Mendelova univerzita v Brně
Provozně ekonomická fakulta

Dolování znalostí z rozsáhlých statistických souborů lékařských dat

Diplomová práce

Vedoucí práce:
doc. Ing. Jan Žižka, CSc.

Vypracoval:
Bc. Elvyn-George Badelita

Brno 2015

Tímto bych rád poděkoval mému vedoucímu práce panu Ing. Janu Žižkovi, CSc., za ochotu a vlídnost při řešení problémů spojených s vypracováním práce a cenných rad a informací. Dále bych rád poděkoval mé rodině za trpělivost a pochopení při tvorbě závěrečné práce.

Čestné prohlášení

Prohlašuji, že jsem tuto práci: **Dolování znalostí z rozsáhlých statistických souborů lékařských dat** vypracoval samostatně a veškeré použité prameny a informace jsou uvedeny v seznamu použité literatury. Souhlasím, aby moje práce byla zveřejněna v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách ve znění pozdějších předpisů, a v souladu s platnou *Směrnicí o zveřejňování vysokoškolských závěrečných prací.*

Jsem si vědom, že se na moji práci vztahuje zákon č. 121/2000 Sb., autorský zákon, a že Mendelova univerzita v Brně má právo na uzavření licenční smlouvy a užití této práce jako školního díla podle § 60 odst. 1 Autorského zákona.

Dále se zavazuji, že před sepsáním licenční smlouvy o využití díla jinou osobou (subjektem) si vyžádám písemné stanovisko univerzity o tom, že předmětná licenční smlouva není v rozporu s oprávněnými zájmy univerzity, a zavazuji se uhradit případný příspěvek na úhradu nákladů spojených se vznikem díla, a to až do jejich skutečné výše.

V Brně dne 25. prosince 2015 _____

Abstract

Badelita, E. G. Knowledge discovery from large statistical sets of medical data. Diploma thesis. Brno: Mendel University, 2015.

Final thesis deals with information-mining from large sets of medical data using methods and machine learning algorithms.

The subject of the theoretical part is machine learning and its distribution, description of the basic data types in data mining, most important classifications and predictions methods, criterion defining the quality of prediction methods, description of data mining methodology and frequently used systems.

The practical part focuses on statistical and informatics survey of provided medical data, appropriate transformation, subsequent design and implementation of experiments using machine learning methods to acquire new knowledge and hidden information and finally interpretation of the results together with conclusions for target groups.

Keywords

Data mining, machine learning, medical data, methods of data mining, CRISP-DM methodology, SPSS

Abstrakt

Badelita, E. G. Dolování znalostí z rozsáhlých souborů lékařských dat. Diplomová práce. Brno: Mendelova univerzita, 2015.

Závěrečná práce se zabývá problematikou dolování informací z rozsáhlých souborů lékařských dat pomocí metod a algoritmů strojového učení.

Předmětem teoretické části je strojové učení a jeho rozdělení, popis základních datových typů v data miningu, nejdůležitější klasifikační a predikční metody, kritéria definující kvalitu predikčních metod, popis data miningové metodiky a nevyužívanějších systémů.

Praktická část se zaměřuje na statistický a informatický průzkum poskytnutých lékařských dat, její vhodnou transformaci, následný návrh a uskutečnění experimentů pomocí metod strojového učení s cílem získat znalosti a skryté informace, a nakonec interpretace získaných výsledků společně s vyvozením závěrů pro cílové skupiny.

Klíčová slova

Data mining, strojové učení, lékařská data, metody data miningu, metodologie CRISP-DM, SPSS

Obsah

1 Úvod

Dnešní svět je zahlcený daty. Množství dat ve světě a v našem životě neustále roste, a nikdo nedokáže určit nebo předpovědět, zda probíhající růst někdy skončí. Všudypřítomné počítače dávají lidem možnost, a často je k tomu také nabádá, velmi jednoduše ukládat věci, které bychom dříve bez přemýšlení vyhodili. Levné disky a on-line uložiště způsobují, že odkládáme rozhodnutí jak s daty naložit a jednoduše si pořizujeme více paměti a ukládáme pro jistotu vše. Všudypřítomná elektronická zařízení nahrávají naše rozhodnutí, volby v supermarketu, osobní finanční zvyky, odchody a příchody. Mezitím co nám internet poskytuje velké množství informací, tak zaznamenává všechna rozhodnutí, která skrze něj podnikáme. S tím, jak se množství dat ve světě zvyšuje, tak se znepokojivě snižuje i míra chápání těchto dat. Ve všech těchto datech leží informace, které jsou potenciálně užitečné, ale které jsou zřídka na první pohled zřejmé a definovatelné.

Data mining slouží k hledání vzorů právě z takových dat. Pro lidstvo to v podstatě nepředstavuje nic nového. Nejlepší kvalitou lidského mozku je schopnost rozpoznávání vzorů a lidé toho taky plně využívají od samého začátku lidského věku. Lovci hledali vzorce v chování zvířat, zemědělci hledali vzorce v pěstování plodin a v počasí, a politici hledají vzorce ve svých voličích. Úkolem analytika je najít smysl v datech, která slouží k objevení vzorů, které řídí fyzický svět a přeměňuje je v teorie, které mohou být využity k predikci toho, co se může stát v nových situacích. Prací interpreta je identifikovat příležitosti, tedy vzorce chování, které mohou být převedeny do výnosného obchodu. Toho všeho lze dosáhnout pomocí metod data miningu.

V data miningu jsou data uložena elektronicky a hledání je automatizované nebo přinejmenším vylepšené počítačem. Data mining je o řešení problémů pomocí analýzy dat, které jsou již uložené v databázích. Je definován jako proces objevování vzorů v datech. Proces je automatický nebo většinou poloautomatický. Objevené vzory musí být smysluplné, tak že vedou k objevení nějaké výhody, většinou nějaké ekonomické výhody. Data mining se nevyužívá pouze v ekonomické sféře k predikci chování zákazníků, ale také v lékařském odvětví. (Witten a spol., 2011)

Prediktivní dolování dat v medicíně se zabývá učením modelů k předpovědění zdravotního stavu pacienta. Modely mohou být využity k podpoře lékařů v diagnóze nebo k terapeutických a monitorovacích úkonům. Data miningové metody v lékařství jsou většinou využívány k analýze retrospektivních dat a poskytují příležitosti k využití velkého množství dat, které se běžně shromažďují každodenním lékařským vyšetřením pacienta.

Medicinská informatika je v konečném důsledku zaměřena na organizování, ukládání a zpracování informací o molekulárních a buněčných procesech, tkání, orgánů, jednotlivců, populace a společnosti k podpoře vhodných rozhodovacích strategií. V takovém složitém scénáři je dostupnost analytických metod a nástrojů k automatizované interpretaci pacientských dat velice zásadní. Schopnost předpovědět zdravotní stav je jistě velkou výzvou pro výzkum medicinské informatiky a klinické medicíny. Predikce se může rozpínat v rozmezí od jednoduchého rozvrstvení populace pacientů na základě známých rizikových faktorů jako je třeba věk nebo životní styl až k predikci důsledků vybraného způsobu léčení nebo vybraného léku na zdraví pacienta. (Bellazi a spol., 2011)

Jeden z předmětů a často řešené problematiky lékařského data miningu je predikce chování onkologických nemocí, čímž se také zabývá tato diplomová práce.

2 Cíl práce

Cílem diplomové práce je seznámit se s problematikou dolování informací a znalostí z rozsáhlých statistických souborů reálných dat pomocí metod a algoritmů strojového učení a zaměřit se na neřízené i řízené induktivní učení.

Dále prozkoumat z informatického a statistického hlediska lékařská data poskytnutá programem „Surveillance, Epidemiology, and End Results Program“, který se dlouhodobě zabývá statistickým zkoumáním rakoviny ve Spojených státech amerických již od roku 1973, a využít poskytnutý manuál a datový slovník k interpretaci atributů. Vybrat podmnožinu dat z konkrétní oblasti statistického zkoumání onkologického onemocnění. Zvolit vhodné softwarové nástroje pro data mining, navrhnout a poté uskutečnit vhodnou transformaci z formy originální.

Poslední částí práce je návrh, odůvodnění a uskutečnění sérii experimentů, jejichž cílem je objevit potenciální skryté informace a znalosti ve zkoumaných datech. Zhodnotit a interpretovat získané výsledky a prodiskutovat přínos získaných znalostí z dat včetně jejich využití pro cílové skupiny. Vyvodit závěry a doporučení.

3 Teoretická část

3.1 Strojové učení

Chceme-li vyřešit problém za pomocí počítače, potřebujeme algoritmus. Algoritmus je posloupnost instrukcí, které je prováděno pro transformaci vstupu na výstup. Pro některé úkoly nemáme přesný použitelný algoritmus – například rozpoznat spam od důvěryhodného emailu. Víme, co je vstupem: emailový dokument, který je zjednodušeně řečeno souborem znaků. Víme, jak by měl vypadat výstup: dvě možné hodnoty, které vyjadřují, zda email je spam či není. Sice nevíme, jak převést vstup na výstup, ale můžeme jednoduše sestavit seznam tisíců a statisíců emailů, o kterých víme, že jsou spamy, a můžeme se „naučit“, co z nich je spamem dělá.

Pokrok ve výpočetní technologii umožňuje ukládat a zpracovat velké množství dat a přistupovat k nim do vzdálených míst za pomocí počítačové sítě. Aplikace metod strojového učení pro velké množství dat se nazývá data mining. Analogie dolování dat spočívá v dolování nerostných surovin, kdy máme k dispozici velké množství zeminy a hrubé nezpracované suroviny, extrahované z dolů, kde výsledkem zpracování vzniká malé množství vzácného materiálu. Podobně jako v data miningu je zpracováno velké množství dat k získání cenných znalostí a informací.

Strojové učení se netýká pouze databází, ale je také spojené s umělou inteligenci. Inteligentní systém je takový systém, který má schopnost se v neustále měnícím prostředí učit a přizpůsobovat. Strojové učení také pomáhá najít správná řešení mnoha problémů, například v rozpoznávání obličejů, rozpoznávání řeči a robotice apod.

Strojové učení umožňuje programovat výpočetní techniku k optimalizaci výkonnosti kritérií za použití vzorových dat nebo předchozích zkušeností. Máme k dispozici model, který je definovaný parametry. Učící proces v tomto případě představuje spuštění programu, který optimalizuje parametry modelu pomocí trénovacích dat nebo předešlých zkušeností. Model může být prediktivní, k vytvoření predikcí využitelných pro budoucnost nebo deskriptivní, k získání znalostí z dat nebo také oboje zároveň. Strojové učení využívá teorii statistiky pro tvorbu matematických modelů, protože smyslem úloh je vytvořit závěr ze vzorku dat. (Alpaydin, 2010)

Základním předpokladem strojového učení je použití množiny pozorovacích dat k odhalení skrytých procesů. Jedná se o velmi všeobecný předpoklad, který je obtížný vměstnat do jednoho rámce. Proto vznikly různé učební paradigmata, která řeší různé situace s různými předpoklady. Nejvíce zkoumaná a nejvíce používaná paradigma je učení s učitelem, i když se nejedná pouze o jediný typ. (Abu-Mostafa a spol., 2012)

3.1.1 Učení s učitelem

Když trénovací data obsahují explicitní příklady toho, co je správným výstupem pro dané vstupy, pak se jedná o učení s učitelem. Představme si například problematiku rozpoznávání ručně psaných cifer, kdy ke každému obrázku s cifrou je specifikováno jakou číslovku každý obrázek představuje. Jedná se o učení s učitelem, protože někdo (učitel) vzal na sebe zodpovědnost podívat se na každou ručně psanou cifru zvlášť a určit správný výstup, v tomto případě jednu z deseti možných hodnot, které mohou nastat.

Z hlediska možných variací hodnot existuje více možností, jak mohou být data prezentována pro učící proces. Například záznamy zákazníků v aplikaci kreditních karet

a hudební hodnocení zákazníků v aplikaci poslechu hudby. Jedná se o situace, kdy data jsou již k dispozici pro využití.

Protokol již připravených data setů je v praxi nejběžnějším případem. Jedná se o aktivní učení, kde data set je získán prostřednictvím dotazů, které vytvoříme. My tedy vybereme konkrétní bod ve vstupní množině a „učitel" vybere výstupní hodnotu pro náš vybraný bod.

Další varianta se nazývá učení on-line, kde data set předkládá algoritmu postupně vzorek po vzorku. Taková situace nastává v případě, kdy máme streamované data, která algoritmus musí zpracovat za chodu. On-line učení je užitečné v případech omezení výpočetní rychlosti a paměťového prostoru, které nám zabraňuje zpracovat celá data najednou. Důležitá je poznámka, že učení on-line lze využit i v jiných paradigmatech a není pouze omezeno na učení s učitelem. (Abu-Mostafa a spol., 2012)

3.1.2 Posílené učení

Učení s učitelem již neplatí v případě, kdy trénovací data již neobsahují explicitní správné výstupy pro každý vstup. Obyčejně se setkáme s přirovnáním dítěte, které se učí nedotýkat se horkého šálku čaje. Zkušenost dítěte by se dala rozdělit na 2 varianty. Pokud se dítě dotkne šálku čaje, tak výsledkem je velká bolest. V případě, že se šálku čaje nedotkne, bude potrestáno mnohem menší „bolestí", v podobě neuspokojené zvědavosti. Dítě se z popsané situace naučí, že je lepší se horkého šálku čaje nedotýkat. Převedeme-li problém na situaci strojového učení, tak v trénovací množině dat se nenachází informace, jak by se dítě mělo zachovat, ale místo toho se zde nachází informace, jaké činnosti dítě vykonalo. Takovéto údaje se využívají k posílení informací sloužící pro rozhodování v podobných situacích. V případě, že trénovací množina neobsahuje cílový výstup, ale pouze některé možné výstupy a ohodnocení vypovídající o tom, jak dobrý ten výstup je, jedná se o posílené učení. (Abu-Mostafa a spol., 2012)

3.1.3 Učení bez učitele

Učení bez učitele se koná bez „dozoru učitele" a můžeme využít pouze vstupní data. Cílem je najít zákonitosti ve vstupu. K dispozici je taková struktura ve vstupním prostoru, kde se některé vzory vyskytují častěji než ostatní a chceme obecně zjistit, co se stane a co ne. Ve statistice se tomu říká „odhad hustoty". Jedna z metod pro stanovení hustoty se nazývá shlukování, kde je cílem najít shluky nebo seskupení vstupů. V případě společnosti s daty minulých zákazníků, tak zákaznická data obsahují demografické informace a rovněž minulé transakce se společností. Společnost může chtít vidět rozdělení profilů svých zákazníků, aby zjistila, jaký typ zákazníků má časté zastoupení. V takovém případě shlukovací model rozděluje zákazníky podobných atributů do stejných skupin – toto se nazývá segmentace zákazníků. (Alpaydin, 2010)

Nicméně tento příklad z praxe dokazuje, že jsme schopni se něco naučit i jen na základě vstupů. Učení bez učitele lze považovat za úlohu hledání vzorů a struktur ve vstupních datech. Zároveň to představuje způsob jak vytvořit vyšší úroveň reprezentace dat. (Abu-Mostafa a spol., 2012)

3.2 Základní datové typy data miningu

Jedním ze zajímavých aspektů data miningu je široká škála datových typů, která jsou k dispozici pro analytický proces. Existují dva typy dat s různou složitostí:

- *Nezávisle orientovaná data*: týkají se obvykle jednoduchých datových typů, jako jsou vícerozměrná data nebo textová data. Tyto datové typy jsou nejjednodušší a nejčastější. V těchto případech datové záznamy nemají žádné specifikované závislosti mezi datovými položkami nebo atributy. Příkladem je data set demografických záznamů o jednotlivcích obsahující jejich věk, pohlaví a poštovní směrovací číslo.
- *Závisle orientované data*: v těchto případech mohou existovat implicitní nebo explicitní vztahy mezi datovými položkami. Například data set sociální sítě obsahující množinu vrcholů, které jsou propojeny soustavou hran. Na druhé straně časová řada obsahuje implicitní závislosti. Například dvě po sobě jdoucí hodnoty získané z teplotního čidla budou pravděpodobně závislé jedna na druhé, protože časový atribut implicitně určuje závislost mezi po sobě jdoucími odečty.

Obecně platí, že závisle orientované údaje jsou náročnější, protože složitosti jsou vytvořené již existujícími vztahy mezi datovými položkami. Takové závislosti je potřeba také začlenit do analytického procesu pro získání smysluplných výsledků.

3.2.1 Nezávisle orientovaná data

Nezávisle orientovaná data jsou nejjednodušší formou dat, jejímž typickým představitelem jsou vícerozměrná data. Tato data obvykle obsahují sadu záznamů. Záznam se také označuje jako datový bod, instance, příklad, transakce, entita, objekt, funkce vektoru nebo proměnná, označení závisí na použité aplikaci. Každý záznam obsahuje sadu polí, které jsou také označovány jako atribut či rozměr. V této práci se setkáme nejčastěji s označením atribut nebo proměnná, které označují tu samou věc. Atributy popisují různé vlastnosti záznamu. Relační databázové systémy byly tradičně navrženy tak, aby běžně pracovaly s tímto druhem dat.

Kvantitativní vícerozměrná data

Představme si data, ve kterých je jméno, věk, pohlaví, rasa a poštovní směrovací číslo. Atribut věku má hodnoty, které jsou číselné a mají přirozené uspořádání. Takový atribut je označován jako kontinuální, číselný nebo kvantitativní. Data, ve kterých jsou všechny atributy kvantitativní, se také označují jako kvantitativní vícerozměrné data. Tento typ je v data miningu nejběžnější a mnoho algoritmů s nimi pracuje. Je zvláště vhodný pro analytické zpracování, protože je s ním ze statistického hlediska mnohem snazší práce. Například střední hodnoty kvantitativních hodnot jsou jednoduše vyjádřitelné (jako průměr těchto hodnot), zatímco u jiných typů dat by byl výpočet složitější.

Kategorická data a data smíšených atributů

Mnoho souborů dat v reálných aplikacích mohou obsahovat kategorické atributy s diskrétními hodnotami. Například atributy jako pohlaví, rasa nebo poštovní směrovací číslo mají diskrétní hodnoty bez přirozeného pořadí.

V případě smíšených atributů se vyskytuje kombinace kategorického a číselného atributu. Data zmíněných atributů jsou smíšená, protože obsahují kombinaci obojího. Atribut pohlaví je speciálním případem, který obsahuje pouze dvě možné hodnoty, ale je možné jim přiřadit umělé pořadí a využít pro ně algoritmy, které jsou vhodné pro číselná data.

Binární data

Binární data lze považovat za zvláštní případ vícerozměrných kategoriálních nebo vícerozměrných kvantitativních dat, ve kterých může každý kategorický atribut mít jednu z maximálně dvou diskrétních hodnot.

Textové data

Textová data mohou být ve formě řetězců nebo také jako vícerozměrná data, v závislosti na tom, jak jsou reprezentována. Každý řetězec je posloupnost znaků nebo slov, který odpovídá dokumentu. Nicméně textové dokumenty jsou zřídka reprezentovány jako řetězce. To je proto, že je obtížné přímo efektivně použít řazení mezi jednotlivými slovy v rozsáhlých aplikacích. V praxi se reprezentace vektorového prostoru používá tam, kde frekvence slov dokumentu jsou využity pro analýzu. V této formě se mohou textová data považovat za vícerozměrné kvantitativní údaje, kde atributy odpovídají slovům a hodnoty odpovídají četnostem těchto atributů. Nicméně tento případ je netradiční.

3.2.1 Závisle orientovaná data

Je to skupina datových záznamů, které mohou být nezávisle zpracovány. V praxi mohou dočasně různé datové hodnoty navzájem implicitně souviset prostorově nebo skrz explicitní vztahy mezi datovými položkami. Znalosti o již existujících závislostech výrazně mění proces data miningu, protože proces data miningu je o hledání vztahů mezi jednotlivými datovými položkami. Existuje několik typů závislostí, které se dělí na implicitní nebo i explicitní:

- *Implicitní závislosti:* závislosti mezi datovými položkami nejsou vysloveně uvedeny, ale je známo, že se vyskytují v dané doméně. Například již zmíněné po sobě jdoucí teplotní hodnoty získané z čidla jsou navzájem pravděpodobně velmi podobné. V případě, že hodnota teploty v určité době je úplně odlišná než hodnota následující, tak jde o velmi neobvyklou situaci, čímž může být docela zajímavá pro proces data miningu. jde o odlišný případ na rozdíl od vícerozměrných dat, kde každý záznam je zpracován jako nezávislá entita.
- *Explicitní závislosti:* Typicky se jedná například o grafová nebo síťová data, kde hrany specifikují explicitní vztahy. Grafy jsou skvělé abstraktní nástroje, které jsou často užívány jako reprezentace k řešení data miningu v kontextu s jinými datovými typy.

Data časových řad

Data spadající do této skupiny obsahují hodnoty, které jsou typicky generované kontinuálním měřením za nějaký časový úsek. Například teplotní čidlo měří hodnoty kontinuálně, oproti tomu elektrokardiogram měří parametry pacientova srdečního rytmu

v určitých úsecích. Taková data mají obyčejně implicitní závislosti zakomponované v získaných hodnotách, které přijímá v průběhu času. Atributy jsou rozděleny na další dva podtypy: kontextuální atributy a behaviorální atributy. Kontextuální atributy mají typicky silný dopad na závislosti mezi behaviorálními hodnotami.

Diskrétní sekvence a řetězce

Diskrétní sekvence mohou být považovány za kategorická analogová data časových řad. Stejně jako v případě údajů časových řad, kontextuální atribut je časové razítko nebo pozice v řazení. Behaviorální atributy jsou kategorické hodnoty, z toho důvodu diskrétní sekvence jsou definovány v podobném duchu jako data časových řad. Hlavní rozdílem od časových řad je, že diskrétní sekvence obsahují kategorické atributy. Diskrétní sekvence jsou často náročné pro algoritmy data miningu, protože nemají hladkou návaznost datových hodnot v časové řadě.

Prostorová data

Konkrétní formou prostorových dat jsou časoprostorová data, která obsahují jak prostorové tak i časové atributy. Přesná povaha údajů rovněž závisí na tom, které z atributů jsou kontextové a které jsou behaviorální. Vyskytují se dvě nejčastější formy časoprostorových údajů:

- *Prostorové a temporální (časové) atributy jsou kontextové:* tento druh dat je vnímán jako přímá generalizace prostorových a temporálních dat. Jsou zvláště užitečná v případě, kdy prostorová a temporální dynamika daného behaviorálního atributu je měřená současně. Příkladem může být měření teploty mořské vody na povrchu hladiny. Teplota je měřená pravidelně v určitém časovém úseku, v tom případě je teplota behaviorálním atributem a temporální atributy jsou kontextuální.
- *Temporální atributy jsou kontextové a prostorové atributy jsou behaviorální:* přesněji řečeno, tento typ dat může být také považován za typ časových řad. Avšak prostorová povaha v chování atributu poskytuje lepší interpretaci a soustředěnější analýzu v mnoha scénářích.

Síťová a grafová data

V síťových a grafových datech hodnoty odpovídají uzlům v síti, zatímco vztahy mezi hodnotami odpovídají hranám v síti. V některých případech jsou atributy asociovány s uzly v síti. Hrana může být orientovaná nebo neorientovaná, to závisí na použité aplikaci. Například síťový graf může obsahovat orientované hrany odpovídající hypertextovým odkazům na stránky, kdežto například přátelství na sociálních sítích je neorientované.

Druhou třídou grafových data-miningových problémů jsou databáze, která obsahují velký počet malých grafů, například chemické sloučeniny.

Síťová data jsou velmi obecnou reprezentací a mohou být využita k řešení mnoha podobných aplikací jiných datových typů. Například vícerozměrná data mohou být převedena na síťová data vytvořením uzlu pro každý záznam v databázi a reprezentací podobností mezi uzly v podobě hran. Taková reprezentace se používá poměrně často pro mnoho podobností v data miningových aplikacích jako je například shlukování. Je možné

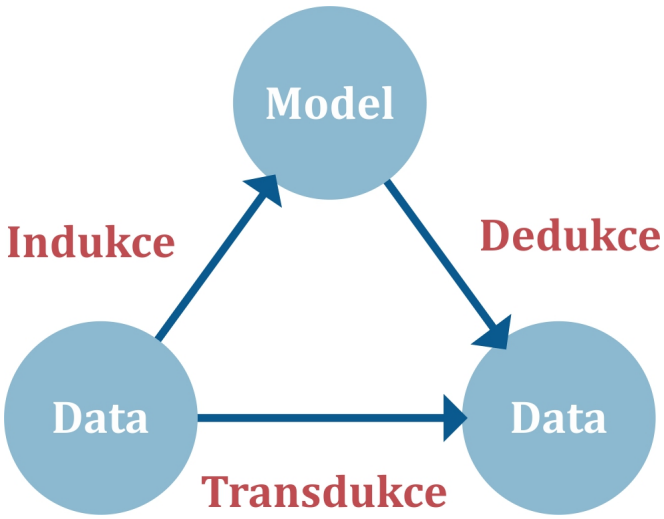
využít algoritmy k zjištění shluků v síťových datech a poté je zpět převést na vícerozměrná data. Některé spektrální shlukovací metody jsou založené na tomto principu. Obecnost síťových dat má svou cenu v podobě náročnějšího vývoje data miningových algoritmů pracující se síťovými daty. (Witten a spol., 2011)

3.3 Klasifikační a predikční metody

Klasifikační a predikční metody pokrývají celou oblast statistiky a získávání dat. Jedná se o nejvíce prověřené, široce používané a nejvíce výnosné data miningové techniky používané v různých oborech, například v medicíně, pro výpočet pravděpodobnosti výskytu onemocnění, k vyléčení pacienta nebo k efektivitě nastavené léčby, ve výzkumu a průmyslu, pro výpočet pravděpodobnosti výskytu určitého jevu, v meteorologii, pro předpověď počasí nebo výskytu znečištění atmosféry, v zemědělství, pro predikci výnosu plodin, v bankách a pojišťovnách, pro výpočet pravděpodobnosti nesplacení závazků zákazníkem, v humanitárních a společenských vědách, k předvídání typů chování, v archeologii, pro datování vyhloubených objektů apod. Existuje velký počet prediktivních technik a algoritmů, které prochází neustálým vývojem. I přesto všechny tyto metody pracují pořád na stejném principu, který se s vývojem nemění a který musí být analytikovi znám, pokud se má zabránit nepatřičnému používání. (Tufféry, 2011)

3.3.1 Induktivní a transduktivní metody

Klasifikační a predikční techniky spadají do dvou skupin. První skupinou jsou induktivní metody, kde se induktivní (trénovací) fáze používá k vytvoření modelu reprezentující shrnutí vztahů mezi atributy a který může být aplikován na nových datech k odvození klasifikace nebo predikce (dedukční fáze). Druhou skupinou technik jsou transduktivní metody mající pouze jeden krok (v případě potřeby se může opakovat), kde je každý atribut klasifikován nebo predikován přímo s odkazem na jiné atributy, které již byly klasifikovány. V případě transduktivních technik nevzniká žádný model.



Obr. 1 Vztah induktivních a transduktivních metod

Nejznámějším zástupcem transduktivních metod je algoritmus k-nejbližších sousedů, který pracuje na principu, kde každý atribut je klasifikován hledáním v množině již klasi-

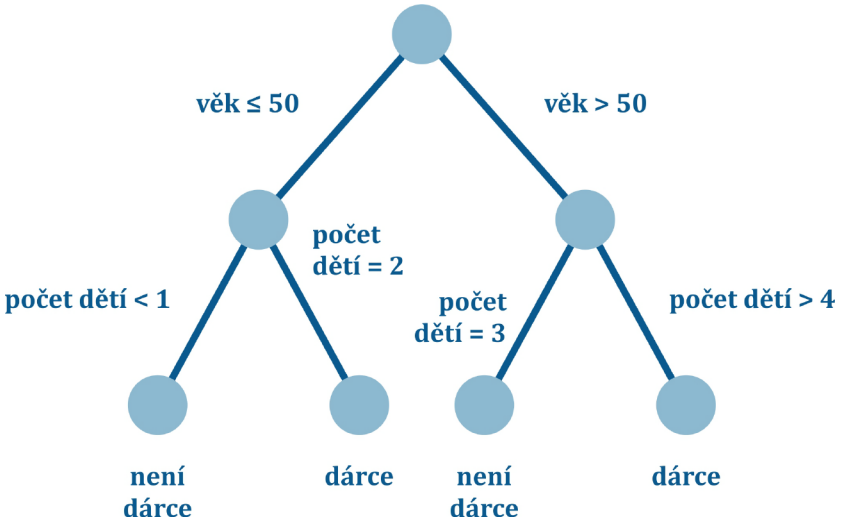
fikovaných atributů ve třídě *k* atributů, které jsou zároveň nejbližšími sousedy na základě euklidovské vzdálenosti nebo i jiných vzdálenostních metrik. V případě induktivních metod se klasifikace a predikce provádí ve třech etapách:

- Etapa tréningu, prováděna na vzorku dat, která je náhodně vybrána a kde klasifikace je již známá.
- Testování na odlišném vzorku dat, které slouží pro kontrolu modelu, jenž je výsledkem trénovací etapy a kde je také již známá klasifikace a také vznikla náhodným výběrem dat. Tato etapa umožňuje vybrat nejlepší modely vytvořené v trénovací etapě a vyhnout se optimistickým předsudkům, které by vznikaly, kdyby testovací množina pocházela ze stejných dat jak ta trénovací.
- Volitelná validační etapa na třetím vzorku dat, kde je klasifikace také známá. Slouží k měření výkonnosti nejlepšího vybraného modelu v předchozích dvou etapách. Smyslem validace je predikce kvality získaných výsledků ve chvíli, kdy je model aplikován.

V praxi trénovací, testovací a validační množina může vznikat ze stejného data setu nebo z rozdělených odlišných data setů. (Tufféry, 2011)

3.3.2 Klasifikace rozhodovacím stromem

Rozhodovací stromy jsou klasifikační metodika, ve které je klasifikační proces modelován za použití množiny hierarchických rozhodnutí na funkci proměnných uspořádané do stromové struktury. Rozhodnutí v konkrétním uzlu stromu označujeme jako rozdělující kritérium, které je obvykle podmínkou na jedné nebo více funkcí proměnných v trénovacích datech. Rozdělující kritérium rozděluje tréninková data do dvou či více částí. Například v případě věku pacienta, kde věk je atributem a rozdělovacím kritériem je věk ≤ 50: pravá větev obsahuje všechny případy pacientů s věkem vyšší jak 50 let a levá větev obsahuje zbylé pacienty mladší než 50 let. Cílem je identifikovat rozdělující kritérium tak, aby byla úroveň míchání proměnných třídy v každé větvi stromu co možná nejnižší. Každý uzel v rozhodovacím schématu představuje množinu datového prostoru definované kombinací rozdělujících kritérií v předchozích uzlech. Rozhodovací strom je obvykle konstruován jako hierarchické rozdělení trénovacích vzorků. Hlavním rozdílem oproti shlukování je, že kritérium dělení v rozhodovacím stromu spadá do učení s učitelem. Mezi rozhodovacími stromy patří algoritmy C5, CART anebo CHAID.



Obr. 2 Zjednodušený příklad struktury rozhodovacího stromu

Cílem rozdělujícího kritéria je maximalizovat rozdělení rozdílných tříd napříč následujících uzlů. Konstrukce rozdělujícího kritéria závisí na povaze základního atributu:

- *Binární atribut*: má jenom jeden typ rozdělení a strom je vždy binární. Každá větev odpovídá jedné z binárních hodnot.
- *Kategoriální atribut*: Pokud kategoriální atribut má *n* rozdílných hodnot, tak existuje více způsobů rozdělení. Jeden možný způsob je použití *n*-násobného rozdělení, ve kterém každá větev rozdělení odpovídá dané hodnotě atributu. Další možností je použití binárního rozdělení testováním každé $2^n - 1$ kombinací kategoriálního atributu a výběrem toho nejlepšího. Nevýhodou této možnosti představuje hodnota *n* v případě, kdy je příliš velká.
- *Číselný atribut*: Pokud číselný atribut *n* obsahuje malé číslo seřazené hodnoty, je možné vytvořit *n*-násobné rozdělení pro každou odlišnou hodnotu. Nicméně pro kontinuální číselné atributy, se toto členění typicky provádí za použití binární podmínky jako je například $x \leq a$ pro atribut hodnoty x a konstanty a. Je potřeba zvážit případy, kdy uzel obsahuje m datových bodů. Proto existuje m možných rozdělujících bodů pro atribut a odpovídající hodnoty, které mohou být stanoveny tříděním dat v uzlu podél tohoto atributu.

Pseudokód obecného trénovacího algoritmu rozhodovacího stromu:

```
Begin
    vyber kořenový uzel obsahující dada set D;
    Repeat
        vyber vhodný uzel ve stromu;
        rozděl vybraný uzel na dva nebo více uzlů na základě předem
        definovaného rozdělujícího kritéria;
    Until žádné další uzly pro rozdělení;
    prořez přetrénované uzly ze stromu;
    označ každý list jeho dominantní třídou;
End
```

Zastavovací kritérium pro růst rozhodovacího stromu je úzce spojen se způsobem prořezávání. Když strom vyroste do konce a každý list stromu obsahuje pouze instanci patřící do konkrétní třídy, výsledný rozhodovací strom docílí 100% přesnosti predikce na trénovacích datech. Nicméně většinou generalizuje hůře na neviditelných testovacích instancích, protože rozhodovací strom se přetrénuje dokonce i na náhodných charakteristikách v trénovacích instancích. Většina z tohoto šumu vzniká kvůli uzlům nižší úrovně, které obsahují menší počet datových bodů. Obecně jednodušší modely (menší výška stromu) jsou mnohem vhodnější než komplexnější modely (vyšší stromy), pokud vykazují stejnou chybovost na trénovacích datech.

Jednou z možností jak snížit míru přetrénování je zastavení růstu stromu dříve. Bohužel neexistuje žádný způsob, jak zjistit, kdy nastane ten správný čas. Z tohoto důvodu platí přirozené pravidlo, které prořezává přetrénované části rozhodovacího stromu a převede vnitřní uzly na listy. K dispozici je mnoho kritérií, které pomáhají s rozhodnutím, jestli má být daný uzel prořezán. Jedna ze strategií je vysloveně sankcionovat složitost modelu za použití principu minimální délky popisu (MDL). V tomto přístupu jsou náklady stromu definované součtem vah jeho chyb a složitosti. Proto je strom konstruován tak, aby optimalizoval spíše náklady než jen chybovost. Problém s tímto přístupem je takový, že nákladová funkce je sama o sobě heuristická, která nefunguje konzistentně dobře v různých datových souborech. Jednodušší a intuitivnější strategií je zadržet malou část trénovacích dat a postavit rozhodovací strom na zbývajících datech. Dopad prořezávání na přesnost klasifikace je pak testován na zadržené množině dat. Pokud prořezání zvýší přesnost klasifikace pak je prořezávání provedeno. Uzly jsou iterativně prořezávány tak dlouho, dokud není možné dále zvyšovat přesnost za pomocí prořezávání. Takový přístup snižuje množství tréninkových dat pro vytvoření stromu, ale dopad prořezávání obecně přebíjí dopad ztráty trénovacích dat. (Aggarwal, 2015)

3.3.3 Klasifikace založená na pravidlech

Rozhodovací strom dělá svou vlastní extrakci příznaků. Jednorozměrný strom používá jenom nezbytné proměnné a poté, kdy je strom postaven, některé příznaky nemusí být vůbec použity. Je důležité poznamenat, že příznaky blíže ke kořeni jsou globálně více důležité. Rozhodovací stromy lze použít k získání příznaků. Další výhodou rozhodovacích stromů je jejich jednoduchá čitelnost a interpretace. Uzly obsahují podmínky, které jsou jednoduše srozumitelné. Každá cesta z kořene do listu odpovídá jednomu pravidlu, které lze sepsat ve formě podmínky IF-THEN. Příkladem algoritmu generujících takové pravidla jsou například C5 pravidla. Například rozhodovací strom na obr. 2 může být přepsán do následujících pravidel:

P1: IF (věk≤50) AND (počet dětí<1) THEN není dárce
P2: IF (věk≤50) AND (počet dětí=2) THEN dárce
P3: IF (věk>50) AND (počet dětí=3) THEN není dárce
P4: IF (věk>50) AND (počet dětí>4) THEN dárce

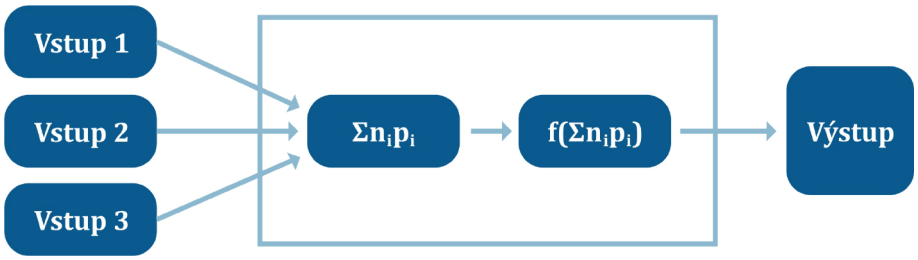
Taková pravidla umožňují získání znalostí, která jsou jednoduše pochopitelná a dovolují expertům ověřit naučený model. Pro každé pravidlo lze také vypočítat procento trénovacích dat, které pravidlo pod sebe zahrnuje. Pravidla odráží hlavní charakteristiky data setu, zobrazují důležité vlastnosti a rozdělující body. V případě klasifikačního stromu

může být více než jeden list přiřazen do stejné třídy. V takovém případě tyto násobné konjunktivní výrazy odpovídající rozdílným cestám mohou být kombinovány jako disjunkce.

Prořezávací pravidla slouží pro zjednodušení a prořezávání podstromu, odpovídají prořezávání pojmů z počtu pravidel ve stejnou dobu. Je možné prořezat pojem z jednoho pravidla, aniž bychom ovlivnili jiné pravidla. Po vykonání procesu prořezávání není již možné složit pravidla zpětně do stromové struktury. (Alpaydin, 2010)

3.3.4 Neuronové sítě

Data mining by neměl podobu, jakou známe dnes bez existence neuronových sítí. Neuronová síť má architekturu založenou na zjednodušeném principu lidského mozku složeného z neuronů, synapsí a nabývá formy v podobě množiny vzájemně propojených jednotek (formálních neuronů). Každá vstupní proměnná odpovídá neuronu první úrovně zvaná vstupní vrstva a každá kategorie kvalitativní proměnné také odpovídá neuronu ze vstupní proměnné. V některých případech, kdy je síť použitá v prediktivních metodách, může existovat jedna nebo více závislých proměnných: každá z nich odpovídá jednomu neuronu v konečné úrovni (výstupní vrstva). Prediktivní sítě spadají do skupiny učení s učitelem a deskriptivní sítě spadají do učení bez učitele.



Obr. 3 Jednotka neuronové sítě (neuron)

Neurony patřící do střední úrovně skryté vrstvy jsou spojené mezi vstupní vrstvou a výstupní vrstvou. V praxi může existovat více skrytých vrstev. Neuron přijme hodnoty na svém vstupu a vrací 0 až *n* hodnot na výstupu. Všechny tyto hodnoty jsou normalizované tak, aby se vyskytovaly mezi hodnotou 0 až 1. Kombinační funkce spočítá první hodnotu z propojených neuronů na vstupu a váhu připojení. V nejvíce používaných typech sítí jde o vážený součet $\sum_i n_i p_i$ vstupných hodnot n_i:

$$y = \sum_{i=1}^{d} n_i p_i + w_0 \qquad (1)$$

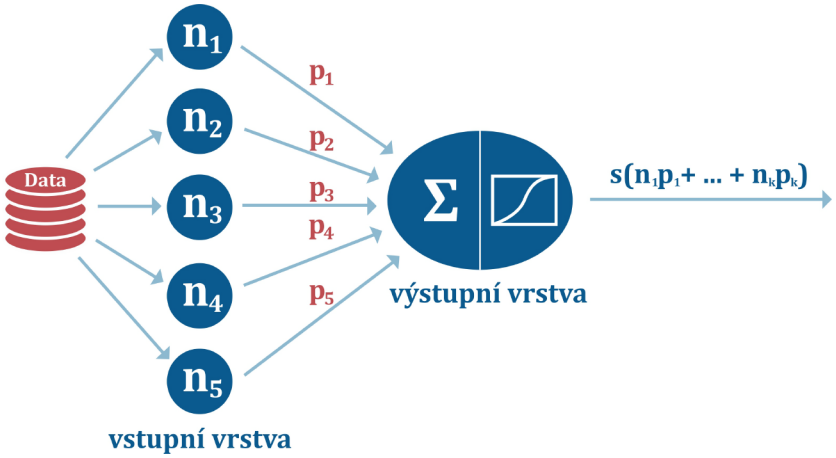
Pro určení výstupní hodnoty se pro tuto hodnotu využívá přenosová funkce. Jednotky ve vstupní vrstvě jsou jednoduché ve smyslu, že nevytvářejí kombinace, ale pouze předávají hodnoty proměnných.

Obecným pravidlem etap implementace predikčních nebo klasifikačních neuronových sítí jsou:

- identifikace vstupních a výstupních dat,
- normalizace těchto dat,
- tvorba sítě s vhodnou strukturou,

- učení,
- testování,
- aplikace modelu generovaného učením,
- denormalizace výstupních dat.

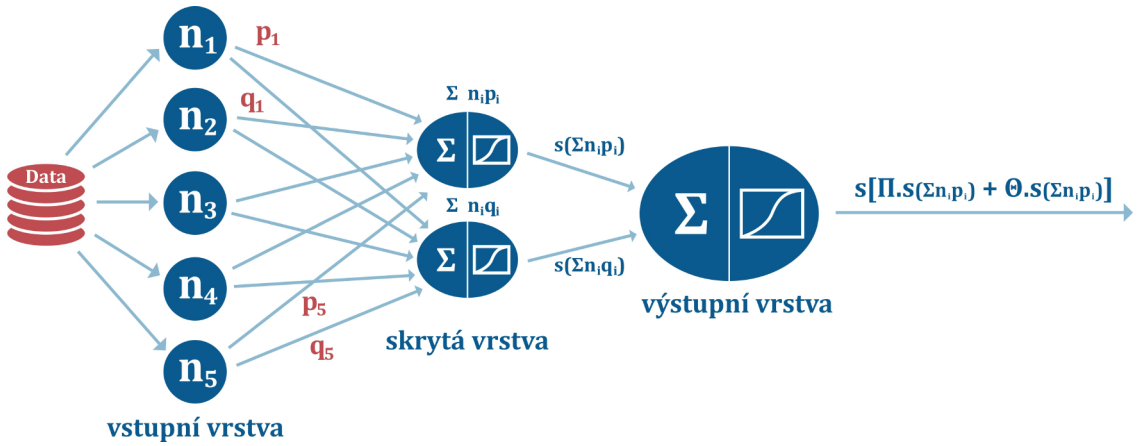
Struktura neuronové sítě, zvaná také jako architektura nebo topologie, se skládá z počtu vrstev neuronů, způsobů jak jsou jednotlivé neurony navzájem propojeny (výběr kombinačních a přenosových funkcí) a váh nestavitelných mechanismů. Volba struktury do značné míry ovlivňuje výsledky, které budou získány, a je tou nejdůležitější částí implementace neuronové sítě. Nejjednodušší strukturou je taková, ve které jsou neurony rozděleny do dvou vrstev: do vstupní a výstupní vrstvy. Každý neuron ve vstupní vrstvě má jeden vstup a jeden výstup, které se rovnají vstupu.



Obr. 4 Neuronová síť bez skryté vrstvy

Výstupní neuron má všechny neurony vstupní jednotky spojené s jejím vstupem za pomocí kombinační funkce a přenosové funkce. Může se vyskytovat více jak jeden výstupní neuron. V případě ilustrovaném na obrázku 4 jde o lineární nebo logistickou regresi závisející na tom, jestli je přenosová funkce lineární nebo logistická a zda jsou váhy sítě regresními koeficienty.

Síla predikce může být posílena přidáním jednoho nebo více skrytých vrstev mezi vstupní a výstupní vrstvy. Ačkoliv se prediktivní schopnost zvyšuje s počtem skrytých vrstev a počtu jednotek v těchto vrstvách musí být tento počet co nejnižší, aby se nestalo, že by neuronová síť pouze ukládala všechny informace z tréninkové množiny dat a nedosáhla žádoucí generalizace, která by zabránila přetrénování vznikající v případě, kdy váhy nutí systém učit se detaily trénovací množiny místo toho, aby objevovala obecné struktury. Toto nastává, když je velikost trénovací množiny příliš malá v poměru komplexnosti modelu, tedy komplexnosti topologie sítě. (Tufféry, 2011)



Obr. 5 Neuronová síť s jednou skrytou vrstvou

Existuje několik druhů modelů neuronových sítí. Mezi hlavní patří vícevrstvý perceptron (MLP) a radiální bázová funkce (RBF).

Obecný trénovací algoritmus pro perceptron pro případ více než dvou tříd by se dal formulovat následovně:

Begin
For i = 1, ... , K **do**
For j = 0, ... , d **do**
$w_{ij} \leftarrow$ rand(-0.01, 0.01)
Repeat
For všechny $(x^t, r^t) \in X$ v náhodném pořadí
For i = 1, ... , K
$o_i \leftarrow 0$
For j = 0, ... , d
$o_i \leftarrow o_i + w_{ij} x^t_j$
For i = 1, ... , K
$y_i \leftarrow \exp(o_i) / \sum_k \exp(o_k)$
For i = 1, ... , K
For j = 0, ... , d
$w_{ij} \leftarrow w_{ij} + \eta (r^t_i - y_i) x^t_j$
Until kovergence
End
(Witten a spol., 2011)

3.3.5 Bayesovské metody

Jedním z nejdůležitějších pojmů v teorii pravděpodobnosti je Bayesův teorém, který se vztahuje na inverzi pravděpodobností pro dvě možné události A a B, podmíněné pravděpodobnosti A vůči B a podmíněné pravděpodobnosti B vůči A. Bayesův vztah pro výpočet podmíněné pravděpodobnosti je tedy formulován jako:

$$P(A/B) = P(B/A)\frac{P(A)}{P(B)} \tag{2}$$

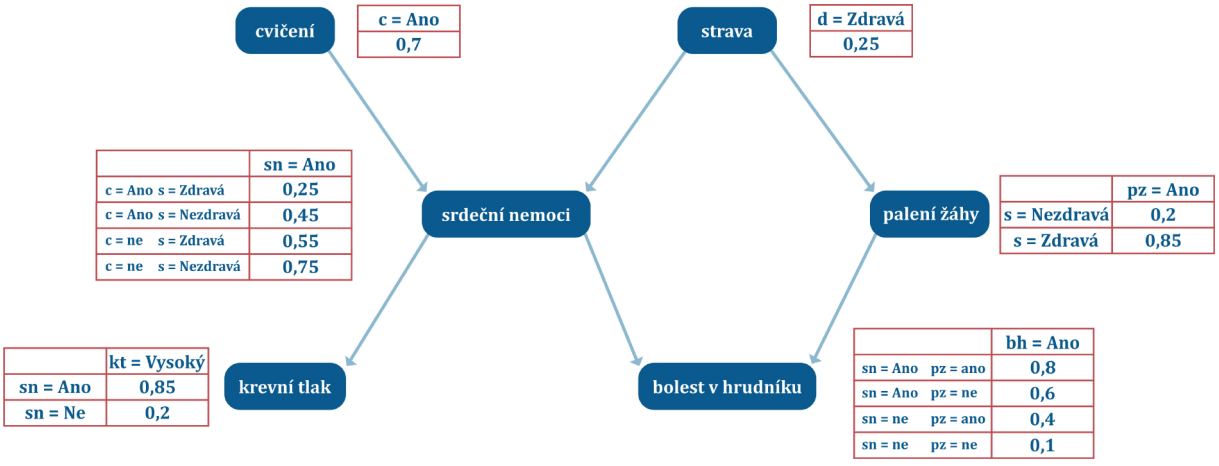
Pokud se jedná o kompletní sadu událostí $\{A_1, A_2, ..., A_n\}$, kde průnik dvou událostí je vždy nulový a sada událostí je rovna všem možným hodnotám, které mohou nastat, pak lze pro každé *i* napsat formuli:

$$P(A_i/B) = P(B/A_i)\frac{P(A_i)}{\sum_{k=1}^{n} P(B/A_k)P(A_k)}, \tag{3}$$

zvanou druhá Bayesovská formule, která je velmi široce používána. Tento základní teorém, přezdívaný Bayesův teorém je platný pro diskrétní a kontinuální rozdělení pravděpodobností a je základním kamenem pro Bayesovskou statistiku. (Tufféry, 2011)

Bayesův naivní klasifikátor je klasifikační metoda, která vychází z Bayesova teorému. Základem naivního klasifikátoru je, že jednotlivá pozorování B_1, ..., B_n jsou podmíněně nezávislá při platnosti hypotézy A. Naivní, protože si je nutné uvědomit, že tento předpoklad je v úlohách reálného světa zřídkakdy splněn. Je sice možné klasifikovat i neúplně popsané vzory, ale aposteriorní pravděpodobnost u chybějících vzorů v trénovací množině je nulová.

Bayesovská síť je acyklický orientovaný graf, kde hrany jsou závislosti mezi náhodnými uzly (veličinami) a ke každému uzlu je přiřazena pravděpodobnostní distribuce ve tvaru *P(uzel|rodič(uzel))*. Každý uzel sítě je očíslován a uspořádán a každý jejich rodič má nižší pořadové číslo. Dále každý uzel *u* je podmíněně nezávislý na všech uzlech s nižším pořadovým číslem s výjimkou svých rodičů a podmíněný rodičem rodičů.



Obr. 6 Bayesovská síť k hledání srdečních onemocnění a pálení žáhy

Diagram na obrázku 6 ilustruje příklad použití Bayesovské sítě pro pacienty se srdečním onemocněním a pálením žáhy. Každý atribut v diagramu je binárního typu (pouze dvě možné hodnoty „ano“ nebo „ne“). Rodičovské uzly srdečního onemocnění odpovídají rizikovým faktorům, které mohou ovlivnit onemocnění (cvičení a strava). Uzloví potomci na druhou stranu odpovídají symptomům nemoci (bolest v hrudníku anebo krevní tlak). Z obrázku lze vyčíst, že pálení žáhy může být výsledkem nezdravé stravy a může vést k bolestem v hrudníku. Vytváření Bayesovské sítě zahrnuje dva kroky:

1) tvorba struktury sítě,
2) odhad pravděpodobnosti hodnot v tabulkách asociovaných s každým uzlem.

Obecný algoritmus pro vytvoření topologie Bayesovské sítě vypadá následovně:

Begin
Označení celkového pořadí proměnných $T = (X_1, X_2, \dots, X_n)$;
For j = 1 to d **do**
$X_{T(j)}$ označí j-té nejvyšší pořadí proměnné T;
$\pi(X_{T(j)}) = \{ X_{T(1)}, X_{T(2)} \dots, X_{T(j-1)}\}$ označí sadu proměnných předcházející $X_{T(j)}$;
Odstranění proměnných z $\pi(X_{T(j)})$, tak aby to neovlivnilo X_j;
Vytvoření oblouku mezi $X_{T(j)}$ a zbývajícími proměnnými v $\pi(X_{T(j)})$
End

Výhodou Bayesovských sítí je umožnění grafického zobrazení znalostí. Mnohdy se Bayesovské sítě označují jako „white box" v porovnání s „black box" neuronových sítí. Výsledky jsou jednoduché pro interpretaci a modifikaci a mohou být použity pro vykreslení závěrů. Jinými slovy, k výpočtu podmiňovacích pravděpodobností sady proměnných sítě poté, co jsou proměnné zkoumány. Závislosti mezi proměnnými mohou být k vykreslení výsledků dokonce použity, i když některé proměnné nemohou být zkoumány. Dedukce může vzniknout na základě nekompletních dat. (Tan a spol., 2006)

3.3.6 Kvality klasifikace a predikce

Přesnost

V klasifikačních technikách chybovost, tedy podíl nesprávné klasifikovaných atributů, musí být co nejnižší. Ukazatele kvality, jako je například Gini index nebo ROC křivka, musí být nejblíž hodnotě 1.

V predikčních technikách existují jiné kvalitativní ukazatelé jako třeba R2 lineární regrese, která musí být také nejblíž hodnotě jedna.

Přesnost nebo také pozitivní prediktivní hodnota je jedna z hlavních ukazatelů kvality predikce a klasifikace v data miningu. (Tufféry, 2011)

Robustnost

Model musí mít co nejnižší závislost na trénovacích datech a musí být schopen vhodně generalizovat na jiných vzorcích dat. Je vyžadována i co nejnižší citlivost na náhodné fluktuaci určitých proměnných a chybějících proměnných. Dokonce pokud se data časem mění, modul by měl být aplikovatelný i na nových datech ještě nějaké rozumné období, které však může být silně zkráceno, pokud nastane nějaká významná změna. Model bude pravděpodobně aplikovatelný kratší dobu, například v odvětví jako jsou mobilní operátoři, které podléhá častějším změnám než třeba oblast bankovnictví. Stabilita by se měla testovat v čase vytvoření modelu, pokud jsou k dispozici zastaralejší data na testování.

Model by neměl být aplikován na proměnných, které bylo obtížné získat nebo které jsou v rámci různých vzorků nestabilní a jsou z rozdílných období. (Tan a spol., 2006)

Stručnost a rozmanitost zpracovaných datových typů

Pravidla modelu by měla být tak jednoduchá jak to je jen možné, to stejné platí i pro počet pravidel. V data miningu se uplatňuje princip Occamové břitvy[1], kdy přídavné komponenty v komplexním modelu způsobují větší šanci, že se zhorší výsledky. Tento princip má za následek lepší srozumitelnost, lepší kontrolovatelnost a vyšší schopnost generalizace na jiných datech než na trénovacích. Stručnost je faktorem robustnosti.

Další obecnou podmínkou je rozmanitost zpracovaných datových typů, protože ne každý algoritmus dokáže zpracovat údaje, které jsou kvalitativní, diskrétní, kontinuální nebo chybějící. (Tufféry, 2011)

Explicitní výsledky

Pravidla by ideálně měla být lehce přístupná a srozumitelná. Když jsou vyjádřená ve formě explicitních podmínek na původních atributech, tak mají dvě výhody: jsou ihned srozumitelná všem uživatelům a jsou jednoduše programovatelná jakýmkoliv programátorem pracujícím s SQL dotazy pro integraci do informačního systému. (Tufféry, 2011)

Rychlost vytvoření modelu

I přesto, že použití daného modelu je vždy relativně rychlé, protože rychlost je jeden z požadavků hlavně u real-time aplikací, může proces trénování modelu trvat příliš dlouho. Důvodem mohou být velká data, která mohou obsahovat statisíce až milióny záznamů. Pokud je k vytvoření modelu potřeba příliš mnoho testů a úprav, tak proces trénování musí být rozumně rychlý. (Tufféry, 2011)

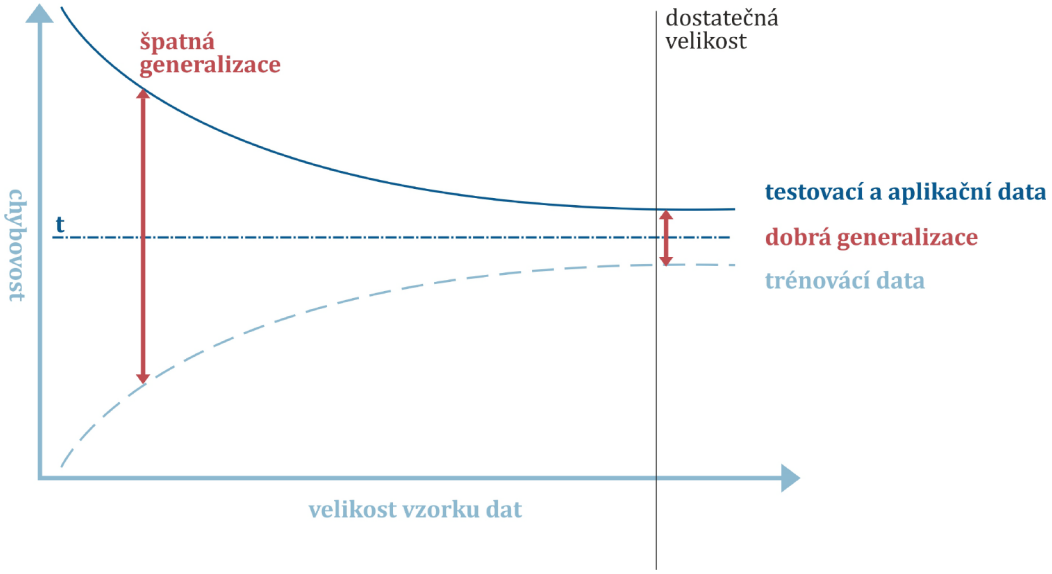
Možnosti nastavení parametrů

V klasifikaci může někdy být užitečné zvýhodnit chyby klasifikace, například z důvodu demonstrace je mnohem důležitější klasifikovat pacienta, který je nemocný než toho, který není nemocný nebo opačně. Užitečná může také být specifikace rozdělení jednotlivců do tříd pro vytvoření predikce. Existuje mnoho možných nastavení parametrů závislých na použité technice a metodě. (Tufféry, 2011)

3.3.7 Generalizace

Malý vzorek trénovacích dat může snadno vykázat nízkou chybovost v trénovací etapě, zatímco by v testovací části vykazovala vysokou chybovost. Důvodem může být špatná generalizace z důvodu nedostatečného trénování. Naopak velká datová trénovací množina může způsobit, že model vypadá během učení méně efektivně, protože nebyl schopen se naučit všechny specifické případy z této více komplexní množiny dat, ale model bude pracovat dobře v testovací fázi, protože má lepší schopnost generalizace.

[1] Occamova břitva je princip logické úspornosti, která říká, že pro vysvětlování určitého jevu má být použito pouze nezbytné množství argumentů či entit. Zbytečné argumenty a entity je třeba eliminovat.



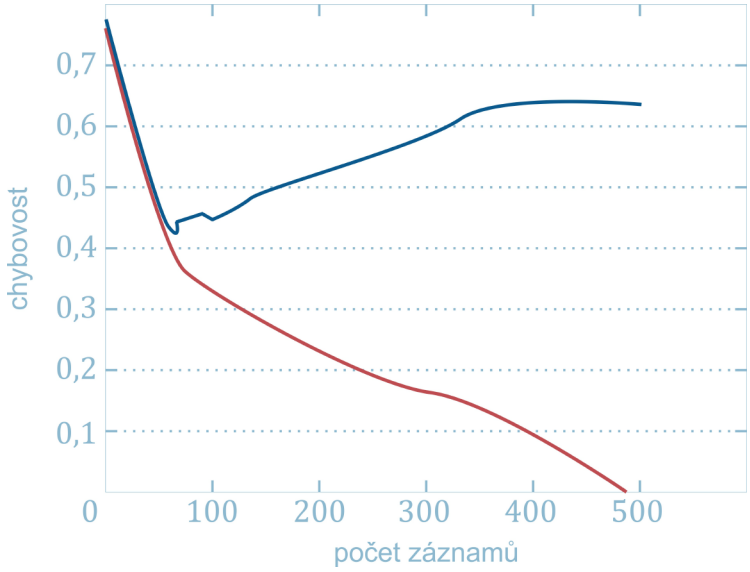
Obr. 7 Chybovost trénování a testování v závislosti na velikosti trénovacích dat

Nicméně chybovost v testovací fázi se nesnižuje do nekonečna zvyšováním velikosti trénovací množiny dat a chybovost v trénovací fázi se také nezvyšuje do nekonečna se zvyšující se velikostí trénovací množiny dat. Trénovací a testovací chybovost se obecně sbíhají do stejného limitu t i pokud se velikost trénovací množiny neustále zvyšuje. Tato vlastnost se v teorii učení nazývá konzistence učení. Neplatí to ve všech učících procesech, ale týká se nejvíce používaných modelů, které patří do skupiny modelů, kde Vapnikova-Červoněnkova dimenze je konečná. Konvergence[2] je průměrná a v praxi pozorovatelná konvergence je nedokonalá, protože chybovost na jednotlivých vzorcích kolísá (vzorkovací citlivost) a testovací chybovost může být dokonce nižší než učící chybovost v konkrétním vzorku. (Tufféry, 2011)

3.3.8 Přetrénování

Jak už bylo naznačeno v předchozí podkapitole, chyby se u klasifikačního modelu obecně rozdělují na trénovací chyby a na testovací chyby. Trénovací chyby, známé taktéž pod názvem resubstituční chyby nebo také zdánlivé chyby, jsou vyjádření počtu špatných klasifikací způsobené na trénovacích datech, kde generalizační chyba je očekávanou chybou modelu na předtím neprozkoumaných záznamech. Je potřeba podotknout, že dobrý klasifikační model nezpracovává dobře pouze jen trénovací data, musí taktéž přesně klasifikovat i záznamy, které jsou nové, a model je ještě neviděl. Jinými slovy model musí mít nízkou trénovací chybovost stejně jako nízkou generalizační chybu. Tato skutečnost je velice důležitá, protože data, které mají příliš dobré výsledky, mohou mít horší generalizaci než model s vyšší chybovostí u trénování. Takovéto situaci říkáme přetrénování modelu.

[2] blížící se vlastní limitě



Obr. 8 Mira trénovacích a testovacích chyb
Zdroj zobrazených výsledků: Tan a spol., 2006

Trénovací a testovací chybovosti jsou vysoké třeba u modelu, který má velmi malý strom. Tuto situaci označujeme naopak jako podtrénování modelu. Vyjadřuje to situaci, kdy se model musí ještě naučit opravdovou strukturu dat. Výsledkem je špatný výkon na trénovacích a testovacích datech. Zvýšením uzlů v rozhodovacím stromu se sníží chybovost, ale ve chvíli, kdy je strom příliš velký, testovací chybovost se začne zvyšovat i přesto, že trénovací chybovost se bude nadále snižovat.

Pro pochopení přetrénovacího fenoménu je potřeba ještě poznamenat, že trénovací chybovost může být snížena zvýšením komplexnosti modelu.

Přetrénování a podtrénování jsou dva patologické jevy, které jsou závislé na komplexitě modelu. Možnou příčinou může být třeba přítomnost datového šumu, nedostatek počtu vzorků nebo mnohonásobná porovnávací procedura. (Tan a spol., 2006)

3.3.1 Vapnikova teorie učení

K vyhodnocení prediktivní kvality modelu můžeme změřit chybu predikce pomocí různých ztrátových funkcí. Nejrozšířenější z nich zahrnuje:

- kvadratickou funkci $L(y, f(x)) = (y - f(x))^2$ pro spojité y,
- funkci $L(y, f(x)) = ½\ ly - f(x)$ pro kvalitativní y s dvěma možnými hodnotami, -1 a +1.

Riziko je definováno jako střední hodnota očekávání ze ztrátové funkce, ale protože společná pravděpodobnost rozdělení X a Y není známa, riziko je pouze ve formě odhadu. Nejběžnějším odhadem je empirický rizikový vzorec:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \qquad (4)$$

nebo

$$\frac{1}{n}\sum_{i=l}^{n}\frac{1}{2}|\,y_i - f(x_i)\,|, \tag{5}$$

kde *n* je velikost vzorku.

Empirické riziko měřené na trénovacích datech má optimistický předpoklad a je obecně nižší než skutečné riziko. Skutečné riziko se nejlépe odhaduje měřením empirických rizik na jiném vzorku, zvaný testovací vzorek, a obvyklým výsledkem je křivka chybovosti podobná té v obrázku 7. Pokud jsou obě křivky blízko určité hodnoty *n,* poté diskriminující síla modelu odpovídá *n* pozorování trénovacího vzorku, který bude pravděpodobně správně generalizovat na dalších vzorcích. V teorii Vladimir Vapnik považoval konvergenci empirických rizik pro trénovací vzorek R_{emp} vůči riziku *R* a demonstroval dva základní výsledky týkající se této konvergence, z nichž jeden se vztahuje k existenci konvergence, zatímco druhý se vztahuje k rychlosti konvergence. (Vapnik, 1998)

VC dimenze

Vapnikova-Červoněnkova dimenze je měřítko komplexity modelu, která je definovaná pro každou skupinu funkcí $R^p \rightarrow R$, která měří oddělovací sílu bodů z R^p. Například lineárně diskriminantní analýza je spojená s lineární funkcí a množinou možných koeficientů definujících skupinu funkcí. Důležitost tohoto konceptu potvrzují dvě Vapnikova zjištění:

- epimirické riziko pro trénovací vzorek R_{emp} modelu konverguje směrem k riziku *R* (model je konzistentní) pouze pokud je VC dimenze konečná,
- pokud je VC dimenze *h* modelu konečná, poté s pravděpodobností chyby α, získáme:

$$R < R_{emp} + \sqrt{\frac{h(\log(2n/h)+1) - \log(\frac{\alpha}{4})}{\mathrm{n}}}\,. \tag{6}$$

Podmínka konečnosti nastavená pro VC dimenzi k zajištění konvergence R_{emp} vůči *R* není triviální. Teoretická hodnota vzorce (6) objevená Vladimirem Vapnikem je její univerzálnost. Může být aplikována na všechny modely bez zvláštního předpokladu na společném rozdělení *x* a *y.* (Vapnik, 1998)

3.4 Metodika Cross-Industry Standard Process for Data Mining

CRoss-Industry Standard Process for Data Mining (dále jen CRIPS-DM), tzv. standardní model procesu dobývání znalostí z databází je metodika popisující univerzální postup, který je aplikovatelný v různých komerčních aplikacích. Metodika vznikla jako výsledek Evropského výzkumného projektu, jejímž cílem bylo vytvořit takovou metodiku, která umožní rychlejší, efektivnější, spolehlivější řešení rozsáhlých úloh s nižšími náklady. Metodika poskytuje navíc postup pro řešení potenciálních problému a úskalí při řešení reálných aplikací na reálných datech. CRISP-DM je výsledkem spolupráce firem SPSS (tvůrce Clementine), OHRA (holandská pojišťovna) a v přední řadě NCR, dodavatelem

datových skladů. Všechny tyto tři firmy spojují bohaté zkušenosti na poli dobývání znalostí z databází.

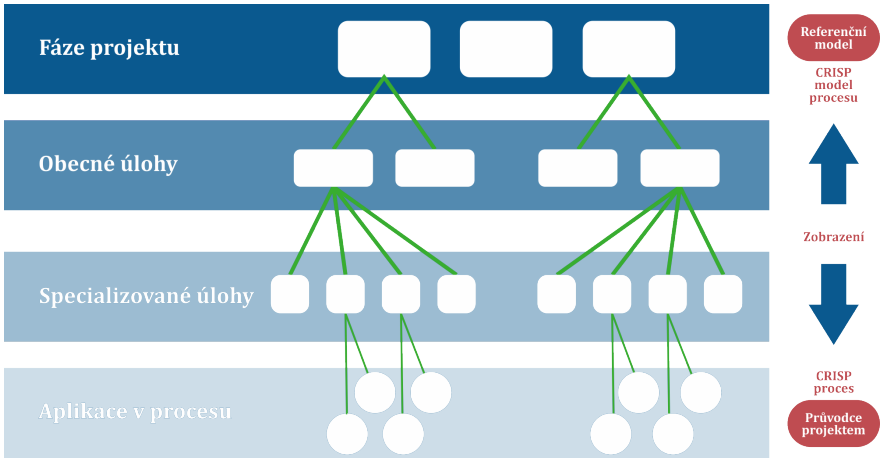
3.4.1 Hierarchie metodiky

První a nejvyšší úroveň, je úroveň rozdělená na fáze projektu. Data mining je rozdělený do fází, kde každou z nich tvoří řada úkolů, která již je na druhé nižší úrovni.

Úroveň obecných úloh je druhou úrovní. Smyslem je pokrýt všechny možné druhy data miningových situací, proto musí být tato úroveň patřičně obecná. Druhá úroveň rozděluje jednotlivé fáze první úrovně na jednotlivé obecné úlohy nezávisle na typu projektu. Hlavním požadavkem je stabilita úloh, protože je důležitá validita modelu pro vývoj, který je v této úrovni ještě nepředvídaný. Úlohy pokrývají celý postup data miningu a všechny možné data miningové aplikace.

Další nižší úrovní jsou specializované úlohy, kde probíhají konkrétní specializované akce, které jsou odvozené z obecných úloh. Nechť příkladem obecné úlohy je úloha nazvaná „vyřešit chybějící hodnoty“. Třetí úroveň by tedy popisovala, zda by se problém chybějících hodnot řešil pomocí výpočtu nových hodnot nebo pomocí eliminace záznamů.

Čtvrtou úrovní jsou aplikace v procesu. Jedná se o záznam všech činností, skutečných výsledků a rozhodnutí. Zde se přistupuje k technické realizaci třetí úrovně.

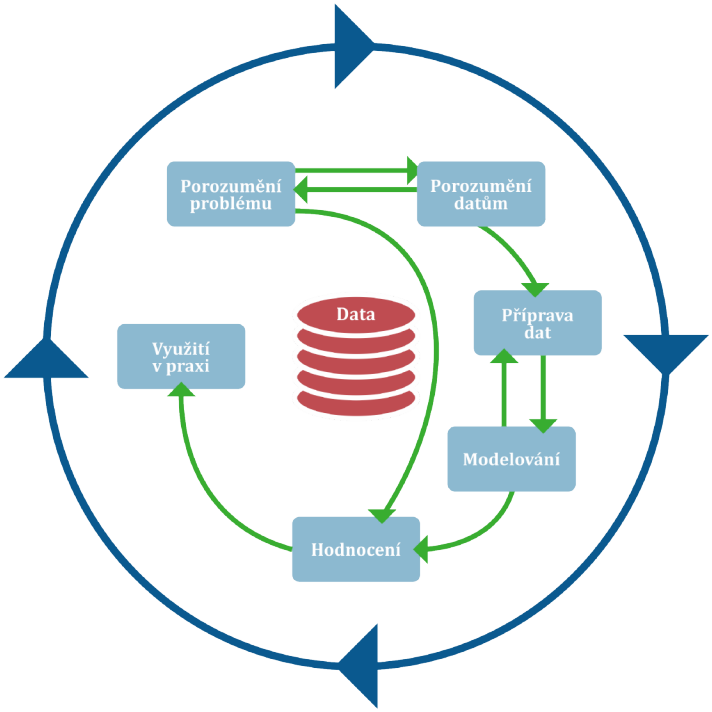


Obr. 9 Hierarchická struktura metodiky CRISP-DM

3.4.2 Fáze metodiky

Fáze metodiky se rozděluje na šest fází, kde pořadí není striktně dané. Výsledek dosažený v jedné fázi má přímý dopad na posloupnost a volbu kroků v následujících fázích. V praxi se stává, že je potřeba se často k některým předchozím fázím vracet. Velký vnější kruh na obrázku vyjadřuje cyklus procesu. Fáze metodiky jsou:

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



Obr. 10 Fáze metodiky CRISP-DM

3.4.2.1 Porozumění problému

Porozumění problému slouží k pochopení cílů řešené úlohy a vymezení požadavků na řešení z manažerského pohledu. Manažerská formulace musí být převedena do zadání reálné úlohy pro data mining. Je potřeba formulovat zadání pro data mining a počítat s možností přeformulovat nebo upřesnit manažerský problém na základě provedených analýz. V této části se obyčejně také hodnotí možná rizika, která mohou nastat, určuje se předběžný plán prací, provádí se inventura zdrojů, náklady a přínos metod KDD. (Petr, 2010)

3.4.2.2 Porozumění datům

Na počátku se musí provést sběr dat, který bude sloužit k dosažení požadovaných cílů. Poté se musí provést analýza dat, která slouží k vytvoření základní představy o datech, rovněž prvotní pohledy do dat, posouzení kvality dat apod. Zjišťují se četnosti zastoupení hodnot atributů, jejich průměry, popřípadě střední hodnoty a extrémy. Doporučuje se využívat různé grafické reprezentační nástroje (grafy apod.). (Petr, 2010)

3.4.2.3 Příprava dat

Patří sem činnosti, které vedou k vytvoření datového souboru, který bude použit jako vstup pro systém dobývání znalostí. Takto vzniklá data by měla mít takovou formu, která bude použitelná pro potřebné analytické algoritmy a měla by obsahovat informace, které jsou potřebné pro úlohu, která se bude řešit. Příprava dat zahrnuje:

- selekci dat,
- transformaci dat,
- čistění dat,

- vytváření dat,
- integrování dat,
- formátování dat.

Jedná se o časově nejnáročnější a nejpracnější fázi metodiky CRISP-DM, která v praxi může zabrat i 80 % celkové času. Jednotlivé kroky této fáze je mnohdy potřeba provádět opakovaně v různém pořadí. (Petr, 2010)

3.4.2.4 Modelování

Analytik má k dispozici celou řadu algoritmů a různých metod k řešení úlohy, avšak je potřeba z nich vybrat pouze ty, které jsou pro danou úlohu nejvhodnější. Ideální situace vypadá tak, že se vyzkouší více metod, vyberou se z nich ty nejlepší a ty se spolu ještě zkombinují. Opět platí pravidlo opakované činnosti, kdy se vyzkouší více algoritmů s různými parametry. Analytik se mnohdy vyskytne v situaci, kdy při aplikování metod a algoritmů zjistí, že je potřeba se vrátit k předchozím fázím a znovu modifikovat data. (Petr, 2010)

3.4.2.5 Vyhodnocení výsledků

Po předchozí fázi, kdy se získají znalosti, které dávají smysl a zdají se být v pořádku z hlediska data miningu, je potřeba výsledky správně interpretovat za pomocí odborníka v oboru a poté výsledky vyhodnotit zda byly splněny cíle, které byly na počátku vytyčeny. V této fázi může být zjištěno, že některé získané znalosti nejsou překvapivé a jiné vypadají zajímavě. Znalosti, které jsou zajímavé, mohou dále být podrobeny dalšímu výzkumu nebo se analytik může vrátit do předchozích fází a provést jiné experimenty. (Petr, 2010)

3.4.2.6 Využití výsledků

Data mining nekončí vyhodnocením výsledků, ale musí se stanovit ještě využití výsledků. Analytik za pomocí odborníka určí, pro jaké cílové subjekty jsou znalosti určené, vyvodí závěry a doporučení. Určí, jestli je potřeba provést další zkoumání, vytvoří se plán nasazení, navrhne se monitorování a údržba, vytvoření reportu a celkové zhodnocení projektu. Znalost se musí předat tak, aby byly využita co nejefektivněji. (Petr, 2010)

3.5 Systémy pro dolování dat a dobývání znalostí

Systémy zabývající se data miningem a dobýváním skrytých znalostí lze rozdělit do třech skupin rozdělených podle objemu dat, s kterým dokážou pracovat.

Nízký (tisíce záznamů):

- NeuralWorks Predict,
- NeuroOne,
- Wizwhy,
- Weka,
- R,
- DataLab.

Vysoký (statisíce až miliony záz.)

- KXEN,
- Oracle Data Mining,
- SPAD,
- IBM SPSS Statistics,
- IBM SPSS Modeler,
- SAS/STAT.

Střední (desetitisíce záznamů) :

- Alice,
- KnowledgeSEEKER,
- KnowledgeSTUDIO,
- C5.0 (Unix), See5 (Win),
- Data Mining Suite,
- CART,
- Polyanalyst,
- TANAGRA,
- JMP,
- S-PLUS.

Dva nejvýznamnější produkty soutěžící na trhu jsou SAS a SPSS, které jsou zároveň také nejvíce využívanými systémy pro dolování dat a znalostí. Oba systémy se specializují na dolování dat z rozsáhlých souborů s vysokým počtem záznamů. (KDnuggets surwey, 2000).

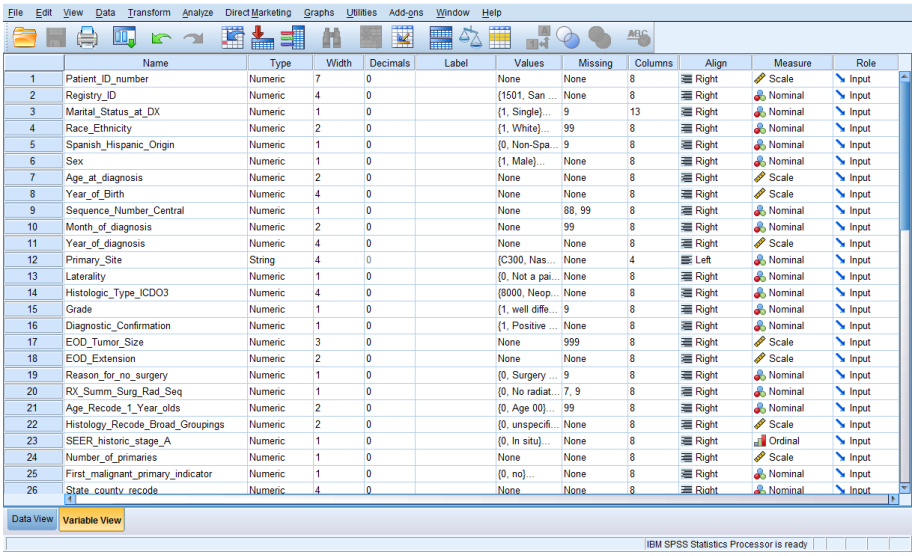
3.5.1 IBM SPSS

Statistical Package for the Social Sciences se poprvé objevil na trhu v roce 1968, odkdy byl velmi často používán v oblastech společenských věd, marketingu a zdravotnictví. Později byl název změněn na IBM Statistics v roce 2009, kdy byla společnost SPSS koupená korporací IBM. Jedná se o nejrozšířenější statistický systém pro PC. Systém disponuje modularitou, která poskytuje možnost postavit systém na míru na základě potřeb uživatele, které se hodí v situacích, kdy je potřeba vytvořit kvalitní průběžné inženýrské a marketingové analýzy, matematicko-statistické aplikace, modelování a data mining. Jednou z hlavních výhod je univerzalita, která zajišťuje pokrytí potřeb v různých sekcích organizace, zabezpečuje kompatibilitu a zjednodušuje přípravu podkladů pro management. (Petr, 2010)

Funkce jsou přístupné pomocí programovacího jazyka čtvrté generace (syntaktický jazyk) a přes grafické rozhraní, které generuje syntaxi neviditelnou pro uživatele, ledaže by uživatel měl požadavek na vkládání vlastních syntaxí v příkazovém řádku. Výhodou grafického rozhraní je její uživatelská přívětivost a jednoduchost. Syntaktický jazyk má ale na druhou stranu výhodu pro mnoho komplexnější operace a k opakování běžných úloh. Část funkcionalit je přístupná pouze pomocí syntaktického jazyka. Příkazy mohou být spuštěny v interaktivním nebo dávkovacím módu. K dispozici jsou také makra pro

automatizaci opakujících se příkazů, které mohou záviset na parametrech. Skriptovací jazyk je k dispozici pro konstrukci dialogových oken.

Okno programu IBM SPPS Statistics se skládá ze záložek „Data Editor“ a „Variable View“, které se podobají tabulkám v Excelu. SPSS příkazy jsou spouštěny řádek po řádku a aktualizují tabulku nebo přidávají výsledky do okna „Output Editor“. Okno editoru výstupů umožňuje uložit provedenou syntaxi s časem provedení. Pokud je potřebné vložit syntaktický příkaz, může tak byt provedeno v okně „Syntax Editor“, které je podobné jako „SAS Editor“ v programu SAS. IMB SPPS Statistics dokáže číst a zapisovat do ASCII souborů, některých druhů databází a do některých tabulek jiných statistických nástrojů.



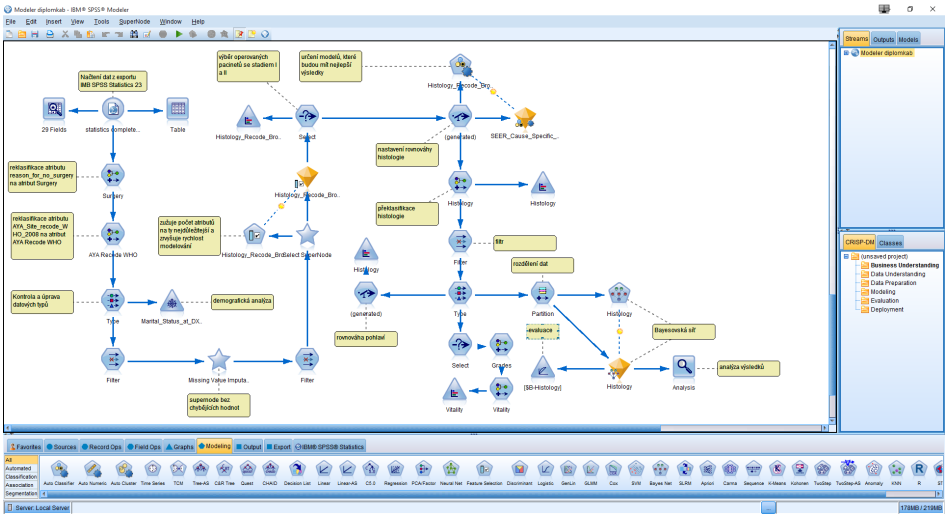
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Patient_ID_number	Numeric	7	0		None	None	8	Right	Scale	Input
2	Registry_ID	Numeric	4	0		{1501, San …	None	8	Right	Nominal	Input
3	Marital_Status_at_DX	Numeric	1	0		{1, Single}…	9	13	Right	Nominal	Input
4	Race_Ethnicity	Numeric	2	0		{1, White}…	99	8	Right	Nominal	Input
5	Spanish_Hispanic_Origin	Numeric	1	0		{0, Non-Spa…	9	8	Right	Nominal	Input
6	Sex	Numeric	1	0		{1, Male}…	None	8	Right	Nominal	Input
7	Age_at_diagnosis	Numeric	2	0		None	None	8	Right	Scale	Input
8	Year_of_Birth	Numeric	4	0		None	None	8	Right	Scale	Input
9	Sequence_Number_Central	Numeric	1	0		None	88, 99	8	Right	Nominal	Input
10	Month_of_diagnosis	Numeric	2	0		None	99	8	Right	Nominal	Input
11	Year_of_diagnosis	Numeric	4	0		None	None	8	Right	Scale	Input
12	Primary_Site	String	4	0		{C300, Nas…	None	4	Left	Nominal	Input
13	Laterality	Numeric	1	0		{0, Not a pai…	None	8	Right	Nominal	Input
14	Histologic_Type_ICDO3	Numeric	4	0		{8000, Neop…	None	8	Right	Nominal	Input
15	Grade	Numeric	1	0		{1, well diffe…	9	8	Right	Nominal	Input
16	Diagnostic_Confirmation	Numeric	1	0		{1, Positive …	None	8	Right	Nominal	Input
17	EOD_Tumor_Size	Numeric	3	0		None	999	8	Right	Scale	Input
18	EOD_Extension	Numeric	2	0		None	None	8	Right	Scale	Input
19	Reason_for_no_surgery	Numeric	1	0		{0, Surgery …	9	8	Right	Nominal	Input
20	RX_Summ_Surg_Rad_Seq	Numeric	1	0		{0, No radiat…	7, 9	8	Right	Nominal	Input
21	Age_Recode_1_Year_olds	Numeric	2	0		{0, Age 00}…	99	8	Right	Nominal	Input
22	Histology_Recode_Broad_Groupings	Numeric	2	0		{0, unspecifi…	None	8	Right	Scale	Input
23	SEER_historic_stage_A	Numeric	1	0		{0, In situ}…	None	8	Right	Ordinal	Input
24	Number_of_primaries	Numeric	1	0		None	None	8	Right	Scale	Input
25	First_malignant_primary_indicator	Numeric	1	0		{0, no}…	None	8	Right	Nominal	Input
26	State_county_recode	Numeric	4	0		None	None	8	Right	Nominal	Input

Obr. 11 IBM SPSS Statistics 23

Umožňuje základní datové funkce jako jsou třídění, agregace, transpozice, sloučení tabulek podobně jako ODS v SAS, ale poněkud v méně flexibilní formě. Dále program dokáže analyzovat data, vytvořit statistickou analýzu, hodnotit kvalitu, pracovat s chybějícími hodnotami, extrémy, měřítky, datovými typy – je to komplexní a snadno použitelná sada nástrojů pro analýzu dat a prediktivní analýzu. Grafika je produkována pomocí R v okně výstupů.

V roce 1998 společnost SPSS koupila ISL, tvůrce data miningového softwaru Clementine, a integroval jejich software do SPSS. Rozhodovací stromy byly do roku 2005 nabízeny samostatně v podobě nástroje Answer Tree, později integrovány do IBM SPSS Statistics ve formě přídavného modulu zvaný jako IBM SPSS Decision Trees. V roce 2009 bylo Clementine přejmenováno na IBM SPSS Modeler. (Tufféry, 2011)

IBM SPSS Modeler je rozsáhlá prediktivní analytická platforma, která je navržena tak, aby využívala prediktivní inteligenci pro rozhodování jednotlivců, skupin, systémů a podniků. Poskytuje celou řadu pokročilých algoritmů a technik, které zahrnují textovou analýzu, analýzu entit, správu a optimalizaci rozhodování.



Obr. 12 IBM SPSS Modeler 17

Systém využívající metodologii CRISP-DM je v současnosti k dispozici ve verzích:

- Gold,
- Premium,
- Professional,
- Personal,
- Predictive Analytics Enterprise.

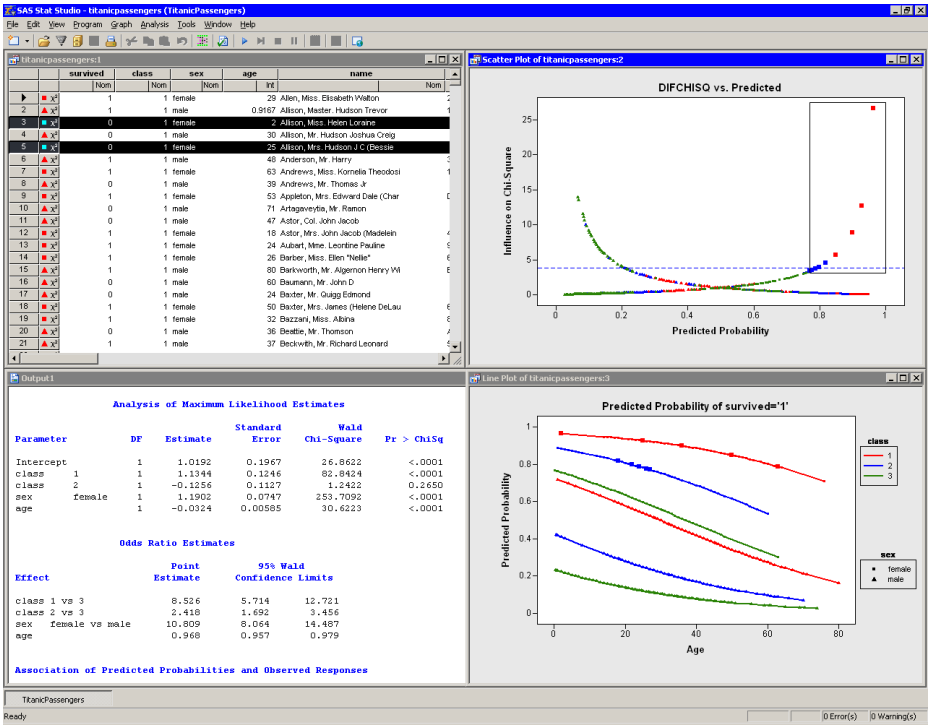
Software využívá vizuálního programování, které je podporováno širokou paletou procesů a uzlů, které se skládají do propojené struktury vytvářející tzv. stream (proud). Procesy a uzly se rozdělují na různé kategorie:

- Sources – procesy pro správu vstupů,
- Record Ops – správa záznamů (select, balance, merge, append, apod.),
- Field Ops – správa atributů (nastavení typů, filtr, reklasifikace, derivace, rozdělení dat, apod.),
- Graphs – grafické vyjádření výsledků podporující lepší interpretaci a prezentaci,
- Modeling – jednotlivé metody a algoritmy data miningu (Neural Net, CART, K-Means, SVM, KNN, Regression apod.),
- Output – uložení a správa výstupů, jejích analýza a hodnocení,
- Export – uzly umožňující export výsledků do různých formátů (Excel, databáze apod.),
- IBM SPSS Statistics – uzly zprostředkující kompatibilitu s IBM SPSS Statistics (vstup i výstup).

Mezi hlavními metodami a algoritmy patří rozhodovací stromy, neuronové sítě, asociace, k-means, regresivní algoritmy apod.

3.5.2 SAS

SAS, celým jménem Statistical Analysis Systém byl vytvořen roku 1974 na sálových počítačích firmy IBM pro zpracovávání velkých objemů dat. Tato vlastnost byla ještě více posílena v roce 1996 implementací paralelní architektury. Jazyk používaný v SAS se úspěšně změnil z Fortanu do Pl/I a následně do C. První verze pro osobní počítače se objevila už v roce 1985, o rok později vyšel první SAS/STAT modul. Podobně jako v případě SPSS se další specializované moduly objevovaly v průběhu let. Některé moduly se více zaměřují na statistiky, například IML (Interactive Matrix Language), STAT, ETS (Econometrics and Time Series), OR (Operational Research) a QC (Quality Control). Ostatní moduly se zaměřují na reporting (AF, EIS). Modul data miningu, Enterprise Miner (hlavní konkurence IBM SPSS Modeleru) se poprvé objevila v roce 1998. V roce 2000 Enteprise Miner získal i možnost grafického programování, který již byl u konkurence implementován. Sedmá verze přinesla novinku s názvem ODS, která umožňovala již neposílat výsledky pouze do okna výstupů, ale také je zapisovat do souborů jako například RTF, PDF, HTML a XML. Od verze 8 bylo možné systém používat i na operačních systémech Windows, Unix a z/OS pro sálové počítače IBM. Verze pro Macintosh, které existovaly již od verze 6, přestaly být podporovány. Následující verze přinesly řadu vylepšení a to v oblastech logistické regrese, nové procedury ROBUSTREG, rozšířením ODS na ODS GRAPHICS apod. (Tufféry, 2011)



Obr. 13 SAS/IML Studio
Zdroj: support.sas.com, 2015.

Podobně jako IBM SPSS Statistics, SAS podporuje L4G programovací jazyk, ale v poněkud v přesnější a efektivnější podobě s více flexibilním a výkonnějším makro jazykem a SQL procedurami. SAS rovněž podporuje maticový jazyk (Interactive Matrix Language).

SAS/IML poskytuje výkonný a flexibilní maticový programovací jazyk v dynamickém prostředí pro programátory, statistiky a výzkumná pracoviště. Software

lze využít od jednoduchých maticových operací po sofistikované statistické metody odhadu. Je vhodný pro nález kořenů, numerické integrace, lineární programování, nelineární optimalizace, boostrap odhadů a simulace.

4 Praktická část

4.1 Porozumění problematice

Při zpracování praktické časti diplomové práce byly využity znalosti popsané v teoretické části. Zejména mezinárodně uznávaná data miningová metodika Cross-Industry Standard Process for Data Mining (CRISP-DM), kterou rovněž podporuje systém pro dobývání znalostí: IBM SPSS Modeler 17. Data pro zpracování poskytl program Surveillance, Epidemiology, and End Results Program zabývající se statistickým zkoumáním rakoviny v USA.

Objem databází od počátku 60. let ve světě a v různých odvětvích neustále roste. Stejně je tomu tak i u medicinských databází, ve kterých se nyní otevírají možnosti vyhledávat závislosti a znalosti, které nejsou na první pohled zřejmé. V současné době existuje více jak 100 podtypů rakoviny, které ovlivňují všechny části lidského těla. Rakovina má na svědomí 13 % všech úmrtí na světě a je zaznamenáno, že v roce 2008 podlehlo na rakovinu 7,6 miliónů lidí. (World Health Organization, 2015)

Jedná se znepokojující počty. Bereme-li v potaz, že existuje vysoký počet různých typů a subtypů onemocnění, kdy každá má jiné příčiny, průběhy a vlastnosti, nabízí se proto využít metody data miningu a statistiky. Získané znalosti z takových dat mohou být využity v boji proti této nemoci.

4.1.1 Definice cílů

Cíle se rozdělují z pohledu National Cancer Institute a z pohledu data miningu. National Cancer Institute poskytuje data na výzkum s cílem obecně snížit břímě rakoviny ve světě. Obecným cílem tohoto data miningu je získat znalost z dat zabývající se statistickým zkoumáním rakoviny v USA. Jelikož data jsou rozsáhlá, je potřeba vybrat podmnožinu dat a zaměřit se na užší okruh problematiky. Konkrétnějším cílem je tedy vytvoření predikčních modelů, které objasní nejdůležitější prediktory mající vliv na uzdravení nebo neuzdravení pacienta a určení nedůležitějších prediktorů ovlivňující rozhodnutí zda pacient půjde či nepůjde na chirurgickou terapii a na základě těchto prediktorů zjistit další skryté informace a znalosti ve zkoumaných datech.

Kritériem úspěchu budou přesnosti uskutečněných predikcí na atributech, které určují přežití pacienta a vykonání chirurgického zákroku.

4.1.2 Posouzení situace

K vypracování tohoto data miningového problému byl využit notebook, statistický software na analýzu dat, data miningový software a tabulkový editor. Celkový seznam využitých prostředků s popisem je uveden v následující inventuře:

Použitý hardware

Notebook:	Asus K53SV
Procesor:	Intel Core i5-2430M Sandy Bridge 2,4 GHz

Operační paměť:	8 GB DDR3-1333
Pevný disk:	HDD 640 GB 5400 ot/min
Grafická karta:	NVIDIA GeForce GT 540M CUDA 2 GB
Monitory:	15,6" integrovaný, 1366 x 768, TFT 22" Philips 227E6EDSD, 1920 x 1080, LED IPS

Použitý software

Operační systém:	Windows 10 Pro 64bitový operační systém
Statistický software:	IBM SPSS Stastistics 23
Data miningový software:	IBM SPSS Modeler 17
Další software:	Microsoft Office 2013 Excel, Word Adobe Acrobat

Terminologie

Onkologie
Obor medicíny, který léči rakovinu. Rozděluje se na chirurgickou, radiační, dětskou a klinickou onkologií. (Bezděková, 2011)

Nádor (tumor, neoplasma, novotvar, blastom)
Jedná se o skupinu abnormálních buněk, které rostou nezávisle na organismu, i když už nepůsobí příčina jejich vzniku. Nádorové buňky se od zdravých liší schopností rychleji se množit a růst. Mají vyšší odolnost, životaschopnost, což je základem jejich převahy nad zdravými a normálními buňkami. (Bezděková, 2011)

Benigní nádory
Jsou nezhoubné nádory, které rostou pomalu, jsou dobře ohraničené, rostou expanzivně (utlačují okolní tkáně z důvodu zvětšování svého objemu), netvoří metastázy a chirurgické odstranění je většinou snadné (vliv lokalizace). (Bezděková, 2011)

Maligní nádory
Zhoubné nádory, které rostou rychle, nejsou ohraničené, rostou infiltrativně (prorůstají do okolí nezávisle na ohraničení orgánu), destruktivně poškozují a ničí tká-

ně. Mají schopnost metastáze krevními nebo lymfatickými cestami a tvoří druhotná ložiska. (Bezděková, 2011)

In situ
Označení pro neinvazivní karcinom. (Bezděková, 2011)

Morfologie/histologie
Jeden z projevů nádoru je jeho morfologická/histologická odlišnost. Zvýšenou schopností růst a množit se způsobuje, že postihnuté orgány jsou zvětšené a liší se tvarem (retrakce, nádorové uzly). Další odchylkou může být i změna zabarvení orgánu. (Bezděková, 2011)

Primární tumor
Ložisko, které vzniklo jako první a které lze většinou při vyšetření jednoduše nalézt. Buňky či skupiny buněk mohou cestovat a vytvářet metastázy - dceřiná ložiska. (Česká onkologická společnost české lékařské společnosti Jana Evangelisty Purkyně, 2015)

Radiační léčba
Je základním druhem léčby, která působí pouze na místo těla, kde se nádor nachází. Na rozdíl od chemoterapie, která působí na nádory v celém organismu, radiační léčba dokáže vyléčit až třetinu všech nádorových onemocnění. (Komplexní onkologické centrum 1. lékařské fakulty Univerzity Karlovy, 2015)

Adenom (adenokarcinom)
Nádor, který vznikl ze žlázového epitelu. Jedná se o epitel, který vytváří žlázy i buňky ve sliznici (buňky, které vytváří hlen v průduškách a ve střevu). Nezhoubný nádor se nazývá adenom a zhoubný jako adenokarcinom. (Česká onkologická společnost české lékařské společnosti Jana Evangelisty Purkyně, 2015)

Spinocelulární/skvamozní karcinom
Jedná se o karcinom z dlaždicového epitelu. Tento epitel je tvořen plochými buňkami a je mechanicky odolný. Dlaždicový epitel vzniká metaplazií v důsledku dlouhodobého dráždění (například cigaretovým kouřem) a vytváří si rohovinové odolné vrstvy. (Česká onkologická společnost české lékařské společnosti Jana Evangelisty Purkyně, 2015)

Stádium rakoviny
Stádium onkologického onemocnění je nejdůležitějším ukazatelem pro předpověď přežití nemocného pacienta.

- Stádium I – II: onemocnění představují relativně malý nádor.
- Stádium III – IVA, B: jedná se o lokálně pokročilý nádor, který se šíří do okolí regionálních lymfatických uzlin.

- Stádium IVC: nevyléčitelné nádory se vzdálenými metastázemi.

Šance přežití v pětiletém horizontu života:

- stádium I: 91%,
- stádium II: 77%,
- stádium III: 32%,
- stádium IVA: 25%,
- stádium IVB: 4%,
- stádium IVC: nevyléčitelné.

(Čapek, 2015)

4.2 Porozumění datům a příprava dat

Lékařské data byla poskytnuta vedoucím této diplomové práce, doc. Ing. Janem Žižkou, CSc., z amerického programu Surveillance, Epidemiology, and End Results Program (dále jen SEER), kterou zaštiťuje National Cancer Institute. Institut má na starosti výzkum rakoviny v USA. Konkrétně se jednalo o lékařská data z období 1973 – 2010.

V průběhu zpracovávání diplomové práce bylo zjištěno, že je veřejně přístupná aktualizovaná verze z období 1973 – 2012. Tato verze obsahuje podstatně více záznamů (3 553 255 pacientů) a má opravené některé nedostatky, které se vyskytovaly ve starších vydáních. Jednou z nich byl například vliv hurikánu Katrina, který měl podstatný vliv na míru mortality mezi pacienty. K získání nových dat bylo zapotřebí pomocí online formuláře zkontaktovat Information Management Services, Inc. a podepsat „SEER Research Data Agreement“, který zajišťuje, že data nebudu dále šířit, nebudou zneužita k identifikování pacientů nebo ke komerčním účelům, či jakýchkoliv jiným účelům kromě statistického a analytického výzkumu.

SEER program se zabývá poskytováním statistických dat onkologických onemocnění ve snaze snížit problém rakoviny v rámci populace USA. SEER shromažďuje údaje o případech rakoviny z různých zdrojů a míst v celých Spojených státech amerických. Sběr dat byl zahájen v roce 1973 s omezeným množstvím registrů a pokračuje v expanzi do dnešní doby ve snaze pokrýt více oblastí a lidské populace. V rámci své výzkumné aktivity, poskytuje speciální studii, zabývající se aktuálními problémy v oblasti prevence a boje proti rakovině a databázi, která pomáhá v průzkumu epidemiologických a zdravotních služeb pro seniory. Toto je v současnosti pouze malá část oblastí působení a výzkumu, kterou podporuje Surveillance Research Program. (National Cancer Institute, 2012)

4.2.1 Vybraná podmnožina

Data se člení na 9 oblastí rozdělené podle místa původu onemocnění. Jako podmnožina dat byla vybrána problematika dýchacích cest, konkrétně plíce a průdušky (Lung and Bronchus). Rakovina plic je v současnosti často řešené téma. V USA je

to druhá nejčastější onkologická nemoc u mužů i žen. U žen druhá po rakovině prsu a u mužů následuje po rakovině prostaty. Rakovina plic má v USA na svědomí 27 % všech úmrtí způsobené rakovinou a tvoří 13 % všech nových případů rakoviny mezi obyvateli. American Cancer Society odhaduje každý rok přibližně 221 200 nových případů onemocnění plic ve Spojených státech (115 610 mužů a 105 590 žen) a 158 040 úmrtí v důsledku tohoto onemocnění. Rakovina plic postihuje hlavně starší obyvatelstvo. Dvě třetiny z nich jsou starší víc jak 65 let a méně než 2 % onemocněných ze všech případů jsou mladší než 45 let. Průměrným věkem je tedy 70 let. (American Cancer Society, 2014)

4.2.2 Prozkoumání dat

Poskytnutý archív obsahoval následující data:

- incidenční výzkumnou databázi s pacienty trpící onkologickým onemocněním z období let 1979 – 2012,
- data 19 věkových skupin posbíraných z různých okresů pomocí SEER registratur a jednotlivé věkové kategorie, dostupná ze statistik sčítání lidu z období 1973 – 2012,
- datové slovníky popisující SEER databázi a populační data.

Datový archív obsahoval podadresáře, ve kterých byla databáze rozčleněna na devět souborů rozdělených dle místa původu onkologického onemocnění:

- prsa,
- tlusté střevo a konečník,
- ostatní trávící orgány,
- ženské pohlavní orgány,
- lymfom všech míst původu a leukémie,
- mužské pohlavní orgány,
- dýchací cesty,
- orgány močových cest,
- všechny ostatní místa původu.

Data týkající se dýchacích cest se nacházela v souboru RESPIR ve formátu TXT, obsahující celkem 624 997 řádků. Každý záznam reprezentoval informace o jednom pacientovi v podobě číselné řady o délce 348 znaků. Data neobsahovala oddělovače. Prázdná místa v řadách upozorňují na chybějící data.

Ukázka nezpracovaných dat:

```
070000080000001502201001081191402041995C3411801238012332104099  99800
070000140000001502201001088188403091973C34998010380103971 70    1999
070000190000001502201001076190202091978C34318070380703921 20-41000---
070000220000001502201002064191002021975C34118140381403311 -0    1999
070000250000001502401002066192402011991C343180103801033210 3085  99800
...
```

4.2.3 Transformace dat

Hrubá nezpracovaná data byla pouze v číselné formě. K datům byl přiložen datový slovník v anglickém jazyce, který popisoval obsažené atributy, s délkami a pozicemi začátků i konců jednotlivých atributů v číselných řadách. V datovém slovníku bylo také popsáno, co každá číselná hodnota reprezentuje. Podle těchto informací bylo možné dohledat reálný význam hodnot atributů. Data v této podobě byla potřeba vhodně transformovat do zpracovatelné formy pro další softwarové nástroje, které v mém případě byli tabulkový editor, statistický a data miningový software. Pro transformaci z originální formy bylo potřeba navrhnout a vytvořit skript na parsování textu, který by na požadovaná místa vložil oddělovače. Pro tvorbu skriptu byl vybrán skriptovací jazyk Python 2.7.10 a vývojové prostředí Spyder 2.3.7.

Ukázka částí kódu:

```
...
print „Zpracovávám text...“
array  = [8,10,1,2,1,1,1,3,4]
array2 = [2,2,4,4,1,4,1,4,1,1, 1,1,3,2,2,1,2,2,13,2,
4,1,1,1,1,3,3,3,2,3 …
array3 = [1,1,1,1,2]
array4 = [1,1,2,1,1,1,1,1,1,1, 1,1,1,1,1,1,1,2]
array5 = [5,4,4]
array6 = [3,3,1]
array7 = [2,2,3,1,1,1,1,2,2,1,1,2,1,5]
array8 =
[5,5,1,1,1,2,2,1,1,1,1,1,1,1,1,2,3,3,3,3,3,3,1,4,1,4,1,1,3,3,3, …
array9 = [1]

inc = 0
file = open("newseer.txt", "w")

with open ("RESPIR.TXT", "r") as f:

      file.write("Patient ID number;Registry ID;Marital Status at DX;
…
      file.write("\n")

      for z in range(624997):
            for i in range(9):
                  value = f.read(array[i])
                  file.write(value)
                  file.write(";")
```

```
                f.read(3)

                for j in range(52):
                value = f.read(array2[j])
                file.write(value)
                file.write(";")

                f.read(1)

                for jb in range(5):
                value = f.read(array3[jb])
                file.write(value)
                file.write(";")

                f.read(2)

                for k in range(18):
                value = f.read(array3b[k])
                file.write(value)
                file.write(";")
...

...

                f.read(5)

                value = f.read(1)
                file.write(value)

                f.read(1)
                file.write("\n")


file.close()
f.close()
```

Skript načte originální TXT soubor s původními číselnými řadami a vytvoří nový TXT soubor, který má v prvním řádku popisy atributů a dále zpracované znaky s vloženými oddělovači reprezentované středníky. Nově vzniklý soubor byl načten pomocí Excelu a ten následně převeden do CSV souboru. CSV soubor byl importován do programu IBM SPSS Statistics ve kterém byla provedena velká část přípravy dat ještě před využití v IBM SPSS Modeler. Před importem bylo třeba přesně definovat formát databáze. Určit, jaké oddělovače jsou použity, kde se nacházejí názvy atributů, jak jsou reprezentovány jednotlivé záznamy apod.

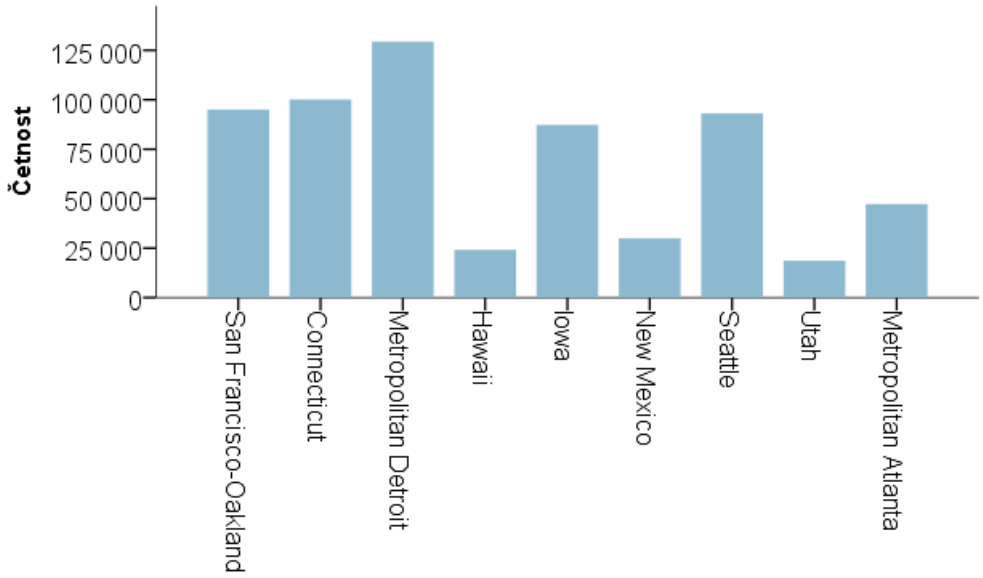
Nově vzniklá databáze obsahovala 624 997 záznamů a 146 atributů. Po prozkoumaní dat a analýzy každého atributu bylo zřejmé, že většina bude muset být vyřazena. Mnoho sloupců neobsahovalo žádná data nebo pro většinu záznamů informace chyběly. Jiné obsahovaly pouze unikátní hodnotu, což by nepřineslo v klasifikaci žádnou informační hodnotu. Určitý počet sloupců byly derivované nebo odvozené data z CS algoritmu, které používá SEER pro vlastní analýzy. Další atributy byly specifické pro jiná onkologická onemocnění (rakovina prsu, žaludku, střev apod.). V důsledku odstranění nepotřebných a nevhodných atributů zbylo celkem 29 atributů použitelných pro vybranou podmnožinu. Ke každému atributu zvlášť

bylo v IBM SPSS Statistics potřeba nastavit datový typ, určit délku atributu, vypsat jakých hodnot může nabývat a ke každé možné číselné hodnotě přiřadit dle datového slovníku jejich reálný význam. Dále bylo potřeba atributům přiřadit roli (vstupní nebo výstupní) k jednomu ze třech způsobů měření (spojitá, ordinální, nominální). V tomto nástroji byla provedena statistická analýza a analýza dat, jejíž výsledek je v následující podkapitole.

4.2.4 Analýza atributů

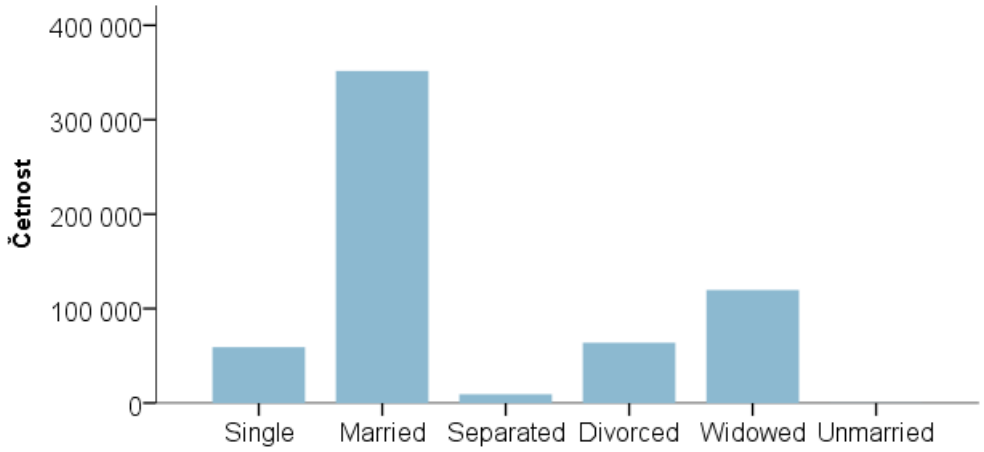
Patient ID Number
Unikátní identifikátor, který byl přiřazen každému pacientovi. Jedná se o primární unikátní klíč.

Registry ID
Unikátní identifikátor přiřazen každé účastněné sběrně dat. Číslo identifikuje registraturu, která posbírala data a zaslala záznam. Nabývá hodnot 1501, 1502 a 1520 až 1547. Každá hodnota reprezentuje název města nebo státu (viz graf níže).



Obr. 14 Frekvenční analýza četnosti atributu Registry ID

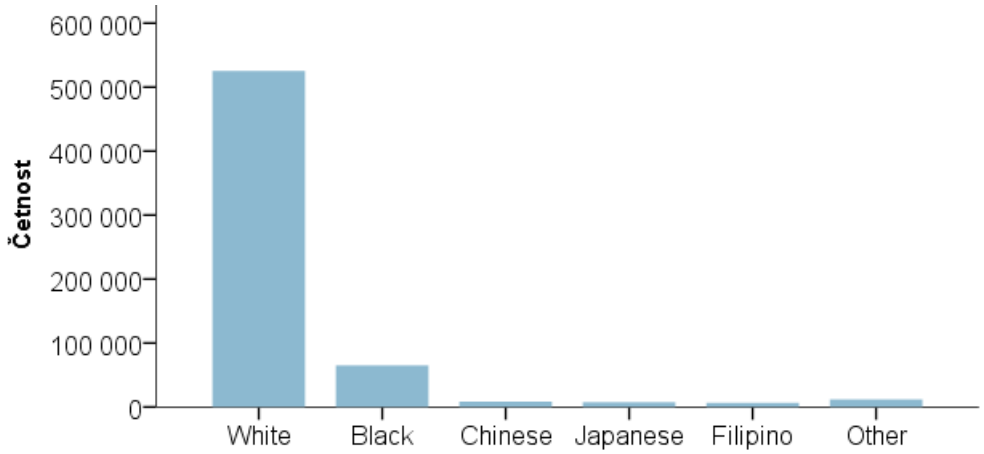
Marital Status at DX
Tento atribut identifikuje rodinný stav pacienta v době diagnózy ohlášeného tumoru. Nabývá hodnot 1 až 6. Včetně 9, která vyjadřuje neznámý rodinný stav.



Obr. 15 Frekvenční analýza četnosti atributu Marital Status at DX

Race / Ethnicity

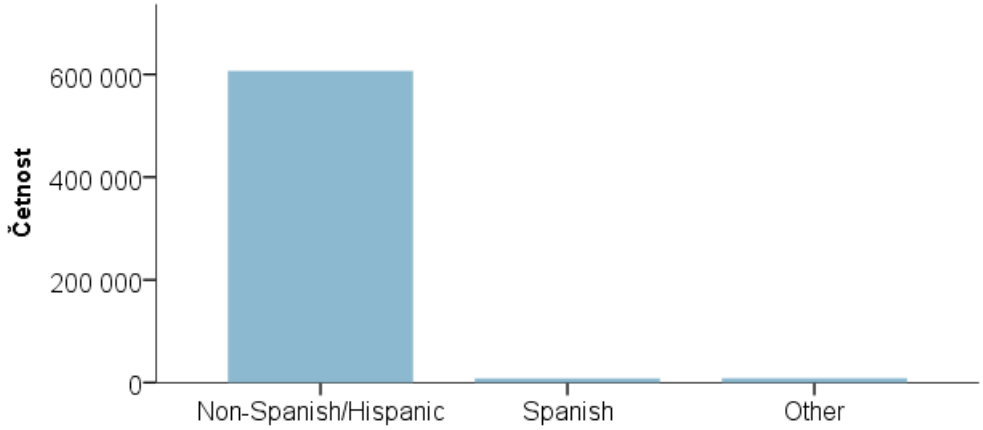
Kód označující rasu nebo etnickou skupinu. U míšenců se upřednostňuje ta rasa, která je rozdílná od rasy bílé. Některé rasy se přidávaly postupně až při pozdějších sběrech dat.



Obr. 16 Frekvenční analýza četnosti atributu Race Ethnicity

Spanish / Hispanic Origin

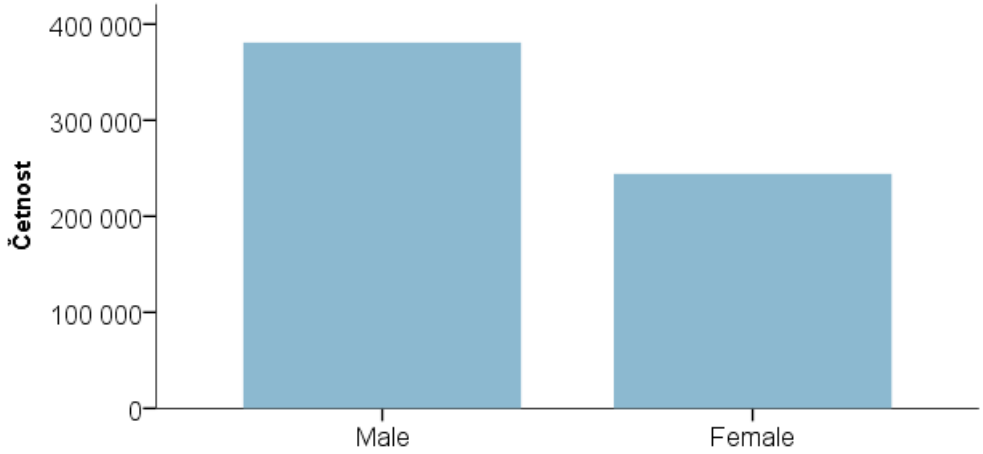
Zaznamenává pacienty se španělským nebo hispánským příjmením anebo ty se španělskými předky. Osoby se španělským nebo hispánským příjmením či předky mohou být jakékoliv rasy.



Obr. 17 Frekvenční analýza četnosti atributu Spanish / Hispanic Origin

Sex

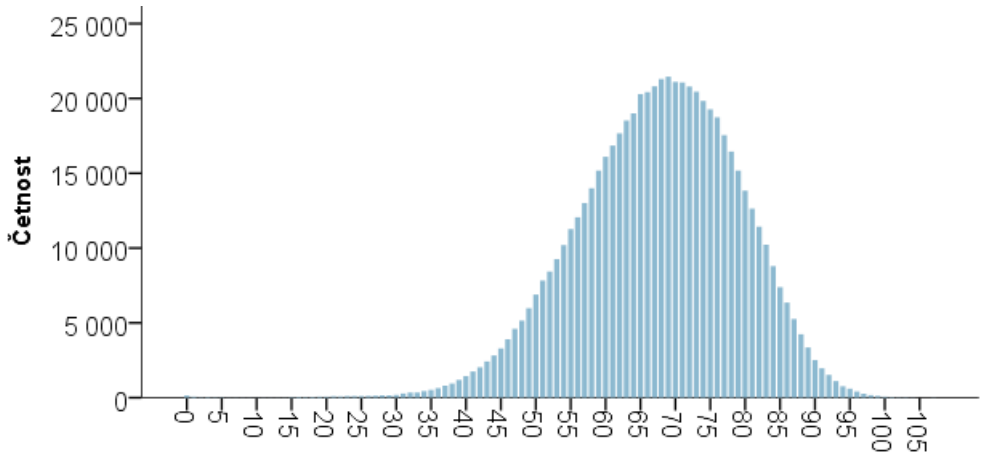
Tato položka identifikuje pohlaví pacienta. Muži mají hodnotu 1 a ženy hodnotu 2.



Obr. 18 Frekvenční analýza četnosti atributu Sex

Age at diagnosis

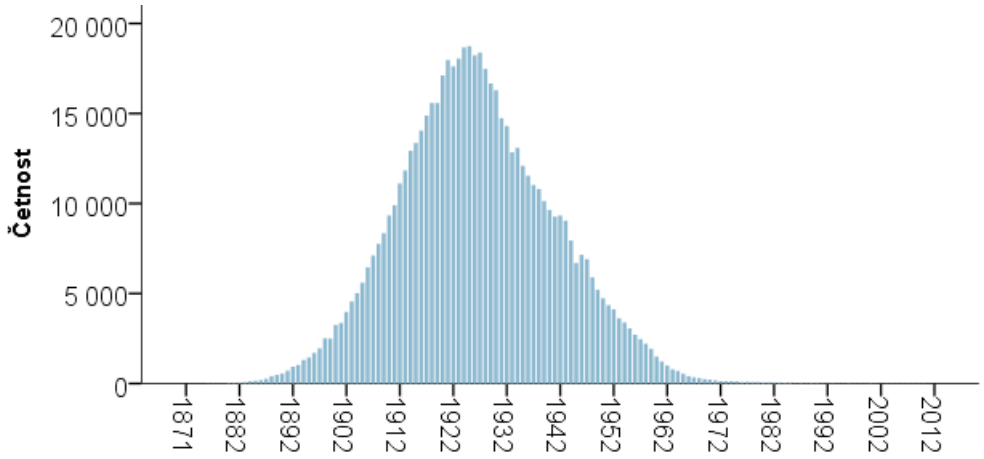
Tento atribut reprezentuje věk pacienta v době diagnózy rakoviny. Kód se skládá z tří číslic, které reprezentují konkrétní věk. Průměrný věk pacientů je 68 let.



Obr. 19 Frekvenční analýza četnosti atributu Age at diagnosis

Year of Birth

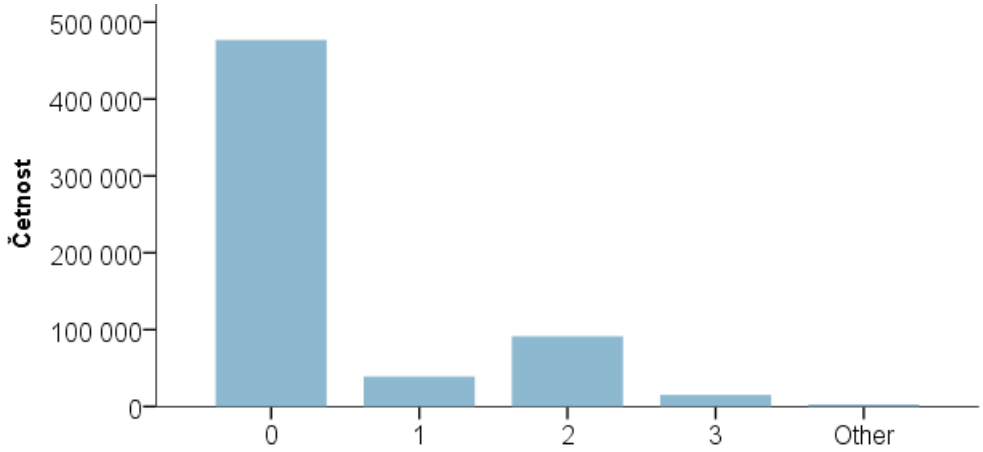
Skládá se ze čtyř cifer, které vyjadřují rok narození pacienta. Průměrným rokem narození je 1926.



Obr. 20 Frekvenční analýza četnosti atributu Year of Birth

Sequence Number – Central

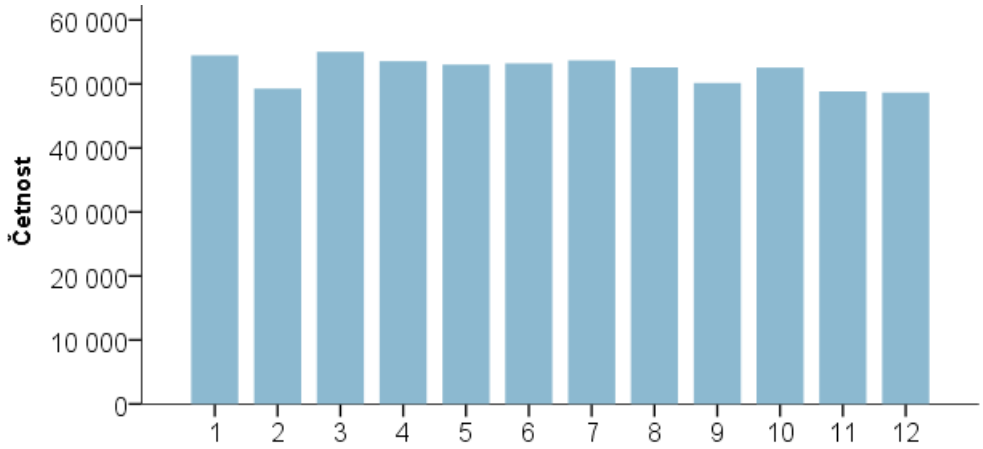
Tento atribut popisuje počet a pořadí všech zaznamenaných maligních, lokálních, benigních a hraničních primárních nádorů, které se vyskytly v průběhu života pacienta. V případě, že pacientovi byl dříve diagnostikován maligní neoplasm (novotar) a následně má druhý nález, tak se sekvenční kód prvního nálezu změní z 0 na 1. Tohle sekvenční číslo počítá i tumory, které byly diagnostikovány i před existencí sběrných míst.



Obr. 21 Frekvenční analýza četnosti atributu Sequence Number - Central

Month of Diagnosis

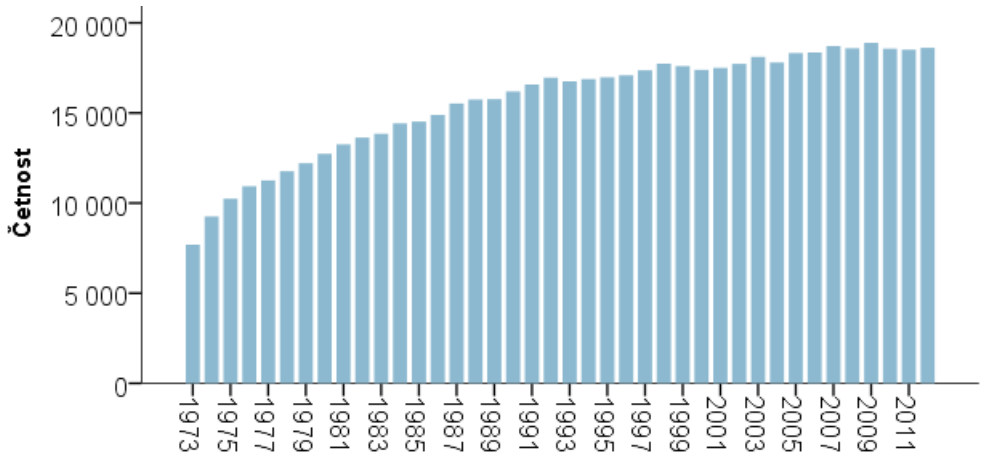
Měsíc diagnózy určuje měsíc, kdy byl tumor poprvé nalezen uznaným medicinským praktikantem ať už klinicky nebo mikroskopicky. Nabývá hodnot 1 – 12.



Obr. 22 Frekvenční analýza četnosti atributu Month of diagnosis

Year of diagnosis

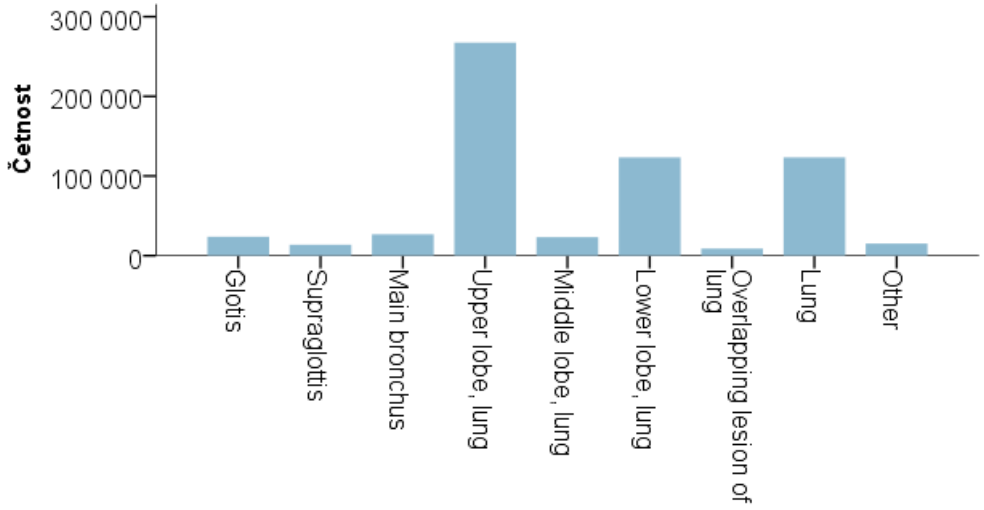
Rok diagnózy je rokem, kdy byl tumor poprvé nalezen uznaným medicinským praktikantem ať už klinicky nebo mikroskopicky.



Obr. 23 Frekvenční analýza četnosti atributu Year of diagnosis

Primary site

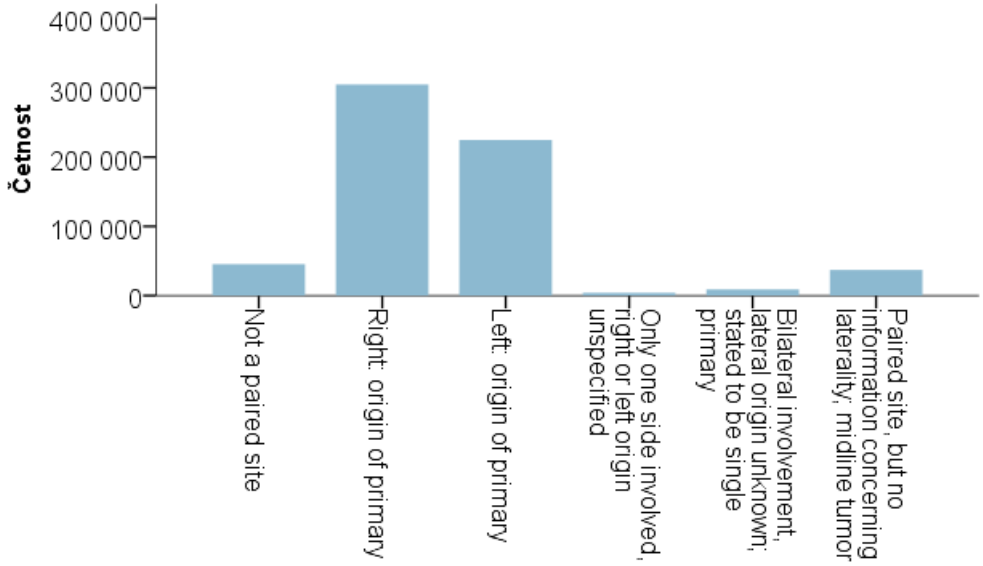
Primary site identifikuje místo, kde primární tumor vznikl. Atribut nabývá hodnot ve tvaru topografických kódů, které jsou definovány ve třetím vydání *International Classification of Diseases for Oncology* (ICD-O-3). Hodnota je uložená bez desetinných čárek.



Obr. 24 Frekvenční analýza četnosti atributu Primary Site

Laterality

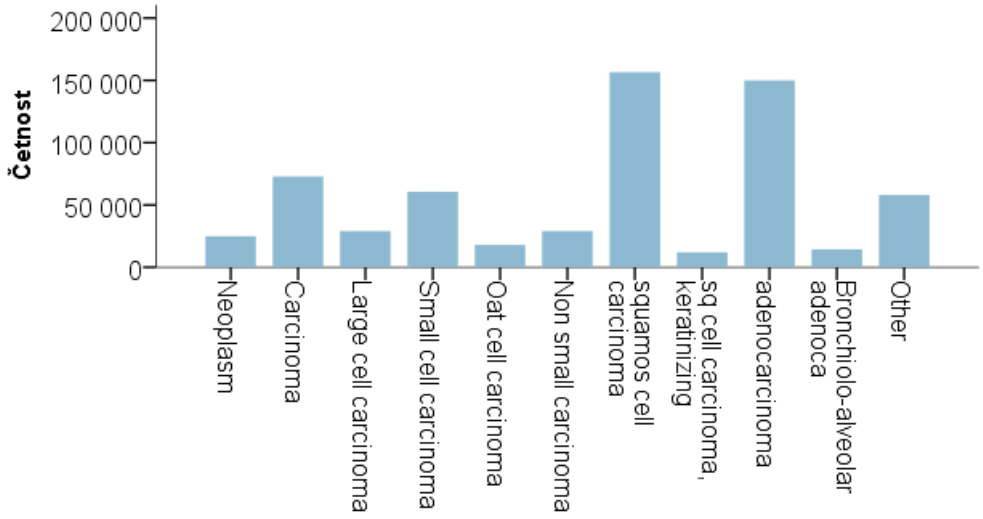
Označení strany párového orgánu nebo bok těla, kde hlášený tumor vznikl.



Obr. 25 Frekvenční analýza četnosti atributu Laterality

Histologic type ICD-O-3

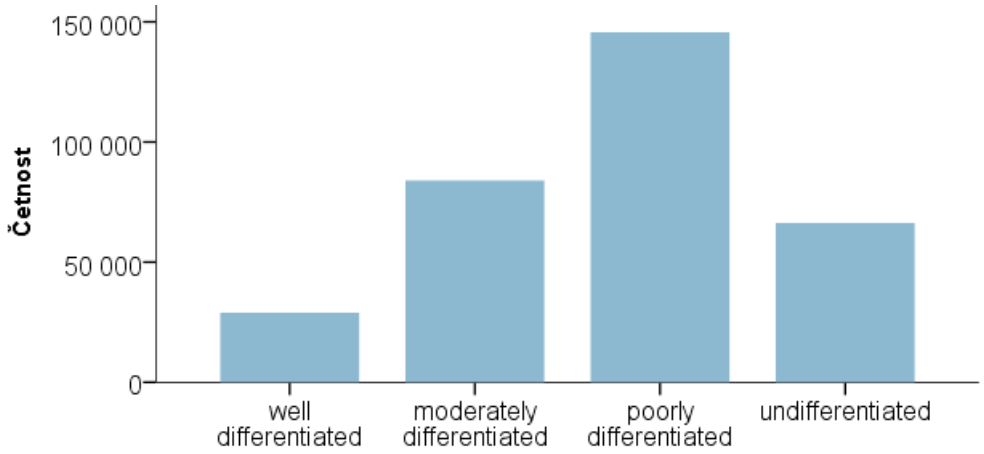
Popisuje mikroskopickou kompozici buněk nebo tkáně primárního tumoru. Typ tumoru nebo histologie jsou základem pro určení způsobu následné léčby rakoviny. SEER program v průběhu času používal několik rozdílných, ale souvisejících kódovacích systémů pro určení morfologie. Proto všechny záznamy byly konvertovány do systému *International Classification of diseases for onkology* třetího vydání (ICD-O-3). Morfologie byla určována pouze mikroskopickým zkoumáním. Tento atribut ovlivňuje průběh a prognózu nemoci.



Obr. 26 Frekvenční analýza četnosti atributu Histologic Type ICD-O-3

Grade

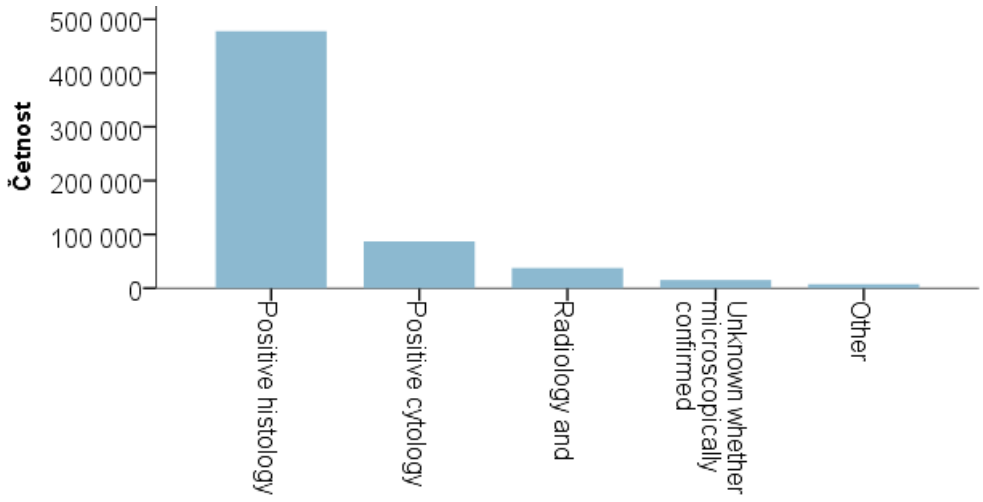
Stupňovací a diferenciační kódy. Hodnoty 1 až 4 jsou definovány v ICD-O-2, 1992. Hodnoty 5 až 8 označují T-cell, B-cell, Null cell a N K cell v lymfomech byly zavedeny programem SEER až počátkem osmdesátých let. Stupeň nádoru vyjadřuje, jak moc abnormální je vzhled rakovinných buněk pod mikroskopem. Indikuje, jak rychle bude pravděpodobně tumor narůstat a šířit se dál. Pokud jsou rakovinné buňky podobné těm normálním, tak jsou „well-differentiated" (dobře ohraničené). Tumory s takovými rakovinnými buňkami mají tendenci růst a šířit se pomalejším tempem než tumory spadající do oblasti „undifferentiated" (neohraničené) nebo „poorly differentiated" (špatně ohraničené), jejichž buňky mají abnormální vzhled nebo postrádají normální tkáňovou strukturu. Na základě těchto a jiných mikroskopických rozdílů mohou doktoři přiřadit stupeň k většině typů rakovin. Faktory určující stupeň tumoru se mohou lišit různými typy onkologického onemocnění. (American Cancer Society, 2014)



Obr. 27 Frekvenční analýza četnosti atributu Grade

Diagnostic confirmation

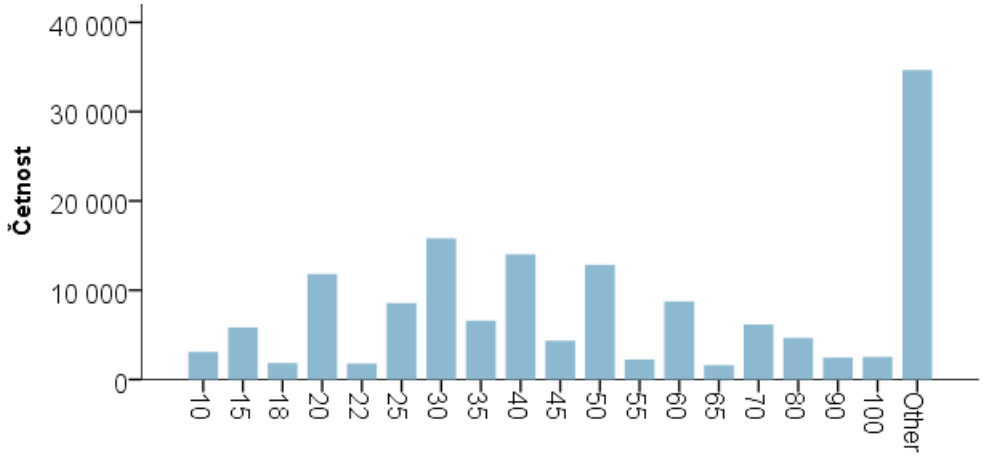
Atribut zaznamenává nejlepší použitou metodu k potvrzení přítomnosti nahlášené rakoviny. Záznamy nejsou limitované pouze na čas první diagnózy, ale určuje nejlepší metodu potvrzení za celou dobu průběhu nemoci.



Obr. 28 Frekvenční analýza četnosti atributu Diagnostic Confirmation

EOD – Tumor Size

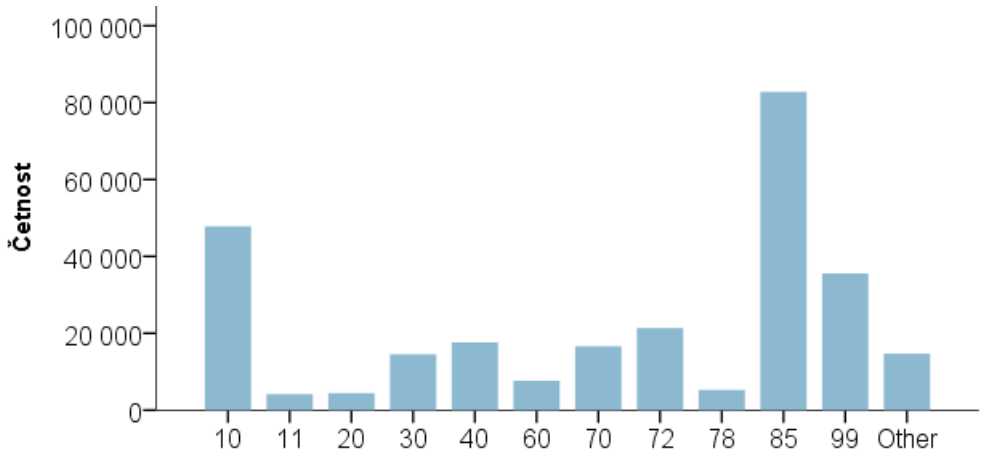
Informace určující největší velikost primárního tumoru v milimetrech. Hodnota 999 určuje neznámou velikost. Tuto informaci obsahují pouze záznamy diagnostikovany mezi lety 1988 – 2003.



Obr. 29 Frekvenční analýza četnosti atributu EOD - Tumor Size

EOD – Extension

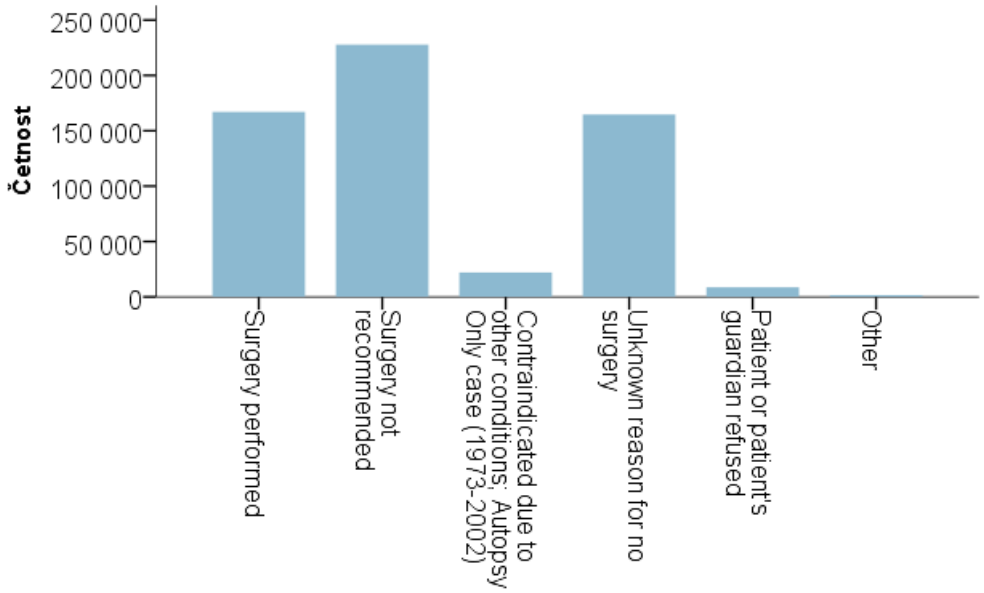
Kód, který dokumentuje nejvzdálenější rozšíření rakoviny od primárního ložiska, buď kontinuální rozšířením nebo metastází. Kód platný pouze pro záznamy vytvořené mezi lety 1988 až 2003.



Obr. 30 Frekvenční analýza četnosti atributu EOD Extension

Reason for no Surgery

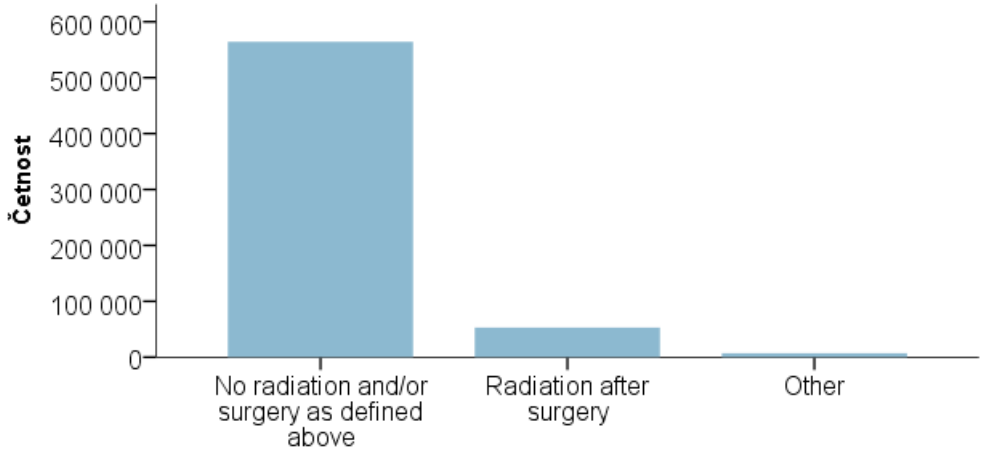
Atribut dokumentuje odůvodnění, proč nebylo přikročeno k chirurgickému zákroku v místě vzniku primárního tumoru.



Obr. 31 Frekvenční analýza četnosti atributu Reason for no surgery

RX SUMM – SURG/RAD SEQ

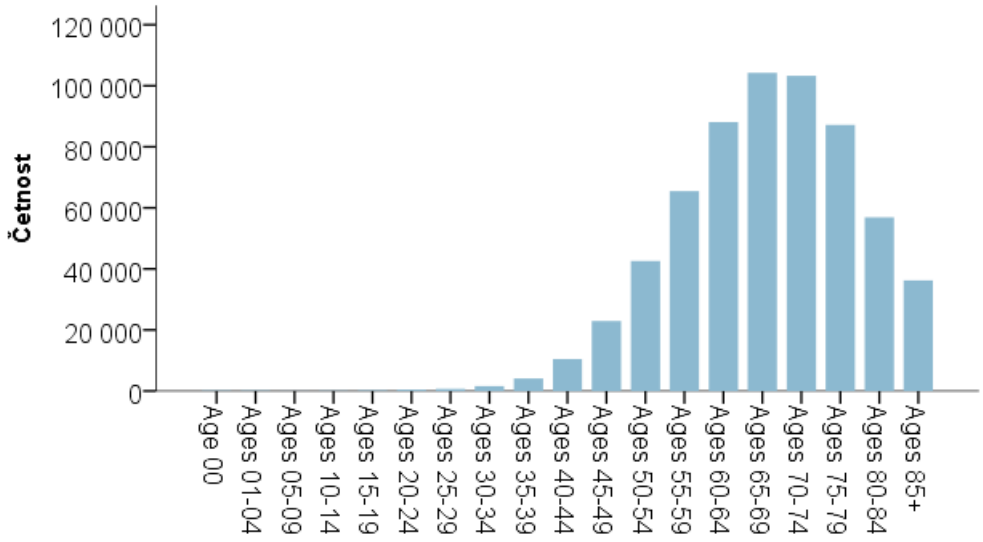
Tento atribut zaznamenává pořadí, ve kterém byla chirurgická a radiační terapie provedena. Pro pacienty, kteří měli jak chirurgickou tak i radiační terapii.



Obr. 32 Frekvenční analýza četnosti atributu RX Summ Surg/Rad Seq

Age Recode < 1 year olds

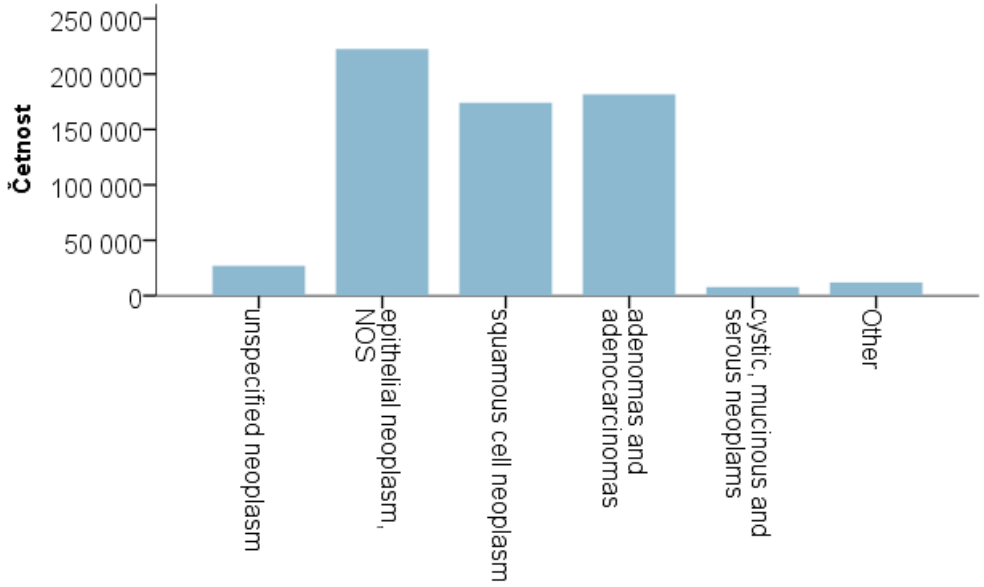
Atribut je založen na základě atributu Age at Diagnosis. Věkové skupiny jsou určeny podle údajů o počtu obyvatel. Rozděleno do 19 věkových skupin.



Obr. 33 Frekvenční analýza četnosti atributu Age Recode < 1 Year olds

Histologic Recode – Broad Groupings

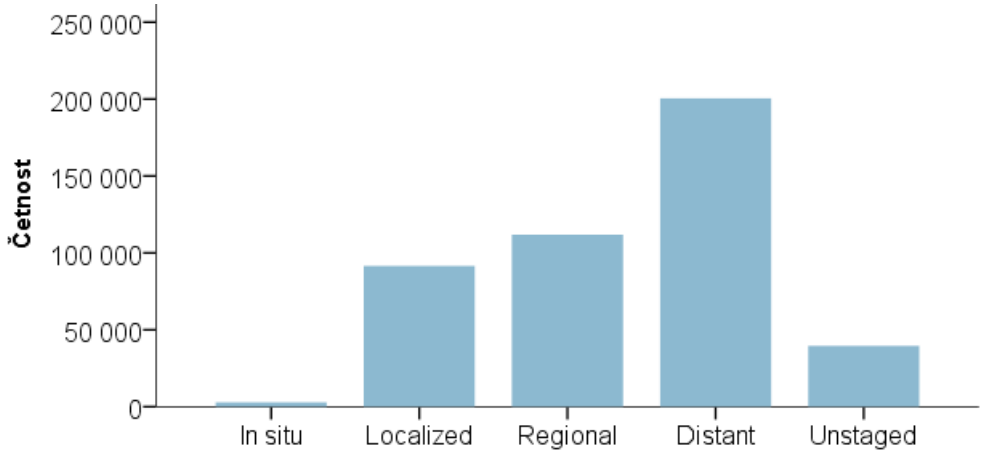
Založeno na atributu Histologic Type ICD-O-3. Specifikuje histologický druh neoplazmy.



Obr. 34 Frekvenční analýza četnosti atributu Histology - Broad Groupings

SEER historic stage A

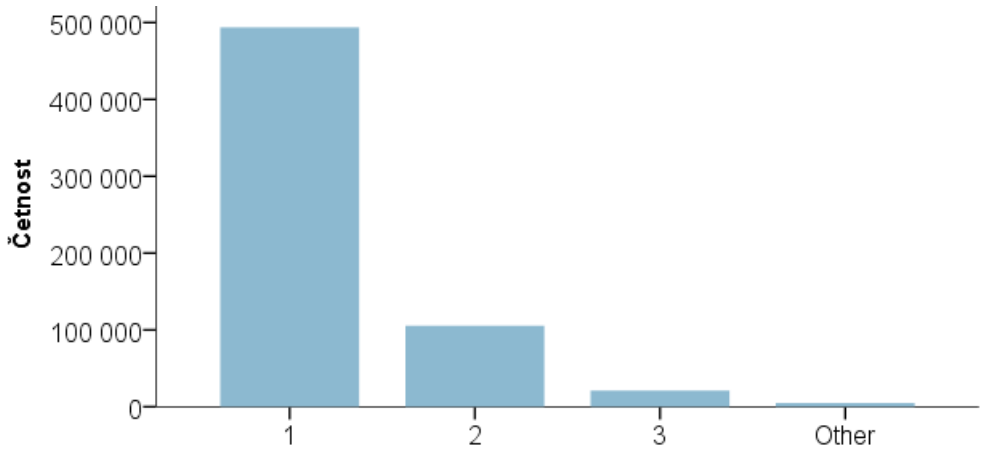
Vyjadřuje zjednodušenou verzi stádia rakoviny: in situ, lokální, regionální, vzdálená a neznámá. Stádium rakoviny je nejdůležitějším faktorem pro předpověď přežití nemocného pacienta. Rakovina plic se u pacientů začíná většinou projevovat až ve fázi IV (distant).



Obr. 35 Frekvenční analýza četnosti atributu SEER historic stage A

Number of Primaries

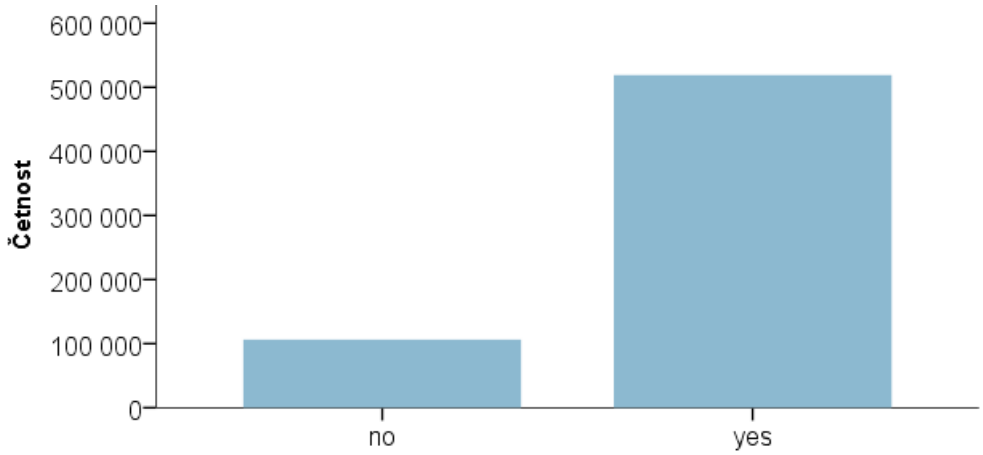
Číslo vyjadřující počet primárních nádorů u jednotlivých pacientů.



Obr. 36 Frekvenční analýza četnosti atributu Number of primaries

First malingnant primary indicator

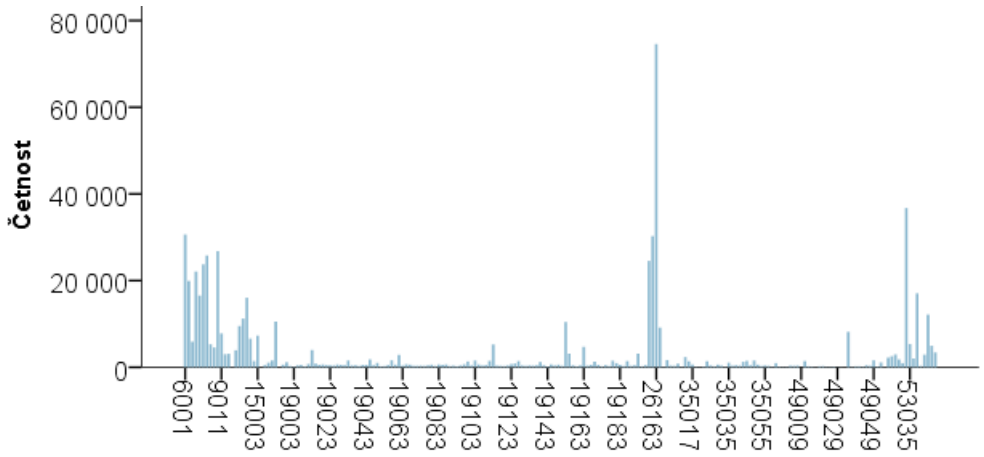
Určení zda jde o maligní nádory. Nádory, které nebyly nahlášené programu SEER, se předpokládají také jako maligní.



Obr. 37 Frekvenční analýza četnosti atributu First malignant primary indicator

State county recode

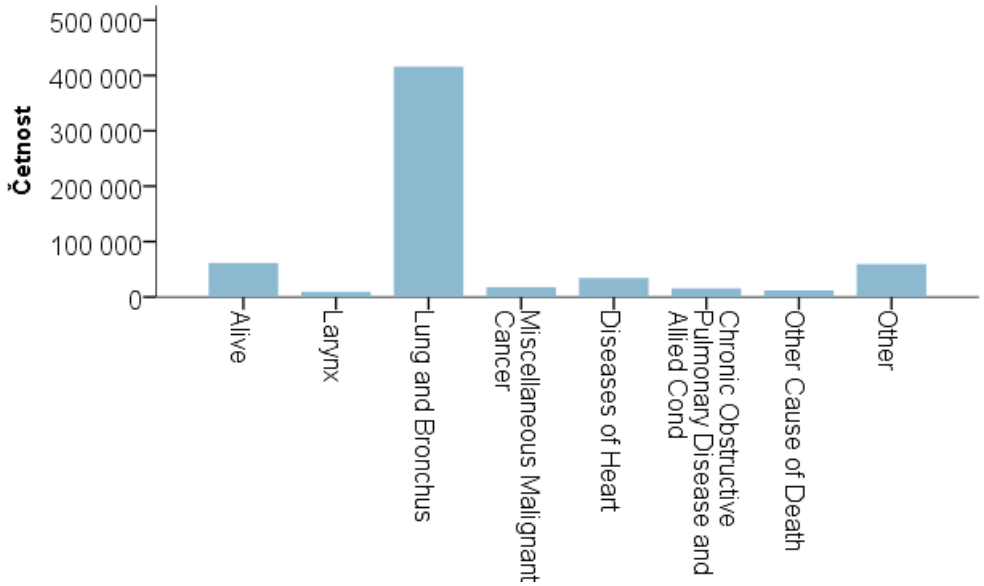
První dvě čísla jsou FIPS kódy definované standardem pro federální zpracování informací charakterizují stát a poslední tři čísla jsou také FIPS kódy, které definují okres.



Obr. 38 Frekvenční analýza četnosti atributu State county recode

Cause of Death to SEER site recode

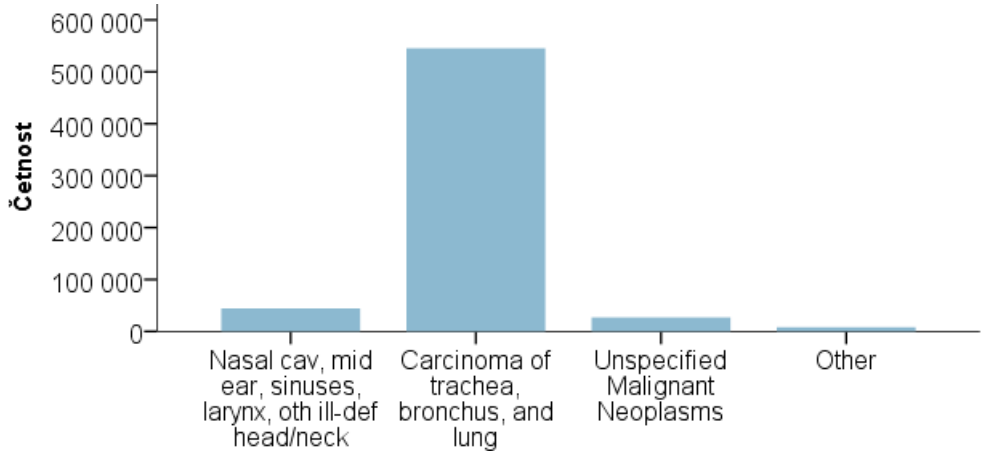
Určení příčiny úmrtí pro případy úmrtí působením rakoviny i pro úmrtí nezpůsobené rakovinou.



Obr. 39 Frekvenční analýza četnosti atributu Cause of Death to SEER site recode

AYA site recode WHO 2008

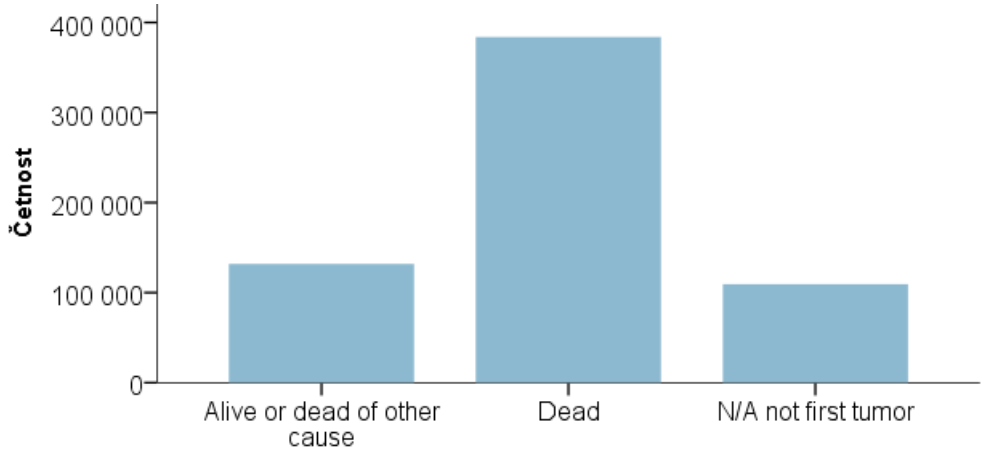
Atribut určený pro analýzu dat hlavně pro dospívající a mladistvé pacienty. Záznam byl aplikován i na dospělé, aby data byla kompletní a mohly být provedeny operace s tímto atributem.



Obr. 40 Frekvenční analýza četnosti atributu AYA site recode WHO 2008

SEER Cause Specific Death Classification

Slouží k určení specifické příčiny přežití. Tato proměnná označuje, zda osoba zemřela na důsledky rakoviny nebo přežila či zemřela z jiných důvodů, nebo to nebyl první nález. Prognóza přežití u rakoviny plic je obecně špatná. První projev nemoci je ve fázi IV, kdy šance na přežití v prvních pěti letech od diagnózy je 15 %. To vysvětluje velký nepoměr hodnot této proměnné.



Obr. 41 Frekvenční analýza četnosti atributu SEER Cause Specific Death Classification

Takto zpracovaná data byla z IBM Statistics exportovaná do souboru formátu .sav a použita jako vstupní soubor v IBM SPSS Modeler, kde byla dále zpracována. V příloze se nachází ukázky vytvořených streamů. Díky využití IBM Statistics nebylo již potřeba vytvářet uzly na počáteční přípravu, analýzu dat a na definování typů, měřítek hodnot a chybějících hodnot vstupního souboru. Vše se importovalo tak, jak bylo nastaveno v předchozích krocích. Ve vstupním souboru bylo potřeba jen určit roli ID atributu pro *Patient_ID_Number*.

4.2.5 Kvalita dat

Součástí porozumění dat je také posouzení kvality dat, která je při dolování dat klíčová. Kvalita dat má významný vliv na rychlost rozhodovacích procesů, rychlost klasifikace a na validitu výsledků. Důležitým ukazatelem kvality dat je množství chybějících hodnot, které se v SPSS posuzují pomocí uzlu „Data Audit“. V původním datovém souboru se nacházely atributy, které byly kompletní i z méně než 20 % záznamů, ty vzhledem ke své vysoké neúplnosti byly pro další práci vynechány. Celkový přehled úplnosti záznamů popisuje následující tabulka Tab. 1, ve které je procentuální vyjádření úplnosti atributů, počet záznamů, nulové hodnoty (tj. unknown values) a prázdné hodnoty. Větší podíl chybějících hodnot byl zaznamenán u atributů *Grade, EOD Tumor Size, EOD Extension, SEER historic Stage A.* Na nejhorší úrovní byl atribut *EOD Tumor Size*, který vyjadřuje velikost nalezeného tumoru. Tento atribut byl kompletní pouze z 23,92 %. Zbytek proměnných obsahovaly víc jak 93 % kompletních hodnot nebo neobsahoval žádnou chybějící hodnotu. Většinou byly atributy v počtu chybějících hodnot na dobré úrovni.

Tab. 1 Hodnocení kvality atributů

Atribut	Úplnost (%)	Validní záznamy	Nulové hodnoty	Prázdné hodnoty
Patient ID number	100.0	624997	0	0
Registry ID	100.0	624997	0	0
Marital Status at DX	96.58	603664	0	21333
Race Ethnicity	99.89	624350	0	647
Spanish Hispanic Origin	99.73	623330	0	1667
Sex	100.0	624997	0	0
Age at diagnosis	100.0	624997	0	0
Year of Birth	99.99	624982	15	0
Sequence Number Central	99.99	624947	0	50
Month of diagnosis	100.0	624997	0	0
Year of diagnosis	100.0	624997	0	0
Primary Site	100.0	624997	0	0
Laterality	100.0	624997	0	0
Histologic Type ICD-O-3	100.0	624997	0	0
Grade	51.98	324887	0	300110
Diagnostic Confirmation	100.0	624997	0	0
EOD Tumor Size	23.92	149553	352587	475444
EOD Extension	43.58	272410	352587	0
Reason for no surgery	94.65	591566	0	33431
RX Summ Surg Rad Seq	99.86	624150	0	847
Age Recode 1 Year olds	99.99	624982	0	15
Histology Recode Broad Groupings	100.0	624997	0	0
SEER historic stage A	71.41	446369	178628	0
Number of primaries	100.0	624997	0	0
First malignant primary indicator	100.0	624997	0	0
State county recode	100.0	624997	0	0
Cause of Death to SEER site recode	100.0	624997	0	0
AYA site recode WHO 2008	100.0	624997	0	0
SEER Cause Specific Death Classification	100.0	624997	0	0

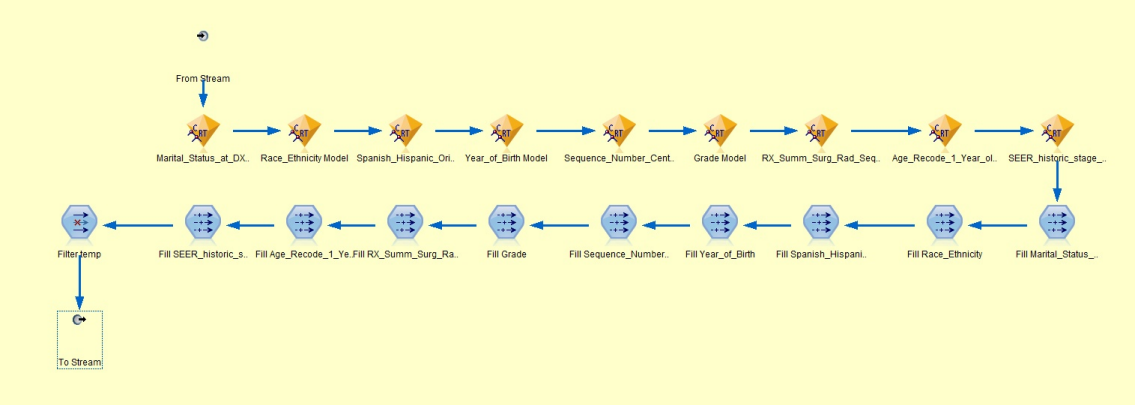
Výsledkem uzlu *Data audit* jsou také extrémní, průměrné, odlehlé hodnoty a medián. Zkoumané data byly ve většině případů nominálního typu, u kterých by byla analýza těchto hodnot bezúčelná. Extrémy a průměrné hodnoty lze analyzovat pouze u kontinuálních atributů, které popisuje následující tabulka.

Tab. 2 Extrémy a průměrné hodnoty kontinuálních atributů

Atribut	Min	Max	Průměr	Median
Age at diagnosis	0	106	67.48	68
Year of Birth	1871	2012	1926	1926
Year of diagnosis	1973	2012	1994	1995
EOD Tumor Size	0	998	43.53	36
EOD Extension	0	99	61.32	72
Number of primaries	1	9	1.26	1

4.2.6 Nahrazení chybějících hodnot

Výpočet byl proveden pomocí algoritmu The Classification and Regression Tree (dále už jen C&R Tree). Tento algoritmus využívá metodu rekurzivního dělení trénovacích záznamů do segmentů s podobnými výstupními hodnotami. C&R Tree začíná zkoumáním vstupních hodnot a hledá nejlepší rozdělení, které měří dle snížení indexu nečistot, které vyplívá z rozdělení. Rozdělení definuje binární strom, který končí ve chvíli, kdy je splněna jedna z kritérií. Dále dopočítává chybějící hodnoty. Před výpočtem lze nastavit velikost vzorku dat, nad kterým výpočet probíhá. Z vyzkoušených možností byly nejlepší výsledky dosaženy s hodnotou 20 % vzorku dat. Pro každý atribut byl zvlášť vypočten model C&R Tree a nové hodnoty byly vloženy do původní tabulky. Výsledkem byl „SuperNode“, který obsahoval čistá a doplněná data. Dopočtem hodnot odpadla povinnost vytvářet filtr pro odstranění atributů s nižší kvalitou.



Obr. 42 Obsah uzlu SuperNode

V dalším kroku bylo potřeba odfiltrovat atribut *Reason_for_no_Surgery* z důsledku předchozí reklasifikace a vytvoření nového zástupného atributu, s kterým se bude dále pracovat. V atributech týkajících se stádia rakoviny (*SEER_historic_stage_A),* v nově vzniklém *Surgery* a *SEER_Cause_Specific_Death_Classification* se nacházely záznamy s neznámými hodnotami. Bylo zapotřebí vytvořit uzly „Select“

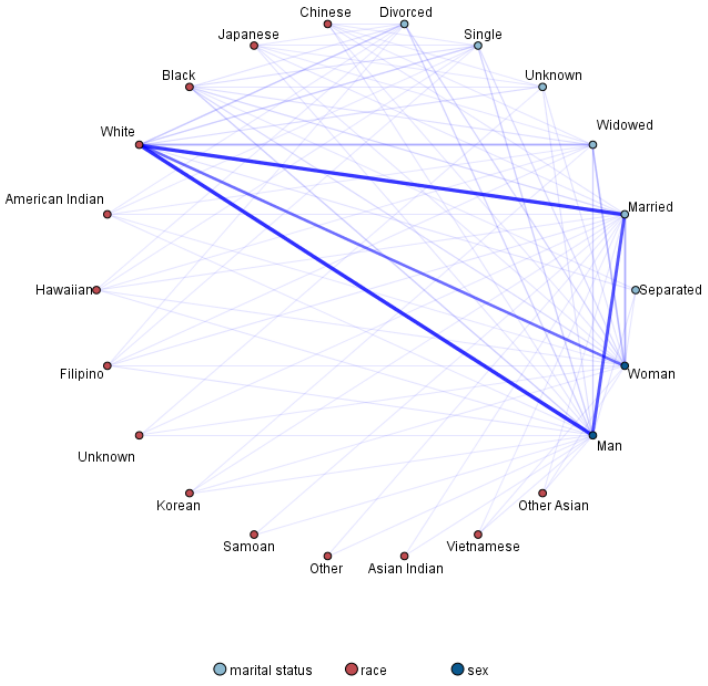
s podmínkami k odstranění takových záznamů. Za pomocí interpretace hodnot v datovém slovníku byly napsány podmínky pro jednotlivé atributy následovně:

- not (SEER_historic_stage_A = 9) and not (SEER_historic_stage_A = 8) and not (SEER_historic_stage_A =0),
- (Surgery = „recommended") and not (Surgery = "unknown") and not (Surgery = „recomended, unknown if done“) and not (Surgery = 9),
- not (SEER_Cause_Specific_Death_Classification = 9).

4.2.7 Prověřování, seřazení a výběr atributů

Pomocí uzlu „Table“ bylo možné si zobrazit kompletní tabulku celého datového souboru. Následně proběhla reklasifikace atributu *Reason_for_no_surgery* na obecnější skupiny*: performed, not recommended, recommended, not performed, recomended (unknown if done) a unknown*, které vycházely z původních hodnot. Výsledkem byl vznik nového atributu *Surgery.* V dalším kroku se pomocí uzlu „Type“ zkontrolovala správnost nastavených atributů, v tomto uzlu byly postupně nastavovány všechny cílové atributy, které byly zkoumány.

S takto připravenými daty bylo možné vytvořit demografickou analýzu zkoumaných pacientů, kteří se podíleli na sběrech dat. Pomocí grafu „Web“ byly zobrazeny nejsilnější vztahy mezi atributy *Marital_Status_at_DX, Sex* a *Race_Ethnicity.* Dle grafu lze vyčíst nejvýznamnější populační skupiny: bílí muži v manželském vztahu následování vdanými ženami bílé rasy.



Obr. 43 Graf demografické analýzy

Řešení problémů dolování dat může zahrnovat někdy stovky až tisíce proměnných, které mohou být potenciálně využity jako vstupní atributy. To by znamenalo velkou spoustu času a úsilí strávenou nad analýzou každého atributu a rozhodování, zda jej zahrnout do modelu. Tímto procesem si musí částečně každý analytik projít, ale může si ušetřit trochu času pomocí algoritmu uzlu „Feature Selection“, který vybere atributy, které jsou nejdůležitější pro danou analýzu. „Feature Selection“ prochází třemi kroky:

- Prověřování: odstraňuje nedůležité a problematické vstupy a záznamy nebo případy, kdy vstupy mají příliš mnoho chybějících hodnot a také příliš mnoho nebo příliš málo variací, aby mohly být užitečné pro modelování.
- Seřazení: řadí zbývající vstupy podle důležitosti.
- Výběr: identifikuje podmnožinu funkcí k použití v dalších modelech například tím, že zachová pouze nejdůležitější vstupy a odfiltruje nebo vyloučí všechny ostatní.

Podle zvolených kritérií uzel vybral ty prediktory, které splňují požadavky na kvalitu dat a zbytek byl vyloučen. Počet atributů byl snížen na 21.

Zvolená kritéria:

- maximum percentage of missing values: 70 %,
- maximum percentage of records in a single category: 90 %,
- maximum number of categories as a percentage of records: 95,
- minimum coefficient of variation: 0,1,
- minimum standard deviation: 0,0.

Výsledkem byl nugget[3], který vyloučil šest atributů: *Age_at_diagnosis*, z důvodu nízké důležitosti, *Year_of_diagnosis* a *Year_of_birth,* z důvodu nízkého koeficientu variace. Atributy *RX_Summ_Surg_Rad_Seq, First_malignant_primary_indicator* a *AYA_Site_recode_WHO_2008* identifikoval jako prediktory s příliš širokou škálou možných hodnot. Při analýze dat bylo zjištěno, že tyto atributy jsou poměrně důležité, proto bylo potřeba se vrátit zpět do reklasifikační části streamu a překlasifikovat tyto tři stávající atributy na nové atributy s novými skupinami hodnot, čímž se snížil počet možných hodnot. Díky definování nových reklasifikačních uzlů tyto atributy po znovuspuštění algoritmu „Feature Selection“ nebyly vyloučeny.

4.2.8 Nastavení rovnováhy

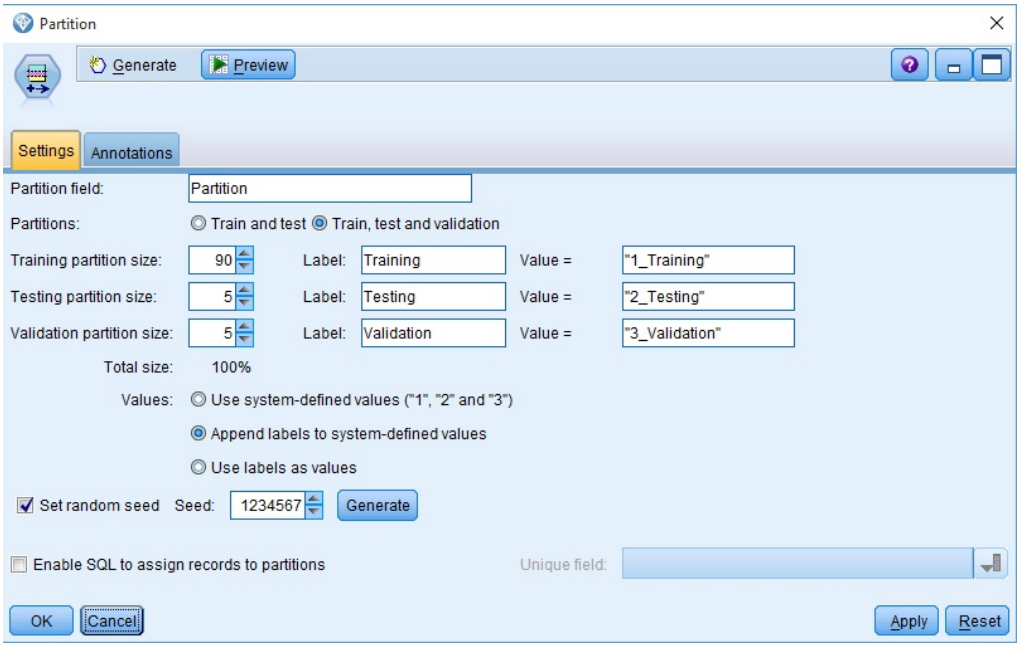
Ve fázi popisu dat si bylo možné všimnout jisté nerovnováhy mezi některými atributy. Například v případě atributu *SEER_Cause_of_Death_Classification* vládne ne-

[3] Nugget je kontejner modelu, jenž představuje soubor pravidel, vzorců nebo rovnic. Představuje výsledky modelu v IBM SPSS Modeler. Hlavním smyslem nuggetu je ohodnocení dat, která je potřebná ke generování predikcí nebo k spuštění další analýzy modelu.

rovnováha mezi pacienty, kteří přežili a ty, kteří nemoci podlehli nebo skonali z jiných než onkologických důvodů. V případě, že se rozhodneme takový atribut použít jako cílový, je důležité napravit nerovnováhu v datasetu. Atribut má pouze 2 hodnoty, do první skupiny spadá 23,81 % záznamů a do druhé skupiny spadá 76,19 % záznamů. Taková situace pro řadu modelovacích technik představuje zásadní problém. Některé modely by měly tendenci učit se pouze z menší skupiny, protože je vzácnější a do určité míry ignorovat vetší skupinu. V případě, že jsou údaje dobře vyvážené, budou mít modely lepší šanci najít vzory, které rozlišují obě skupiny. Poměrně jednoduché řešení se skrývá v podobě uzlu *„Balance".* Vyvažování se provádí duplikací a následným vylučováním na základě podmínek, které se zadají. Záznam, který nesplňuje žádnou podmínku, prochází dále. Proces funguje na základě duplikace nebo vyřazení záznamů, důsledkem je ztráta původní sekvence dat, proto se všechny operace a derivace dat provádí před aplikováním tohoto uzlu. Tento uzel byl použit v případě všech cílových atributů, které byly v průběhu práce vyzkoušeny. Tím byla za cenu nižší přesnosti (průměrné v řádech jednotek procent) dosažena podstatně vyšší validita výsledků. Nerovnováhy byly zjišťovány pomocí distribučního grafu, na jehož základě byly sepisovány vstupní podmínky pro uzel „Balance*".*

4.2.9 Rozdělení dat

Uzel „Partition" se používá k vytvoření oddílů, které rozděluje data do samostatných podskupin pro trénování, testování a validaci. Použitím jednoho vzorku na vytvoření modelu a dalšího vzorku k testování, lze získat údaj, který vypovídá o tom, jak dobře bude model generalizovat nad rozsáhlejšími daty. Trénovací, testovací a validační data byla rozdělena v poměru 90:5:5. Ze všech vyzkoušených poměrů vracela tato konfigurace nejlepší výsledky.



Obr. 44 Konfigurace uzlu Partition

4.3 Modelování

Z pohledu strojového učení se predikce na těchto cílových atributech řadí do skupiny učení s učitelem. „Učitel“ poskytuje systému přímou informaci o požadovaném chování. Atributy (proměnné) se rozdělují na cílové a na vstupní (necílové). Cílové atributy, tj. atributy predikované jsou v tomto případě atributy:

- *SEER_Cause_of_Death_Classification*,
- *Surgery*,
- *Histology_recode_Broad_Groupings*.

První atribut určuje, zda pacient přežil nebo nepřežil onkologickou nemoc, druhý atribut určuje, zda pacient byl či nebyl chirurgicky operován a poslední atribut je typ histologie. Ostatní atributy jsou vstupní, tedy necílové. Na základě znalostí, které specifikují, jaké proměnné jsou cílové a jaké vstupní, lze odvodit znalosti, které jsou potřebné pro klasifikaci nových objektů. Tato skupina objektů je označována jako trénovací množina a ověření schopnosti získaných znalostí se provádí nad množinou testovací. Dále bylo využito množiny dat, která se nazývá validační a slouží k optimalizaci parametrů. Všechny tyto tři množiny jsou disjunktní.

Na každém cílovém atributu bylo vyzkoušeno více modelů, pokaždé byl vybrán ten, který prokázal nejlepší výsledky. Výběr modelů byl uskutečněn na základě vyhodnocení pomocí uzlu „Auto Classifier“, který sestavuje seznam modelů, které se nejvíce hodí pro aktuální zkoumaná data. Výběr je prováděn na principu úspěšnosti predikce navrhovaných modelů. U každého vytvořeného modelu byla

vyzkoušena různá kombinace konfigurací a parametrů, následně byla vybrána ta, která měla nejpřesnější, nejzajímavější a zároveň validní výsledky.

4.3.1 Rozhodovací strom C5.0

Prvním aplikovaným algoritmem byl C5.0, který se řadí do skupiny rozhodovacích stromů s jednoduchou uživatelskou interpretací výsledků. Pro práci je nutné určit pouze jeden cílový atribut kategoriální úrovně měření (nominal, categorical, flag nebo ordinal), v tomto případě byl pro modelování zvolen atribut, který vyjadřuje příčinu úmrtí pacienta (*SEER_Cause_Specific_Death_Classification).* Atribut nabývá pouze dvou hodnot:

- Alive or dead of other cause,
- Dead.

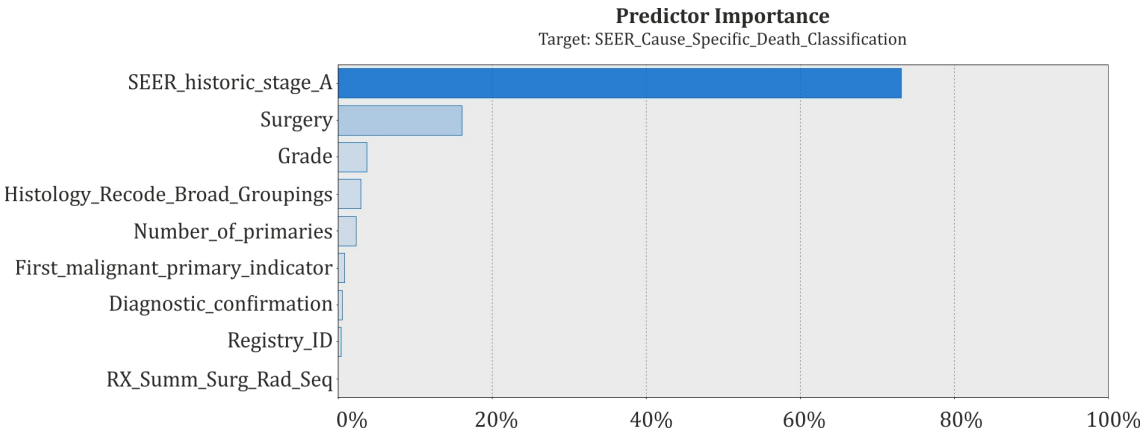
Výsledkem je seřazení prediktorů dle důležitosti, které vedou k dvěma zmíněným stavům pacienta. Všechny vstupní atributy musí mít definovanou úroveň měření. Cílový atribut může být libovolné úrovně měření.

Algoritmus byl nastaven tak, aby používal pouze trénovací data k tvorbě modelu a tvořil samostatné modely pro každou možnou hodnotu vstupních atributů, které jsou určené jako rozdělovací atributy. Dále algoritmus byl nastaven tak, aby spojoval symbolické hodnoty, které mají podobné vzory s ohledem na výstupní atribut. V nastavení byla zvolena možnost křížové validace, která zjišťuje, jak moc bude model za využití statistické analýzy ovlivňovat nezávislé vzorky dat. Počet modelů využitých pro křížovou validaci byl nastaven na hodnotu 10. Rovněž byl povolen mód „Expert“ z důvodu nepřehledného výsledného rozhodovacího stromu, který byl příliš velký. Mód „Expert“ umožňuje nastavit míru prořezávání, která zmenšuje velikost a zvyšuje stručnost rozhodovacího stromu. Dále byl nastaven minimální počet potomků na hodnotu 2, která zabraňuje přetrénování modelu. Model taktéž využíval metody „globálního prořezávání“, který odstraňuje slabé podstromy a metody „prosetí atributů“, které analyzují užitečnost prediktorů ještě před tvorbou samotného modelu a vedou k eliminaci těch, které jsou označeny jako irelevantní.

Nastavení modelu:

- use partitioned data: true,
- calculate predictor importance: true,
- calculate raw propensity scores: false,
- calculate adjusted propensity scores: false,
- use weight: false,
- output type: Decision tree,
- group symbolics: true,
- use boosting: false,

- number of trials: 10,
- cross-validate: true,
- mode: Expert,
- pruning severity: 100,
- minimum records per child branch: 2,
- winnow attributes: true,
- use global pruning: true,
- use misclassification costs: false.



Obr. 45 Důležitost prediktorů algoritmu C5.0 Decision Tree

Další C5.0 model se stejnou konfigurací byl vytvořen pro účely vygenerování rozhodovacích pravidel, které obsahují předpisy pro predikci cílového atributu. Bylo získáno celkem 29 pravidel. 16 pravidel pro pacienty, kteří přežili onemocnění a 13 pravidel pro ty, kteří nemoci podlehli. Kompletní seznam všech pravidel lze nalézt v příloze.

4.3.2 Neuronová síť

V případě použití neuronových sítí může mít cílový atribut nastavený libovolný typ úrovně měření. V tomto případě byl jako cílový atribut nastaven *Surgery,* který je typu nominal, a určuje, zda byl či nebyl proveden chirurgický zákrok. Nabývá třech možných hodnot:

- Performed,
- Not performed,
- Not recommended.

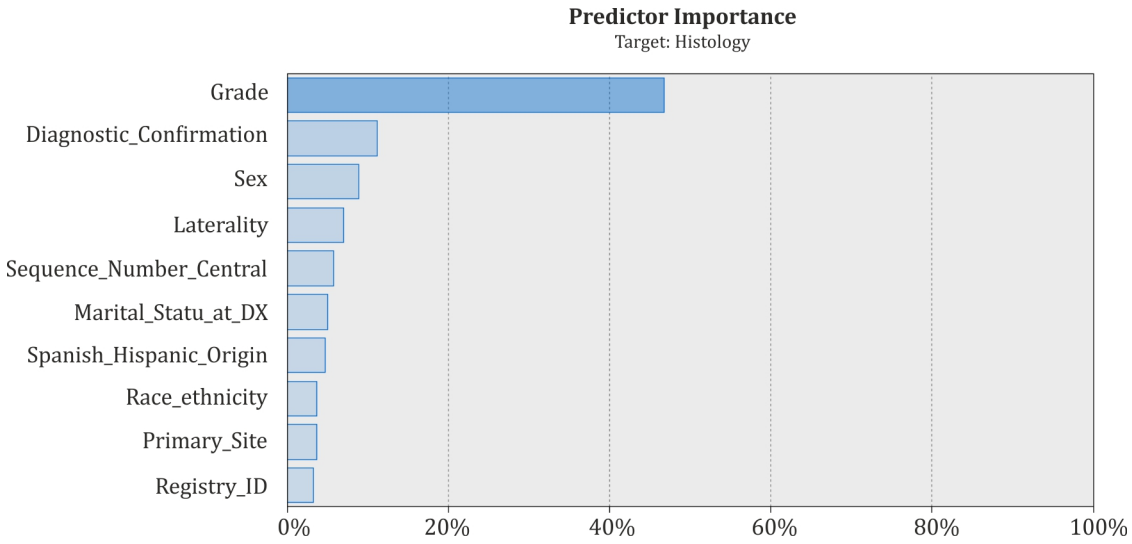
Všech ostatních 22 atributů, které slouží jako vstupní proměnné, mohou být také libovolné úrovně měření.

Algoritmus neuronové sítě byl nastaven tak, aby vytvořil nový standardní vícevrstvý perceptron (dále jen MLP), který dle doporučení IBM vykazuje lepší vý-

sledky než radiální bázová funkce (RBF). Testování na vlastních datech toto tvrzení potvrdilo. Model měl nastavený automatický počet skrytých vrstev. MLP síť poskytuje možnost nastavit ukončující pravidlo, které zabraňuje přetrénování modelu. Na výběr je mezi časovým omezením trénovacího času, maximálním počtem trénovacích cyklů a dosažením minimální přesnosti modelu. Z vyzkoušených metod se nejvíce osvědčilo časové omezení 15 minut. Byla vyzkoušena metoda „bagging", která náhodným výběrem s opakováním vytváří několik trénovacích modelů, na kterých paralelně probíhá učení a metoda „boosting", která generuje posloupnost modelů, které mají za cíl získat přesnější predikce. Obě metody s různými ukončovacími pravidly vedly k přetrénování modelu. To bylo důvodem k vytvoření standardního modelu. V pokročilých nastaveních je možné určit, jakým způsobem se budou zpracovávat chybějící a nulové hodnoty ve vstupních atributech. Tato problematika byla již vyřešena ve fázi přípravy dat pomocí C&R Tree algoritmu, tudíž to již nebylo nutné řešit. Důležitost prediktorů je zobrazena na následujícím obrázku.

Nastavení modelu:

- what do zou want to do? Build a new model,
- what is your main objective?: Build a new model,
- neural network model: Multilayer Perceptron (MLP),
- hidden Layers: Automatically compute number of units,
- use maximum training time (per component model): true,
- minutes: 15,
- customize number of maximum training cycles: false,
- use minimum accuracy: false,
- overfit prevention set(%): 30.0,
- replicate Results: true,
- random seed: 229176228,
- missing values in predictors: Delete listwise.



Obr. 46 Důležitost prediktorů algoritmu Neural Network

4.3.3 Bayesovská síť

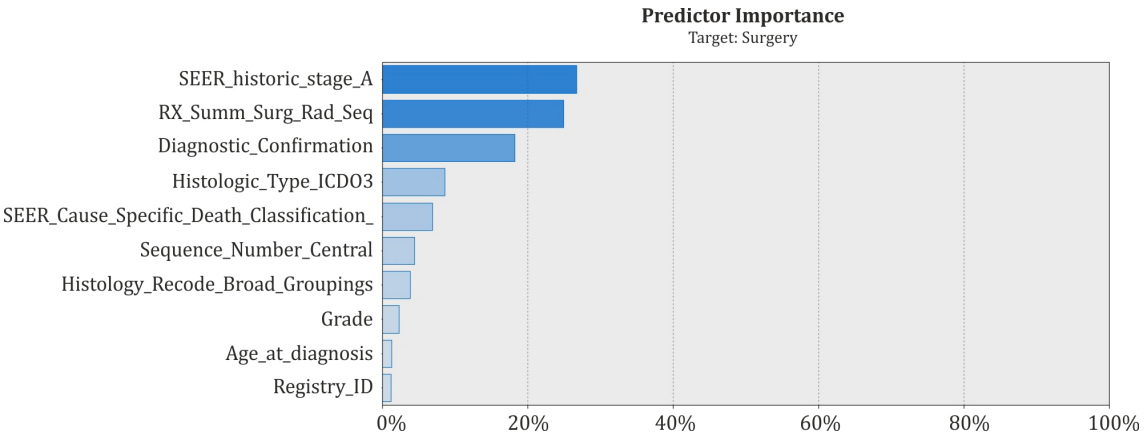
Požadavkem pro vytvoření modelu Bayesovské sítě je kategorický, nominální, ordinální nebo binární cílový atribut. Cílový atribut v tomto případě je nastaven na *Histology_recode_Broad_Groupings*, který představuje překlasifikaci atributu *Histologic_Type_ICDO3,* který musel být pro výpočet vyloučen ze vstupních atributů, jinak by model měl kvůli této proměnné 100% přesnost predikce. Cílová hodnota nabývá 38 možných hodnot, z kterých pouze 4 měly dominantní zastoupení, z toho jedna vyjadřovala neznámý histologický typ a zbylých 34 hodnot mělo zanedbatelné zastoupení:

- epithelial neoplasm, NOS,
- squamous cell neoplasm,
- adenomas and adenocarcinomas.

Vstupní proměnné mohou být jakéhokoliv typu. V konfiguraci modelu byla vybrána možnost sestavení modelu pro každou možnou hodnotu vstupních atributů, které jsou specifikované jako dělící atributy. Byla upřednostněna struktura typu TAN (The Tree Augmented Naive Bayes model), která vytváří jednoduchou Bayesovskou síť a zlepšuje standardní naivní model. Umožňuje závislost každého prediktoru na dalším prediktoru kromě cílového atributu, čímž se zvyšuje přesnost. Dále byla vybrána „Maximální pravděpodobnosť", která se používá pro velké data sety. V expertním módu bylo přeskočeno ošetření nekompletních záznamů. Dále byl proveden test nezávislosti, kde se posuzuje, zda párová pozorování dvou proměnných jsou na sobě nezávislá. Typ testu nezávislosti byl nastaven na pravděpodobnostní poměr, úroveň významnosti na 0,01, maximální velikost podmiňovacího setu na hodnotu 5 a maximální počet vstupů ponechán na defaultních 10.

Nastavení modelu:

- use partitioned data: true,
- calculate predictor importance: true,
- calculate raw propensity scores: false,
- calculate adjusted propensity scores: false,
- use frequency field: false,
- continue training existing model: false,
- structure type: TAN,
- include feature selection preprocessing step: true,
- parameter learning method: Maximum likelihood,
- mode: Expert,
- use only complete records: true,
- append all probabilities: flase,
- independence test: Likelihood ratio,
- significance level: 0,01,
- maximal conditioning set size: 5,
- maximum number of inputs: 10.



Obr. 47 Důležitost prediktorů algoritmu Bayesian Network

4.4 Hodnocení

Každý model byl podroben hodnotícímu uzlu „Analysis“, který sestavuje tabulku znázorňující přesnost modelu na trénovacích datech, na kterých při modelování probíhá učení a ověřování na datech testovacích a validačních.

Ke každému modelu byla provedena evaluace pomocí uzlu „Evaluation“. Výstupem tohoto uzlu je „Response chart“ obsahující křivky kumulativní četnosti, základní křivky a nejlepší křivky. Základní (červená) křivka vyjadřuje nejhorší možný případ a nejlepší (modrá) křivka značí nejlepší případ. Graf je opět rozdělen na trénovací, testovací a validační část. Jedná se o počet procent záznamů, které jsou správnými predikcemi. Vypočítáno jako:

$$\frac{X}{N} \cdot 100\%, \tag{7}$$

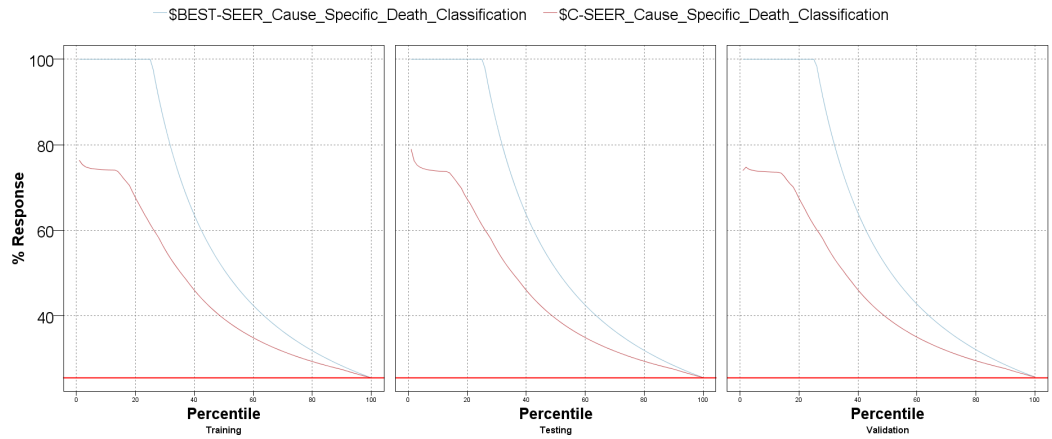
kde *N* je počet hitů (validních predikcí) v kvantilu a X je počet záznamů v kvantilu.

4.4.1 Rozhodovací strom C5.0

Byla zjištěna 81,92% správnost predikce na trénovacích datech (377 416 správných záznamů), 81,62% správnost na testovacích datech (21 001 správných záznamů) a 81,91% správnost na validačních datech (21 197 správných záznamů). V ≈18 % případů model predikoval přežití pacienta chybně, zařazením do nesprávné třídy.

Tab. 3 Analýza modelu C5.0 na trénovacích, testovacích a validačních datech

Partition	Training		Testing		Validation	
Correct	377 416	81,92%	21 001	81,62%	21 197	81,91%
Wrong	83 307	18,08%	4 730	18,38%	4 682	18,09%
Total	460 723		25 731		25 879	



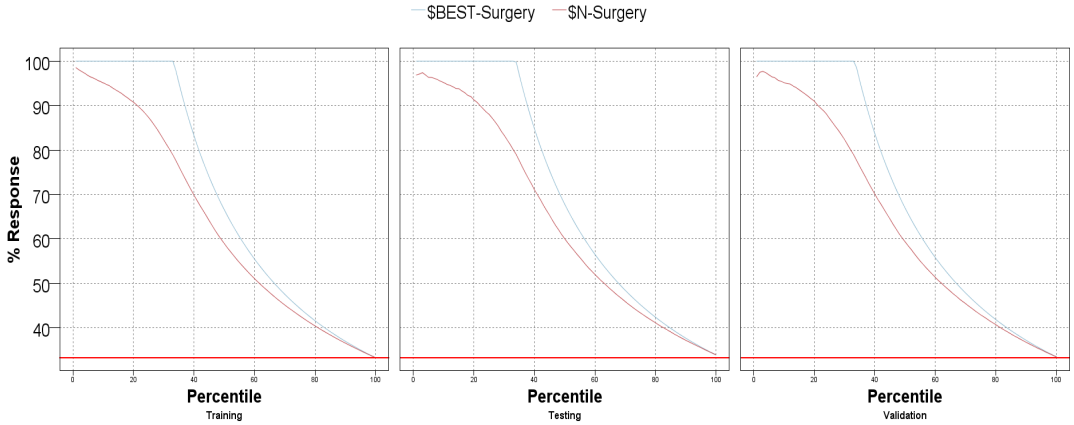
Obr. 48 Response chart modelu C5.0 Decision Tree

4.4.2 Neuronová síť

Neuronová síť získala 67,69% správnost predikce na trénovacích datech (311 760 správných záznamů), 68,03% správnost na testovacích datech (17 498 správných záznamů) a 67,7% správnost na validačních datech (17 515 správných záznamů). V ≈32 % případů model predikoval provedení chirurgického zákroku špatně, zařazením do nesprávné třídy.

Tab. 4 Analýza modelu Neural Network na trénovacích, testovacích a validačních datech

Partition	Training		Testing		Validation	
Correct	311 760	67,69%	17 498	68,03%	17 515	67,7%
Wrong	148 810	32,31%	8 224	31,97%	8 357	32,3%
Total	460 570		25 722		25 879	



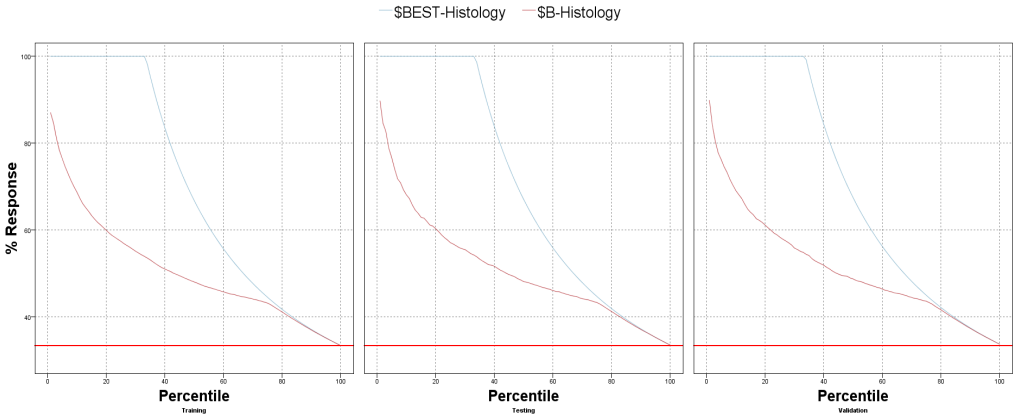
Obr. 49 Response chart modelu Neural Network

4.4.3 Bayesovská síť

Byla zjištěna 61,68% správnost predikce na trénovacích datech (162 236 správných záznamů), 61,77% správnost na testovacích datech (9 051 správných záznamů) a 61,84% správnost na validačních datech (9 170 správných záznamů). V ≈38 % případů model predikoval provedení chirurgického zákroku špatně, zařazením do nesprávné třídy.

Tab. 5 Analýza modelu Bayesian Network na trénovacích, testovacích a validačních datech

Partition	Training		Testing		Validation	
Correct	162 236	61,68%	9 051	61,77%	9 170	61,84%
Wrong	100 800	38,32%	5 601	38,23%	5 659	38,16%
Total	263 036		14 652		14 829	



Obr. 50 Response chart modelu Bayesian Network

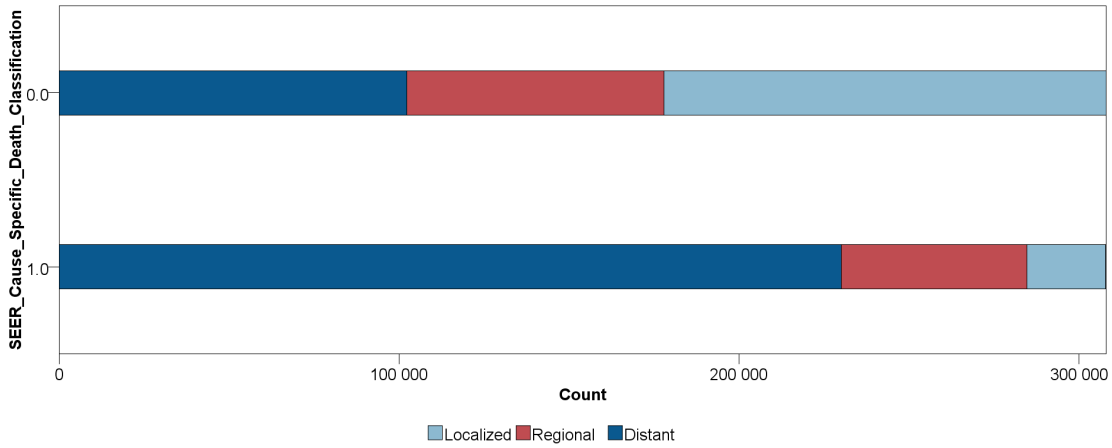
4.5 Interpretace výsledků

V předchozím kroku byly provedeny výpočty nad poskytnutými daty. Kompletní analýza se neomezuje pouze na výpočtech. Pro užitečnost výsledků musí být provedena interpretace výsledků a jejích věcná slovní formulace. Tuto etapu obvykle provádí analytik za pomocí odborníka, v tomto případě lékaře, protože se jedná o přelom mezi jazykem matematiky a informatiky i věcnou problematikou. V rámci této práce nebyla využita interpretace lékaře, ale vlastní interpretace za pomocí získaných vědomostí z nastudované problematiky plicního onkologického onemocnění. Postupně jsou interpretovány výsledky všech vygenerovaných modelů.

4.5.1 Rozhodovací strom C5.0

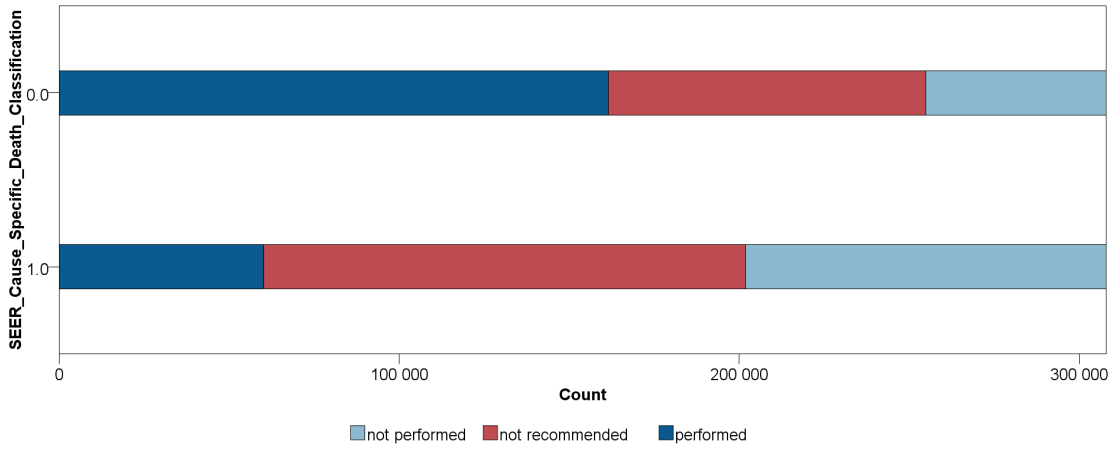
Rozhodovací strom C5.0 určil nejdůležitější prediktory pro příčinu úmrtí zkoumaných pacientů. Třemi nejsilnějšími faktory, které ovlivňují mortalitu pacientů s rakovinou plic, je vliv stádia rakoviny (73 %), skutečnost, zda byl či nebyl proveden chirurgický zákrok (16 %) a stupeň rakovinných buněk (4 %). V závěsu následují faktory: výsledky histologie (3 %), počet primárních nálezů (2 %), skutečnost zda jde či nejde o maligní nádor (1 %) atd.

Sestavením distribučního grafu závislosti příčiny smrti na vliv stádia rakoviny, kde 0 reprezentuje stav „živý nebo smrt z jiné než onkologické příčiny“ a 1 „mrtvý“, zjistíme patrné rozdíly mezi dvěma skupinami pacientů. Pacienti, kteří skonali v důsledku onemocnění rakoviny, mají výrazně vetší podíl případů, u kterých byl prokázán nález vzdálených metastáz rakoviny plic. U těchto onemocněných je dle výzkumů šance na přežití velmi malá. Naopak pacientům, kterým se podařilo překonat svoji nemoc nebo nakonec zemřeli v důsledků jiných než onkologických příčin, mají větší podíl lokálních nádorových nálezů v těle, u nich je šance přežití po včasné léčbě poměrně vysoká. Co se týče regionálních nálezů, tj. pokročilých lokálních nálezů rozšiřující se do regionálních lymfatických uzlin: u první skupiny bylo nalezeno o 20 808 více případů. Tato skupina má 32% až 77% šanci na přežití. Graf potvrzuje skutečnost, že se první projevy rakoviny plic u pacienta projevují až ve fázi IV (distant), kdy prognózy již nejsou příznivé.



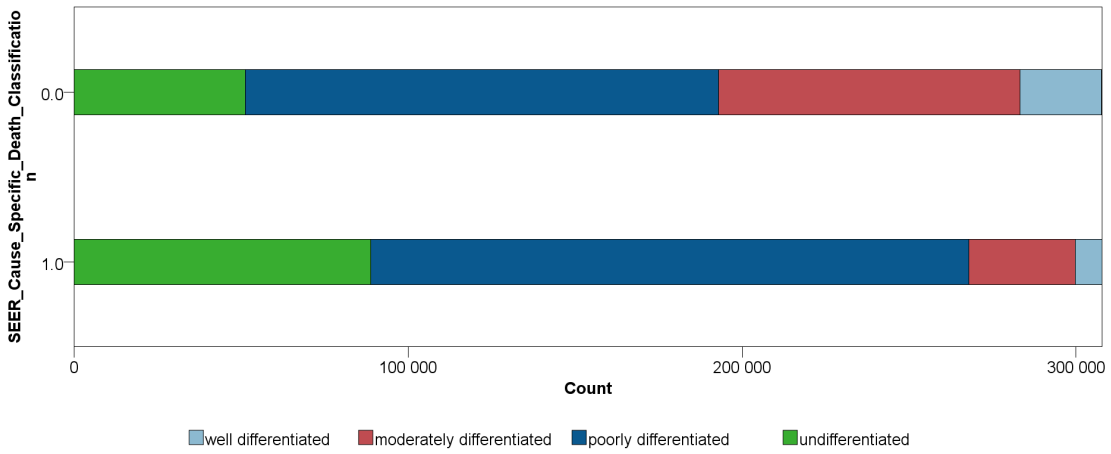
Obr. 51 Graf závislosti příčiny úmrtí na stádium rakoviny

Druhým nejdůležitějším faktorem na přežití pacienta je vliv chirurgického zákroku. Z grafu lze vyčíst, že provedený chirurgický zákrok má významný vliv na zotavení pacienta. V případech, kdy nebyl zákrok vůbec proveden, nebo nebyl z jiných důvodů doporučen, byla šance na přežití podstatně menší.



Obr. 52 Graf závislosti příčiny úmrtí a chirurgické terapie

Třetí nejsilnější vliv na zkoumaný cílový atribut má stupeň rakovinných buněk. Většina zesnulých spadá do skupiny pacientů se špatně ohraničenými rakovinnými buňkami nebo do skupiny s neohraničenými buňkami. Buňky zařazené do těchto skupin stupňů mají abnormální vzhled nebo postrádají normální tkáňovou strukturu. Tendence rychlejšího růstu a šíření jsou vyšší, tím pádem šance na přežití je menší.



Obr. 53 Graf závislosti příčiny úmrtí na stupni rakovinných buněk

Z vygenerovaných pravidel je možné vyčíst přesné podmínky, pomocí kterých lze určit, do jaké třídy pacient spadá. Příklady vygenerovaných pravidel:

if Histologic_Type_ICDO3 = 8240
and Grade = 2
and Number_of_primaries <= 1
and Surgery = performed
then 0

Interpretace: ookud histologie prokázala, že se jedná o středně ohraničený (stupeň 2) maligní rakovinný tumor, počet primárních nálezů se rovná nebo je menší než 1 a chirurgický zákrok byl uskutečněn, poté pacient přežil (nebo zemřel z jiných než onkologických důvodů).

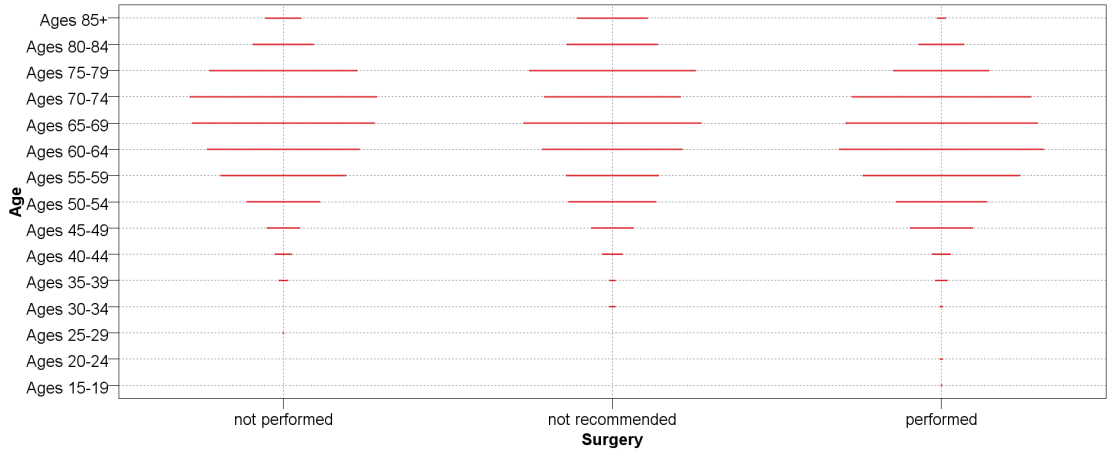
if Grade = 2
and Histology_Recode_Broad_Groupings = 2
and Surgery = not performed
then 1

Interpretace: ookud jsou rakovinné buňky středně ohraničené (stupeň 2), jedná se o skvamozní karcinom dlaždicového typu a chirurgický zákrok nebyl proveden, tak pacient skonal.

4.5.2 Neuronová síť

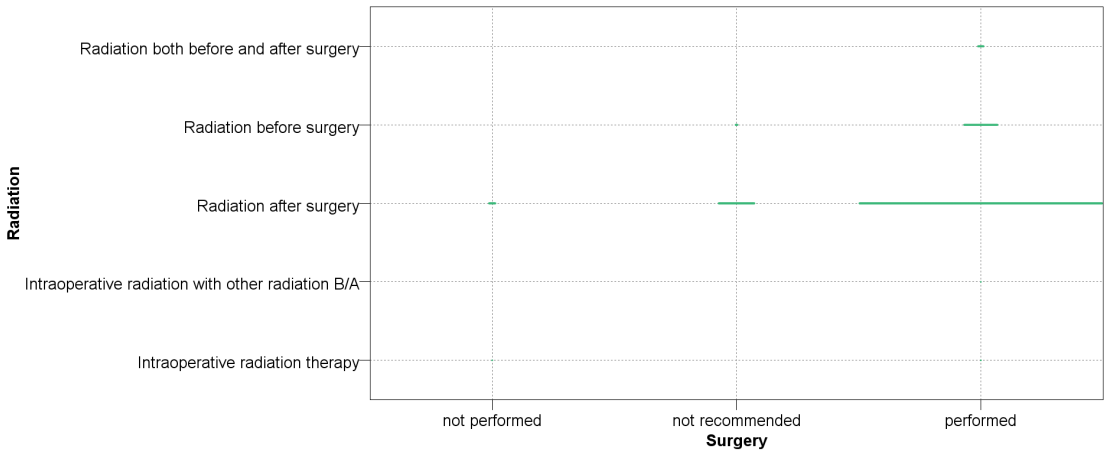
Nejdůležitějšími kritickými faktory, které dle neuronových sítí ovlivňují, zda pacient půjde či nepůjde na operaci nebo jestli operace není pacientovi doporučena, je věk pacienta při diagnóze (15 %), vztah chirurgické terapie a radiační terapie (11 %), nejlepší metoda diagnózy rakoviny (8 %), počet primárních nálezů (8 %), typ histologie (5 %), stádium rakoviny (5 %) atd.

Před sestavením 2D-bodového grafu bylo potřeba překlasifikovat atribut *Age_at_diagnosis.* 2D-bodový graf i distribuční graf potřebují nečíselné, respektive nespojité atributy. Vytvořený nový atribut *Age* (typu string a úroveň měření nominal) byl rozdělen do 15 věkových skupin, každá s rozpětí 5 let. Poté bylo možné sestavit graf závislosti věku pacienta v čas diagnózy a chirurgie. Na grafu lze vysledovat mírný rozdíl ve věkových skupinách. Pacienti, na kterých byl proveden chirurgický zákrok, patří do mladší věkové kategorie. Největší zastoupení je ve skupině 60 – 64 let, kdežto ve skupinách, kterým chirurgický zákrok nebyl doporučen, skupina 65 – 70 let a pacienti, kteří nepodstoupili zákrok, měli největší zastoupení ve věku 65 – 74 let. Pacienti, kterým nebyl zákrok doporučen má navíc silnější zastoupení věkových skupin 80+. Zjištění, že hlavním předpokladem pro zákrok je věk pacienta, potvrzují i pravidla, která mají stanovené i sdružení „British Thoracic Society“ a „Society of Cardiothoracic Surgeons of Great Britain and Ireland“ pro výběr pacientů vhodných pro podstoupení chirurgického zákroku. Na prvním místě je to věk následován plicní funkcí a kardiovaskulární zdatností.



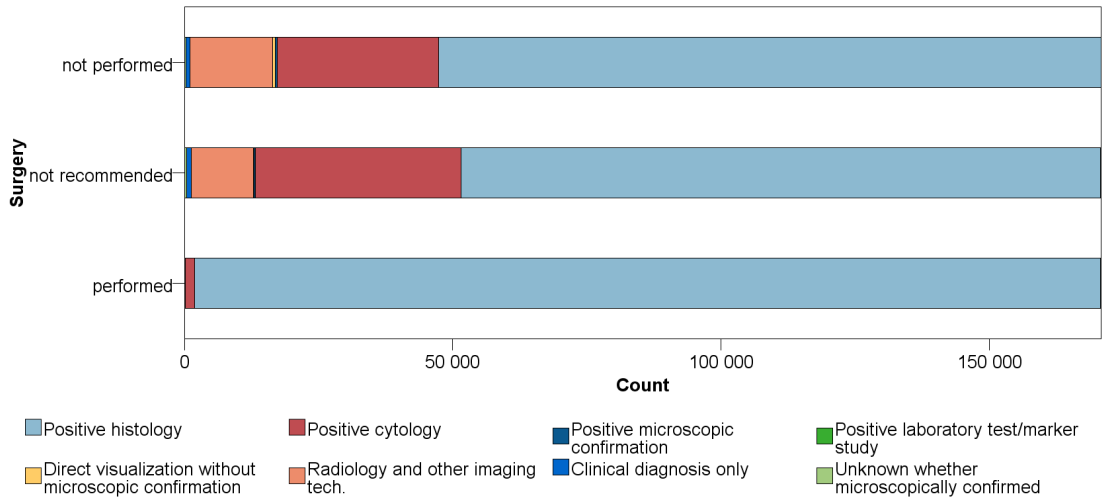
Obr. 54 Graf závislosti věku a chirurgické terapie

Dalším nejvýznamnějším atributem je proměnná, která zaznamenává, zda pacient podstoupil radiační terapii. Většina pacientů, kteří podstoupili chirurgickou terapii, podstoupili poté i radiační terapii, aby zabránili návratu rakoviny. Pouze malá část z nich se léčila radiací před chirurgickým zákrokem a zlomek z nich se léčil radiací před zákrokem i po zákroku.



Obr. 55 Graf závislosti radiační terapie a chirurgické terapie

Třetím nejvýznamnějším prediktorem byla nejlepší metoda diagnózy rakoviny. Jde o nejlepší metodu, pomocí které byl nález u pacienta potvrzen. Pacienti, kteří byli operováni, patřili z 98,9 % do skupiny těch pacientů, u kterých se nález prokázal pozitivní histologií. Jde tedy o nálezy, které se potvrdily zkoumáním buněčné struktury. Ostatní pacienti, kteří se neléčili chirurgickou terapií, mají větší podíl nemocných, kteří byli prokázáni pozitivní cytologií, tedy metodou, která studuje anatomii buněk, fyziologií a chování ve vztahu k životním funkcím a těch, kteří byli prokázáni radiologií, metodou využívající ionizujícího záření (CT plic).



Obr. 56 Graf závislosti chirurgické terapie a nejlepší metody diagnózy

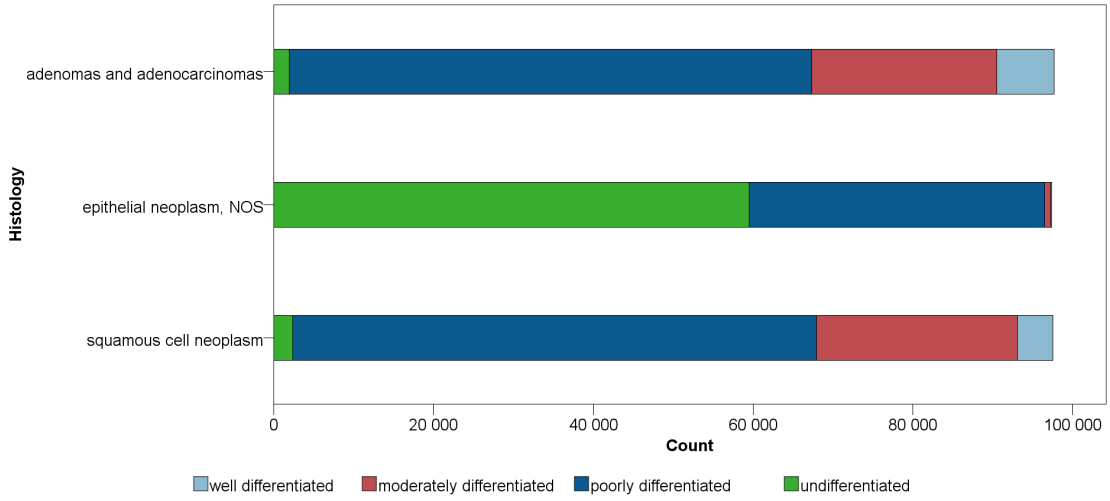
4.5.3 Bayesovská síť

Datová položka *Histlologic_Type_Broad_Groupings* popisuje mikroskopickou kompozici buněk anebo tkáně primárního nálezu, která rozděluje nádory do subtypů.

Typ histologie je základem pro určení stádia rakoviny a určení dalšího postupu léčení a ovlivňuje prognózu i celkový průběh nemoci.

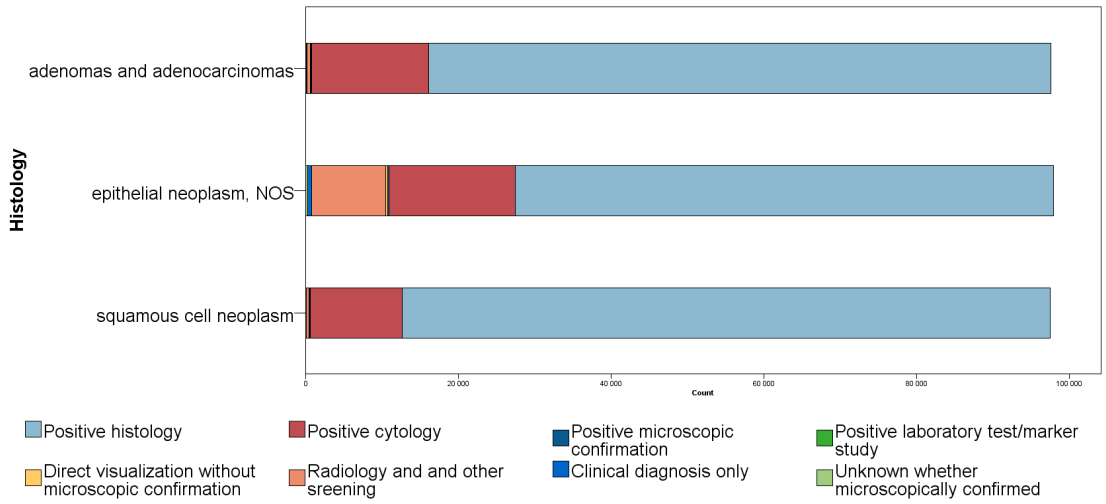
Nejdůležitějšími faktory, které dle Bayesovské sítě ovlivňují typ histologie, jsou věk, stupeň nádorových buněk (47 %), nejlepší metoda diagnózy rakoviny (11 %), pohlaví (9 %), párovost orgánu (7 %), počet nálezů (6 %) atd.

Bayesovská síť určila atribut *Grade* jako nejdůležitější prediktor histologického typu. *Grade* určuje stupeň diferenciace rakovinných a zdravých buněk. Model objevil velké zastoupení tumorů s neohraničeným buňkami (undifferentiated) nebo špatně ohraničenými (poorly undifferentiated) u pacientů s epitelovými nádory (epithelial neoplasm). Epitelové nádory vznikají nekontrolovaným šířením epitelových buněk. Zde si lze odvodit určitou závislost s neohraničenými a špatně ohraničenými buňkami, které mají také tendenci růst a šířit se dál rychleji než ostatní stupně. Poměry zastoupení stupňů nádorů u adenokarcinomů (adenomas and adenocarcinomas) a spinocelulárních karcinomů (squamous cell neoplasm) jsou přibližně stejné.



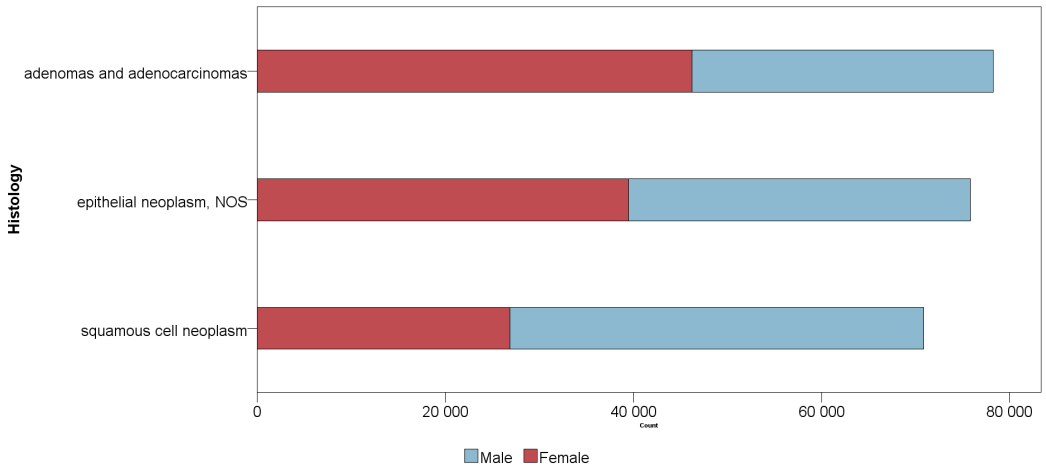
Obr. 57 Graf závislosti histologického typu a stupně tumoru

U nejlepší metody diagnózy byla většina pacientů ošetřena pomocí pozitivní histologie a menší část, poměrově stejný počet pacientů, pomocí pozitivní cytologie. Patrný rozdíl je opět vidět u pacientů s epitelovými nádory. Na úkor pozitivní cytologie se zde nachází část pacientů, kteří byli vyšetřeni i diagnostickou radiologií. Může se jednat o rentgenové, ultrazvukové či vyšetření magnetickou rezonancí.



Obr. 58 Graf závislosti histologického typu a nejlepší metody diagnózy

Obecně rakovinou plic trpí více muži než ženy. Současné trendy naznačují výrazně se snižující počet mužských případů a mírný nárůst případů mezi ženami. Co se týče rozdílu histologického typu napříč pohlavím po vyvážení celkového poměru počtu mužů a žen, muži převládají v subtypu spinocelulárních nádorů (squamous cell neoplasm). Na druhou stranu ženy trpí více nádorovým subtypem vznikajícím z buněk zabývající se hlenovou sekrecí, zhoubnými (adenomas) i nezhoubnými (adenocarcinomas). Poměr mužů a žen u nádoru epitelových žláz je skoro vyrovnaný (53:47).



Obr. 59 Graf závislosti histologického typu a pohlaví pacienta

4.6 Získané znalosti

V modelovací a interpretační části byly získány výsledky, z kterých je možné na první pohled rovnou vytěžit zřejmé znalosti. Jedná se o znalosti, které jsou již známé i mezi laickou veřejností, jsou to znalosti, které si lze logicky odvodit:

- šance na vyléčení pacienta nejvíce ovlivňuje stádium rakoviny,
- vzdálené a metastazované nálezy jsou hůře léčitelné a mají negativní vliv na prognózy pacienta,
- nádory v místě primárního nálezu jsou lépe léčitelné,
- první projevy rakoviny plic se většinou projevují ve fázi IV (distant),
- operovaní pacienti mají vyšší šanci na přežití,
- špatné ohraničené a neohraničené buňky (vyšší stupně nádoru) mají negativní vliv na prognózy pacienta,
- podmínky pro chirurgickou terapii rakoviny plic splňují spíše mladší věkové skupiny,
- rakovinou plic více trpí muži než ženy.

Vzhledem ke skutečnosti, že poskytnutá data jsou rozsáhlá, bylo by možné dalším zkoumáním a zkoumáním i jiných cílových atributů získat takovýchto znalostí bezpočet. Další znalosti lze získat z vygenerovaných pravidel rozhodovacího stromu, kde jsou vepsané konkrétní podmínky pro určení šancí pacienta na uzdravení.

Tato kapitola se bude více zaměřovat na znalosti, které nejsou na první pohled tak zřejmé a pro jejich získání je potřebné se více ponořit do problematiky onkologických onemocnění a spojit dohromady výsledky více modelů.

4.6.1 Vliv postchirurgické radiační léčby na prognózy pacienta

U modelu C5.0 bylo zjištěno, že důležitými prediktory ovlivňující přežití pacienta byly atributy *Surgery* a *SEER_historic_stage_A,* tedy fáze stádia rakoviny. Pokud se omezíme pouze na ty pacienty, kteří podstoupili chirurgickou terapii a zároveň na ty, kteří byli pouze ve stádiu I nebo II, podmínkou:

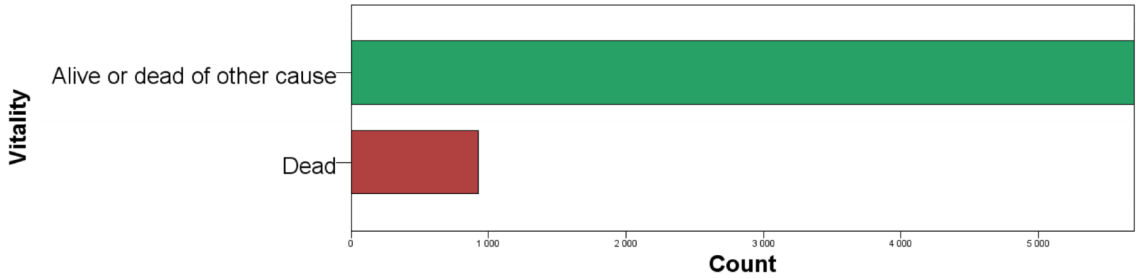
- Reason_for_no_surgery = 0 and SEER_historic_stage_A = 1 or SEER_historic_stage_A = 0

a spojíme-li tyto poznatky se zjištěním týkajícího se radiační terapie a dalším omezením na pacienty, kteří navštěvovali radiační terapie až po chirurgickém zákroku:

- RX_Summ_Surg_Rad_Seq = 3,

zjistíme, že tito pacienti měli mnohem větší šanci na vyléčení než ostatní pacienti a to o celých 71,94 %. Z toho vyplývá, že pacientům, kterým byl chirurgicky odstraněn nemalobuněčný nádor, který se nestačil rozšířit dále do okolních orgánů,

se podařilo následnou radiační léčbou zabránit opětovnému návratu rakoviny a vyhubením karcinogenních zbytků po chirurgickém zákroku.



Obr. 60 Mortalita operovaných pacientů stádia I nebo II po následném podstoupení radiační léčby

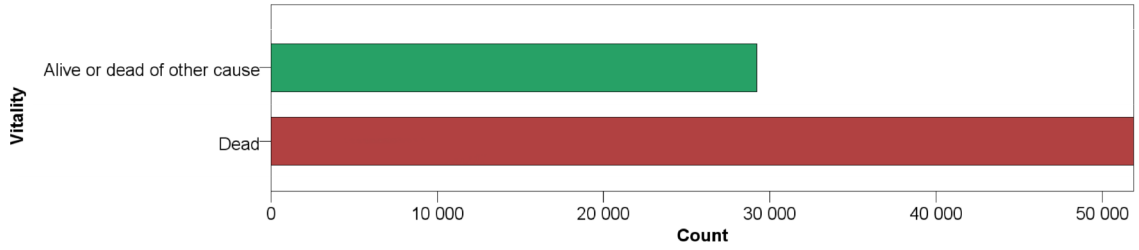
Po chirurgickém odstranění plicního nádoru ve fázích stádia I nebo II se může stát, že na okrajích zákroku zůstanou zbytky rakovinných buněk označující se jako „pozitivní okraje“. V tomto případě následuje sekundární zákrok k odstranění těchto zbytků za doprovodu pooperačních radiačních léčeb s cílem zabránit opětovného rakovinného bujení (American Cancer Society, 2015). Data prokazují, že tento radiační postup má pozitivní dopad na míru přežití pacientů.

4.6.2 Vliv pozitivní cytologie na prognózy pacienta

Model Neural Net odhalil významnou roli diagnostické metody při rozhodování o chirurgickém zákroku. Pacienti, kteří se neúčastnili zákroku, nebo jim zákrok nebyl doporučen, měli vetší podíl pozitivní cytologie jako nejlepší metodu diagnózy. Byli vybráni pouze pacienti s pozitivní cytologii a u nich byla analyzována mortalita:

- Diagnostic_confirmation = 2.

Bylo zjištěno, že tyto pacienti mají horší prognózu než pacienti, kteří měli negativní cytologii nebo u nich byl nález prokázán jinou metodou (pozitivní histologií aj.)



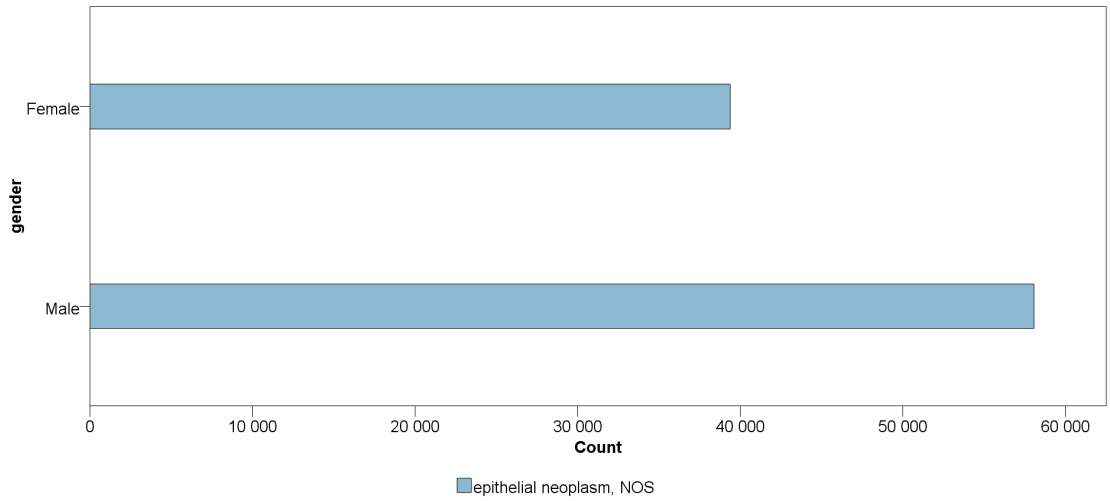
Obr. 61 Mortalita pacientů prokázaných pozitivní cytologií

Tyto vydolované znalosti potvrzují studii, která proběhla v Japonsku v roce 2001 na objasnění prognóz pacientů trpících nemalobuněčnými karcinomy plic

bez pleurální efuze, u kterých byly intrapleurální rakovinné buňky detekovány pomocí cytologického vyšetření získané bezprostředně po thorakotomii (otevřené plicní biopsii). Analyzovaná data byla posbírána mezi lednem roku 1985 a prosincem 1994 organizací Japan Clinical Oncology Group z celkem 15 institucí. Bylo zkoumáno celkem 1890 pacientů, z kterých 142 (7,8 %) mělo pozitivní cytologii. U těchto pacientů v době diagnózy již byla známa prognóza s šancí přežití jen 30 %. Výsledkem studie je špatná pětiletá prognóza míry přežití u pacientů s pozitivní cytologií. (Ichinose Y a spol., 2001)

4.6.3 Rozdíl nádorových subtypů mezi pohlavím

Při rozdělení pacientů na subtypy nádorů byly zjištěny rozdíly mezi pohlavím. U mužských pacientů převládají spinocelulární nádory. Jedná se o nádory, které vznikají ve vnitřku dýchacích cest v plících. Většinou se nacházejí ve středu plic blízko průdušek a jsou většinou přímým důsledkem kouření. (American Cancer Society, 2014)



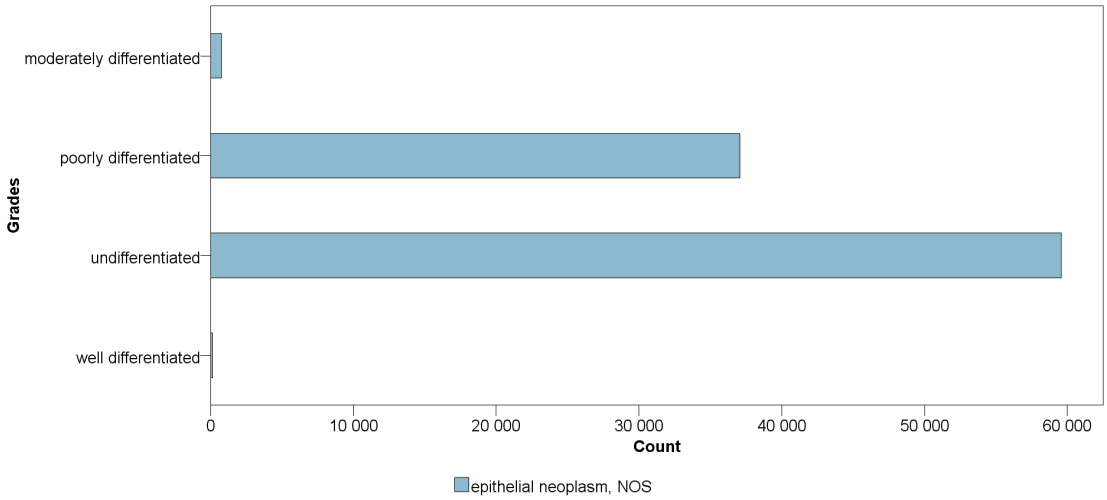
Obr. 62 Výskyt spinocelulárního nádoru mezi muži a ženami

Získaná znalost poukazuje na fakt, že se mezi muži vyskytuje více kuřáků než mezi ženami. Tuto skutečnost podporuje statistika zveřejněna americkou organizací „Centers for disease control and prevention“, která říká, že je ve spojených státech amerických každý pátý muž (20,5 %) a každá šestá žena (15,3 %) kuřákem. Kouření se projevuje na buněčném složení nádorů, které je rozhodující pro určení nádorových subtypů.

4.6.4 Vliv kouření na stupeň nádoru

V předchozí podkapitole bylo objasněno, že spinocelulární nádory vznikají většinou u kuřáků. Prostřednictvím Bayesovské sítě byl zkoumán i vztah nádorového subtypu a stupně nádoru. Omezením se pouze na spinocelulární nádory zjistíme, že většina těchto nádorů jsou buď neohraničené, nebo špatné ohraničené. Výskyty

dobře ohraničených a středně ohraničených nádorů jsou zanedbatelné. Dle analýzy je možné si odvodit, že kouření má oproti jiným příčinám rakoviny plic pravděpodobně větší vliv na tvorbu takových nádorů, které mají tendenci rychleji růst a rychleji se šířit dál, tj. nádory, které mají horší prognózu a šanci na vyléčení. Jinými slovy: kouření způsobuje agresivnější typ rakoviny plic než jiné příčiny rakoviny.



Obr. 63 Stupně spinocelulárních nádorů

4.7 Využití znalostí v praxi

Hlavní cílem dolování dat bylo získat znalosti týkající se určení prognóz přežití pacienta, zjistit jaké proměnné mají největší vliv na vyléčení rakoviny plic, poté zjistit jaké proměnné nejvíce ovlivňují, zda pacient je či není poslán na chirurgický zákrok a další znalosti o plicní rakovině obecně. Pro splnění tohoto cíle bylo vyzkoušeno více modelovacích technik, z kterých byly vybrány 3 nejúspěšnější a nejzajímavější. Znalosti získané data miningem v oboru lékařství jsou využívány většinou ve třech skupinách zainteresovaných stran:

- pacienti a spotřebitelé,
- poskytovatelé (vláda, nemocnice, farmaceutické organizace, lékárny, kliniky, lékařská ordinace, výzkumné centra a další),
- plátci (zaměstnavatelé, pojišťovny, odpory apod.).

Takový typ výsledků je spíše využitelný pro výzkumná centra a okruh lékařů zabývajících se onkologickými nemocemi. Takto získané znalosti mohou poskytnout určitou představu, kam mohou svůj výzkum dále směřovat. Mohou sloužit jako kritérium nebo hledisko při posuzování dalšího vývoje zkoumání onkologických nemocí. Přihlédneme-li na fakt, že toto není práce studenta lékařské fakulty, tak by podle mého názoru bylo potřeba takovéto výsledky podrobit intepretaci skutečné-

ho lékařského odborníka (i přesto, že většina získaných znalostí byla potvrzena zmíněnými výzkumy nebo cizím zkoumáním), který posoudí korektnost a validitu výsledků, určí potenciální využití znalostí a může znalosti poslat na další vědecký výzkum, na kterém se podílí řada jiných odborníků. Pokud by se znalosti potvrdily jako pravdivé a použitelné v lékařské praxi, může dle nich lékař mít lepší představu o dopadech jednotlivých složek rakoviny na pacienta a podle nich rozhodnout o dalším postupu. Například podle pravidel rozhodovacích stromů může lékař rozhodnout, jaké má pacient šance na přežití a podle toho přizpůsobit další postup v léčbě onkologického onemocnění.

Lékař může dále například podle takovýchto znalostí vytvořit představu o tom, jaký má pacient-kuřák pravděpodobně subtyp nádoru, do jaké třídy patří a dokáže si ho předběžně ještě před výsledky testů či biopsie zařadit do určité kategorie.

Obecně řečeno mohou takové predikce podpořit diagnostické, léčebné nebo monitorovací úkony lékaře. Přiřazení pacientů s rakovinou do nádorových subtypů může sloužit k výběru různých terapeutických strategií. Terapeutická předpověď se vztahuje k výběru nejvhodnější léčby pacienta, což je druh predikce, který je velmi častý v klinickém kontextu ať už pro plánování léčby pomocí léků nebo pro prognózy chirurgických zákroků.

5 Závěr

V teoretické části diplomové práce byly popsány základní pojmy a úskalí data miningu. Dále byly popsány poznatky důležité pro realizaci praktické části, tj. techniky, metodiky a nástroje data miningu. V průběhu praktické části bylo vyzkoušeno více metod a algoritmů data miningu s různými cílovými atributy. Jejich popis v teoretické části byl omezen pouze na ty, které přinesly ty nejlepší a nejzajímavější výsledky. Bylo postupováno systematicky dle světově uznávané metodiky CRISP-DM, která dokumentuje všechny nezbytné kroky data miningu a procesu dobývání znalostí. K splnění cíle byl vybrán, mezi analytiky často využívaný a výkonný IBM SPSS specializující se na soubory rozsáhlých dat.

Hlavním cílem diplomové práce nebylo získat nějaké překvapivé nebo přelomové znalosti z oblasti onkologických nemocí, ale spíše dokázat pomocí metod data miningu, že z takových dat je možné vydolovat nějaké znalosti, které jsou smysluplné a v praxi ověřitelné.

Ověřilo se, že časově nejnáročnější fázi celé praktické části byla příprava a transformace dat, která si vyžádala důslednou analýzu atributů, která byla zásadní pro orientaci v datech, pochopení problematiky a počáteční komunikaci s technickou podporou programu SEER ohledně možností transformace a zpracování dat. Byla navržena řada experimentů, která byla postupně vyzkoušena s cílem objevit potenciální skryté informace a znalosti.

Ohledně výsledků úspěšnosti predikcí modelů nejlépe skončila metoda rozhodovacího stromu C5.0, která byla schopna správně predikovat 81,92 % pacientů, což by se dalo hodnotit jako dobrý výsledek predikce. Hůře dopadly prediktivní sítě a to konkrétně neuronová síť s 67,69% správnosti predikce a Bayesová síť s 61,68% správnosti predikce, což se nedá považovat za moc dobrý výsledek. I tak dopadly mnohem lépe než jiné metody, které byly vyzkoušeny, ale nebyly zahrnuty v práci. Příčinou nižší úspěšnosti predikce mohou být chybějící data, která byla sice doplněná výpočtem, ale i přesto absence reálných dat má na výsledky velký vliv. Při odstranění všech záznamů s chybějícími daty a ponecháním pouze úplných záznamů byly výsledky dokonce ještě horší v důsledku příliš nízkého celkového počtu úplných záznamů.

Některé získané znalosti, které jsou popsané v práci se shodují s obecně potvrzenými znalostmi, jenž se dneska obyčejně v praxi využívají. Nicméně věcný názor lékařského odborníka by byl jistě velmi zajímavý.

Takovéto získané znalosti se nejvíce hodí pro cílové skupiny, ve kterých patří výzkumné centra zabývající se onkologickými nemocemi a také pro jednotlivé lékaře, k podpoře diagnostických, léčebných nebo monitorovacích úkonů.

6 Literatura

Knižní literatura

ABU-MOSTAFA, YASER S, MALIK MAGDON-ISMAIL A HSUAN-TIEN LIN. L*earning from data: a short course*. s.l.: AMLBook, 2012, xii, 201 s. ISBN 978-1600490064.

AGGARWAL, CHARU C. *Data mining: the textbook*. Cham: Springer, 2015, xxix, 734 stran. ISBN 978-3-319-14141-1.

ALPAYDIN, ETHEM. *Introduction to machine learning*. 2nd ed. Cambridge, Mass.: MIT Press, c2010, xl, 537 p. Adaptive computation and machine learning. ISBN 026201243x.

ICHINOSE Y., TSUCHIYA R., YASUMITSU T., YAMATO Y., KOIKE T., NAKAGAWA K., TADA H., YOKOI K., NAGAI K., et al. *Prognosis of non-small cell lung cancer patients with positive pleural lavage cytology after a thoracotomy*: results of the survey conducted by the Japan Clinical Oncology Group [online]. Tokyo, Japonsko, 2001, 2015-11-18 [cit. 2015-11-18]

TUFFERY, STÉPHANE. *Data mining and statistics for decision making*. Hoboken, NJ.: Wiley, 2011, xxiv, 689 p. Wiley series in computational statistics.

PETR, PAVEL. Data Mining. Vyd. 3. Pardubice: Univerzita Pardubice, 2010-, ^^^sv. ISBN 978-80-7395-325-6.

TAN, PANG-NING, MICHAEL STEINBACH A VIPIN KUMAR. *Introduction to data mining*. Boston: Pearson Addison Wesley, c2006, xxi, 769 p. ISBN 0-321-32136-7.

VAPNIK, VLADIMIR NAUMOVIČ. *Statistical learning theory*. New York, N.Y.: John Wiley & Sons, c1998, xxiv, 736 s. Adaptive and learning systems for signal processing, communications, and control. ISBN 0471030031.

WITTEN, IAN, EIBE FRANK A MARK A HALL. *Data mining: practical machine learning tools and techniques*. 3rd ed. Burlington: Morgan Kaufmann, c2011, xxxiii, 629 s. Morgan Kaufmann series in data management systems. ISBN 9780123748560.

BELLAZZI, RICCARDO, FULVIA FERRAZZI A LUCIA SACCHI. Predictive data mining in clinical medicine: a focus on selected methods and applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery [online]. 2011, 1(5): 416-430 [cit. 2015-12-09]. DOI: 10.1002/widm.23. ISSN 19424787. Dostupné z: http://doi.wiley.com/10.1002/widm.23

Internetové zdroje

AMERICAN CANCER SOCIETY. Treatment choices by stage for non-small cell lung cancer. [online]. Atlanta, Georgia, 2015-02-10 [cit. 2015-11-18]. Dostupné z: http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-treating-by-stage

AMERICAN CANCER SOCIETY. What are the key statistics about lung cancer? [online]. Atlanta, Georgia, 2014-08-15 [cit. 2015-11-18]. Dostupné z: http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-key-statistics

AMERICAN CANCER SOCIETY. Tumor Grade [online]. Atlanta, Georgia, 2014-08-15 [cit. 2015-11-18]. Dostupné z: http://www.cancer.gov/about-cancer/diagnosis-staging/prognosis/tumor-grade-fact-sheet

AMERICAN CANCER SOCIETY. What is non-small cell lung cancer? [online]. Atlanta, Georgia, 2014-08-15 [cit. 2015-11-18]. Dostupné z: http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-what-is-non-small-cell-lung-cancer

BEZDĚKOVÁ, MILADA, MUDR. Definice a vlastnosti nádorů. PASP: patologie pro speciální pedagogy [online]. Praha, 2011 [cit. 2015-11-24]. Dostupné z: http://pasp.upol.cz/index.php?option=com_content&view=article&id=54:definice-a-vlastnosti-nador&catid=12:nadory&Itemid=4

ČAPEK, KAREL. Rakovina plic. IRakovina [online]. Praha, 2015 [cit. 2015-11-24]. Dostupné z: http://www.irakovina.cz/rakovina-plic/

ČESKÁ ONKOLOGICKÁ SPOLEČNOST ČESKÉ LÉKAŘSKÉ SPOLEČNOSTI JANA EVANGELISTY PURKYNĚ. Primární tumor. Linkos [online]. Brno, 2015 [cit. 2015-11-24]. Dostupné z: http://www.linkos.cz/slovnicek/primarni-tumor-1/

KDNUGGETS SURWEY: Data Mining Community's Top Resource, The Popularity of Data Analysis Software: R vs SAS vs SPSS. [online]. 2000 [cit. 2015-11-29]. Dostupné z: http://www.kdnuggets.com/2011/03/data-analysis-software-popularity.html

KOMPLEXNÍ ONKOLOGICKÉ CENTRUM: 1. LÉKAŘSKÉ FAKULTY UNIVERZITY KARLOVY: Radiační léčba. (KOC) [online]. Praha, 2015 [cit. 2015-11-24]. Dostupné z: http://www.koc.cz/pro-verejnost/typy-lecby/radiacni-lecba/

NATIONAL CANCER INSTITUTE: Surveillance, Epidemiology, and End Results Program , About the SEER Program. [online]. Bethesda, 2012 [cit. 2015-11-28]. Dostupné z: http://seer.cancer.gov/about/

WORLD HEALTH ORGANIZATION. 10 facts on cancer.[online]. Paris, 2015 [cit. 2015-11-24]. Dostupné z: http://www.who.int/features/factfiles/cancer/facts/en

/index1.html

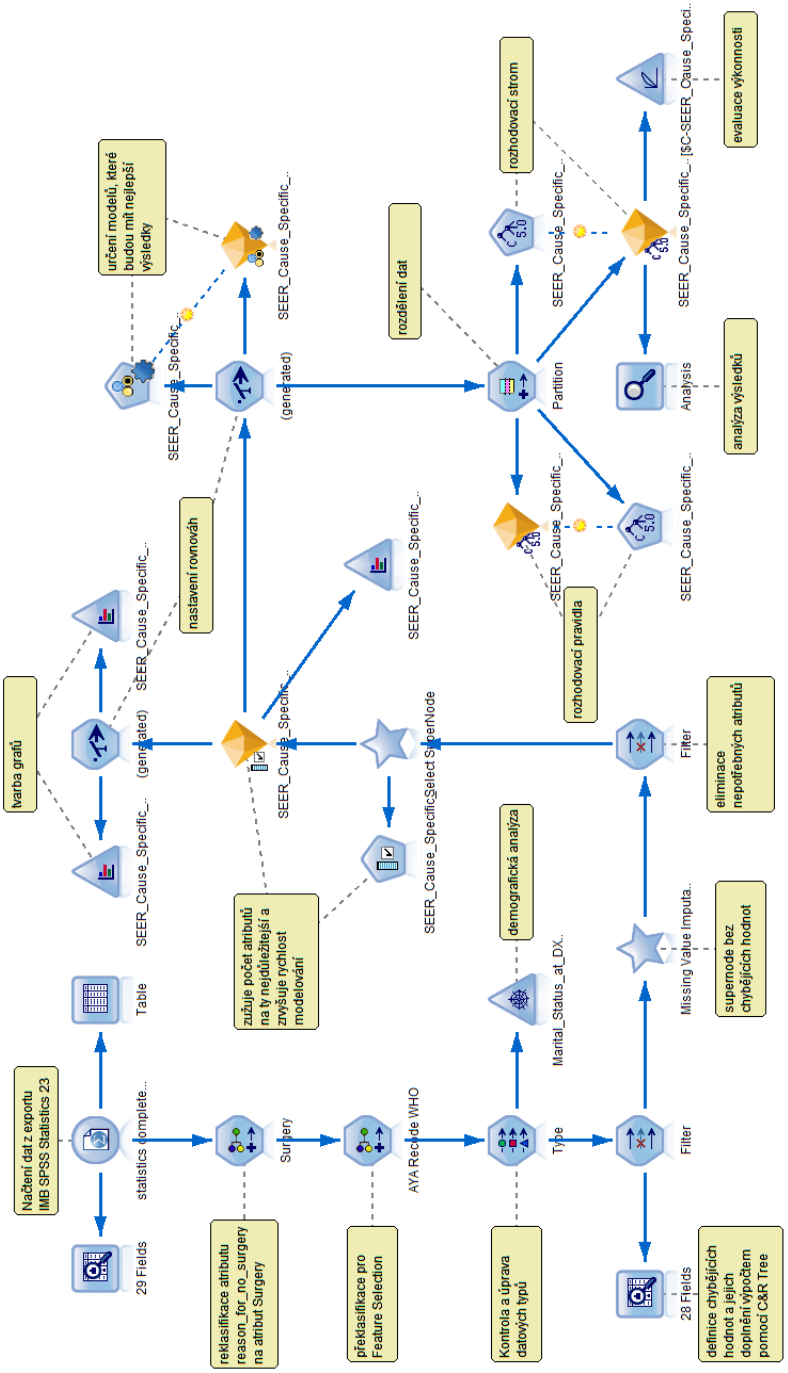
7 Seznam obrázků

8 Seznam tabulek

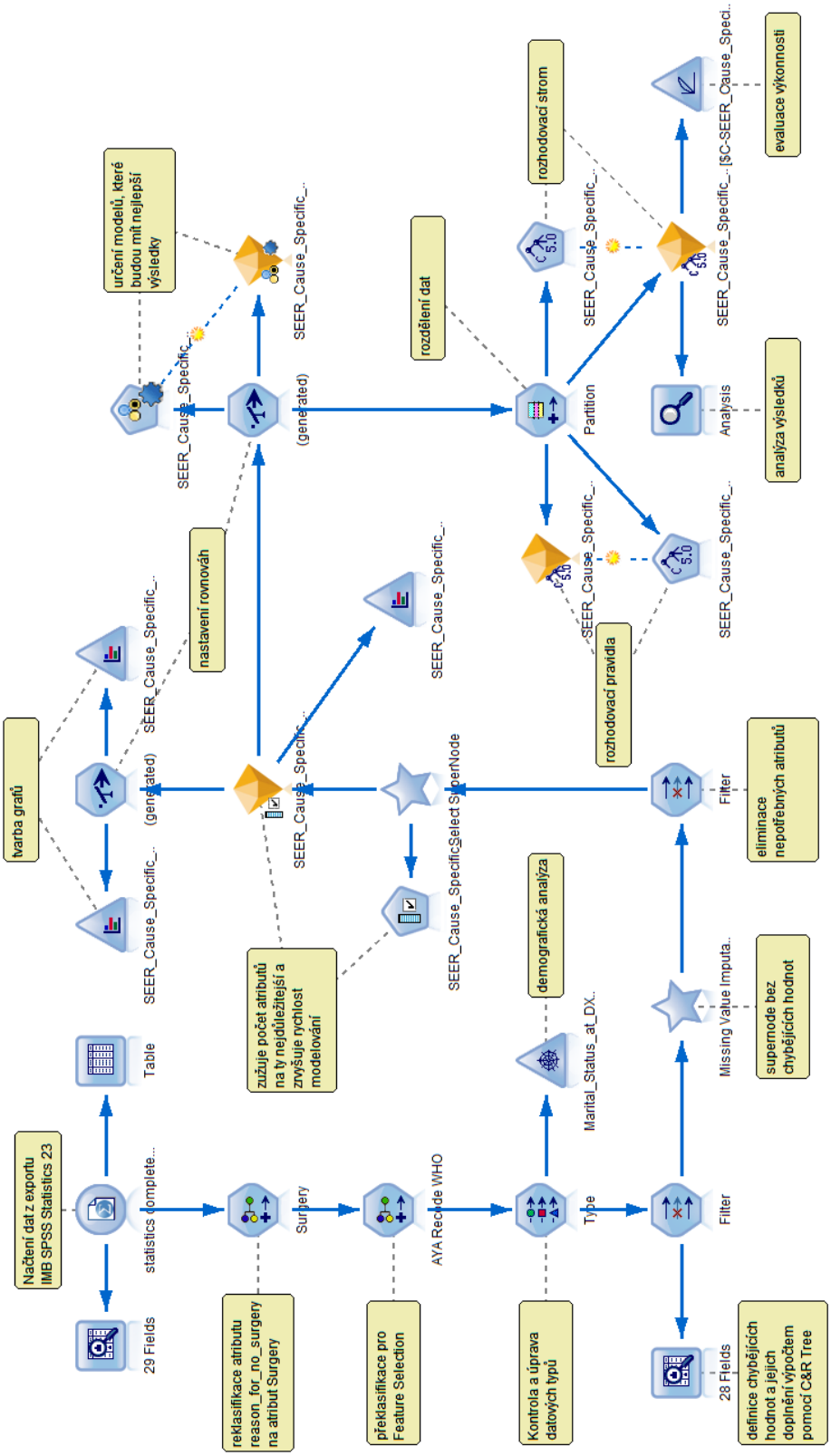
Přílohy

A Vytvořené streamy

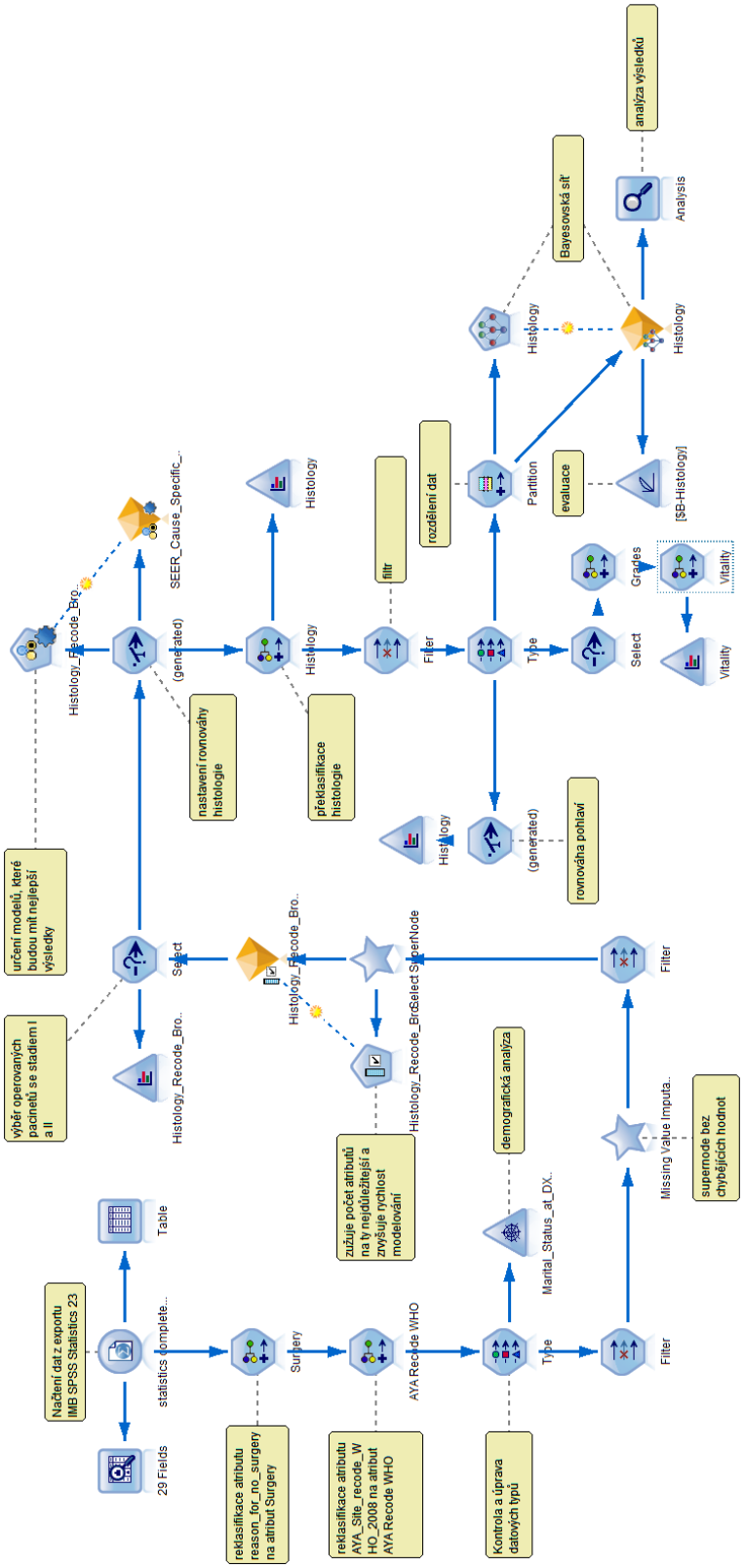
Na následujících obrázcích, které jsou exportovány ze systému IBM SPSS Modeler 17, se nacházejí jednotlivé streamy - pro každý model jeden. Jsou na nich zobrazeny jednotlivé vazby, závislosti a operace mezi okomentovanými uzly.



Obr. 64 Stream rozhodovacího stromu a rozhodovacích pravidel uzlu C5



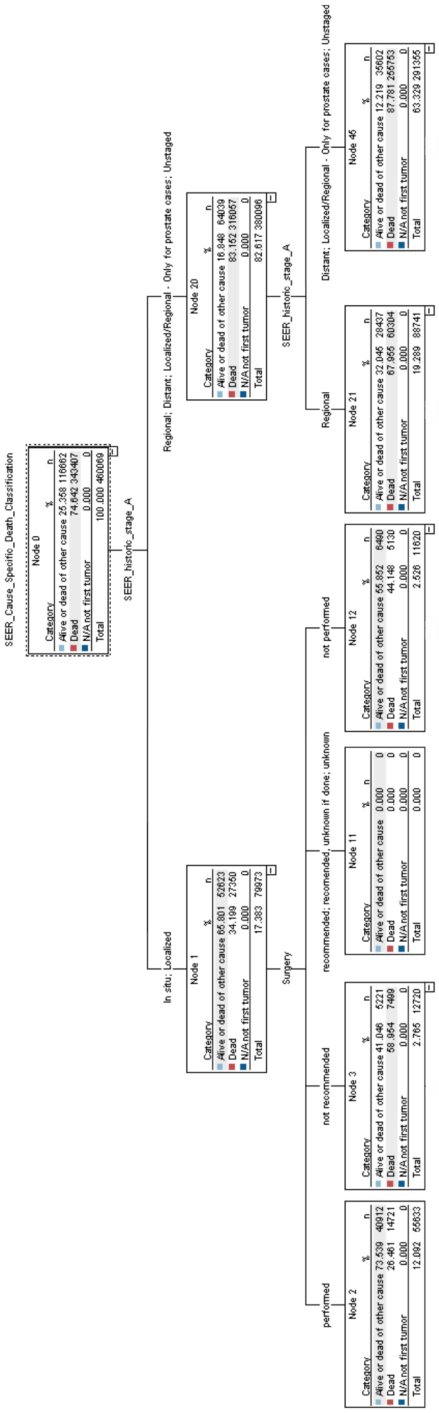
Obr. 65 Stream modelu neuronové sítě



Obr. 66 Stream modelu Bayesovské sítě

B Rozhodovací strom

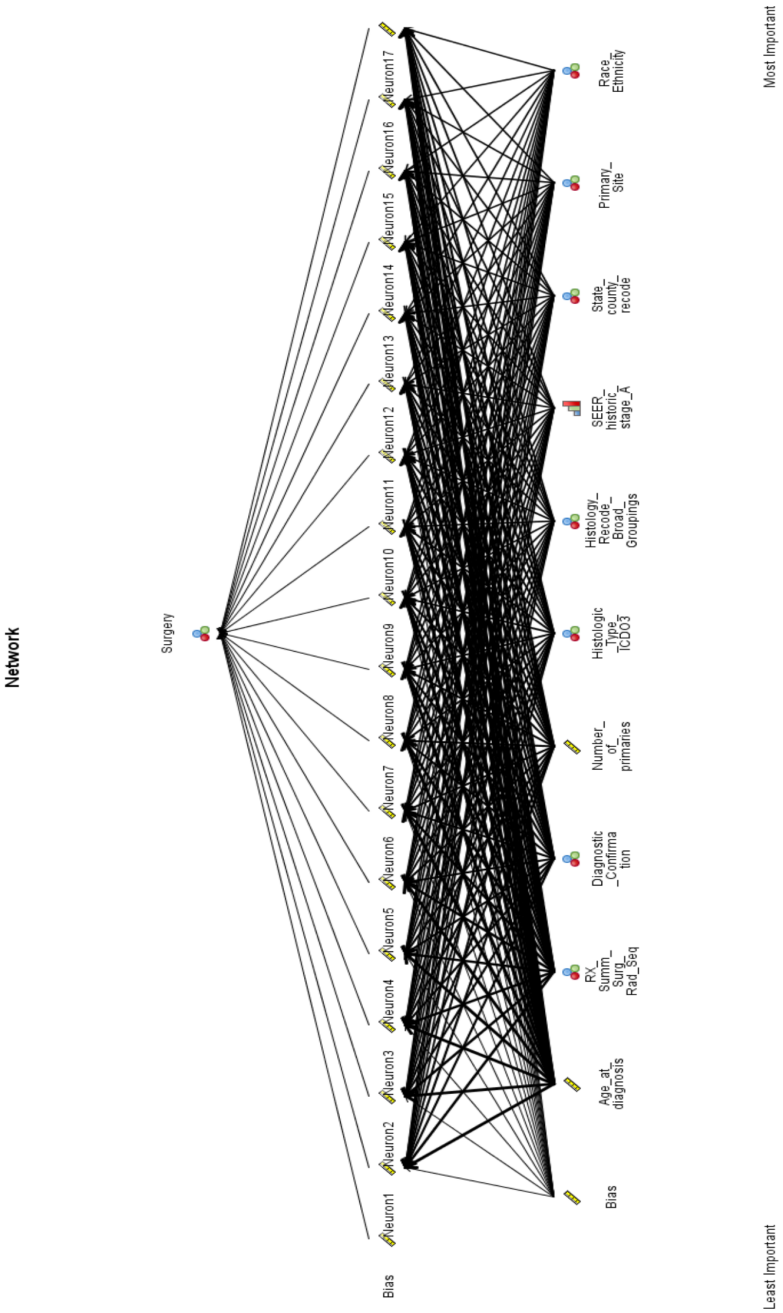
První tři úrovně rozhodovacího stromu C5.0.



Obr. 67 Rozhodovací strom uzlu C5

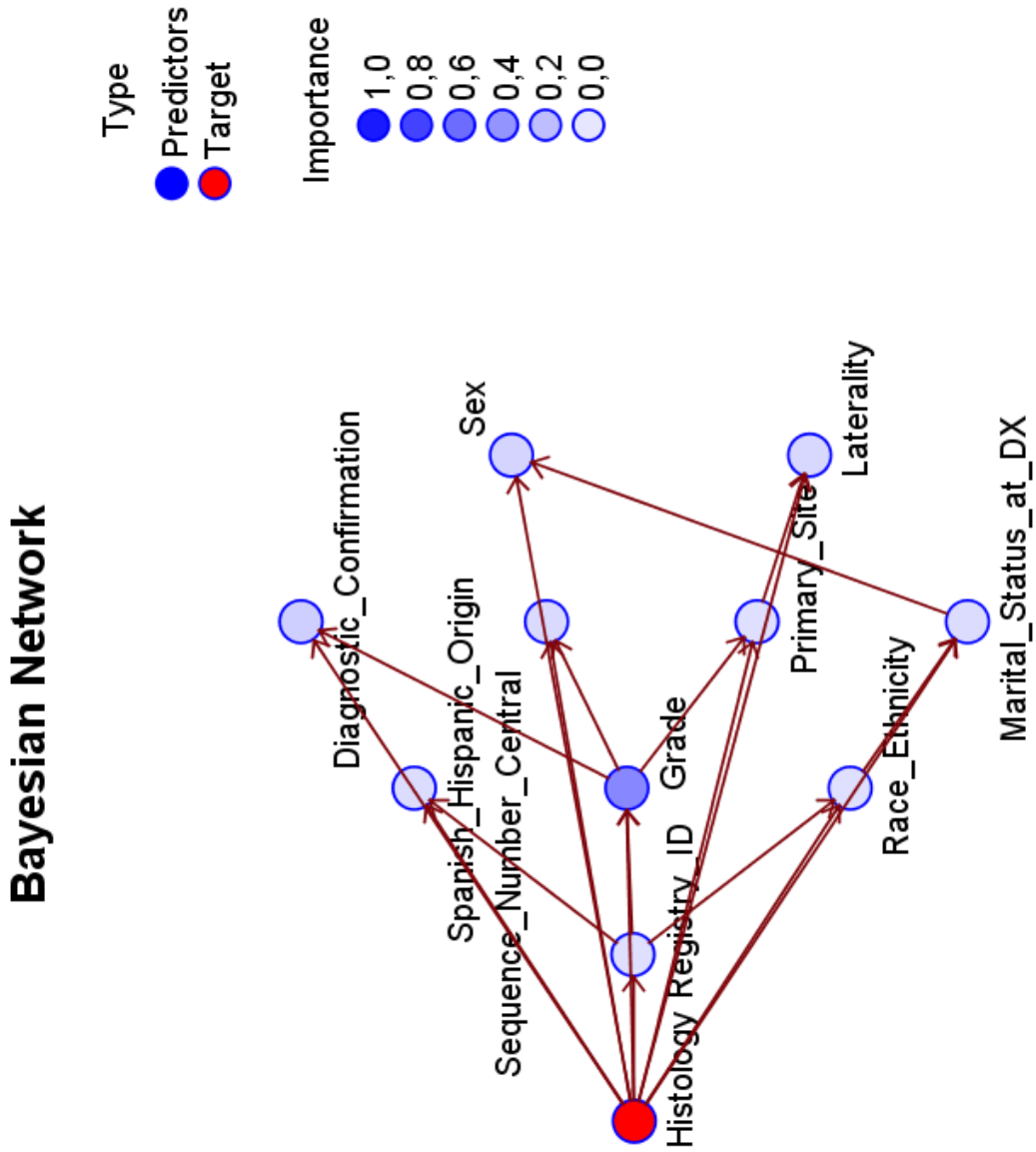
C Neuronová síť

Schéma vytvořené neuronové sítě obsahuje 17 neuronů v jedné skryté vrstvě. Vzniklý model je typu MLP, kde cílovým atributem je proměnná *Surgery*.



Obr. 68 Neuronová síť

D Bayesovská síť



Obr. 69 Bayesovská síť

E Rozhodovací pravidla

Kompletní seznam rozhodovacích pravidel vygenerované uzlem C5.0.

Rules for 0 - contains 16 rule(s)

Rule 1 for 0.0
if Histology_Recode_Broad_Groupings = 3,000
and SEER_historic_stage_A in [0.000 1.000]
then 0.000

Rule 2 for 0.0
if SEER_historic_stage_A in [4.000 8.000 9.000]
and First_malignant_primary_indicator = 0,000
then 0.000

Rule 3 for 0.0
if Histology_Recode_Broad_Groupings = 17,000
and SEER_historic_stage_A in [0.000 1.000]
then 0.000

Rule 4 for 0.0
if Histology_Recode_Broad_Groupings = 10,000
and SEER_historic_stage_A in [0.000 1.000]
then 0.000

Rule 5 for 0.0
if Histology_Recode_Broad_Groupings = 9,000
and SEER_historic_stage_A in [0.000 1.000]
then 0.000

Rule 6 for 0.0
if Grade = 1,000
and SEER_historic_stage_A in [0.000 1.000]
then 0.000

Rule 7 for 0.0
if Grade = 2,000
and Histology_Recode_Broad_Groupings = 5,000
and SEER_historic_stage_A in [0.000 1.000]
then 0.000

Rule 8 for 0.0
if Grade = 2,000
and Diagnostic_Confirmation = 2,000
and Histology_Recode_Broad_Groupings = 5,000
then 0.000

Rule 9 for 0.0
if Grade = 2,000
and RX_Summ_Surg_Rad_Seq = 3,000
and Surgery = not recommended
then 0.000

Rule 10 for 0.0
if Sequence_Number_Central = 1,000
and SEER_historic_stage_A in [0.000 1.000]
then 0.000

Rule 11 for 0.0
if Sequence_Number_Central = 1,000
and RX_Summ_Surg_Rad_Seq = 3,000
and SEER_historic_stage_A = 2,000
and Surgery = not recommended
then 0.000

Rule 12 for 0.0
if Grade = 1,000
and RX_Summ_Surg_Rad_Seq = 0,000
and First_malignant_primary_indicator = 1,000
and Surgery = performed
then 0.000

Rule 13 for 0.0
if SEER_historic_stage_A in [0.000 1.000]
then 0.000

Rule 14 for 0.0
if Grade = 2,000
and RX_Summ_Surg_Rad_Seq = 0,000
and Surgery = performed
then 0.000

Rule 15 for 0.0
if Grade = 2,000
and Diagnostic_Confirmation = 2,000
then 0.000

Rule 16 for 0.0
if RX_Summ_Surg_Rad_Seq = 0,000
and Surgery = performed
then 0.000

Rules for 1 - contains 13 rule(s)

Rule 1 for 1.0
if SEER_historic_stage_A = 2,000
and First_malignant_primary_indicator = 0,000
then 1.000

Rule 2 for 1.0
if Histology_Recode_Broad_Groupings = 21,000
and Surgery = not performed
then 1.000

Rule 3 for 1.0
if Registry_ID = 1526,000
and Surgery = not performed
then 1.000

Rule 4 for 1.0
if Histology_Recode_Broad_Groupings = 11,000
and Surgery = not performed
then 1.000

Rule 5 for 1.0
if Diagnostic_Confirmation = 1,000
and SEER_historic_stage_A in [4.000 8.000 9.000]
and First_malignant_primary_indicator = 1,000
then 1.000

Rule 6 for 1.0
if Registry_ID = 1523,000
and Surgery = not performed
then 1.000

Rule 7 for 1.0
if Grade = 4,000
and Surgery = not recommended
then 1.000

Rule 8 for 1.0
if Histology_Recode_Broad_Groupings = 1,000
and Surgery = not performed
then 1.000

Rule 9 for 1.0
if Sequence_Number_Central = 0,000
and Grade = 3,000
and Surgery = not recommended
then 1.000

Rule 10 for 1.0
if Histology_Recode_Broad_Groupings = 1,000
and Surgery = not recommended
then 1.000

Rule 11 for 1.0
if SEER_historic_stage_A in [2.000 4.000 8.000 9.000]
and First_malignant_primary_indicator = 1,000
then 1.000

Rule 12 for 1.0
if Histology_Recode_Broad_Groupings = 11,000
and Surgery = not recommended
then 1.000

Rule 13 for 1.0
if RX_Summ_Surg_Rad_Seq = 3,000
and SEER_historic_stage_A in [2.000 4.000 8.000 9.000]
and Surgery = performed
then 1.000

Default: 1