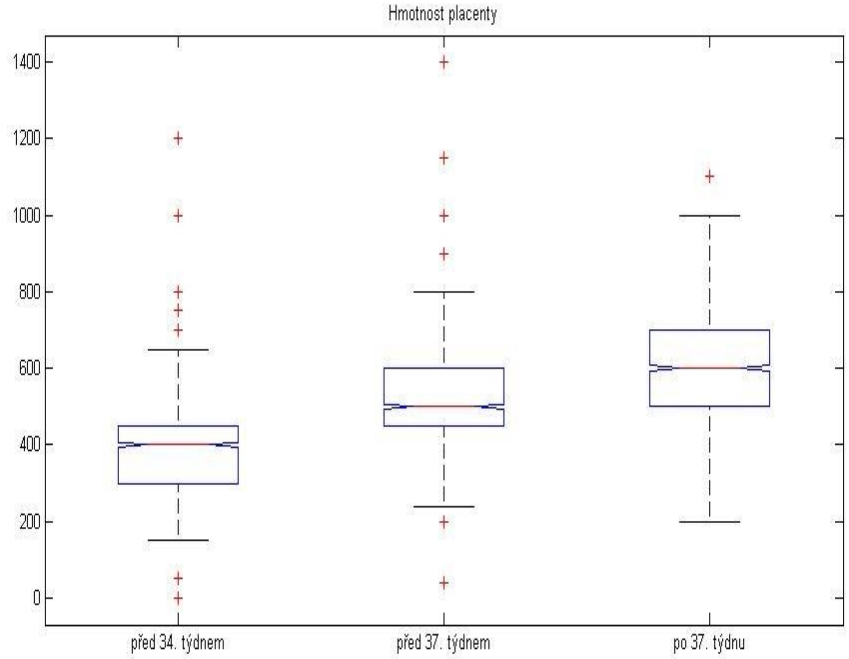
Obrázek 18: : Boxplot výška novorozence: Kruskal – Walis - statisticky významný rozdíl je mezi všemi třemi výběry



Obrázek 19: Boxplot věk matky: Kruskal – Walis - statisticky významný rozdíl je mezi porodem po 37. Týdnu a oběma předčasnými porody



Obrázek 20: Boxplot pH novorozence: Kruskal – Walis - statisticky významný rozdíl je mezi porodem po 37. Týdnu a oběma předčasnými porody



Obrázek 21: Boxplot hmotnost placenty: Kruskal – Walis – mezi výběry není statisticky významný rozdíl

5.6 Výsledky testu dobré shody

Testem dobré shody jsme testovali rizikové faktory pro dystokii ramének (Tabulka 13), předčasný porod (Tabulka 14), císařský řez (Tabulka 15), deceleraci (Tabulka 16) a hypoxii (Tabulka 17).

Nulová hypotéza pro dystokii ramének:

- Podíl dystokie ramének je stejný u obou pohlaví.

Jelikož p hodnota testu vyšla menší než 0,05, zamítáme nulovou hypotézu a můžeme tedy říct, že pohlaví má vliv na výskyt dystokie ramének.

Tabulka 13: Dystokie: Chí test - u příznaků, jejichž p-hodnota je <0,05, jsou rozdíly signifikantní.

příznaky		počet		p-hodnota
		dystokie	bez dystokie	
pohlaví	chlapci	49	28975	< 0,05
	dívky	29	27496	

Nulové hypotézy pro předčasné porody:

- Podíl předčasných porodů je stejný u matek se SAG (streptokok skupiny B) a bez SAG.
- Podíl předčasných porodů je stejný u žen různých národností.
- Podíl předčasných porodů je stejný u žen s různým rodinným stavem.

Nulové hypotézy o SAG a státní příslušnosti nezamítáme. Předčasný porod je ovlivněn rodinným stavem.

Tabulka 14: Předčasný porod: Chí test - u příznaků, jejichž p-hodnota je <0,05, jsou rozdíly signifikantní.

příznaky		počet		p-hodnota
		předčasný porod	normální porod	
SAG	ano	724	9104	0,051
	ne	3208	37106	
státní příslušnost	CZ	3765	43960	0,393
	SK	77	927	
	UA	22	356	
	VN	29	354	
	ostatní východ	35	537	
	ostatní západ	4	67	
rodinný stav	svobodná	1256	12071	< 0,05
	vdaná	2430	31874	
	rozvedená	237	2149	
	vdova	7	81	

Nulové hypotézy pro císařský řez:

- Podíl císařských řezů je stejný u matek se SAG a bez SAG.
- Podíl císařských řezů je stejný porodů s abnormální rotací plodu a bez ní.
- Podíl císařských řezů je stejný u novorozenců s pupečníkem kolem krku a bez něj.
- Podíl císařských řezů je stejný u porodu s nepoměrem pánve ženy a plodu a bez něj.

Císařský řez je ovlivněn všemi tetovanými prediktory - SAG, abnormální rotací, pupečníkem kolem krku a nepoměrem pánve ženy a plodu.

Tabulka 15: Císařský řez: Chí test - u příznaků, jejichž p-hodnota je <0,05, jsou rozdíly signifikantní.

příznaky		počet		p-hodnota
		císařský řez	vaginální porod	
SAG	ano	1169	8660	< 0,05
	ne	6300	33979	
abnormální rotace	ano	0	444	< 0,05
	ne	7831	48842	
pupečník kolem krku	ano	57	689	< 0,05
	ne	7774	48153	
nepoměr rozměrů pánve ženy a plodu	ano	290	91	< 0,05
	ne	7541	48751	

Nulové hypotézy pro deceleraci:

- Podíl decelerace je stejný u novorozenců s hypoxií a bez ní.
- Podíl decelerace je stejný u novorozenců s pupečníkem kolem krku a bez něj.

Oba příznaky vyšly signifikantní, výskyt decelerace je ovlivněn hypoxií a pupečníkem kolem krku.

Tabulka 16:Decelerace: Chí test - u příznaků, jejichž p-hodnota je <0,05, jsou rozdíly signifikantní.

příznaky		počet		p-hodnota
		decelerace	bez decelerace	
hypoxie	ano	72	1271	< 0,05
	ne	3862	51352	
pupečník kolem krku	ano	100	646	< 0,05
	ne	3834	52007	

Nulové hypotézy pro hypoxii:

- Podíl hypoxie je stejný u matek s preeklampsií a bez preeklampsie.
- Podíl hypoxie je stejný u matek s diabetem a bez diabetu.
- Podíl hypoxie je stejný u porodu s nepoměrem pánve ženy a plodu a bez něj.

Opět všechny příznaky vyšly signifikantní, výskyt hypoxie je ovlivněn preeklampsií, diabetem, a nepoměrem pánve ženy a plodu.

Tabulka 17: Hypoxie: Chí test - u příznaků, jejichž p-hodnota je <0,05, jsou rozdíly signifikantní.

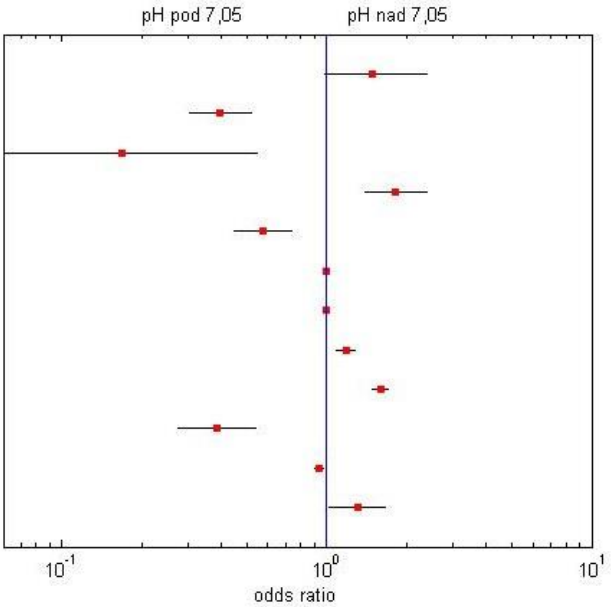
příznaky		počet		p-hodnota
		hypoxie	bez hypoxie	
preeklampsie	ano	23	1321	< 0,05
	ne	90	55239	
diabetes	ano	15	302	< 0,05
	ne	1329	55027	
nepoměr rozměrů	ano	17	364	< 0,05
	ne	1327	54965	

5.7 Výsledky logistické regrese

Pomocí logistické regrese jsme zkoumali, jaké faktory jsou významné pro výskyt císařského řezu a nízkého pH novorozence. Obě logistické regrese jsme nejprve počítali se 62 nezávisle proměnnými (pohlví, věk matky, analgetika..). Podle p hodnot z výstupu funkce gml() jsme výběr nezávislých proměnných zúžili na ty, které byly statisticky významné.

Výsledky logistické regrese pro pH zobrazuje Obrázek 22. Nízké pH je nejvíce ovlivněno poruchou placenty, dále spojeno s decelerací (přechodné zpomalení srdeční frekvence plodu), indukovaným porodem a porodem kleštěmi. Nepatrný vliv má také výška novorozence. Normální pH se vyskytuje u porodů s čirou plodovou vodou, také je ovlivněno dirupcí, Apgar skórem v 1. a 10. minutě a pohlavím – normální pH mají spíše chlapci. Porodní doby nemají na hodnotu pH vliv.

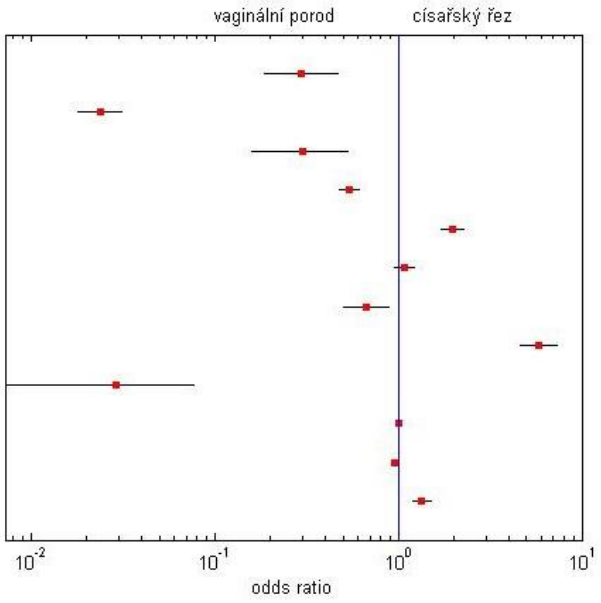
příznak	OR	IS (2,5% - 97,5%)	
čirá plodová voda	1.491	0.978	2.379
decelerace	0.396	0.305	0.517
porucha placenty	0.169	0.060	0.546
dirupce	1.815	1.391	2.383
indukce	0.573	0.445	0.737
třetí doba porodní	0.998	0.992	1.005
druhá doba porodní	0.990	0.985	0.997
apgar suma 10.min	1.181	1.085	1.284
apgar suma 1.min	1.594	1.495	1.700
porod kleště	0.383	0.276	0.540
výška novorozence	0.935	0.896	0.973
pohlaví (muž)	1.305	1.017	1.673



Obrázek 22: Forest plot – pH: Poměr šancí je v grafu vyznačen čtverečkem, který je protnut horizontální čarou představující 95% interval spolehlivosti. Pokud se některé intervaly spolehlivosti protínají se svislou linií v bodě 1, ukazuje to, že při dané hladině významnosti je vliv rizika nulový. Jestliže interval leží v levé části grafu, pak tento příznak ovlivňuje nízké pH.

Obrázek 23 shrnuje výsledky logistické regrese pro císařský řez. Ten je ovlivněn polohou plodu, provádí se, když je plod natočen koncem pánevním. Dále na něj má vliv pohlaví, častěji se císařským řezem rodí chlapci. Graf také ukazuje, že při císařském řezu jsou podávána analgetika a naopak není podáván oxytocin. Císařský řez není prováděn u porodů kleštěmi, u porodů, kde je placenta bez infarktu nebo kde má novorozenec pupečník kolem krku. U vaginálního porodu dochází k dirupci (protržení plodových obalů - vyvolání odtoku plodové vody) a také k inhalaci plodové vody.

příznak	OR	IS (2,5% - 97,5%)	
placenta bez infarktu	0.297	0.187	0.471
oxytocin	0.024	0.018	0.031
pupecnik kolem krku	0.304	0.158	0.531
dirupce	0.540	0.477	0.611
analgetika	1.989	1.723	2.294
SAG	1.088	0.960	1.232
inhalace	0.674	0.506	0.886
poloha koncem panevnim	5.825	4.625	7.357
porod kleště	0.029	0.007	0.077
hmotnost novorozence	1.000	1.000	1.000
výška novorozence	0.955	0.919	0.992
pohlaví (muž)	1.345	1.200	1.509



Obrázek 23: Forest plot – císařský řez: Poměr šancí je v grafu vyznačen čtverečkem, který je protnut horizontální čarou představující 95% interval spolehlivosti. Pokud se některé intervaly spolehlivosti protínají se svislou linií v bodě 1, ukazuje to, že při dané hladině významnosti je vliv rizika nulový. Jestliže interval leží v pravé části grafu, pak je tento příznak spojen s císařským řezem

Závěr

Cílem práce bylo provést analýzu porodnického datového souboru pomocí statistických dataminingových metod. Data mining je v současné době jedním z nejpoužívanějších nástrojů pro analýzu dat.

Teoretická část diplomové práce nás seznámila s pojmy dobývání dat (data mining) a dobývání znalostí z databází a s další terminologií týkající se těchto procesů. Také zde byly popsány motodiky, které se v dobývání znalostí z databází používají. Dále se věnuje tématu data miningu v medicíně a pracím s podobným tématem.

V praktické části byl analyzován datový soubor z porodnického modulu nemocničního informačního systému Fakultní nemocnice Brno. Bylo postupováno podle jednotlivých fází metodiky CRISP-DM popsaných v teoretické části práce. Ve fázi porozumění datům byla provedena vizualizace a popisná statistika a z nich získány základní informace o datech. Výstupem fáze přípravy dat byly upravené datové soubory, které neobsahovaly nulové hodnoty a atributy s osobními údaji. V rámci modelování byla u všech dat otestována normalita pomocí jednovýběrového Kolmogorova-Smirnovova testu. Vzhledem k tomu, že žádná data nepocházela z normálního rozdělení, byly pro další práci vybrány a realizovány následující neparametrické statistické testy: Wilcoxonův dvouvýběrový rank sum test, Kruskal – Wallisův test, Wilcoxonův párový test, Spearmanův test nezávislosti, test dobré shody, logistická regrese. V další fázi metodiky CRISP-DM byly shrnuty výsledky všech výše zmíněných testů.

Při testování jsme se nejvíce zaměřili na pH novorozence, císařský řez a předčasný porod.

Dvourozměrná analýza pomocí Wilcoxonova dvouvýběrového rank sum testu, ukázala, že rozdílné hodnoty pH mají chlapci a dívky a také, že hodnotu pH ovlivňuje dystokie ramének. Vícerozměrná analýza pomocí logistické regrese potvrdila vliv pohlaví na hodnotu pH, avšak vliv dystokie se neprokázal. Logistickou regresí jsme dále zjistili,

že nízké pH je nejvíce ovlivněno poruchou placenty, dále je spojeno s decelerací, indukovaným porodem, porodem kleštěmi a nepatrný vliv má také výška novorozence.

Při analýze způsobu porodu bylo zjištěno, že 18 % všech porodů bylo ukončeno císařským řezem. Vývoj četnosti císařských řezů měl rostoucí tendenci od roku 2004 do roku 2010, v dalších letech začal počet mírné klesat. Test dobré shody prokázal statisticky významný rozdíl mezi vaginálním porodem a císařským řezem, u příznaků: abnormální rotace plodu, pupečník kolem krku, přítomnost streptokoka skupiny B a nepoměr pánve ženy a plodu. Logistickou regresí se potvrdil vliv streptokoka a pupečníku kolem krku na způsob porodu. Dále se ukázalo, že císařský řez je ovlivněn polohou plodu, pohlavím a jsou při něm častěji podávána analgetika.

Analýza délky těhotenství ukázala, že 3 % žen rodí před 34. týdnem těhotenství, 4 % žen před 37. týdnem a 93 % žen po 37. týdnu. Vývoj počtu předčasných porodů od roku 2004 do roku 2013 nevykazuje rostoucí ani klesající charakter, hodnoty kolísají okolo 3 % u předčasného porodu před 34. týdnem a okolo 4-5 % u porodů mezi 34. a 37. týdnem. Spearmanovým testem nezávislosti jsme zjistili, že s rostoucí délkou těhotenství roste hmotnost i výška novorozence. Wilcoxonův dvouvýběrový rank sum test ukázal na statisticky významný rozdíl mezi délkou těhotenství u porodů s dystokií ramének a bez dystokie. Testem dobré shody byl zjištěn vliv rodinného stavu na délku těhotenství. A pomocí testu Kruskal – Walis jsme zjistili, že délka těhotenství má vliv na hmotnost, výšku a pH novorozence a dále, že věk matky ovlivňuje délku těhotenství.

Výsledky práce je možné využít především jako základ pro budoucí hlubší analýzu dat, která by mohla sloužit jako efektivní nástroj pro získávání zajímavých informací z porodnické databáze.

Zásadním problémem, se kterým jsme se při analýze potýkali, byl sběr dat. Ne každý lékařský pracovník vyplnil všechny potřebné informace. Doporučením by mělo být zavedení jednotného systému, který by nedovolil žádnou položku vynechat.

Protože se v databázi vyskytovaly také nesmyslné hodnoty (nulová hmotnost/výška, příliš nízký věk otce, atd.) bylo by vhodné vytvoření integritních omezení. Integritní omezení se může týkat jednotlivých hodnot vkládaných do polí databáze (například hmotnost a výška novorozence nesmí být 0), či může jít o podmínku, která je kombinací hodnot v některých polích jednoho záznamu (například datum narození otce nesmí být pozdější než datum narození dítěte).

Dále by bylo dobré zaznamenávat více údajů o rodičích, které by mohly vést k zajímavým informacím (např. zda má na dítě a porod vliv kouření matky, BMI matky atd.).

Data z porodnického modulu NIS představují velký potenciál, a proto by bylo vhodné použít pro analýzu těchto dat i jiné dataminingové metody.

Literatura

1. ALBERICO, Salvatore et al. The role of gestational diabetes, pre-pregnancy body mass index and gestational weight gain on the risk of newborn macrosomia: results from a prospective multicentre study. BMC Pregnancy and Childbirth. 2014, svazek 14, č. 1, s. 23.
2. ALIYU, Muktar H. et al. The risk of intrapartum stillbirth among smokers of advanced maternal age. *Archives of Gynecology and Obstetrics*. 2008, svazek 278, číslo 1, s. 39-45.
3. BERKA, Petr. Aplikace systémů dobývání znalostí pro analýzu medicínských dat. In: *EuroMISE* [online]. 2001 [cit. 2014-05-11]. Dostupné z:http://euromise.vse.cz/kdd/index.php?page=uvod
4. BERKA, Petr. *Dobývání znalostí z databází*. Vyd. 1. Praha: Academia, 2003, 366 s. ISBN 80-200-1062-9.
5. BLÁHA, Jan. Porodnická anestezie v Česku. *Lékařské listy* [online]. 2014, č. 6 [cit. 2014-06-11]. Dostupné z: http://zdravi.e15.cz/clanek/priloha-lekarske-listy/porodnicka-anestezie-v-cesku-471722
6. CAVALLI, Adriana Schüler et al. Relationship between maternal physical activities and preterm birth: risk factors for a life-threatening condition. *Environmental Health and Preventive Medicine*. 2001, svazek 6, č. 2, s. 74-81.
7. CIOS, Krzysztof J. a G. William MOORE. Uniqueness of medical data mining. *Artificial Intelligence in Medicine*. 2002, svazek 26, s. 1-24.
8. CZEIZEL, A. E. et al. Oral phenoxymethylpenicillin treatment during pregnancy: results from a prospective multicentre study. *Archives of Gynecology and Obstetrics*. 2000-4-26, svazek 263, č. 4, s. 178-181.
9. CZEIZEL, A. E. et al. Reproductive outcome after exposure to surgery under anesthesia during pregnancy: results from a prospective multicentre study. *Archives of Gynecology and Obstetrics*. 1998-8-17, svazek 261, č. 4, s. 193-199.
10. DOHNAL, Luděk. *Štatistické metódy pre klinickú epidemiológiu a laboratórnu prax: Chybějící a odlehlé hodnoty*. Košice: Aprilla, 2008. ISBN 978-80-89346-00-4.
11. FAYYAD, Usama, Gregory PIATETSKY-SHAPIRO a Padhraic SMYTH. From Data Mining to Knowledge Discovery in Databases. *American Association for Artificial Intelligence*. 1997, s. 37-54.

12. FERREIRA, Duarte, Abílio OLIVEIRA a Alberto FREITAS. Applying data mining techniques to improve diagnosis in neonatal jaundice. *BMC Medical Informatics and Decision Making*. 2012, svazek 12, č 1.

13. FREDERICK, Ihunnaya O. et al. Pre-pregnancy Body Mass Index, Gestational Weight Gain, and Other Maternal Characteristics in Relation to Infant Birth Weight: results from a prospective multicentre study. *Maternal and Child Health Journal*. 2008, svazek 12, č. 5, s. 557-567.

14. HARIZOPOULOU, Vicentia C. et al. Maternal physical activity before and during early pregnancy as a risk factor for gestational diabetes mellitus: results from a prospective multicentre study. *Acta Diabetologica*. 2010, svazek 47, č. 1, s. 83-89.

15. HUPTYCH, Michal. Získávání znalostí z dat. [Přednáška]. Praha: ČVUT, 14.03.2010. Dostupné z: http://bio.felk.cvut.cz/~huptycm/Vyuka/IKTZ_prednasky/Dataming.pdf

16. CHAMPMAN, Pete. CRISP-DM 1.0: Step-by-step data mining guide. [online]. 2000 [cit. 2014-04-20]. Dostupné z: http://www.the-modeling-agency.com/crisp-dm.pdf

17. CHEN, Jianxin, et al. A Comparison of Four Data Mining Models: Bayes, Neural Network, SVM and Decision Trees in Identifying Syndromes in Coronary Heart Disease. *Artificial Intelligence in Medicine*. 2002, č. 26, 1-2, s. 1274.

18. IANVINDRASANA, Jimison. Clinical Data Mining: a Review. *IMIA Yearbook of Medical Informatics* [online]. 2009 [cit. 2014-04-20]. Dostupné z:http://www.researchgate.net/profile/Henning_Mueller2/publication/38035298_Clinical_data_mining_a_review/file/e0b4952613a39a4203.pdf

19. LONDERO, A. P. et al. Ultrasonographic assessment of cervix size and its correlation with female characteristics, pregnancy, BMI, and other anthropometric features: results from a prospective multicentre study. *Archives of Gynecology and Obstetrics*. 2011, svazek 283, č. 3, s. 545-550.

20. MEIXNEROVÁ, Marcela. Prediktory předčasného porodu. *Postgraduální medicína* [online]. 2007, č. 1 [cit. 2014-05-12]. Dostupné z:http://zdravi.e15.cz/clanek/postgradualni-medicina/prediktory-predcasneho-porodu-285075

21. MÉZL, Martin. *Pokročilé dolování v datech v kardiologii*. Brno, 2009. Diplomová práce. Vysoké učení technické v Brně.

22. NOVICK, Danielle M. et al. Representativeness of obstetric patients who participate in perinatal depression research: findings from the Women's Mental Health and Infants Program (WMHIP) integrated dataset. *Archives of Women's Mental Health*. 2014, svazek 17, č. 2, s. 97-105.

23. ORTEGA-GARCÍA, Juan A. et al. Head circumference at birth and exposure to tobacco, alcohol and illegal drugs during early pregnancy: results from a prospective multicentre study. *Child's Nervous System*. 2012, svazek 28, č. 3, s. 433-439.

24. PARKER, Margaret G et al. Prepregnancy body mass index and risk of preterm birth: association heterogeneity by preterm subgroups. *BMC Pregnancy and Childbirth*. 2014, svazek 14, č. 1, s. 153-158.

25. PAVELKA, František. Aplikovaná statistika. 1. vyd. Brno: VUT v Brně, 2000, 131 s. ISBN 80-214-1545-2.

26. SAJJA, Sunitha. *Data mining of medical datasets with missing attributes from different sources*. Youngstown, 2010. Doctoral dissertation. Youngstown State University.

27. SINDOS, Michael et al. Ruptured ectopic pregnancy: risk factors for a life-threatening condition.*Archives of Gynecology and Obstetrics* [online]. 2009, svazek 279, č. 5, s. 621-623.

28. SOLEIMANIAN, Farhad, Peyman MOHAMMADI a Parvin HAKIMI. Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study. *International Journal of Computer Applications*. 2012, č. 6, s. 21-26.

29. SVOBODOVÁ, Kristýna. Data Mining. In: WikiKnihovna [online]. 2012 [cit. 2014-05-14]. Dostupné z: http://wiki.knihovna.cz/index.php/Data_Mining

30. WEBB, D.A. a J CULHANE. Time of day variation in rates of obstetric intervention to assist in vaginal delivery. In: *Journal of epidemiology and community health*. 2002. ISSN 1470-2738.

31. YOO, Illhoi et al. Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *Journal of Medical Systems* [online]. 2012, svazek 36, č. 4, s. 2431-2448.

32. ZVÁRA, Karel. *Regrese*. Vyd. 1. Praha: Matfyzpress, 2008, 253 s. ISBN 978-80-7378-041-8.

33. ZVÁROVÁ, Jana. *Základy statistiky pro biomedicínské obory*. 2., dopl. vyd. Praha: Karolinum, 2011, 219 s. Biomedicínská statistika, 1. ISBN 978-802-4619-316.