Mendelova univerzita v Brně
Provozně ekonomická fakulta

Dolování znalostí z rozsáhlých statistických souborů lékařských dat

Diplomová práce

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

Bc. Jan Cupal

Brno 2016

Tímto bych rád poděkoval mému vedoucímu práce panu doc. Ing. Janu Žižkovi, CSc. za ochotu a vlídnost při řešení problémů a dále za poskytnutí cenných informací a rad. 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/a 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/a, 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 21. prosince 2015 _____

Abstract

Cupal, J. Data mining in major statistical data sets of medical data. Thesis. Brno, 2015. In the theoretical part, we introduce the concepts that have close to the issues. These concepts then will be used in the analysis of medical data and for selecting appropriate algorithms that we will use in area of data mining. The results will be tested and based on the results, that we get by algorithms, will be made conslusions. The practical part will run in a programs SPSS Modeler and SPSS Statistics of company IBM.

Keywords

Thesis, analysis, data, data mining, cancer, research

Abstrakt

Cupal, J. Dolování znalostí z rozsáhlých statistických souborů lékařských dat. Diplomová práce. Brno, 2015. V teoretické části práce se seznámíme s pojmy, které úzce souvisí se zkoumanou problematikou. Tyto poznatky poté využijeme při analýze lékařských dat a k vhodnému zvolení algoritmů, které použijeme při dolování znalostí z dat. Získané výsledky otestujeme a na základě těchto výsledků zhotovíme závěr. Praktická část bude probíhat v programu od společnosti IBM SPSS Statistics a programu IBM SPSS Modeler.

Klíčová slova

Diplomová práce, analýza, data, data mining, rakovina, výzkum

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

V současné době je lidstvo zahlceno obrovským množstvím dat, které nejde jednoduše pochopit. Proto patří dolování znalostí z dat k jednomu z nejrychleji rostoucím odvětvím výzkumu.

V každém oboru poznání se shromažďuje nepřeberné množství dat. V oboru lékařství jsou to data například o záznamech jednotlivých pacientů, o provedených chirurgických zákrocích nebo například o průběhu pooperačních pozorování. Pokud chce být podnik v jakémkoliv oboru (ekonomika, pojišťovnictví, bankovnictví a další), konkurence schopný, předvídat aktuální trendy, je potřeba tyto data jistým způsobem zkoumat. Na základě těchto dat poté vzniknou informace a na základě informací vzniknou pro podnik znalosti, které dokáží schopní manažeři přeměnit v účelnou výhodu proti konkurenci.

Technologie data mining se tedy zabývá extrakcí skrytých prediktivních informací z rozsáhlých databází dat. Jedná se o relativně novou technologii s obrovským potenciálem, pomocí které se mohou společnosti zaměřit na nejdůležitější informace obsažené v jejich datových skladech. Oblast data mining nám umožňuje předvídat budoucí trendy a chování trhu, umožňuje společnostem být více proaktivní a umožňuje dělat rozhodnutí na základě znalostí získaných z dat. Nástroje dolování dat umí dát odpověď na tradiční obchodní otázky, které byly dříve časově velmi náročně řešitelné. Tyto nástroje prohledávají databáze a hledají skryté vzory, předpovídající informace, které i experti mohou jednoduše přehlídnout, neboť nejsou přímo očekávané.

V současné době již většina společností disponuje obrovským množstvím dat. Proto může být technika dolování dat implementována ihned na existující software a hardware platformy. Tato technika poté zvýší hodnotu stávajících informačních zdrojů a může být integrovaná s novými produkty a systémy.

Aplikování data mining může být kdekoliv, kde je potřebné dělat rozhodnutí založeném na důkazech. Rozmanitost aplikování je velká, ale mezi nejčastější místa můžeme zahrnout například předpovídání prodeje, vědecké zkoumání, herní průmysl, sporty nebo získávání daných zákazníků. Předpovídání prodeje (trhu) přitom spadá do kategorie vůbec prvního aplikování data miningu. V herním průmyslu můžeme využít metody pro předpovídání v otázce, který zákazník má největší potenciál ke koupi dané hry. Ve sportu poté pro zjištění, který hráč či tým, má nejvyšší potenciál k dosažení výhry (sázkové kanceláře). Dalším uplatněním data miningu může být například otázka, jak získat zákazniky a jakým způsobem identifikovat potenciální zákazníky, kteří mají největší pravděpodobnost, že budou reágovat na poskytnout nabídku.

2 Cíl práce

Práce je zaměřena na automatizované nalézání vzorů, pravidel, resp. dalších prezentací znalosti v reálných medicínských datech většího rozsahu poskytnutými programem "Surveillance, Epidemiology, and End Results Program", SEER, z USA z let 1973-2010. Cílem je tedy návrh a realizace přípravy dat do formy vhodné pro aplikaci algoritmů a metod data-mining. Dále návrh a realizace experimentů vhodnými algoritmy a interpretace výsledků spolu s vyvozením patřičných závěrů o možnostech a vhodnosti vyzkoušených postupů.

S poskytnutými daty bude dále pracováno v software od společnosti IBM, a to SPSS Modeler a SPSS Statistics.

3 Základy oboru dolování z dat a jeho historie

Pro pochopení zkoumané problematiky je potřeba seznámit se se základními pojmy, které úzce souvisí s pojmem data mining.

V první části se seznámíme s pojmy jako jsou data, informace, metadata a znalosti. Jejich propojení je poté znázorněno na obrázku 1. V další části bude popsán data mining obecně, historie tohoto oboru a metodologie, které nám definují postup při řešení nějakého problému. Dále budou popsány hlavní druhy učících metod, které se používají v data miningu.

3.1 Data

Data jsou jakékoliv materiálně zaznamenané vědomosti, zkušenosti nebo výsledky zpozorovaných procesů a obecné reality. Jakákoliv data poté slouží jako zdroj, ze kterého se vytváří informace. Data můžeme získat pozorováním, zápisem nebo například měření.

Všude kolem nás je spoustu dat a každým dnem počet těchto dat ikrementuje. Jednotlivá data rozdělujeme na tvrdá a měkká, kde měkká data vyjadřují subjektivní názory lidí a tvrdá jsou jasně definovaná. (Gillenson, 2012)

3.2 Informace

Pojmem informace se zabývalo již velké spektrum vědců, badatelů či filozofů. Je odvozený z latinského slova informatio, které původně sloužilo pro popsání vtištění nějaké formy, tvaru či utváření. Nicméně až Claude Shannon definoval informaci jako statistickou pravděpodobnost určitého signálu či znaku, který je na vstupu určitého systému. Pokud je pravděpodobnost znaku malá, tak má daný znak větší informační hodnotu. Tudíž tím, že systém signál zpracoval, tak se dostal na nižší úroveň nejistoty, tzv. entropii. Byl tedy více uspořádán. Tímto procesem lze poté charakterizovat míru neurčitosti přijímajícího systému.

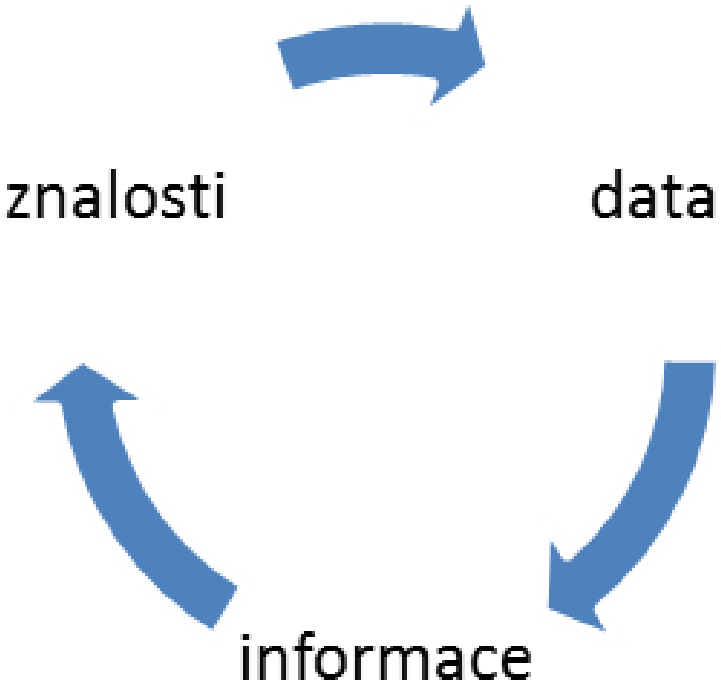
Obecně je informace velmi široký pojem, který je používán v různých významech a oborech poznání. Z laického pohledu můžeme informaci chápat jako údaj o prostředí, jeho stavu a procesech v něm probíhajících. Pojem informace však není v žádném případě možné zaměnit za data, neboť data jsou nositelem informace.

Každá informace by měla byt pravdivá, srozumitelná, včasná, relevantní a etická. (Gála, 2006)

3.3 Metadata

Metada jsou strukturované informace, které popisují, vysvětlují nebo jinak usnadňují získání, používání nebo řízení jakéhokoliv informačního zdroje. Metadata

tedy slouží k bližšímu popisu prostých dat. Často jsou metadata nazývané data o datech nebo informace o informaci. (Understanding metadata, 2004)



Obr. 1 Data, informace, znalost
Zdroj: inflow.cz, 2015

3.4 Znalost

Znalostí se rozumí to, co jednotlivec získá z poskytnuté informace. Jedná se tedy o řetězec vztahů a souvislostí mezi daty a informacemi nad danými daty. Vztah a postup, jakým vzniká informace a znalost, je znázorněn na obrázku 1. Obecně je znalost získáná souvisejícími poznatky a zkušenostmi, které byly nabyty praxí či studiem. Znalost se tedy rovná součtu informace, abstrakce, vztahu, zdůvodnění a aplikace. Nad znalostí poté figuruje moudrost. (Bureš, 2007)

3.5 Data mining

Termín Data mining zahrnuje více oblastí zájmu jako například Statistická analýza, Umělá inteligence, Strojové učení či Vývoj rozsáhlých databází. Statistická analýza je přitom deduktivní metoda, která hledá vztahy v množině dat. Dedukce je potom tzv. Aristotelův proces[1] detailní analýzy dat, výpočtu počtu metrik a formulování

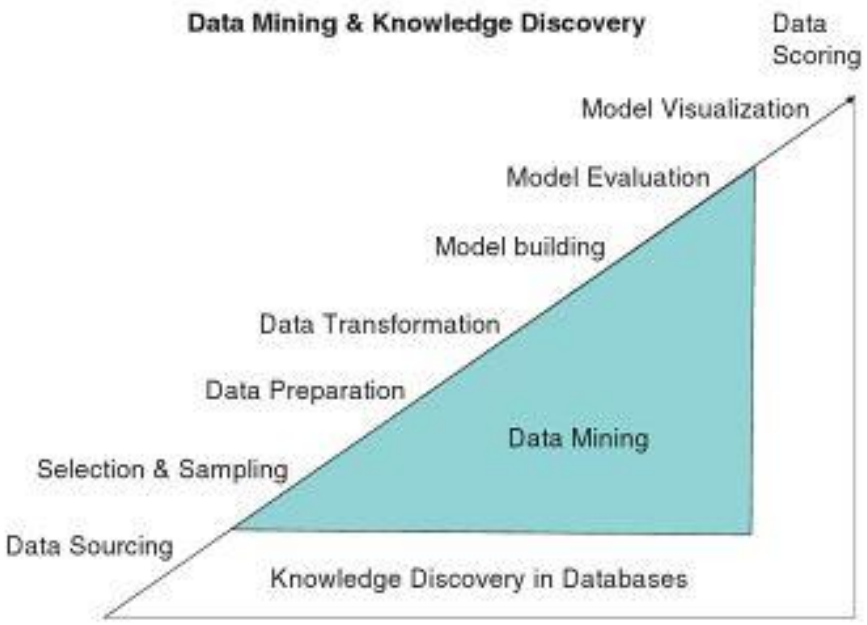
[1] Aristoteles věřil, že realita může být rozeznána pouze pomocí toho, co je dotknutelné. Věřil, že nejvyšší úroveň intelektu, může být studována pouze pomocí dotknutelného světa kolem nás. Celek byl pro něho složený pouze z jednotlivých dotknutelných věcí. (Nisbet, 2009)

nějakého závěru založeném výhradně na matematických metrikách. Oproti tomu techniky umělé inteligence a strojového učení (neuronové sítě, rozhodovací stromy) jsou induktivní metody. Tyto techniky hledají v datové množině spíše skryté, slabé nebo nejasné vazby. Induktivné metody jsou tzv. Platonické procesy[2], při kterých jsou odvozeny závěry založené nejen na základě vstupních dat.

Samostatný pojem Data mining může být definován několika způsoby, které se liší primárně v jejich pohledu na aspekty data miningu. Mezi první definice, které popisovaly dolování dat, můžeme zahrnout definici "The non-trivial extraction of implicit, previously uknown, and potentially useful information from data" (Non-triviální extrakce implicitních, dříve neznámých a potenciálně užitečných informací z dat), kterou pronesl pan Frawley v roce 1991. (Nisbet, 2009)

Během vývoje oboru bylo nezbytné rozlišit mezi předcházejícími aktivitami statistického modelování a širšími aktivitami objevování znalostí z dat. Pro popsání tohoto vývoje si můžeme zavést tři pojmy: Statistické modelování, Data mining a Objevování znalostí. Termín statistické modelování používalo pouze parametrické statistické algoritmy k seskupování nebo predikci výstupů nebo událostí založených na prognostických proměnných. Jak již bylo popsáno výše, tak Data mining používal algoritmy strojového učení k nalezení vztahů mezi datovými elementy v rozsáhlé chaotické množině dat, což vedlo k dosažení většího přínosu v podobě diagnóz nebo profitu. Termín Objevování znalostí v sobě zahrnoval celý proces přistupu k datům, exploraci dat, připravu dat, modelování, nasazení modelu a monitorování tohoto modelu. (IBM, 2014)

[2] Platón věřil, že realita není rovna jen součtu dotknutelných věcí kolem nás, ale i něčemu, co je za těmito dotknutelnými věcmi. Pro Plata byla tedy realita větší, než realita, co představoval Aristoteles. (Nisbet, 2009)



Obr. 2 Data mining a KDD
Zdroj: Nisbet, 2009

Jak docházelo k vývoji technologií, tak docházelo i k vývoji oboru Data mining. V roce 1996 poté pan Fayyad představil novou více specifickou definici Data Miningu "Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potential useful, and ultimately understamble patterns in data." (Získávání znalostí v databázích je netriviální proces identifikování validních, nových, potencionálně použitelných a nakonec srozumitelných vzorů v datech). Jak je z definice čitelné, tak se v této době zaměřil fokus již více na vzory v datech než na informaci v obecném smyslu. Vzory jsou přitom velmi těžké získat, neboť jsou slabé a těžko rozlišitelné a mohou být získány pouze pomocí algoritmů, které mohou vyhodnotit nelineární vztahy mezi prognostickými proměnnými a jejich cíli. Pro zjištění těchto vzorů můžeme použít například analytické nástroje jako rozhodovací stromy nebo neuronové sítě, které umožňují analýzu nelineárních vzorů v datech snadněji, než je možné pomocí parametrických statistických algoritmů. To je zapříčiněno tím, že algoritmy ve strojovém učení se učí způsobem shodným s lidským, a to dělat na základě příkladů, nikoliv pomocí počítání metrik založeném na průměrech a rozdělení dat.

Celý proces data mining si lze představit na obrázku číslo 2. Začíná se fází sběru a vzorkování(Selection & Sampling), následují fáze přípravy dat (Data preparation) a transformace dat (Data Transformation). Jakmile jsou data upravena, dojde k sestavení modelu (Model Building). Jako poslední fáze u data miningu probíhá fáze

ohodnocení modelu (Model Evaluation). KDD[3] však k těmto fázím přídával ještě fázi získávání dat (Data Sourcing) a fáze, kde došlo k vizualizaci dat (Model Visualizition) a k hodnocení dat (Data Scoring). Proces KDD tedy kombinoval matematické přístupy k objevování zajímavých vztahů v datech a takto vytvořený model poté používal k analýze jiné datové množiny.

Síla oboru Data miningu je oproti tradičním statistickým metodám v tom, že statistické metody používaly minulé informace k popsání budoucího stavu systému, zatímco Data mining používá minulé informace ke konstrukci vzorů založených nejen na vstupních datech, ale také na logických důsledcích těchto dat. (Nisbet, 2009)

3.5.1 Historie oboru Data mining

Počátky hledání vzorů v datech se datují již do 6. století před naším letopočtem, kdy bylo vynalezeno bambusové počítadlo abacus v Číně. Ve starověké Číně a Řecku pomáhala statistika shromažďovat informace vládcům států, kteří se na základě těchto informací rozhodovali ve fiskálních a vojenských záležitostech.

V 16. a 17. století se staly velmi populární hazardní hry mezi bohatými lidmi, což vedlo k mnoha otázkám, s jakou pravděpodobností mají šanci na výhru. K oboru pravděpodobnost se v tomto období věnovali významní matematici, a proto došlo v následujících letech k významným výzkumům v oborech matematika a statistika.

V 18. století byly vyvinuty dva obory statistiky, a to Bayesova a klasická statistika. Pro Bayesovu statistiku byla pravděpodobnost, že nastane nějaká událost v budoucnosti rovna pravděpodobnosti minulých výskytů událostí. Oproti tomu klasická statistika, která vycházela z matematických prací Gausse a Laplace, měla jako základ tzv. společnou pravděpodobnost. Klasická statistika tedy hledá na základě dat vlastností náhodné veličiny.

V 19. století se dostávala pravděpodobnost například i do oboru biologie. V tomto století byly také vyvinuty koncepty regrese a korelace pro analyzování genetických dat. Pravděpodobnost se dostávala i do sociálních věd a byl poprvé vyvinut parametrický model. (History of machine learning, 2013)

První náznaky aktivit, které dnes označujeme jako data mining, se objevily v 60. letech 20. století s rozvojem počítačové techniky. V tomto případě se jednalo o využívání regresní analýzy či prvních rozhodovacích stromů. Nicméně se jednalo spíše o procesy, které probíhaly na akademické půdě. Zde byly poprvé použity termíny jako "Data Fishing[4]" nebo "Data Dredging[5]".

V následujících letech probíhal velký progres ohledně rychlosti zpracování instrukcí v počítači, a také se zvětšovala dostupná paměť počítače, a tak bylo umožněno první systematické využití data miningové metodologie v praxi. V této době však nebylo stále umožněno plné využití data miningu. Jednalo se spíše o hledání korelací v datových souborech, což vedlo k nebezpečí, že se objeví jen

[3] KDD – Knowledge discovery in databases (Získávání znalostí v databázích)

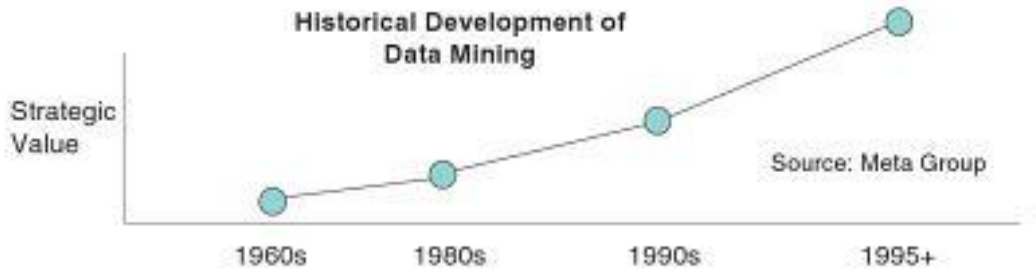
[4] Data Fishing – Výlov dat

[5] Data Dredging – Bagrování dat

nahodilé fluktuace v datech bez možnosti zobecnění výsledků, tudíž praktického využití.

Obrat přišel počátkem 90. let. Ve zmiňované době již byly představeny metody, které umožňovaly vyhnout se zmíněnému nebezpečí. Nastala velká poptávka od komerčních organizací, neboť již nestačily výsledky, které byly získány pouhými tabulkovými metodami. Jednalo se o aplikace především v přímém marketingu, finančnictví, telekomunikací a také například internetového prodeje. V této době byl poprvé objeven termín "Data Mining" a byl použit v oblasti databází. V historii se používaly další termíny pro označení práce s rozsáhlými daty. Mezi ně patřily například "Data Archaelogy[6]", "Information Harvesting[7]", "Information Discovery[8]" a "Knowledge Extraction[9]". Poslední z těchto termínu se stal velmi populární v oblasti umělé inteligence a strojového učení. Zatímco pojem data mining se stal velmi populární v oblasti obchodu.

V současné době je již široká nabídka software pro účely data miningu a touto oblastí se zabývá také spoustu komerčních firem.



Obr. 3 Historický vývoj oboru Data mining
Zdroj: Nisbet, 2009

Obecně můžeme říci, že data mining nevznikl jako nová akademická disciplína ze studií na univerzitách. Jednalo se o logický krok, kdy docházelo k větší poptávce po odpovědích na obchodní otázky a kdy všichni chtěli dosáhnout v budoucnu lepších obchodních výsledků než v minulých letech. Tento vývoj je znázorněn v tabulce číslo 1. V prvním řádku jsou popsány jednotlivé kroky vývoje a v dalších řádcích poté znázorněny obchodní otázky, technologie a charakteristiky typické pro daný vývoj. Na obrázku číslo 3. je poté zobrazen historický vývoj data miningu, kde na ose y je zobrazena strategická hodnota, kterou v různých letech umožňovaly dosažené výsledky oboru data mining. (Nisbet, 2009)

[6] Data Archeology – Archeologie dat

[7] Information Harvesting – Sklizeň informace

[8] Information Discovery – Objevování informace

[9] Knowledge Extraction – Extrakce znalostí

Tab. 1 Historický vývoj oboru Data mining

Kroky vývoje	**Sbírání dat**	**Přístupy k datům**	**Datové sklady a podpora rozhodování**	**Data mining**
Obchodní otázky	"Jaký byl můj celkový příjem minulý rok?"	"Co se prodávalo v Ohiu minulý březen?"	"Co se prodávalo v Ohiu minulý březen?" – zaměřeno na Dayton	"Co bude prodáváno v Dayton příští měsíc?"
Dostupné technologie	Počítačové pásky a disky	Relační databáze a SQL	Datové sklady Multidimenzionální databáze	Algoritmy víceúčelových masivních databází
Charakteristiky	Dodání statických souhrných minulých dat	Dodání dynamických minulých dat na úrovni záznamů	Dodání dynamických minulých dat na několika úrovních	Dodání prospektivních proaktivních informací

Zdroj: Nisbet, 2009

3.5.2 Metodologie oboru Data mining

Metodologie pochází z řeckého slova methodos, což v překladu znamená sledování či stopování a jedná se o vědní disciplínu, která se zabývá metodami, jejich tvorbou a aplikací. V oblasti data mining se můžeme setkat také s termínem a popisem metodologie.

Data mining obecně zahrnuje velkou šíři metod a způsobů práce a je tedy velmi obtížné podat jednoznačný návod k postupu. Všechny metody ovšem čerpají jednotlivé kroky z obecného postupu, který je popsán níže.

1. Rozhodnout se, zda-li se jedná o krátkodobý projekt, kterým potřebujeme odpovědět na otázku, či otázky nebo se jedná o dlouhodobý projekt.
2. Získat množinu dat, která se má použít k analýze. Tento krok často zahrnuje náhodný výběr vzorků z rozsáhlé databáze zachycující záznamy pro účel analýzy. Může se jednat o spojení dat z různých databází, které mohou být interní nebo externí.
3. Dalším krokem je prozkoumání, pročištění a upravení dat. Jedná se o verifikaci dat, které splňují podmínky pro daný projekt. Je potřeba pokládat si otázky jako například: "Jak by měly být řešeny chybějící údaje?" "Jsou hodnoty v rozumném rozsahu?" "Dávají nám informaci, kterou od nich očekáváme?" "Jsou v datech zřejmé hranice (jakého typu jsou data)?" Dále je potřeba zajistit konzistenci v definici jednotlivých polí, mezi jednotlivými měřeními a časových údajích.

4. Redukce dat. Tento úkol může zahrnovat rozdělení datové množiny do trénovací, validační a testovací množiny dat. Můžeme toho docílit eliminací nepotřebných proměnných, transformací proměnných nebo například vytvořením nových proměnných. Při tomto kroku je nezbytné ujistit se, zda víme, co každá proměnná znamená a zda je rozumné zařadit ji do cílového modelu.
5. 5.krokem je popsání úkolů dolování dat jako například klasifikace, predikce či shlukování. Jedná se o předělání kroku číslo 1 do více specifické statistické otázky.
6. Vybrání techniky, algoritmu, který bude použit k analýze a k dosažení potřebných výsledků.
7. Spuštění algoritmu pro splnění úkolů. Jedná se o typicky iterativní proces, kdy se zkouší více variant a často se používá více variant u jednoho algoritmu. Pokud máme zpětnou vazbu od jednotlivých algoritmů, tak můžeme data upravit k dosažení lepších výsledků.
8. Předposledním krokem je interpretace výsledků algoritmů. Tímto krokem se rozumí vybrání nejlepšího algoritmu pro nasazení, a pokud je možné, tak algoritmus otestovat na testovacích datech pro zjištění, jak výkonný je. Algoritmus by měl být otestován na validačních datech pro upravující procesy. V tomto kroku se stává validace součástí ověřovacího procesu.
9. Posledním krokem je nasazení modelu. Proces zahrnuje integraci modelu do operačního systému a spuštění na reálných záznamech k produkování rozhodnutí nebo různých akcí. (Schmueli, 2010)

V následující části se blíže podíváme na dvě metodologie, které vznikly v 90. letech 20. století. Jedná se o metodologie firmy SAS SEMMA a firmy SPSS, a to metodologie CRISP-DM.

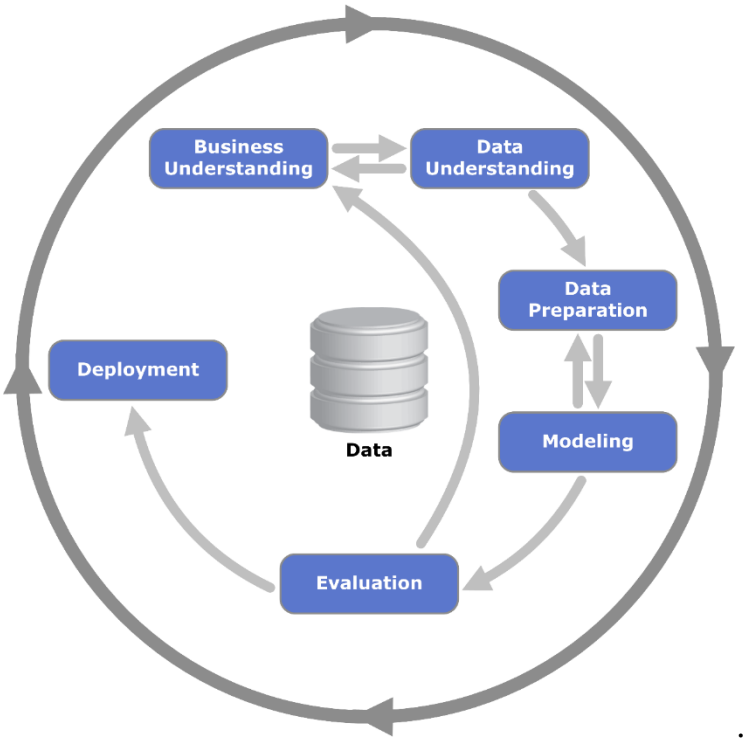
SEMMA je zkratka ze slov Sample, Explore, Modify, Model and Assess, což ve volném překladu znamená vzorek, průzkum, modifikace, model a posouzení. Jedná se o seznam postupných kroků vyvinutých SAS Institute Inc. SEMMA je často považována za obecnou metodiku dolování dat.

- Sample – proces začíná se vzorkováním dat. Množina dat by měla být dostatečně velká, aby obsahovala dostatečné informace k načtení, ale dostatečně malá, aby byla efektivně využitá.
- Explore – tato fáze zahrnuje porozumění údajů objevování očekávané a neočekávané vztahy mezi proměnnými, a také abnormality, s pomocí vizualizace dat.
- Modify – fáze obsahuje metody pro výběr, vytváření a transformování proměnné v přípravě na modelování dat.
- Model – V této fázi je kladen důraz na použití různých modelovacích technik na připravovaných proměnných s cílem vytvořit modely, které poskytují požadované výsledky.

- Assess – Poslední fází je posouzení, ve které dochází k vyhodnocení výsledků modelování ukazující spolehlivost a užitečnost vytvořených modelů. (Refaat, 2007)

CRISP-DM je zkratka z anglických slov Cross Industry Standard Process for Data Mining (Postup pro dolování znalostí z dat skrz průmyslové odvětví) a značí procesní model v oblasti Data mining, který popisuje běžně používané přístupy, které jsou používané k řešení problémů.

Tento proces má 6 hlavních fází. Sled těchto fází není přesně daný a pohybuje se tam a zpět mezi různými fázemi. Šipky na obrázku procesního diagramu označují nejdůležitější a časté závislosti mezi jednotlivými fázemi. Vnější kruh na obrázku 4 pak symbolizuje cyklické povahy dolování dat. Přínosy v průběhu procesu mohou vyvolat nové, často kladené obchodní otázky a následné procesy budou mít přínos ze zkušeností těch předcházejících.



Obr. 4 Metodologie CRISP-DM
Zdroj: IBM, 2015

- Business understanding (pochopení obchodních cílů) – tato počáteční fáze se zaměřuje na pochopení cílů a požadavků projektu z obchodního hlediska a poté převádí tyto znalosti do definice problémů dolování dat. V této fázi je také navržen předběžný plán k dosažení požadovaných cílů.
- Data understanding (porozumění datům) – fáze začíná s počátečním shromažďováním a výnosy s aktivitami s cílem seznámit se s daty, identifikovat

problémy kvality dat, objevovat první pohledy do dat nebo detekovat zajímavé podmnožiny pro vytváření hypotéz pro skryté informace.

- Data preparation (příprava dat) – fáze přípravy dat se vztahuje na všechny aktivity, které vedou k vytvoření finální množiny dat z počátečních holých dat. Tato fáze se provádí vícekrát a v řádném předepsaném pořadí. Zahrnuje tabulky, záznamy, výběr atributů, jejich transformaci a čištění dat pro modelovací nástroje.
- Modeling (modelování) – jsou vybrány a použity různé modelovací techniky a jejich parametry jsou kalibrovány k optimalizaci hodnot. Typicky existuje několik technik pro stejný typ problému data mining. Některé techniky mají specifické požadavky na formu dat.
- Evaluation (zhodnocení výsledků) – v této fázi projektu se objevují modely, které mají vysokou kvalitu z hlediska analýzy dat. Před konečným nasazením modelu je potřeba model důkladně posoudit a zkontrolovat kroky, které nám řeknou, zda model dosahuje obchodních cílů.
- Deployment (další rozvoj) – vytváření modelu není obecně koncem projektu. Dokonce i když účel modelu zvyšuje znalosti dat, budou muset být znalosti prezentovány a organizovány způsobem, který zákazník může využít. V závislostech na požadavcích může být zaváděcí fáze stejně jednoduchá, jako generování výsledků nebo stejně komplexní jako implentování opakováného hodnocení dat nebo celkový proces data miningu. (IBM, 2014)

3.6 Strojové učení

Strojové učení je vědní disciplína, která zkoumá konstrukci a studii algoritmů, které se mohou učit z dat. Takovéto algoritmy pracují způsobem sestavení modelu, který je založený na vstupních datech a používá předpovědi a rozhodnutí spíše než pouze explicitně naprogramované pokyny.

Strojové učení lze považovat za podoblast informatiky a statistiky. To má silné vazby na umělou inteligenci a optimalizaci, které poskytují metody, teorie a aplikace domény. Strojové učení je tedy obor v rozsahu výpočetních úloh, kde explicitní navrhování a programování, založené na pravidlech algoritmů, je nemožné. Jako příklady strojového učení můžeme zahrnout filtrování nevyžádané pošty nebo například optické rozpoznávání znaků. Termín strojové učení je někdy zaměňován s dolováním dat, s tím, že se zaměřuje více na oblast prozkoumání analýzy dat.

Toto odvětví rozdělujeme na základě způsobů učení do čtyř skupin. Učení s učitelem, učení bez učitele, kombinace učení s učitelem a bez učitele (semi-supervised learning) a zpětnovazební učení. (Witten, 2005)

3.6.1 Základní druhy úloh

- Klasifikace je základní úlohou ve strojovém učení, která nám přiřazuje vstupní data do správné výstupní skupiny – klasifikuje. Tyto třídy mohou být definovány pomocí obchodních pravidel, hranicemi tříd nebo matematickou funkcí. Klasifikace může být založena na vztahu mezi známou třídou přiřazení

a charakteristikou entity, která má být klasifikována. K tomuto účelu máme danou trénovací množinu dat, která vznikla pozorováním. Tyto jednotlivá pozorování jsou analyzována do množiny klasifikovatelných vlastností, které jsou známy jako nezávislé proměnné (features). Tyto vlastnosti můžou být kategoriální, ordinální, celočíselné nebo reálné. Jako klasický příklad problému klasifikace je udáván problém přiřazení diagnózy danému pacientovi, podle jeho pozorovaných charakteristik. (Nisbet, 2009)

- Regrese je funkce používaná v oboru data mining k předpovídání čísel, teploty, vzdálenosti, zisku nebo například prodeje v následujících letech. K tomuto předpovídání slouží regresní model, který by mohl být například použit pro odhad hodnoty nějakého domu. Odhad je poté založen na počtu pokojů, velikosti domu, lokaci a dalších faktorů. Úloha regrese začíná s datovým modelem, ve kterém je známa cílová hodnota, která má být dosažená. Ve zmíněném příkladu odhadu hodnoty domu, by mohl být model založen na pozorovaných datech pro domy v různých letech či jiných časových úsecích. Na hodnotu domu může mít vliv také stáří pokojů, oblast kolem domu (škola, úřady, nákupní centra) a tak dále. V tomto případě by poté byla hodnota domu brána jako cílová hodnota a ostatní atributy jako predikátory. Regresní algoritmus odhaduje hodnotu cíle jako funkci predikátorů pro každý případ v datech. Tyto vztahy mezi predikáty a cílem jsou vypočteny v modelu, který může být poté aplikován na různá data, kde nejsou známé cílové hodnoty. Regresní modely jsou testovány a výsledkem je poté rozdíl mezi předpovídanou hodnotou a očekávanou. Jak již bylo zmíněno výše, tak je množina dat rozdělena do dvou množin, kde jedna slouží pro vybudování modelu (natrénování) a druhá pro testování modelu. Techniky regrese můžeme rozdělit na lineární a nelineární. Lineární techniky můžeme použít v případech, kdy vztah mezi predikáty a cílem můžeme aproximovat lineárně (přímkou). Oproti tomu u nelineárních technik nemůžeme popsat vztah lineární funkcí. Regresní modely jsou používány v trendových analýzách, podnikovém plánování, marketingu nebo například pro finanční předpověď. (Regression Oracle, 2015)
- Shlukovací analýza hledá shluky datových objektů, které jsou v jistém smyslu podobné jeden druhému. Členové jednoho shluku jsou více podobné ostatním členům svého shluku než členům jiného shluku. Cílem je tedy najít kvalitní shluky tak, že mezi jednotlivými shluky je podobnost nízká, ale vnitřní podobnost v jednotlivých shlucích je vysoká. Oproti klasifikaci shlukování rozděluje data do skupin, které nebyly dříve známé. Je tedy zřejmé, že oproti klasifikaci, kde se děje rozdělení dat do předem známých skupin dat (tříd), nemá shlukovací model předem známý cíl. Techniku shlukování lze také dobře použít pro detekci anomálií nebo odhlehlých hodnot. Shlukování se používá v mnoha oblastech strojové učení jako například rozpoznávání vzorů, analýza obrazů a bioinformatiky. Shluková analýza není sama o sobě algoritmus, ale obecně úloha k řešení. Řešení může být dosáhnuto různými algoritmy, přičemž se tyto algoritmy liší svým pojetím. Každý algoritmus má totiž jinak definováno, co je představováno pod pojmem shluk a jak je co nejefektivněji najít. Volba

vhodného shlukovacího algoritmu a parametrů nastavení, závisí na individuální množině dat a na zamýšleném využití výsledků shlukování. Shlukovací algoritmy mohou být charakterizovány hierarchicky (uskupení objektů do hierarchické struktury, kde hierarchie může být tvořena shora dolů nebo zdola nahoru), dělícími příčkami (rozdělení datových objektů do daného počtu shluků. Shluky jsou poté vytvořeny tak, aby se optimalizovala objektivní kritéria, jako je například vzdálenost) nebo například lokálními bázemi (uskupení sousedních datových objektů do skupin založených na místních podmínkách). (Clustering Oracle, 2015)

3.6.2 Učení s učitelem

Učení s učitelem, někdy nazývané také konceptuální učení, je úlohou strojového učení, kde se jedná o vyvozování funkcí z označených trénovacích dat. Trénovací data jsou množinou trénovacích vzorů (příkladů). Každý z těchto vzorů je pak složen z párů, které se skládají ze vstupního objektu a požadované výstupní hodnoty. Učení s učitelem je tedy algoritmus, který analyzuje trénovací data a produkuje nad nimi požadovanou funkci, která se používá pro mapování nových vzorů (příkladů).

Každý algoritmus má svůj optimální scénář a algoritmus Učení s učitelem není výjimkou. Ideální scénář bude umožňovat algoritmu správné rozdělení neviditelných instancí do popsaných tříd. Tento požadavek znamená, že algoritmus nesmí být přeučen nad trénovacími daty. Pokud by tato situace nastala, tak by nám poskytoval nejasné a nepravdivé výstupy (výsledky). Musí být tedy zachována generalizace od trénovacích dat do neviditelných instancí.

Aby bylo dosaženo cíle, popsaného výše, je potřeba provádět následující kroky:

- V prvním kroku je potřeba určit správně typ trénovacích dat. Při analýze ručně psaného textu to může být například jednotlivý znak, celé slovo nebo například celý řádek textu.
- Dále je potřeba nastavit celou množinu trénovacích dat. Musíme správně spárovat vstupní slovo s požadovaným výstupem. Jako příklad můžeme uvést analýzu příspěvku na sociální síti Twitter, kde bychom zkoumali, zda daný příspěvek je pozitivní či negativní. Manuálně bychom mohli shromáždit několik příspěvků, z nich vybrat negativní slova a poté bychom mohli na datech říci, že pokud obsahuje příspěvek nějaké slovo ze shromážděné množiny negativních slov, tak je příspěvek negativní, v opačném případě pozitivní či neutrální.
- Dalším krokem je určení učící funkce. Přesnost této funkce je silně závislá na tom, jak je reprezentován vstupní objekt. Obecně je vstupní objekt přeměněn do vstupního vektoru. Každý vstupní vektor je poté popsán vlastnostmi. Počet těchto vlastností by neměl být příliš velký, ale měly by obsahovat dostatek informací, abychom mohli předpovědět výstup s co možná největší pravděpodobností. Zmíněné doporučení se uvádí především kvůli výpočetní a prostorové náročnosti algoritmů.
- Ve 4. kroku je potřeba určit, jaký učící algoritmus bude použit.

- V tomto kroku se již provádí běh algoritmu nad shromážděnou trénovací množinou dat. U některých algoritmů můžeme nastavit parametry výpočtu, abychom docílili optimalizaci výkonu. Tyto parametry upravujeme například na základě validační množiny či na základě kros-validace.

Posledním krokem je vyhodnocení přesnosti učící funkce. Veškeré měření by mělo byt prováděno na testovací množině dat, která je disjunktní s trénovací množinou dat. (Schmueli, 2010)

3.6.3 Učení bez učitele

Učení bez učitele zkoumá, jak se systémy mohou naučit reprezentovat určitý vstup vzorů způsobem, který odráží statistickou strukturu celkové kolekce vstupních vzorů. Problém učení bez učitele je snaha najít skryté struktury v neoznačených datech. V tomto případě není žádná zpětná vazba, zda daný algoritmus pracuje správně či ne. Neexistují žádné explicitní výstupy nebo prostředí, ve kterém by bylo znázorněné spojení s každým vstupem. Neexistuje žádná chyba či zpětný signál, podle něhož by se vyhodnotilo potenciální řešení. Agent v tomto případě přináší spíše předchozí zkreslení aspektů struktury vstupu, které by měly být zachyceny na výstupu. Zmíněný popis je hlavním rozdílem oproti učení s učitelem.

Metody učení bez učitele se používají v mnoha technikách, které se snaží shrnout a vysvětlit klíčové vlastnosti dat. Hodně těchto metod je poté založeno na metodách oboru dolování znalostí z dat, a to především pro předzpracování dat.

U učení bez učitele existuje několik přístupů. Je to například shlukování, skryté Markovovy modely nebo například analýza nezávislých komponent. (Schmueli, 2010)

3.6.4 Částečné učení s učitelem

Částečné učení s učitelem je třídou úlohy učení s učitelem a technik, které používají neoznačená data pro trénink. Typickým příkladem je malá množina označených dat a velké množiny neoznačených dat. Toto učení tedy spadá mezi učení s učitelem a učení bez učitele. Mnoho výzkumů ve strojovém učení rozhodlo, že neoznačená data s použitím s malou množinou označených dat, mohou produkovat značené zlepšení v přesnosti učení.

Použití označených dat pro učící problém často požaduje začlenění lidského faktoru. Jedná se například o přepsání audio nahrávky nebo například stanovení 3D struktury proteinu atd. Kvůli tomuto problému můžou náklady spojené s označováním vstupních dat stanovit trénovací proces neproveditelným. Zatímco získávání neoznačených dat je relativně nenákladné. V takových situacích je ideálním případem použití částečného učení s učitelem, kdy je použit jak teoretický zájem ve strojovém učení tak lidský faktor. (Zhu, 2009)

3.6.5 Posilovací učení

Posilovací učení je typ strojového učení a tudíž i odvětvím umělé inteligence. Jedná se o učení, které nám říká, co dělat a jak mapovat situace na akce takovým způsobem, aby se maximalizovala odměna signálu. Agentovi není řečeno, které akce mají být použity, ale místo toho objevuje akce, které mají největší vliv na získanou odměnu signálu. Tyto akce nemusejí mít vliv pouze na okamžitou odměnu, ale také na další situace a všechny následující odměny. Posilovací učení je charakteristické právě těmito dvěma znaky – vyhledávání metodou pokus-omyl a získáním odměny.

Posilovací učení však není charakterizováno pouze učícími metodami, ale také učícím problémem. Každá metoda, které je vhodná pro řešení problému, se poté považuje za metodu posilovacího učení. Základní myšlenkou je co nejjednodušeji zachytit nejdůležitější aspekty reálného problému učícího agenta ve spojení s okolím za účelem dosáhnutí požadovaného cíle. Učící agent musí být schopný snímat stav prostředí v určitém rozsahu a musí být schopen přijmou opatření, které mají vliv na stav prostředí. Přitom musí zahrnovat tři základní aspekty – pocit, akci a cíl.

Od učení s učitelem se posilovací učení liší v možnosti učení z interakce. Učení s učitelem není dostačující pro učení se z interakce, neboť se učí na základě příkladů, které byly poskytnuty učiteli. V interaktivních problémech je často nepraktické získat příklady požadovaného chování, které jsou korektní a reprezentují všechny situace, ve kterých agent musí jednat. V těchto situacích je vhodné použití posilovacího učení, neboť jeho agent je schopný se učit z vlastních zkušeností.

V současné době je jedním z největších výhod použití posilovacího učení v tom, že dokáže propojit obor umělé inteligence a jiné strojírenské disciplíny. Posilovací učení je možné použít ve spoustech aplikací, především díky schopnosti generalizovat problém. Problémy řešené v oblasti posilovacího učení je možné mapovat do rozhodovacích problémů. Tato vlastnost vede k možnostem aplikovat stejnou teorii do mnoha rozdílných domén specifických problémů s malým úsilím. V praxi je v současné době odvětví posilovacího učení používáno nejvíce v ovládání robotických paží, v robotické navigaci nebo například k hledání největší efektivnosti kombinace motorů. Jako příklad použití si můžeme uvést příklad, kdy se mobilní robot snaží rozhodnout, zda má vstoupit do nové místnosti za účelem shromáždění více odpadků nebo se snažit najít cestu zpět do své dobíjecí stanice. Rozhodnutí je přitom založené na tom, jak rychle a snadno se mu podařilo najít dobíjecí stanici v minulých případech. (Sutton, 1998)

4 SEER, analyzovaná data a proces přípravy dat

V hlavní části této práce se seznámíme s programem, který se zabývá sběrem dat ohledně rakoviny. Dále si představíme data sloužící k analýze, následně dojde k jejich podrobnému popsání. Dalším krokem bude popsání vlastní analýzy, která bude provedena v modelovacím programu IBM SPSS Modeler a také popsání dosažených výsledků.

4.1 Program SEER

Zkratka SEER pochází z anglických slov Surveillance, Epidemiology and End Results a ve volném překladu se jedná o Dohled, Epidemilogie a Výsledky a jedná se o program Národního rakovinového institutu v USA. Hlavním cílem programu je poskytování informací o statistice rakoviny za účelem snížení zátěže rakoviny mezi populací žijící v USA.

Program obsahuje registry, ve kterých jsou uložena data již od 1. ledna 1973. První data ohledně rakoviny byla uložena ve státech Conecticut, Iowa, New Mexiko, Utah, Hawai a v metropolitních oblastech Detroit a San Francisko. V následujících letech byla přidána data z oblasti Atlanta a oblasti ze Seattle. V roce 1978 byly přidány převážně černošské a indiánské země z oblastí Georgia a Arizony. Do roku 1990 byly přídany ještě státy New Orleans, Louisiana, New Jersey a Puerto Rico. Dále jsou v tomto programu shromažďována data z oblasti Aljašky.

Od roku 1992 došlo k rozšíření sběru dat i na menšiny, a to především na Hispánce, přidáním krajů Los Angeles a čtyř krajů v San Francisco. V roce 2001 byly přidány oblasti Kentucky a zbývající kraje ze státu Kalifornie. Dále byly do programu opět přidány, po vyřazení v roce 1977 a 1989, státy New Jersey a Louisiana a v roce 2010 byly již zahrnuty do programu všechny kraje ze státu Georgia.

V některých státech dochází k financování programu nejen pomocí Národního rakovinového institutu, ale také financování v kombinaci s Centrem pro řizení nemoci a prevence. Tato kombinace financování je u států Kentucky, Greater California, New Jersey, Louisiana a Greater Georgia. Rozdělení financování je ukázáno na obrázku 5. (About the SEER Program, 2015)

Obr. 5 Financování programu SEER
Zdroj: SEER registries, 2015

Zmíněné registry se dále seskupují pro analýzy. Pro analýzy se používá rozdělení do následujících 5 registrů.

- SEER 9 registry – v tomto registru jsou státy Atlanta, Conecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco, Seattle a Utah. V registru jsou data dostupná pro případy diagnostikované od roku 1973 a od roku 1974 byly registrované případy i pro San Francisco a od roku 1975 i pro stát Seattle.
- SEER 11 registry – registr vznikl v roce 1992 a obsahuje státy popsané výše plus státy Los Angeles a San Jose.
- SEER 13 registry – registr vznikl ve stejném roce jako registr SEER 11 a obsahuje stejné země plus země Georgii a Alasku.
- SEER 17 registry – registr vznikl v roce 2000 a obsahuje stejné země jako jsou obsaženy v registru SEER 13 a země Greater California, Kentucky, Louisiana a New Jersey. Data z roku 2005, v měsících červen až prosince, jsou vyloučena z registru a jsou popsána v jiném souboru. Vyloučení proběhlo z důvodu hurikánů Katrina a Rita, které měly velký dopad na populaci.
- SEER 18 registry – jedná se o poslední registr, ve kterém jsou seskupována data pro analýzy. Tento registr se skládá ze stejných států jako jsou obsaženy v registru SEER 17 a ještě stát Greater Georgia. Data v tomto souboru obsahují všechny případy od roku 2000 až do současnosti s výjimkou dat, které jsou ze zemí zasažených hurikány Katrina a Rita. (Registry Groupings for Analyses, 2015)

S daty v těchto registrech je manipulováno pomocí Systému řízení dat. Tento system zařizuje importování dat, editaci, propojování, integraci a reportování. Dále je v programu SEER zabezpečeno zvyšování kvality dat a její udržování pomocí části programu, které se touto oblastí přímo zabývá. Jejím cílem je:

- zavádění, provádění a řízení projektu zvyšování kvality,
- identifikování standardů a systematické měření daných norem,
- rozvíjení měřících a výpočetních nástrojů,
- posuzování dosažených výsledků a výkonostních měření,
- vyvíjení strategií, které zlepšují zmíněnou kvalitu,
- posouzení potřeb a satisfakce registru,
- sledování kvality dat,
- poskytování vzdělávání a školení k zvýšení kvality dat. (SEER Data Management System, 2015)

4.2 Analyzovaná data

Analyzovaná data pochází z jednoho z registru, který byl popsán výše, a to SEER 18. Jedná se o registr, ve kterém jsou uložena data ze všech států, které mají data v programu SEER. V tomto registru došlo k rozšíření v oblasti identifikace rasy pacienta a hispánské etniky. Data dále pokrývala přibližně 27,8 % veškeré populace v USA, kdy je tento údaj založen na sčítání lidu, které proběhlo v roce 2010.

Poskytnutá data programem mají danou stromovou strukturu. V kořenové složce dochází k rozdělení na složku obsahující data o populaci a na složku, ve které jsou data ohledně incidence. Tento adresář obsahuje data od počátku spuštění programu, až do současné doby. V adresáři populace jsou složky, ve kterých jsou údaje o různých rasách (Hispánci, běloši a černoši). Dále jsou data rozdělena na údaje podle daných let (registrů) a na poslední úrovni jsou rozdělena do dvou textových souborů s názvy singleages[10] a 19agegroups[11]. V textovém souboru singleages jsou údaje za každého člověka, který byl uveden v programu zvlášť a v textovém souboru 19agegroups jsou údaje o populaci rozděleny do 19 skupin podle věku osoby, která byla léčená ohledně různého onemocnění rakoviny. V druhé složce v tomto adresáři jsou údaje ohledně incidence. Data jsou rozdělena podle států a jednotlivých registrů. V každém registru jsou textová data. Každý jeden textový soubor obsahuje údaje o jednom druhu rakoviny. Jsou zde tedy údaje o rakovině prsu, tlustého střeva a konečníku, ostatních trávicích orgánů, ženských genitálií, leukémie a lymfatického onemocnění, mužských genitálií, dýchacích orgánů, močových cest a všech ostatních druhů rakoviny. Jednotlivá data jsou poté popsána v přiloženém souboru seerdic.pdf. Pro naše účely a hlavní část praktické práce jsme si zvolili data ohledně rakoviny prsu.

[10] Singleagess – jednotlivé roky

[11] 19agegroups – 19 věkových skupin

Textový soubor singleages.txt a 19agegroups.txt obsahuje 19 atributů, které popisují každého člověka, či skupinu lidí, kteří byli zaneseni do námi analyzovaného registru. Jelikož je daný registr anonymní a nesmí obsahovat žádné konkrétní informace, tak není v žádném souboru uvedeno jméno či blíže specifikované osobní údaje jedince. Jednotlivé atributy v souboru jsou:

- Year (rok) – atribut je složen ze čtyř numerických znaků, které udávají rok, kdy se daná osoba narodila.
- State postal abbreviation (státní poštovní zkratka) – jedná se o dvouznakový atribut, který určuje stát, ze kterého pochází daná osoba.
- State FIPS[12] code – atribut je složen ze dvou alfabetických znaků. Byl publikován Federal Information Processing Standard v roce 1987 a sloužil k identifikaci amerických států a dalších podobných oblastí. V roce 2008 byl nahrazen standardem ANSI INCITS[13] 38.2009.
- County FIPS code – atribut je složen ze tří numerických znaků a byl definován a popsán stejnou organizací jako atribut výše.
- Registry (registr) – jedná se o identifikaci registru a daný atribut je složen ze dvou numerických atributů. Například 21 = Hawai nebo 42 = Kentucky).
- Race (rasa) – atribut "rasa" obsahuje různé informace, které záleží na tom, zda byl záznam pořízen od roku 1973 do roku 1992 nebo od roku 1992 do současnosti. Pokud byl záznam pořízen od roku 1973, tak identifikuje, zda je záznam spjatý s bělochem, černochem či jinou rasou. Pokud byl záznam pořízen od roku 1992, tak identifikuje, zda je záznam spjatý s bělochem, černochem, americkým indiánem či asiatem. V našem případě budeme pracovat se záznamy od roku 1992 až do roku 2010, takže budeme identifikovat, zdali je osoba běloch, černoch či jiné rasy.
- Origin (původ) – tento atribut byl aplikován na pořízená data až od roku 1992. Identifikuje, zda-li je či není daná osoba hispánec. Pokud nebyl atribut vyplněn, tak je u tohoto atributu příznak "Not applicable".
- Sex (pohlaví) – jak z názvu daného atributu vyplývá, tak se jedná o atribut, který popisuje, zda je osoba žena (2) či muž (1). Daný atribut je důležitý, neboť můžeme dosahovat velmi zajímavých výsledků, pokud budeme data zkoumat z hlediska pohlaví.
- Age (věk) – jedná se o atribut, který udává věk jednotlivé osoby nebo skupinu, do které spadá více různých stáří. Atribut nám říká, kolik bylo dané osobě let, když se začala léčit na různý druh rakoviny. Daný atribut je složen ze dvou numerických znaků. Označení se liší podle toho, zdali se jedná o textový soubor, který obsahuje skupinu různých stáří nebo jednotlivce. Pokud se jedná o soubor skupin, tak má tento atribut hodnoty od 00 do 18, kde poslední hodnota (18) obsahuje skupinu lidí, kteří měli věk větší než 85 let. Oproti tomu soubor, který

[12] FIPS – Federal information processing standards (federální standardní zpracování informací)

[13] ANSI INCITS – American National Standards Institute International Committee for Information Technology Standards

obsahuje jednolivé stáří osob, má hodnoty od 00 do 85, kde jednotlivé dvojice čísel udávají přesný věk nemocného.

- Population (populace) – atribut je složen z 10 znaků a popisuje populaci.

Abychom lépe pochopili obsažená data v daném registru, je nutné blíže popsat i textový soubor "BREAST.txt". Soubor obsahuje 143 atributů, které popisují jednotlivé instance dat. Soubor obsahuje mnoho atributů, které obsahovaly pouze jednu hodnotu, či neměly, po bližším prozkoumání, vliv na zkoumanou problematiku. Mezi jednotlivé atributy patří Patient ID, Registry ID, Marital Status, Race/Ethnicity, Spanish/Hispanic Origin, NHIA Derived Hispanic Origin, Sex, Age at Diagnosis, Birthdate, Birthplace, Sequence number, Month of Diagnosis, Year of Diagnosis, Primary Site, Laterality, Histology, Behavior, Histologic Type, Behavior Code, Grade, Diagnostic confirmation, Type of Reporting source, EOD-Tumor size, EOD-Extension, EOD-Extension Prost path, EOD-Lymph node involv, Regional nodes positive, Regional nodes examined, EOD-old 13 digit, EOD-old 2 digit, EOD-old 4 digit, Coding system for EOD, Tumor Marker 1-3, CS Tumor size, CS Extension, CS Lymph nodes, CS Mets at DX, CS Site-specific factor 1-6, CS Site-specific factor 25, Derived AJJC-6T-M, Derived AJJC-6 Stage GRP, Derived SS1997, Derived SS2000, Derived AJJC-Flag, Derived SS1997-Flag, Derived SS2000-Flag, CS Version input original, CS Version derived, CS Version input current, RX Summ-Surg prim site, RX Summ-Scope reg LN sur, RX summ-Surg OTH REG/DIS, RX Summ-reg LN Examined, Reconstruction-First Cource, Reason for no surgery, RX Summ-Radiation, RX-SUM-Rad to CNS, RX Summ-surg/rad seq, RX Summ-Surg type, RX Summ–surg site 98-02, RX Summ-Scope reg 98-02, RX Summ-Surg OTH 98-02, SEER record number, Over-ride Age/site/morph, Over-ride seqno/dxconf, Over-ride sit/lat/seqn, Over-ride surg/dxconf, Over-ride site/type, Over-ride Histology, Over-ride Report source, Over-ride Ill-define site, Over-ride Leuk/Lymphoma, Over-ride Site/Behavior, Over-ride site/EOD/dx dt, Over-ride site/lat/EOD, Over-ride site/lat/morph, SEER type of follow-up, age recode<1 year olds, Site recode ICD-0-3/WHO 2008, Recode ICD-O-2 TO 9, Recode ICD-O-2 to 10, ICCC site recode ICD-O-3/WHO 2008, ICCC site rec extended ICD-O-3/WHO 2008, Behaviour recode for analysis, Histology recode—broad groupings, Histology recode—brain groupings, CS schema v0204, Race recode(white, black, other), Race recode(W, B, AI, API), Origin recode NHIA (Hispanic, Non-hisp), SEER historic stage A, AJCC stage 3rd edition (1988-2003), SEER modified AJCC stage 3rd ED (1988-2003), SEER sumarry stage 1977, SEER sumarry stage 2000, Number of primaries, First malignant primary indicatior, State-county recode, Cause of death to seer site recode, COD to site REC KM, Vital status recode, IHS link, Summary stage 2000 (1998+), Aya site recode/WHO 2008, Lymphoma subtype recode/WHO 2008, SEER cause specific death classification, SEER other cause of death classification, CS tumor size ext/eval, CS lymph nodes eval, CS mets eval, Primary by international rules, ER status recode breast cancer (1990+), PR status recode breast cancer (1990+), CS Schema—AJCC 6th ED (previously called V1), CS site specific factor 8, 10, 11, 13, 15, 16, Lymph vascular invasion, Survival months, Survival Months flag, Survival monts–presumed alive, Survival monts–presumed alive flag, Insurance recode (2007+), Derived AJCC-7 T-

M, Derived AJCC-7 stage GRP, Adjusted AJCC 6th T (1988+), Adjusted AJCC 6th N (1988+), Adjusted AJCC 6th M(1988+), Adjusted AJCC 6th Stage (1988+).

Z těchto atributů jsme pro naši práci odstranili ty, které nebyly důležité anebo ty, které obsahovaly pouze jednu hodnotu nebo hodnoty, které se nevztahovaly k rakovině prsů. Tyto odstraněné atributy jsou popsány v samostatné kapitole níže a je zde také uveden důvod nezařazení daných atributů pro vlastní analýzu v použitém programu.

4.2.1 Zpracování dat

V prvním bodě zpracování dat, jsme museli zpracovat textový soubor "BREST.txt", který obsahoval jednotlivé záznamy onemocnění rakoviny prsů. Obecně byla data v tomto souboru holá, prostá a na první pohled neposkytovala žádné informace. Pro úpravu dat byl použit skript napsán v programovacím jazyce Perl. Tento skrip je zobrazen na obrázku 6. Pomocí tohoto scriptu byl vytvořen soubor typu csv[14]. Po vytvoření souboru csv, bylo potřeba projít data a zkontrolovat, zda došlo ke správnému rozdělení atributů. K jednotlivým atributům bylo potřeba doplnit záhlaví. Toto záhlaví jsme doplnili na základě souboru seerdic.pdf, který byl přiložen s daty.

[14] Comma-separated values = Soubor, ve kterém jsou odděleny jednotlivé hodnoty čárkou.

```
#!/usr/bin/perl

use warnings;
use strict;
local $/ = "\r\n";
my $source=shift;

if(open(FILE, "$source")){
  while( my $line = <FILE>){
    #my $line=<FILE>;
    chomp $line;
    my $i=0;
    my $j=0;
    my @array=(4,2,2,3,2,1,1,1,2,10);
    #print "Scalar is-".scalar(@array)."\n";
    foreach $a (@array){
      print substr($line,$i,$a);
      #print "$i-$a\n";
      #print $j;
      if($j<(scalar(@array)-1)){
        print ","}
      $i=$i+$a;
      if($i==193){$i=198;}
      if($i==211){$i=217;}
      if($i==224){$i=225;}
      if($i==250){$i=254;}
      $j++;
    }
    print "\n";
  }
  close(FILE)
}else {die "File couldn't open \n";}

```

Obr. 6 Skript v programovacím jazyce Perl

Jak již bylo zmíněno výše, pro účely práce byla inkludována data pouze ohledně rakoviny prsů z období posledních 10 let. Soubor obsahoval údaje od roku 2000 do roku 2010 pro státy Greater California, Kentucky, Louisiana, New Jersey a Greater Georgia s tím, že od roku 2005 do roku 2010 bylo potřeba nahradit data za stát Louisiana ze souboru yr2005.lo_2nd_half.txt. Toto nahrazení bylo potřeba udělat kvůli ovlivnění populace ve zmíněném státu hurikány Katrina a Rita. Takto upravený soubor byl poté předložen jako vstup programu IBM SPSS Statistics verze 13.

V programu SPSS Statistics byla data zpracována následujícím způsobem. V první řadě byl načten upravený zdrojový soubor. Tento zdrojový soubor byl tvořen upraveným souborem csv. a měl již vymazány nepotřebné atributy, které jsou popsány v samostatné kapitole níže. V programu SPPS Statistics bylo potřeba určit, jakým způsobem jsou v souboru uložena data a jak rozdělit jednotlivé atributy. Velká výhoda programu SPSS Statistics je, že automaticky rozezná typ proměnných, ale pokud je potřeba, tak tento typ můžeme jednoduše změnit v záložce Variable View. Typy jednotlivých atributů byly změněny na základě informací poskytnutých v souboru seerdic.pdf a také na webových stránkách programu SEER. V této záložce je také možné definovat, zdali je atribut vstupní či cílová proměnná. Další velkou

výhodou je, že jsme mohli jednoduše přiřadit jednotlivým numerickým hodnotám, které reprezentovaly atributy, jejich popisek. Tohoto docílíme ve sloupci "Values", kdy je vždy potřeba určit numerickou hodnotu a k té odpovídající popisek. Takto jsme prošli všech 72 různých atributů. Tento krok je značnou výhodou, pokud chceme v dosažených výsledcích analýzy lépe číst a vyznat se v tom, co nám který atribut reprezentuje. Soubor měl také několik atributů, které obsahovaly prázdné záznamy. Prázdné hodnoty byly nahrazeny hodnotami, které jsme si sami určili. Další možností bylo, že bychom smazali záznamy, které obsahovaly prázdné hodnoty, nicméně bychom tak přišli zhruba o jednu třetinu všech záznamů, a tak jsme upřednostnili zmíněné nahrazení určenou hodnotou. Prázdné záznamy byly u atributů zobrazených níže, kde je určena také hodnota, která sloužila k nahrazení chybějící hodnoty.

- CS tumor size – 666
- CS extension – 666
- CS lymph nodes – 666
- CS mets eval – 66
- CS site-specific factor 1 – 66
- CS site-specific factor 2 – 66
- CS site-specific factor 3 – 66
- CS site-specific factor 4 – 666
- CS site-specific factor 5 – 666
- CS site-specific factor 6 – 666
- Derived AJCC-6 T – 66
- Derived AJCC-6 N – 66
- Derived AJCC-6 M – 66
- Derived AJCC-6 Stage GRP – 66
- Derived SS1997 – 6
- Derived SS2000 – 6
- RX Summ-surg oth reg/dis – 6
- CS tumor size ext/eval – 7
- CS Lymph nodes eval – 7
- CS mets eval – 7
- Insurance recode – 8

Po tomto kroku jsme už soubor jen uložili do příslušného formátu (v našem případě se jednalo o formát .sav) a práce s programem skončila.

Dalším bodem při zpracování dat byla eliminace proměnných, které nemají vliv na zkoumanou problematiku. Eliminace proběhla na základě informací poskytnutých v souboru seerdic.pdf a také na základě informací, které jsme získali z webových stránek programu SEER. Atributy, které nebyly zařazeny do samostatného zpracování v programu IBM SPSS Modeler, jsou popsány v kapitole níže. U každého eliminovaného atributu je uveden popis a samostatný důvod nezařazení.

4.2.2 Nezařazená data

Po bližším prozkoumání jsme do části zpracování v programu SPSS Modeler, nezařadili 71 atributů. V této části bude popsáno 10 zajímavých atributů, které byly z finálních dat vyřazeny a zbylých 61 atributů je uvedeno v příloze práce.

- Histology (histologie) (92-00) ICD[15]-O-2 – atribut popisuje morfologii v průběhu času. Program SEER začal popisovat a zaznamenávat morfologii již v roce 1977 a od tohoto roku použil několik různých kódovacích systémů. Obvykle jsou analyzovaná data morfologie omezena jen na mikroskopicky potvrzené případy. V našem případě se jedná o verzi ICD-O-2, která ale obsahovala data převedená z novější verze ICD-O-3, která ve zkoumaných datech je zařazena a jednotlivé záznamy měly tudíž stejné hodnoty.
- Behavior (chování) code ICD-O-3 – atribut Behavior code ICD-O-3 reprezentuje chování jednotlivého nálezu a jeho kódy jsou zařazeny v atributu Histology Type ICD-O-3, který popisuje blíže morfologii. Tento atribut není zařazen v datech, která jsou použita v programu SPSS Modeler z důvodu, že hodnoty v něm obsažené jsou shodné s hodnotami, které jsou v atributu Behavior recode for analysis.
- EOD[16]—Tumor size (velikost nádoru) – tato položka je součástí 10-ti místného kódu EOD. Jedná se o atribut, který popisuje velikost tumoru pro případy do roku 2004. Pro naše účely nebylo vhodné zařadit tuto položku do finálních vstupních dat, neboť obsahovala velmi mnoho prázdných záznamů. Místo toho je ve finálních datech atribut CS[17] Tumor size, který reprezentuje velikost tumor pro případy od roku 2004 až do současnosti.
- EOD—Extension (rozšíření) – datová položka EOD-Extension není zařazena ve finálních datech ze stejných důvodů jako atribut EOD-Tumor size. Atribut nám popisuje nejdelší rozšíření nádoru z počátečního místa a jedná se buď o souvislé prodloužení či o vzdálené metastázy. V našem případě jsou tyto informace inkludovány v atributech CS Extension a CS Mets at DX.
- Tumor marker 1 (pozice nádoru) – atribut nám neposkytoval žádné bližší informace, neboť popisoval jen případy do roku 2003 a nesloužil pro případy rakoviny prsů. Ve finálních datech je nahrazen tento atribut atributem CS Site-specific Factor 1, který slouží pro popis rakoviny prsů z hlediska výskytu estrogen receptorů.
- CS Version input original (verze vstupního originálu) – tato položka udává číslo verze použité pro počáteční kód CS pole. První dvě číslice představují hlavní číslo verze, další dvě číslice představují drobné změny verze a poslední dvě číslice znamenají i méně významné změny, například opravy typografických chyb, které ne-ovlivňují kódování nebo odvození výsledků. Atribut však nebyl zařazen do finálních dat, neboť neposkytoval žádnou bližší informaci.

[15] ICD-O – Internation Classification of Disease for Oncology

[16] EOD – Extent of disease

[17] CS – Collaborative stage

- RX Summ[18]-Scope Reg LN[19] Sur (rozsah lymfatických uzlin) – atribut slouží pro definování rozsahu regionálních lymfatických uzlin, které byly chirurgicky odstraněny. Popisuje postup odstranění, biopsii nebo aspiraci regionálních lymfatických uzlin provedených během počátečního zpracování nebo prvním cyklu léčby na všech možných zařízeních. V našich datech byly poskytnuty pouze dvě možné varianty – prázdné nebo s hodnotou 9, která odpovídala neznámým nebo neaplikovaným případům. Prázdná pole potom odpovídala případům před rokem 2003.
- Reconstruction-first course (1998-2002) (rekonstrukce v prvním kurzu) – Program SEER začal shromažďovat informace k této položce jen pro rakovinu prsu a jen pro rekonstrukci, která začala v rámci prvního cyklu léčby. Tato položka je k dispozici pouze pro případy, které jsou diagnostikovány v letech 1998 až 2002. Nezařazení daného atributu ve finálních datech je právě z důvodu zaznamenávání informace pouze do roku 2002 a atribut tedy obsahuje spoustu prázdných polí.
- Survival months (přežité měsíce) – Atribut vytvořený kompletními datumy, zahrnující dny, a proto se může lišit od doby přežití vypočítané z let a měsíců. Atribut je popsán hodnotami 000-9998, značícími počet měsíců a 9999, kdy nebyla tato doba uvedena. Atribut nebyl zařazen do finálních dat, protože po bližším prozkoumání neposkytoval žádné další informace.
- Derived AJCC[20]-7 Stage GRP (odvozená fáze stádia) – jedná se o položku, ve které je obsažena AJCC "Stage Group" složka, která je odvozena pomocí algoritmu CS, z CS kódovaných polí. Tato položka je efektivní pro případy diagnostikovaných od roku 2010. V našem případě nebyla položka zařazena do finálních dat z důvodu velkého množství prázdných záznamů.

4.2.3 Zařazená data

Po redukování celé množiny vstupních dat eliminováním nepotřebných či nedůležitých atributů, jsme vytvořili finální množinu, která slouží jako vstup do programu IBM SPSS Modeler. Jedná se o nejdůležitější atributy, které byly popsány v dokumentaci SEER, a které jsme blíže zkoumali v programu Microsoft Office Excel 2013 a v programu IBM SPSS Statistics. V této části jsou popsány nejzajímavější atributy a zbylé atributy jsou popsány v příloze práce.

- Registry ID (stát) – jedná se o unikátní identifikaci osoby v registru, kde na základě tohoto ID jsou přiřazena data populace. Atribut má délku 10 numerických znaků a jako příklad si můžeme uvést ID pro Kentucky, které má hodnotu 0000001542.

[18] Summ – shrnutí

[19] LN – lymph nodes (lymfatické uzliny)

[20] AJCC – American Joint Comittee on Cancer (Americký výbor zabývající se rakovinou)

- Marital Status (manželský stav) – tato datová položka identifikuje stav pacienta v době, kdy u něj byl potvrzen výskyt tumoru. Stav může být buď svobodný, ženatý, žijící odděleně, rozvedený, ovdovělý nebo neznámý stav.
- Race/Ethnicity (rasa/etnika) – dvouznakový numerický atribut, který upřednostňuje rasy, které nespadají k bělochům. Jedná se převážně o smíšené rasy. U atributu si můžeme všimnout, že všechny možnosti nebyly efektivně v datech použité. Jako příklad si můžeme uvést hodnotu 01, která odpovídá bělochům anebo například hodnota 05, která odpovídá Japoncům.
- Sequence number–central (pořadové centrální číslo) – atribut slouží k identifikaci a popisuje počet a pořadí všech zaznamenaných zhoubných, in situ, benigních a hraničních primárních nádorů, které se vyskytly v průběhu životnosti pacienta. Pořadové číslo se může v průběhu životnosti pacienta měnit. Pokud byl jednotlivec dříve diagnostikován s jediným potvrzeným zhoubným nádorem a následně je diagnostikován s druhým potvrzeným zhoubným nádorem, sekvenční kód pro první záznam se změní z 00 na 01. Atribut je složen ze dvou číselných znaků, které značí počet nalezených nádorů. Značení začíná od 00 do 88, kde první polovina slouží k označení zhoubných nebo in-situ nádorů a od hodnoty 60 se atribut používal k označení non-maligních tumorů.
- Sex – tento atribut určuje, zda-li je osoba muž či žena (1, 2).
- Primary Site (primární místo) – položka označuje místo, ze kterého pochází primárně nádor. Atribut je určen pomocí Mezinárodní klasifikace nemocí pro onkologii, třetí vydání (*International Classification of Diseases for Oncology, third edition)* (ICD-O-3) pro topografické kódy. Případy diagnostikované od roku 1977 do roku 1991 byly kódovány pomocí Mezinárodní klasifikace nemocí pro onkologii, 1976 Edition (ICD-O-1976). Pro případy diagnostikované před rokem 1977, bylo použito kódování *Manual of Tumor Nomenclature and Coding*, *1968* (MOTNAC). Všechny případy mezi lety 1973-1991 byly strojově převedeny na ICD-O-2 kódování bez úplného ručního přezkoumání.
- Laterality (lateralita) – lateralita popisuje stranu spárovaného orgánu nebo části těla, na kterém byl potvrzený původ tumoru. Atribut se začal používat od roku 1.1.2004 až doposud. Atribut se používá pro vybrané invazivní, benigní a hraniční primární mozkové a nádory CNS.
- Diagnostic confirmation (diagnostické potvrzení) – druh potvrzení diagnózy nám říká, jakým způsobem byl u pacienta potvrzen nález onemocnění rakoviny. Druhy potvrzení se rozdělují na mikroskopické potvrzení, nemikroskopické potvrzení nebo neznámé, kde byl výskyt potvrzen až po smrti pacienta.
- Type of reporting source (typ zdrojového souboru) – položka reprezentuje druh zdroje, souboru, který nejlépe popisuje průběh onemocnění. Nemusí to být nutně originální soubor, ale je potřeba chápat tento atribut jako položku, která reprezentuje dokument, který nejlépe popisuje onemocnění a průběh léčby. Jedná se o atribut, který je vyjádřený jedním numerickým znakem. Druhy reportu jsou ústavní nemocnice, onkologická či ozařovací zdravotnické centra, laboratoře, lékařské kanceláře, úmrtní listy nebo ostatní nemocnice.

- Regional nodes positive (regionální pozitivní uzliny) – atribut ukazuje přesný počet regionálních lymfatických uzlin vyšetřených patologem, které bylo zjištěno, že obsahují metastázy.
- CS Tumor size – jak z názvu vyplývá, jedná se o atribut, který nám udává velikost tumoru. Atribut je dostupný od roku 2004. Případy před rokem 2004 mohou být konvertovány a přidány nové kódy, které nebyly k dispozici před aktuální verzí CS. Různé případy jsou popsány v tabulce 2.

Tab. 2 CS Tumor size

Kódy	**Popis**
000	Kód označuje stav, když nebyl nalezen tumor
001 - 988	Exaktní velikost v mm
989	989 mm nebo větší
990	Mikroskopická zaměření nebo jen ohniska. Není dána žádná velikost zaměření
991	Popsány jako menší než 1 cm
992	Popsány jako menší než 2 cm
993	Popsány jako menší než 3 cm
994	Popsány jako menší než 4 cm
995	Popsány jako menší než 5 cm
996 - 998	Nové velikostní kódy
999	Neznámý, velikost není uvedena, nebylo uvedeno v záznamu pacienta
888	Není aplikován

- CS extension – Informace o rozšíření nádoru. K dispozici od roku 2004 až do současnosti. Dřívější případy mohou být převedeny a dodány nové kódy, které nebyly k dispozici pro použití před aktuální verzi CS. Dřívější připady se měnily tak, že se z dvouznakového označení stalo tříznakové tak, že se přidala 0 zprava. Výjimkou je označení 99, které bylo změněno na 999 a slouží pro neznámé rozšíření.
- CS Lymph nodes (lymfatické uzliny) – Informace o zapojení lymfatických uzlin. K dispozici pro případy od roku 2004. Dřívější případy mohou být převedeny a dodány nové kódy dodal, které nebyly k dispozici pro použití před aktuální verzi CS.
- CS Mets at DX – Informace o vzdálených metastázách. K dispozici pro případy od roku 2004. Dřívější případy mohou být převedeny a dodány nové kódy, které nebyly k dispozici pro použití před aktuální verzi CS.
- CS Site-specific factor 1, 2, 3, 4, 5, 6 – Pro tyto atributy platí stejná definice. Každý CS faktor site-specific (SSF) je závislý na daném schématu. Ty mohou poskytnout informace potřebné k fázi případu, klinicky relevantní informace nebo prognostické informace. Je k dispozici pro různá období a schémata v závislosti na nastavení standardních požadavků. Dřívější případy mohou být

převedeny a dodány nové kódy, které nebyly k dispozici pro použití před aktuální verzi CS.

- Derived SS2000 (odvození shrnutí fáze) – datová položka je odvozená ze "SEER Summary Stage 2000" (shrnutí fáze) z algoritmu CS. Atribut je efektivní u případů diagnostikovaných roku od 2004.
- RX Summ—Surg prim site (operace primárního místa) – atribut Surgery prim site popisuje chirurgický zákrok, který odstraňuje anebo ničí tkáň primárního místa, který byl proveden jako součást počátečního zpracování, nebo v prvním cyklu léčby.

Tab. 3 RX Summ – Surg prim site

Kód	**Popis**
00	Žádný, nebyl proveden žádný chirurgický zákrok primárního místa. Byl diagnostikován jen na základě pitvy
10-19	Zničení nádoru. Žádný patologický vzorek ani není známo zda nějaký patologický vzorek existuje
20-80	Resekce, patologický vzorek
90	Byl proveden chirurgický zákrok na primární místa, ale není známá žádná informace o typu chirurgického zákroku
98	Zvláštní kódy pro hemopoetické[21], retikuloendoteliální[22], imunoproliferativní[23], myeloproliferativní[24] onemocnění; špatně definované místa; a neznámé primární volby, výjimkou je úmrtní list
99	Není známo, jestli byl provede zákrok. Pouze na základě úmrtního listu

- Reason for no surgery (důvod pro neprovedení operace) – tato datová položka dokumentuje důvod, proč nebyla provedena operace na primárním místě výskytu rakoviny.

[21] Hemopoetické – tvorba červeného krevního barviva

[22] Retikuloendoteliální – imunitní onemocnění

[23] Imunoproliferativní – jedná se o nemoc těžkých řetězců alfa

[24] Myeloproliferativní – onemocnění kostní dřeně

Tab. 4 Důvod pro neprovedení operace

Kód	Popis
0	Operace byla provedena
1	Operace nebyla doporučena
2	Kontradikována vzhledem k jiným podmínkám. Jediným případem je pitva
5	Pacient zemřel před doporučenou operací
6	Neznámý důvod pro neprovedení operace
7	Operace byla odmítnuta pacientem
8	Doporučena. Není však známo, jestli byla provedena
9	Neznámo.

- RX Summ – Radiation (radiace) – v atributu jsou zahrnuty různé druhy radiace, které byly provedeny v prvním cyklu léčby.
- RX Summ – surg/rad seq (sekvence operace/radiace)- Toto pole zaznamenává pořadí, ve kterém byly chirurgie a radiační terapie aplikovány pro ty pacienty, kteří byli určeni pro chirurgický zákrok i pro ozařování.

Tab. 5 Pořadí operace a ozařování

Kód	Popis
0	Ozařování ani operace nebyla provedena
2	Ozařování před operaci
3	Operace před ozařováním
4	Ozařování před i po operaci
5	Předoperační ozařovací terapie
6	Předoperační ozařovací terapie s jiným druhem ozařovaní před i po operaci
9	Neznámé pořadí, ale obě dvě mětody byly aplikovány

- Age recode <1 year olds (věk) – Proměnná age recode je založena na věku v době stanovení diagnózy (single-years věkové kategorie). Seskupení používaná v této proměnné se stanoví na základě věkových skupin v údajích o počtu obyvatel. Tento záznam má 19 věkových skupin v proměnné age recode (<1 rok, 1-4 roky, 5-9 let, ..., 85+ let). Hodnoty v tomto atributu jsou dvouznaková čísla, která začínají na hodnotě 00, která odpovídá rokům < 1. Následuje dalších 19 hodnot, kde prvních 18 odpovídá věkovým skupinám a poslední, která má hodnotu 99, odpovídá záznamům, kde nebyl věk znám.

- ICCC[25] site recode ICD-O-3/WHO 2008 – jedná se o záznam, který nám udává informace o místě/histologii, a který se používá hlavně pro analýzu dat u dětí. Záznam byl aplikován na všechny případy bez ohledu na věk, aby mohlo dojít k věkovému srovnání s těmito uskupeními (ICD-O-3 a WHO 2008). Lze poznamenat, že diagnostikované případy před rokem 2001 nebyly kódovány v ICD-O-3 a byly převedeny na ICD-O-3 z ICD-0-2, a případ nemusí mít stejnou specifičnost jako případy po roce 2000, které byly kódované přímo pod ICD-O-3.
- Number of primaries (počet nádorů) – hodnota u položky je založena na celkovém počtu nádorů v těle osoby. Položka je volitelná a hodnota je stejná u všech nádorů u osoby.
- First malignant primary indicator (indikátor prvního zhoubného nádoru) – atribut je založen na všech nádorech v SEER. Nejsou zaznamenávány tumory, které nejsou maligní. 0 = ne, 1 = ano.
- Cause of death to SEER site recode (případ smrti) – tato položka byla vytvořena k vypočítání několik nových validních ICD-10 kódů a zahrnuje jak případy smrti zapříčiněné rakovinou, tak i případy smrti nazaviněné rakovinou.
- Vitalstatus recode (životní stav) – Každý pacient, který zemře po sledovaném dnu, je překódován na živý k danému dnu. 1 = živý, 4 = mrtvý.
- IHS[26] link – Incidenční soubory jsou pravidelně spojeny se soubory Indian Health Service (IHS) k identifikaci původních Američanů. Záznam o původu využívá informace z tohoto oboru a používá se pro určení, zda je člověk domorodý Američan nebo ne.
- AYA[27] site recode/WHO 2008 – atribut obsahuje informace o místě/ histologii nádoru, který se používá hlavně pro analýzu údajů o dospívajících a mladých dospělých. Záznam byl aplikován na všechny případy bez ohledu na věk, aby mohlo dojít ke srovnání mezi těmito věkovými skupinami.
- SEER cause-specific death classification (klasifikace specifického případu úmrtí) – vytvořen pro použití v příčinách specifického přežití. Tato proměnná označuje, že osoba zemřela na příčinu rakoviny nebo přežila či zemřela na jiné příčiny. Hodnota 0 = přežila nebo zemřela na jiné příčiny, 1 = zemřela, 9 = N/A nebyl to první nádor.
- SEER other cause of death classification (klasifikace jiného případu smrti) – tento atribut označuje, že člověk zemřel na jiné příčiny, než na příčinu rakoviny, pro kterou se léčil. Hodnota 0 = přežil nebo zemřel v důsledku rakoviny, 1 = zemřel, 9 = N/A není to první tumor.
- Primary by international rules (nález pomocí mezinárodních pravidel) – atribut byl vytvořen pomocí IARC[28] primárních pravidel. Pravidla nezahrnují benigní nádory nebo in-situ nádory, které nejsou v oblasti močového měchýře. Nebyla

[25] ICC – International Classification of Childhood Cancer

[26] IHS – Indian Health Service – Indické zdravotnictví

[27] AYA – Adolescents and Young Adults – dospívající a mladí dospělí

[28] IARC – International Agency for Research on Cancer – mezinárodní agentura pro výzkum rakoviny

změněna žádná informace o nádoru v žádném záznamu. Hodnoty v atributu jsou 1 – ne, 2 – ano a 9 – vyloučen z IARC primárního algoritmu kvůli chování.

- ER[29] Status recode breast cancer (1990+) (záznam stavu rakoviny prsu) – atribut vytvořený kombinací informací z atributu Tumor marker 1 (1990-2003), s informacemi z CS site-specific factor 1 (2004+). S tím, že toto pole je prázdné pro případy, kdy nebyla zjištěna rakovina prsu a u případů diagnostikovaných před rokem 1990. Hodnoty u tohoto atributu jsou 1 – pozitivní, 2 – negativní, 3 – hraniční, 4 – neznámý, 9 – ne 1990+ prsa.
- Insurance recode (2007+) (pojištění) – Atribut popisující druh pojištění. U tohoto atributu je potřeba dbát zvýšené opatrnosti při jeho zpracování, neboť většina pacientů, kteří byli ve věku 65 let a starší v době diagnózy, kteří byli klasifikováni jako "Nepojištěný", nebo mají "Soukromé pojištění", nebo mají stav "Pojištění neznámé" byli Medicare způsobilé, zatímco u pacientů diagnostikovaných před svým 65 rokem nebyly. Hodnoty jsou 1 = nepojištěný, 2 = Medicaid, 3 = pojištěný, 4 = pojištěný/nespecifikovaný, 5 = pojišťovací status neznámý, 9 = nedostupný.
- Adjusted AJCC 6th T (1988+) (upravená fáze tumoru) – atribut je vytvořený sloučením z EOD třetí edice a CS informacích o nemoci. Atribut je dostupný pouze pro schéma prsů a od roku 1988. Pro pochopení atributu si můžeme uvést pár příkladů, jak docházelo a dochází ke změnám v datech. TX[30] klesá u většiny schémat v čase. Kromě celkového poklesu klesá počet TX spíše mezi roky 2003 a 2004, kdy začal CS. Jak počet uvedených TX časem klesá, tak ostatní kategorie ukazují nárůst, které jsou adekvátně upraveny klesajícím TX. Velikostní kategorie pro rok 2004 a dále mají podkategorie, za účelem klasifikovat případy, které nemají konkrétní informaci o velikosti, jako je například hodnota "stated as T1[31]". Hodnota "Any T with Mets[32]" není kategorie AJCC 6th T, ale byla vyvinuta, protože starší případy s M1[33] přecházely do TX M1 a vlastní T kategorie byla ztracena. Aby došlo k podobným definicím, tak případy M1 pro roky 2004+, které byly založené na prodloužení, byly odebrány z individuální T kategorie a byly umístěny do "Any T with Mets". Podkategorie T mohou být odstraněny, pokud jsou v rozporu mezi EOD a CS nebo nemohou být definovány pro EOD. Upravené TX jsou kombinací TX plus všech nezařazených případů podle stádia. Pro nezařazené případy podle stádia došlo k rozporu mezi EOD a CS případech v tom, že případy CS mohly mít známou T a N složku s MX, zatímco EOD případy měly mít alespoň jednu neznámou složku. Jednotlivé kategorie T neexistují pro případy, které měly fázi zařazenou jako neznámou nebo M1, což odpovídá obvykle fázi IV. Kategorie T je upravená NX[34],

[29] ER – estrogen receptor

[30] TX – označení velikosti a rozšíření tumoru

[31] Stated as T1 – uvedeno jako T1, kde T1 označuje velikost a rozšíření tumoru

[32] Any T with Mets – nějaký tumor s metastázemi

[33] M0, M1, MX – označení velikosti a typu metastáz

[34] N0, N1, N2, N3, NX – označení velikosti a typu uzlin

pokud se informace mapuje do stáze neznámé. Pokud jsou potřebné kategorie T a N pro případy s AJCC 6. ed. neznámé fáze, použijí se proměnné "Adjusted AJCC T, 6th ed (2004+)" a "Derived AJCC N, 6th ed (2004+)", které jsou k dispozici pouze pro 2004+.

- Adjusted AJCC 6th N (1988+) (upravená fáze uzlin) – atribut je vytvořený sloučením z EOD třetí edice a CS informací o nemoci. Atribut je dostupný pouze pro schéma prsů a od roku 1988. Pro lepší pochopení atributu si můžeme uvést několik příkladů, jak se data měnila nebo mění. Počet NX klesá u většiny schémat v čase. Kromě celkového poklesu je zřejmé, že největší pokles nastal mezi roky 2003 a 2004, kdy začal CS. Jako vždy, když dojde k poklesu v jedné kategorii, je následné zvýšení v jedné nebo více jiných kategoriích. "Any N with Mets[35]" není kategorie AJCC 6th N, ale byla vyvinuta proto, že starší případy EOD s M1 přecházely do NX M1 a vlastní N kategorie byla ztracena. Aby došlo k podobným definicím, tak případy M1 pro roky 2004+, které byly založené na prodloužení, byly odebrány z individuální N kategorie a byly umístěny do "Any N with Mets". Podkategorie N mohou být odstraněny, pokud jsou v rozporu mezi EOD a CS nebo nemohou být definovány pro EOD. Kategorie N v AJCC 6th je složitější než jednoduché kódy používané v EOD. Jedním z problémů u této kategorie je, že EOD nezahrnuje všechny kombinace kategorie regionálního uzlu a neodděluje je klinicky od patologické informace. Podkategorie N0, N1, N2 a N3 mohou být dostupné v odvozeném AJCC 6th N kódu, ale pro srovnatelnost s EOD musí být podkategorie udržovány v hlavních N kategoriích N0, N1, N2 a N3, pokud jsou kategorie dostupné pro roky 1988 až 2003. Tento proces je proveden pouze tehdy, když nejsou použity konkrétní dílčí kategorie N0, N1, N2 nebo N3 k určení fáze. Podkategorie N mohou být odstraněny, pokud jsou v rozporu s EOD a CS nebo nemohou být definovány pro EOD(např. N2a a N2b mohou být převedeny do N2). Upravené NX jsou kombinací NX plus všech nezařazených případů podle stádia. Pro nezařazené případy podle stádia došlo k rozporu mezi EOD a CS případech v tom, že případy CS mohly mít známou T a N složku s MX, zatímco EOD případy měly mít alespoň jednu neznámou složku (TX a NX). Dalším způsobem, jak se lze na data podívat, je, že pro CS (2004+) byly kódovány nezávisle hodnotou T a N. Kategorie N, N0, N1, atd neexistují pro případy, které měly neznámou fázi nebo ekvivalent M1 do fáze IV. Kategorie N je NX upravena, pokud se informace mapuje do neznámé fáze – upravený NX. Pokud jsou potřebné kategorie T a N pro případy s AJCC 6th ed. neznámé fáze, použijí se proměnné "Derived AJCC T[36], 6th ed (2004+)" a "Derived AJCC N[37], 6th ed (2004+)", které jsou k dispozici pouze pro 2004+.
- Adjusted AJCC 6th M (1988+) (upravená fáze metastáz) – atribut je vytvořený sloučením z EOD třetí edice a CS informacích o nemoci. Atribut je dostupný

[35] Any N with Mets – nějaké uzliny s metastázemi

[36] Derived AJCC 6th T – odvození stavu tumoru

[37] Derived AJCC 6th N – odvození stavu uzlin

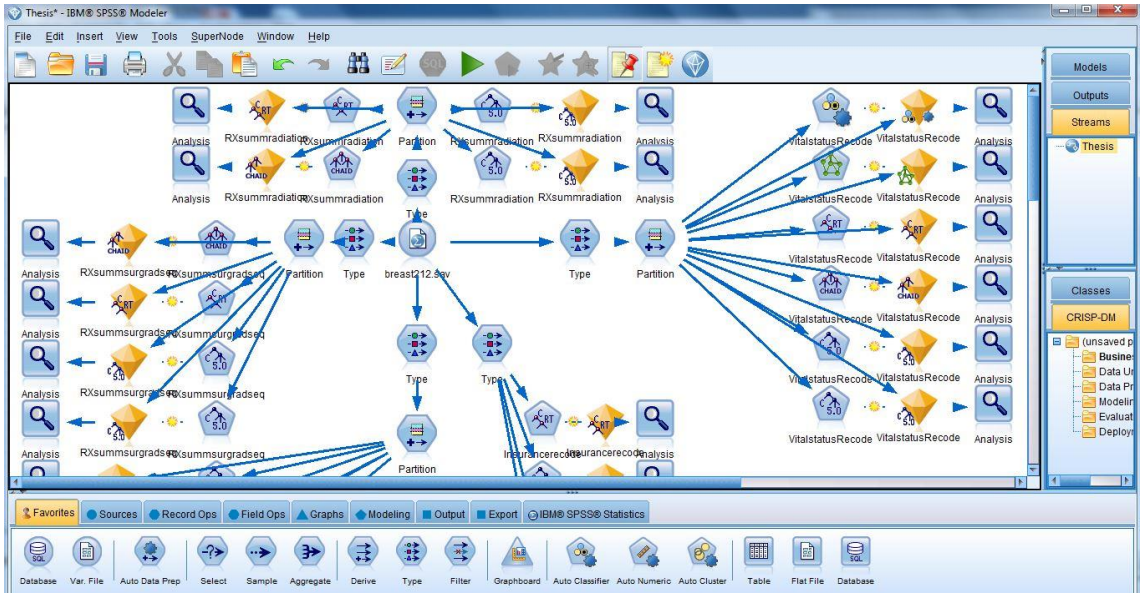
pouze pro schéma prsů a od roku 1988. Můžeme zde uvést některé příklady. Jedna z možností, kterou může daný atribut obsahovat, je hodnota MX. MX se snižuje pro většinu schémat v průběhu času. Kromě celkového poklesu je také zřejmé, že došlo k nejvíce změn v letech 2003 až 2004, kdy začal CS a některé z těchto změn můžou být způsobeny tím, že MX je kódováno nezávisle na T nebo N pro CS, ale ne pro EOD. V datech je také vidět, že kolem roku 2010, MX ukazuje pokles a zvyšuje se M0. Tyto změny jsou dány tím, že lékaři a zapisovatelé rakoviny se dozvěděli, že v příští verzi AJCC budou případy MX považovány za M0. Záznamy "no distant metastases" se měnily v průběhu času, což vede k trvalému růstu M0 a následným poklesem v MX, který vyústil ve více představených případech. Tyto změny nelze vyřešit mezi EOD a CS. Opatrnosti je třeba při porovnávání M0 a MX v průběhu času.

Jakmile jsme měli definován zdrojový soubor, popsána data a uložený soubor v potřebném datovém typu, tak jsme mohli přejít k vlastnímu sestavování modelu.

4.3 Model v program IBM SPSS Modeler

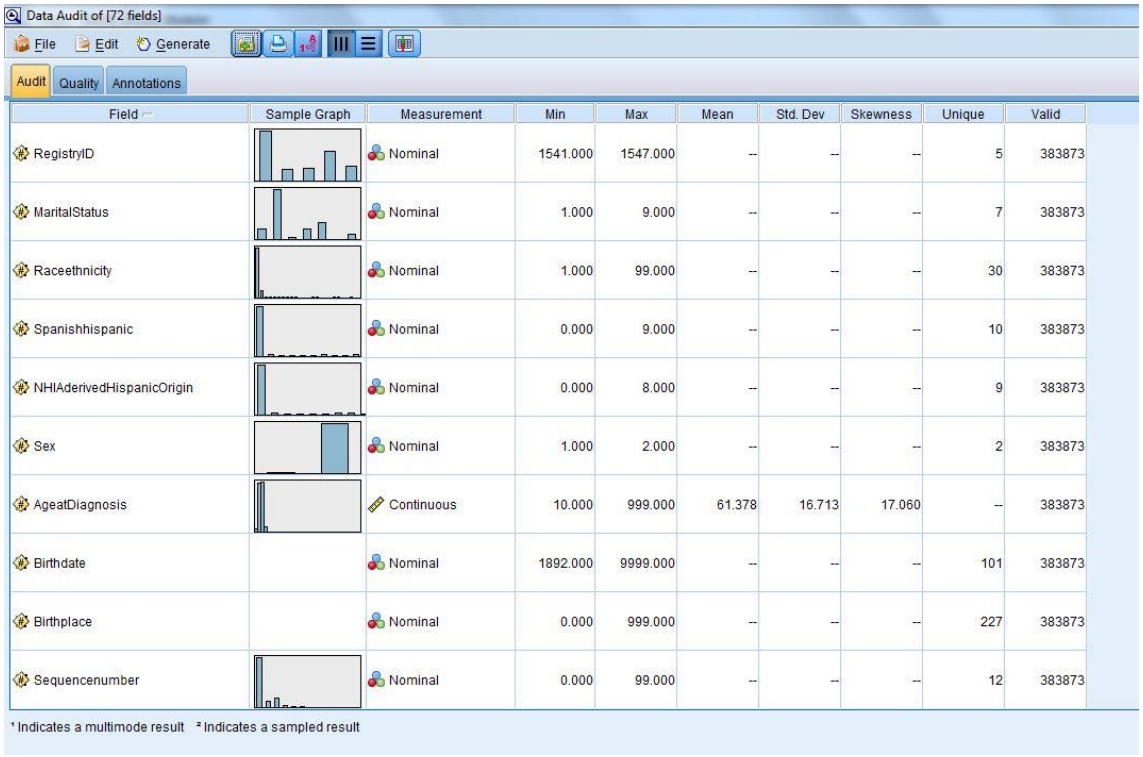
V kapitole, kde jsme popisovali metodologii oboru data mining, jsme se seznámili s krokem sestavení modelu. V této kapitole se blíže seznámíme se zmíněným krokem.

Na začátku sestavování modelu bylo potřeba upravit zdrojový soubor. Toto upravení bylo popsáno v předešlé kapitole a bylo provedeno pomocí skriptu v jazyce Perl. Následovala úprava samostatných dat. V programu IBM SPSS Modeler bylo jako prvním krokem potřeba načíst zdrojová data. K tomuto účelu slouží modul, ze záložky "Sources" a to "Statistics file", jelikož jsme měli data uložena ve formátu .sav. Data jsme si poté zobrazili modulem "Table" a poté použili modul "Type", kde je možné zobrazit datový typ atributů. Datový typ již byl definován v programu SPSS Statistics, a proto zde došlo k překontrolování a k možnému definování vstupních a výstupních proměnných. Samostatný model je zobrazen na obrázku číslo 7. Jak je vidět na zmíněném obrázku, tak byla provedena analýza dat pro 3 definované atributy – Vitalstatus recode, RX Summ-radiation, RX Summ-surg/rad seq. Dále je na obrázku vidět, že byly použity čtyři základní algoritmy, a to rozhodovací stromy C5, CART a CHAID a neuronová síť. Rozhodovací stromy jsou popsány blíže v samostatné podkapitole. Neuronová síť byla použita u analyzování každého algoritmu, nicméně dosahovala velmi špatných výsledků. Největší možná přesnost, které bylo dosaženo, je 36 % přiřazení do správné hodnoty a 64 % do špatné hodnoty. Z tohoto důvodu nebyla ani popisována v samostatné kapitole. Nicméně jednotlivé výstupy Neuronové sítě jsou znázorněny v příloze.



Obr. 7 Model v programu IBM SPSS Modeler

Další moduly, které byly v modelu použity, jsou "Data audit", "Partition" a "Analysis". Modul "Data audit" slouží pro zobrazení všech použitých proměnných a zobrazení různých vlastností proměnných. Na obrázku 8 je vidět vždy název proměnné, jednoduchý graf, který nám zobrazuje počet jednotlivých hodnot obsažených v proměnné, typ proměnné, minimální a maximální hodnotu, střední hodnotu, počet unikátních hodnot obsažených v atributu nebo například počet všech validních záznamů.

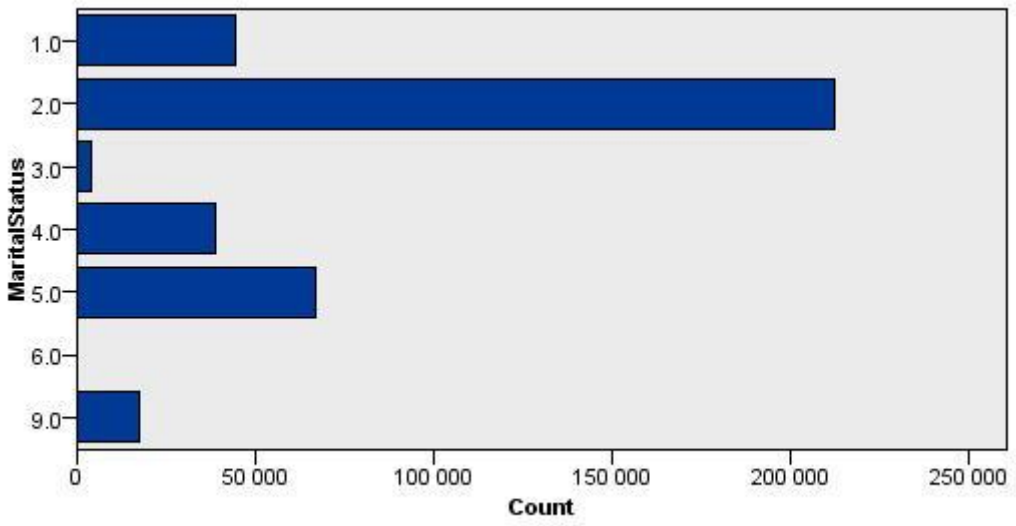


Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
RegistryID		Nominal	1541.000	1547.000	--	--	--	5	383873
MaritalStatus		Nominal	1.000	9.000	--	--	--	7	383873
Raceethnicity		Nominal	1.000	99.000	--	--	--	30	383873
Spanishhispanic		Nominal	0.000	9.000	--	--	--	10	383873
NHIAderivedHispanicOrigin		Nominal	0.000	8.000	--	--	--	9	383873
Sex		Nominal	1.000	2.000	--	--	--	2	383873
AgeatDiagnosis		Continuous	10.000	999.000	61.378	16.713	17.060	--	383873
Birthdate		Nominal	1892.000	9999.000	--	--	--	101	383873
Birthplace		Nominal	0.000	999.000	--	--	--	227	383873
Sequencenumber		Nominal	0.000	99.000	--	--	--	12	383873

[1] Indicates a multimode result [2] Indicates a sampled result

Obr. 8 Modul "Data audit"

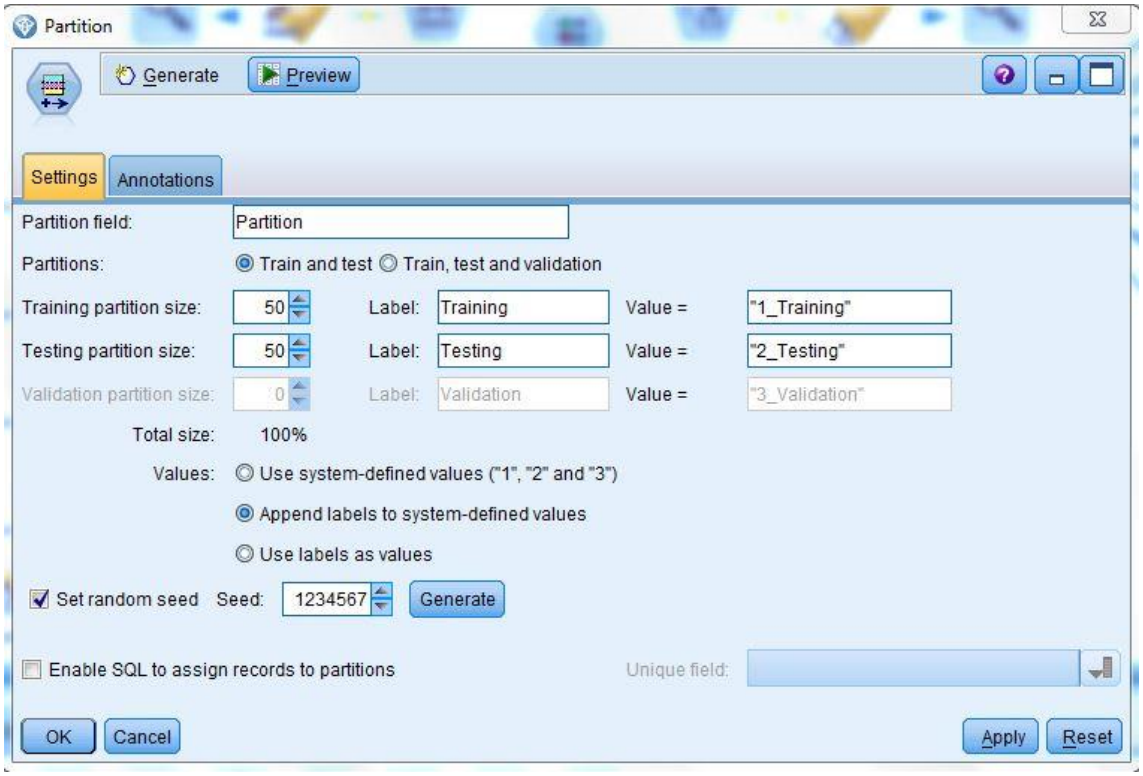
Jako příklad si můžeme uvést atribut "Marital status", který nám říká, jaký stav měla osoba pod daným záznamem. Po zobrazení grafu uvidíme hodnoty obsažené v proměnné a jejich počet. Jak je vidět na obrázku 9, tak největší počet záznamů mělo manželský stav ženatý a nejmenší počet poté rovnu žijící odděleně. Dále je možné si všimnout, že atribut je typu "nominal", má minimální hodnotu 1, maximální hodnotu 9, počet unikátních hodnot je roven 7 a validních záznamů je pro tento atribut 383373. Takto je možné si zobrazit každou proměnnou a zkontrolovat, jestli je potřeba daný atribut zahrnout do výsledných dat či nikoliv.



Obr. 9 Marital status – graf

Dalším použitým modulem byl modul "Partition". Modul "Partition slouží k rozdělení vstupních hodnot do trénovací a testovací množiny dat. Proces nastavení daného modulu je zobrazen na obrázku 10. U daného modulu můžeme nastavit název modulu, jestli chceme generovat jen trénovací a testovací data nebo trénovací, testovací a validační data. V našem případě bylo vždy pracováno jen s trénovací a testovací množinou. V dalším kroku je potřeba určit jak velké chceme jednotlivé množiny. Pro naše účely bylo optimální defaultní rozdělení, které je stanoveno na rozdělení 50 % všech dat do trénovací množiny a 50 % všech dat do testovací množiny. Toto rozdělení probíhá náhodně, což jsme si zobrazili modulem "Table". Dále je možné určit, jaký chceme název a předponu u jednotlivých množin. Znovu bylo zachováno defaultní rozdělení – trainng, 1_training a testing, 2_testing. V posledním kroku je možné určit, jak chceme definovat hodnoty u jednotlivých množin. Znovu bylo zachováno defaultní nastavení – append labels to system-defined values. Poté už jen zbývalo potvrdit změny.

Posledním modulem, který je v modelu použit a nejedná se o samostaný algoritmus, je modul "Analysis". Tento modul slouží k analýze, jak přesný byl daný algoritmus. V modulu můžeme nastavit, co vše si přejeme zobrazit. Pro naše účely bylo zachováno defaultní nastavení modulu. Dále bylo důležité zaškrnout možnost rozdělení na základě jednotlivých oddílů, kdy jsme poté mohli vidět poskytnutou přesnost predikce jak na trénovacích, tak na testovacích dat. Jak již bylo zmíněno výše, tento produkt byl použit v našem případě k určení přesnosti daných algoritmů, a proto jeho jednotlivé výsledky budou popsány u každého algoritmu zvlášť.



Obr. 10 Modul "Partition"

4.3.1 Použité algoritmy

V následující kapitole budou popsány základní algoritmy, které byly použity v nástroji SPSS Modeler.

- C5 – Algoritmus C5 spadá do množiny algoritmů strojového učení, které pracují na základě klasifikace a jedná se o metodu známou jako rozhodovací stromy. Funguje na principu učení s učitelem, kdy algoritmus musíme naučit rozhodovat pomocí vstupní množiny dat tzv. trénovací množiny. Rozhodovací stromy obecně pracují na principu, že na vstupu dostanou množinu případů S. Poté roste strom pomocí algoritmu rozhoduj a panuj následujícím způsobem:
    - Pokud všechny případy množiny S patří do stejné třídy nebo je množina S malá, je strom listem označený pomocí nejfrekventovanější třídy v S.
    - Jinak se vybere test založený na jednom atributu s dvěma nebo více výstupy. Provede se test kořene stromu s jednou větví pro každý výstup testu, část třídy S se rozdělí do korespondujících podmnožin S1, S2, …., podle tohoto výstupu se pro každý případ aplikují stejné procedury rekurzivně na každou podmnožinu.

Atributy C4.5 mohou být číselné či nominální hodnoty a pomocí těchto atributů jsou rozděleny testované výstupy. Pro číselné atributy A jsou {A≤h, A>h}, kde prahová hodnota h je nalezena pomocí seřazené množiny S na základě hodnoty A a vybírání mezi úspěšnými hodnotami maximalizovaného kritéria výše. Atribut A s diskrétními

hodnotami má defaultní jeden výstup pro každou hodnotu, ale má volbu, která umožňuje seskupovat hodnoty do dvou či více podmnožin s jedním výstupem pro každou podmnožinu. Inicializační strom je poté potřeba prořezat, abychom předešli přeučení. Prořezávací algoritmus je založen na pesimistickém odhadu míry chyby spojené s množinou N případů. (Wu, 2008) Prořezávání se vždy provádí od listů ke kořeni. Odhaduje se chyba na listu s N případy a E chyba je N-krát pesimistická míra chyby, která je uvedena výše. Pro každý podstrom přidává algoritmus C5 odhadovanou chybu větve a porovnává ji s odhadovanou chybou uzlu. Pokud je tato hodnota větší, tak je podstrom nahrazen listem, pokud ne, tak je na podstrom použit algoritmus prořezávání. Stejným způsobem C5 kontroluje odhadovanou chybu při nahrazování podstromu nějakou větví a když se nahrazení jeví jako prospěšné, tak je strom upraven. Přitom by měl být proces prořezávání dokončen jedním průchodem stromu.

Konstrukce stromu C5 je rozdílná od konstrukce dalšího rozhodovacího stromu CART v několika bodech:

- Testy v CART stromech jsou vždy pouze binární, ale ve stromech C5 umožňují testy 2 a více výstupů.
- CART používá pro hodnocení testů Gini různorodý index, zatímco C5 používá pro hodnocení testů kritéria založená na dosažených informací
- Prořezávání probíhá ve stromech CART pomocí Cost-complexity modelu, kde jsou parametry odhadované pomocí cross-validation a ve stromech C5 používá jednoduchý algoritmus založený na binomických limitech důvěry.

Hlavní nevýhoda algoritmů C5, je potřeba požadovaného počtu CPU času a paměti. (Wu, 2008)(Classification and Regression trees, 2015)

- CART – Zkratka C&RT znamená Classification and regression tree. Jedná se o rekurzivní oddělující metodu, která buduje klasifikační a regresní stromy pro predikci spojité závislé proměnné (regrese) a kategorické proměnné prediktoru (klasifikace). Klasický C&RT algoritmus byl propagován pány Breiman, Friedman, Olshen, & Stone v roce 1984. Algoritmus C&RT rozhodovací strom je binární rekurzivní oddělující procedura schopná zpracovávat spojité a nominální hodnoty, a to jak cíle, tak predikátory. Data jsou zpracovávána v jejich holé podobě, kde není žádný začátek či konec. Stromy rostou do maximální velikosti bez použití nějakých pravidel zastavování, ale používají zpětné prořezávání. Mechanismus algoritmu je určen k produkci více než jednoho stromu. Produkuje více sekvenčních vnořených prořezaných stromů, z nichž jsou všechny kandidáti pro optimální strom. Tento optimální strom je zvolen na základě vyhodnocení predikční výkonnosti každého stromu v sekvenci prořezávání, kde výkonnost stromu je měřena vždy na základě nezávislosti testovaných dat a výběr stromu postupuje pouze po vyhodnocení testu založenému na datech. Pokud nemáme k dispozici žádné testovací data a nebyl proveden proces cross-validace, bude poté C&RT trvat agnosticky na tom, který strom v sekvenci je nejlepší. To je na příklad v ostrém kontrastu s metodami rozhodovacích stromů jako C5, který generuje preferované modely

na základě měření provedeným na trénovacích datech. Mechanismus algoritmu C&RT zahrnuje automatické volitelné balancování tříd, automatickou správu chybějících hodnot a umožňuje nákladově citlivé učení, dynamické funkce konstrukce a odhad pravděpodobnosti stromu. Balancování tříd je metoda, která je používaná u algoritmu stromu C&RT. V defaultním klasifikačním módu C&RT vždy počítá třídy frekvence v nějakém uzlu vzhledem k třídě frekvence v kořenu stromu. To je ekvivalentní s automatickým vyvažováním dat do balančních tříd a zajišťuje, že strom zvolen jako optimální, minimalizuje vyváženou třídu chyby. Vyvažování je implicitní při výpočtu všech pravděpodobností a zlepšení a nevyžaduje uživatelský zásah. Hlavní výhody stromů C&RT jsou jednoduchá prezentace výsledků a také, že metody stromů jsou neparametrické a nelineární. Jednoduchost výsledků je výhodou nejen pro účely dosažení rychlé klasifikace nového pozorování, ale může také přinést často mnohem jednoduší model pro vysvětlení výsledků, proč je pozorování klasifikováno nebo předpovídáno daným způsobem. Stromové metody C&RT jsou zvláště vhodné pro úlohy data miningu, kde je často málo předem známých znalostí ani žádná koherentní množina teorií nebo předpovědí, které by popisovaly proměnné a vztahy mezi těmito proměnnými. (Wu, 2008) (Classification and Regression Trees, 2015)

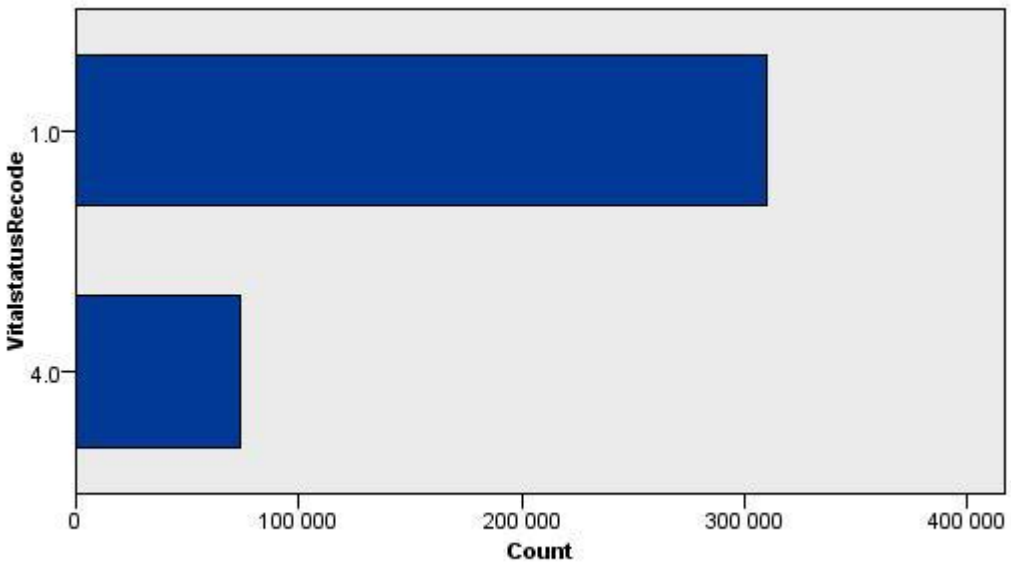
- CHAID – kratka CHAID znamená Chi-squared Automatic Interaction Detector, což lze přeložit jako automatický detektor chi-kvadrát interakcí. Jedná se o jednu z nejstarších stromovou klasifikační metodu, která byla původně navržena již v roce 1980 panem Kass. Strom CHAID buduje nebinární stromy, založené na relativně jednoduchém algoritmu, který se hodí k analýze velkých datových skladů. Stejně jako algoritmus CART, konstruuje CHAID stromy, kde každý uzel identifikuje rozdělující podmínku pro získání optimální předpovědi nebo klasifikaci. Z tohoto důvodu může být použit k řešení jak na problémů regrese tak na problémů klasifikace. (Claus, 2005) Algoritmus CHAID pracuje ve třech základních krocích: příprava predikátorů, sloučení kategorií a volba rozdělení proměnné. V prvním kroku algoritmus rozdělí spojité proměnné do více přibližně stejných kategorií. Pokud se jedná o proměnnou typu nominalní nebo kategoriální, tak je toto rozdělení již hotovo. V dalším kroku algoritmus cyklicky prochází predikátory a stanoví pro každého pár predikátorů, které jsou alespoň významně odlišné s respektem k nezávislým proměnným. Pokud algoritmus řeši klasifikační problém, bude tento úkon spracovávat vypočítáváním Chi-kvadrát testu. Tento krok se opakuje pro každé dvojice, které vzniknou sloučením již projitými dvojcemi. V případě, že statistická významnost pro příslušnou dvojici prediktivních kategorií je významná, poté bude vypočítána Bonferonniho upravena p-hodnota pro množiny kategorií pro příslušný predikátor. Jako třetím krokem je výběr rozdělení proměnné s nejmenší upravenou p-hodnotou, která přinese nejvýznamnější rozdělení. (CHAID analysis, 2015)

4.4 Analyzované atributy

Po analyzování vstupních dat, sestavení modelu a popisu jednotlivých bloků modelu bylo potřeba definovat, jaké atributy budou blíže analyzovány. V našem případě došlo k výběru 3 cílových atributů. Výstupní atributy jsou Vitalstatus recode, RX Summ-radiation, RX Summ-surg/rad seq. Jednotlivý postup analyzování každého atributu bude popsán v samostatné kapitole a bude obsahovat úpravu vstupních dat, nastavení jednotlivých atributů a popis získaných výstupů.

4.4.1 Vitalstatus recode

Atribut Vitalstatus recode byl již blíže popsán v kapitole věnující se zařazeným atributům do finálních dat. Nicméně v tomto bodě si ukážeme kolik a jaké hodnoty atribut obsahoval. Obsah atributu je znázorněn na obrázku číslo 11. Jak je zřejmé, tak u tohoto atributu převažovala hodnota 1, která nám značí, zda pacient žil až do sledovaného dne vyřazování. Oproti tomu hodnota 4, která značí, že pacient zemřel do sledovaného dne vyřazování, tvoří zhruba jen jednu čtvrtinu všech záznamů.



Obr. 11 Vital status recode

Dalším krokem, který bylo potřeba udělat, než došlo k vlastní práci s algoritmy je definování vstupních a výstupních proměnných, které budou vstupovat do daného algoritmu. Jako výstupní atribut byl nastaven zmíněný Vitalstatus recode a jako vstupní byly ponechány všechny ostatní. Jak však pozdější analýza ukázala, bylo potřeba u atributů SEER cause-specific death classification, SEER other cause of death classification a Cause of death to SEER site recode jejich roli na "none". Toto

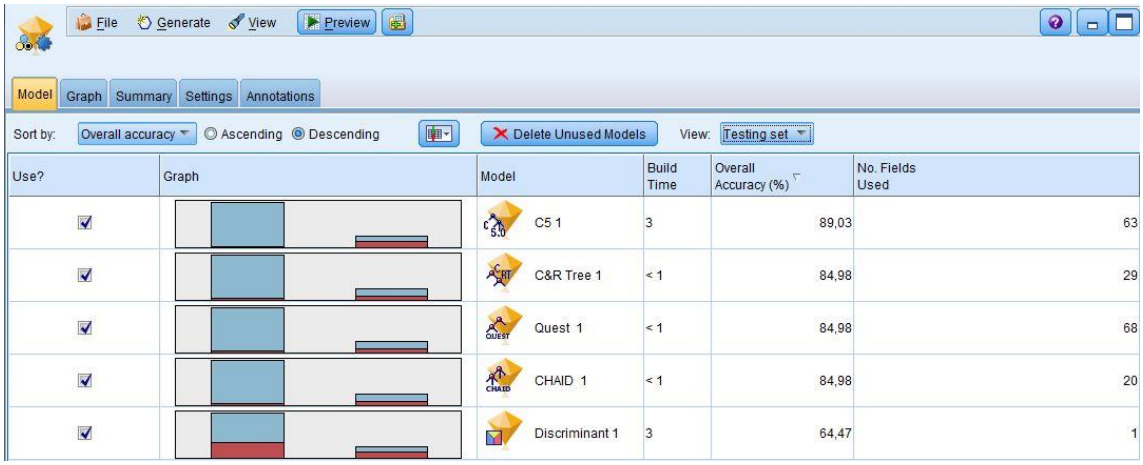
nastavení je vidět na obrázku 12 a k tomutto kroku došlo z důvodu velkého ovlivnění výstupního atributu těmito proměnnými.

OriginrecodeNHIAhispanicNonhisp	Nominal	0.0,1.0		None	Input
SeerhistoricstageA	Nominal	0.0,1.0,2.0,4.0,9.0		None	Input
Numberofprimaries	Nominal	1.0,2.0,3.0,4.0,5.0,6.0,7.0,10.0		None	Input
firstmalignantprimaryindicator	Nominal	0.0,1.0		None	Input
causeofdeathtoseersiterecode	Nominal	0.0,20020.0,20030.0,20050.0,20060.0,...		None	None
VitalstatusRecode	Nominal	1.0,4.0		None	Target
HISLink	Nominal	0.0,1.0		None	Input
AYAsiterecode	Nominal	13.0,17.0,18.0,21.0,22.0,23.0,25.0,28.0...		None	Input
SeerCausespecifiDeathclass	Nominal	0.0,1.0,9.0		None	None
SeerOtherCauseofDeathclass	Nominal	0.0,1.0,9.0		None	None
CSTumorSizeExt	Nominal	0.0,1.0,2.0,3.0,5.0,6.0,8.0,9.0	*	None	Input
CSLymphNodeseval	Nominal	0.0,1.0,2.0,3.0,5.0,6.0,8.0,9.0	*	None	Input
CSMetsEval	Nominal	0.0,1.0,2.0,3.0,5.0,6.0,8.0,9.0	*	None	Input
Primarybyinternational	Nominal	0.0,1.0,9.0		None	Input
ERStatusBreastCancer	Nominal	1.0,2.0,3.0,4.0		None	Input
PRStatusrecodeBreastcancer	Nominal	1.0,2.0,3.0,4.0		None	Input
Insurancerecode	Nominal	1.0,2.0,3.0,4.0,5.0	*	None	Input

Obr. 12 Typ atributu Vital status recode

Po definování vstupních a výstupních atributů, byly data nastavena na vstup modulu "Partition". Výstup modulu "Partition" poté sloužil jako vstup do jednotlivých algoritmů.

V prvním kroku byl spuštěn algoritmus "Auto Classifier". Tento algoritmus slouží k automatické analýze dat a k poskytnutí seznamu algoritmů, které se nejlépe hodí k použití na daná data. V našem případě bylo potřeba změnit defaultní nastavení algoritmu následujícím způsobem. Došlo ke změně v počtu modelů, které nám jsou předloženy jako výstup a také došlo ke změně ve velikosti procent u modelů, které se nemají použít. Defaultní nastavení bylo 80 %, které bylo změněno na 60 %, abychom dostali na výstup, co nejvíce algoritmů k následnému použití. Výstup algoritmu je znázorněn na obrázku 13 a ukazuje nám potřebnou dobu k sestavení modelu, celkovou přesnost daného modelu a počet nepoužitých proměnných.



Obr. 13 Vital status – Auto classifier

Na základě doporučení, které nám bylo poskytnuto jsme použili následující 4 algoritmy – C5, CHAID, C&R strom a neuronovou síť.

Neuronová síť byla použita, i když nebyla doporučena algoritmem Auto classifier. Její použití je z důvodu ověření, zda program poskytuje pravdivé informace. Během analyzování dat neuronovou sítí se však potvrdily schopnosti algoritmu "Auto classifier", neboť neuronová síť poskytovala velmi špatné výsledky. Jako příklad si můžeme ukázat výstup modulu "Analysis", kterému sloužil jako vstup výstup z agoritmu. Jak je vidět na obrázku 14, tak během porovnávání predikovaných výstupů se ukázalo, že algoritmus dosahuje pouze 36,62 % správné přesnosti predikování, zda daná osoba přežila, či zemřela. Oproti tomu je vidět, že během stejného porovnání dosahoval algoritmus 63,38 % špatné predikce.

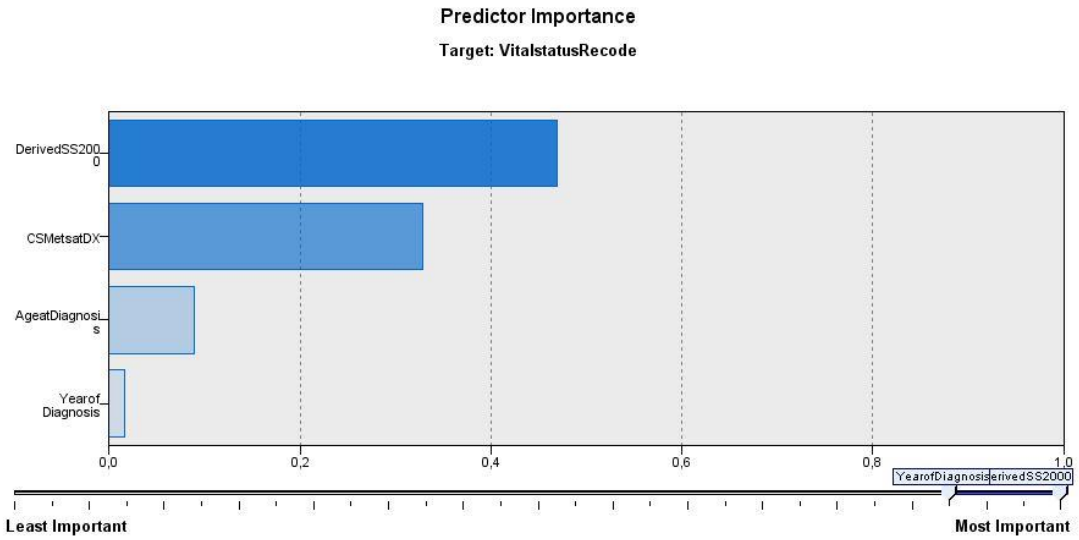
Comparing $N-VitalstatusRecode with VitalstatusRecode

'Partition'	1_Training		2_Testing	
Correct	70 004	36,49%	70 331	36,62%
Wrong	121 820	63,51%	121 718	63,38%
Total	191 824		192 049	

Obr. 14 Vital status recode – Neuronová síť

Rozhodovací strom C&RT byl použit k analýze dat na základě poskytnutých informací z programu "Auto classifier". Při nastavování algoritmu zůstala všechna nastavení v defaultním režimu. Bylo požadováno zobrazení důležitosti u predikovaných proměnných, zobrazení stromu a také zobrazení základního rozdělení. Na obrázku 15 je vidět výsledek algoritmu z hlediska důležitosti jednotlivých predikovaných proměnných. Na obrázku jsou vidět 4 nejvýznamnější proměnné, které ovlivňují výstup algoritmu. Nejvýznamnější proměnnou byla určena Derived SS2000, která popisuje chování nálezu. Druhou nejvýznamnější proměnnou je CS Mets at DX, která nám říká informaci ohledně vzdálených metastází. Třetí nejvýznamnější proměnnou, z hlediska predikce, je uvedena

proměnná Age at diagnosis, která nám udává věk pacienta, kdy mu byla diagnostikována rakovina prsu. Jako poslední nejvýznamnější proměnnou byla určena proměnná Year of diagnosis, která nám říká, v jakém roce byla u daného pacienta diagnostikována rakovina prsu.



Obr. 15 Vital status recode – C&RT

Dalším získaným výstupem algoritmu byl model stromu. Jelikož model stromu je příliš rozvětvený a jeho největší výška byla rovna 5, tak je přiložen v příloze. Nicméně nám ukázal, že pokud měli pacienti nález, který byl in-situ, lokalizovaný a lokální, kde se vykytovaly pouze lymfatické uzliny a nález byl diagnostikován u pacientů mladších než 73 let, kteří měli diagnostikován nález v roce 2000 až 2004 a měli fázi onemocnění, která byla reprezentována tím, že tumor byl in-situ, nebyly potvrzeny žádné vzdálené metastázy, ale v nálezu bylo zahrnuto 10 a více lymfatických uzlin, jich 853 zemřelo a 678 přežilo stanovený den. Lze tedy říci, že ze všech pacientů, kteři splňovali tato kritéria, byla u 55,7 % diagnostikována smrt a u 44,3 % pacient nezemřel. Tento počet pacientů odpovídá 1,14 % všech záznamů, které byly obsaženy v testovacích datech.

Na výstup algoritmu C&RT byl připojen modul "Analysis", neboť bylo potřeba zjistit, s jakou celkovou přesností predikce algoritmus pracuje. Po ukončení procesu algoritmu, jsme získali výsledek, který je ukázán na obrázku 16. Jak je na výstupu vidět, tak algoritmus poskytoval predikci s 84,87% správnou přesností na testovaných datech a predikci s 15,13% špatnou přesností na testovaných datech. Jelikož nám algoritmus předložil nízkou důležitost u jednotlivých proměnných sloužících k predikci, došlo k obavám, zda bude poskytovat správné výsledky. Nicméně modulem "Analysis", jsme potvrdili správné použití algoritmu a ukázali, že může dosahovat velmi dobrých výsledků nad těmito daty.

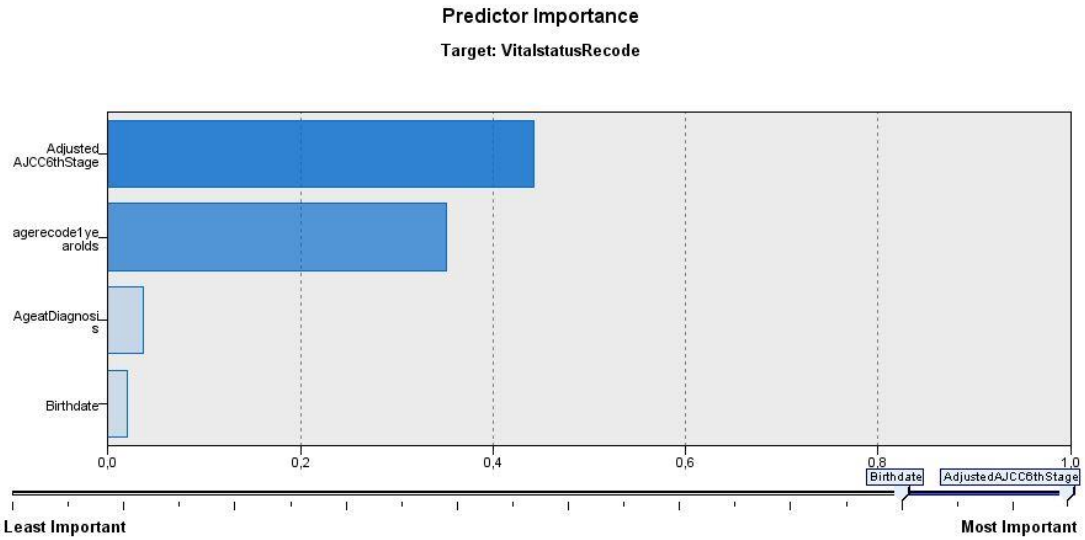
Results for output field VitalstatusRecode

Comparing $R-VitalstatusRecode with VitalstatusRecode

'Partition'	1_Training		2_Testing	
Correct	163 020	84,98%	162 988	84,87%
Wrong	28 804	15,02%	29 061	15,13%
Total	191 824		192 049	

Obr. 16 Vital status recode – C&RT Analysis

Dalším použitým algoritmem u tohoto atributu byl algoritmus rozhodovací strom CHAID. Algoritmus byl použitý na základě získaného výstupu algoritmu Auto classifier. U algoritmu zůstala všechna nastavení v defaultním stavu, jen byl znovu kladen důraz na poskytnutí důležitosti jednotlivých proměnných, které sloužily k predikci. Na obrázku 17 je vidět výstup algoritmu z hlediska požadované důležitosti proměnných. Je zde vidět, že jako nejdůležitější proměnná byla určena fáze stádia onemocnění, druhou nejdůležitější proměnnou byla určena věk, který je rozdělen do 19 skupin, další důležitou proměnnou je zde věk pacienta v době diagnostikování onemocnění a jako čtvrtou nejdůležitější je určena proměnná rok narození.



Obr. 17 Vital status recode – CHAID

Kromě významných proměnných byl také získán výstup v podobě modelu stromu. Jelikož model stromu je příliš rozmanitý a jeho největší výška byla rovna 5, tak je přiložen v příloze. Tento druh výstupu nám ukázal, že pokud měli pacienti fázi onemocnění 1 (nádor byl mamograficky nebo xerograficky diagnostikován, ale není určena přesná velikost, nádor nebyl klinicky hmatný, nebyly potvrzeny žádné uzliny, žádné metastázy), spadali do věkové skupiny 4 (věk 15 – 19), 8 (věk 35 – 39), 9 (věk 40 – 44), 10 (věk 45 - 49), 11 (věk 50 – 54) nebo 99 (věk neznámý),

onemocnění bylo diagnotikováno v roce 2003 až 2006, nález ER byl pozitivní a sekvence použití radiace a operace byla 3 (provedení radiace po operaci), 4 (radiace před i po operaci), 5 (intraoperativní radiace), 6 (intraoperativní radiace s jiným typem operace před nebo po operaci) nebo 9 (neznámá sekvence), tak byla diagnostikována smrt u 57 pacientů a 2609 pacientů nemoc přežilo, což odpovídá 1,39% všech záznamů v datech, které byly obsaženy v testovacích datech. Na základě tohoto poznatku lze říci, že všem pacientům, kteří tato kritéria splňovaly, byla diagnostikována smrt u 4,5 % a v 95,5 % případů pacient neumřel. Všechna tato kritéria splňovalo 1,39 % (2666 pacientů) všech pacientů, zanesených v testovacích datech.

Jelikož nám algoritmus neposkytoval důležitost u jednotlivých proměnných větší než 50 %, tak bylo potřeba ověřit, zda poskytuje pravdivou predikci či nikoliv. Na výstup algoritmu byl tedy připojen modul "Analysis". Po ukončení procesu jsme získali informace na obrázku 18. Jak je na tomto obrázku vidět, tak algoritmus poskytoval predikci s 83,66% správnou přesností na testovaných datech a predikci s 16,34% špatnou přesností na testovaných datech. Tímto bylo ukázáno, že algoritmus CHAID poskytuje správné informace a že může dosahovat velmi dobrých výsledků nad těmito daty.

Results for output field VitalstatusRecode

Comparing $R-VitalstatusRecode with VitalstatusRecode

'Partition'	1_Training		2_Testing	
Correct	160 590	83,72%	160 668	83,66%
Wrong	31 234	16,28%	31 381	16,34%
Total	191 824		192 049	

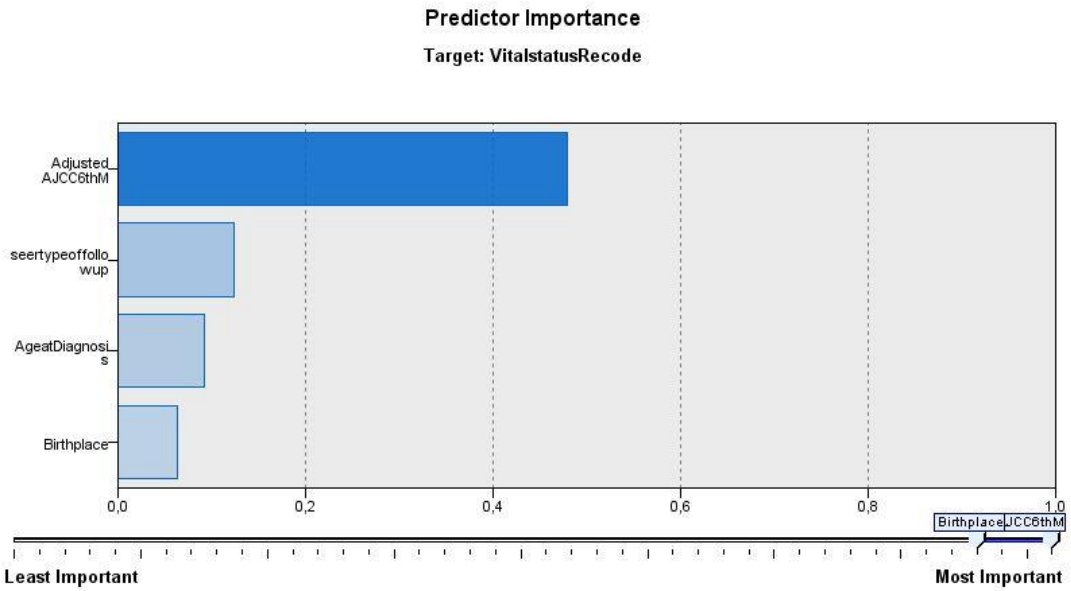
Obr. 18 Vital status recode – CHAID Analysis

Posledním použitým algoritmem byl rozhodovací strom C5. Strom C5 byl použit na základě doporučení z algoritmu Auto classifier a v našem modelu byl použit algoritmus 2x, jednou pro vygenerování důležitosti proměnných použitých pro predikci a pro vygenerování stromu a podruhé byl použit pro vygenerování jednotlivých pravidel, které nám blíže popisují závislosti jednotlivých proměnných.

U obou případů bylo potřeba nastavit výstup algoritmu. V prvním případě byl nastaven výstup jako rozhodovací strom a v módu expert bylo nastaveno prořezávání stromu na hodnotu 95 (maximální hodnota 100). Po tomto nastavení byl spuštěn daný algoritmus a výsledek je zobrazen na obrázku 19. Na tomto obrázku můžete vidět, že jako nejvýznamnější proměnná byla určena velikost a stupeň vzdálených metastáz, následovala proměnná, která určovala, zda byl daný pacient aktivně sledován či nikoliv. Třetí nejvýznamnější proměnnou byl určen věk v době diagnózy a jako čtvrtou nejvýznamnější proměnnou bylo místo narození pacienta.

Dalším výstupem byl v tomto případě model stromu, ale jelikož byl model značně rozvětven a jeho maximální výška byla rovna 10, tak je přiložen v příloze a zde si ukážeme jeden z mnoha příkladů, který nám tento strom ukázal. Pokud byli

pacienti aktivně sledování programem SEER, měli stupeň a velikost metastáz rovnu 0 (nebyly potvrzeny žádné metastázy), byli starší než 72 let, měli pozitivně diagnostikovanou histologii, nález měli diagnostikován v roce 2004, narodili se v Oklahomě, jejich ER nález byl pozitivní a měli nález na levém prsu a pokud byli starší než 73 let, tak byla smrt diagnostikována ve 100% případů (12 případů). Pokud však byli mladší než 73 let, tak byla pravděpodobnost 100 % úmrtí (2 případy).



Obr. 19 Vital status recode – C5

Druhé použití algoritmu C5 sloužilo pro vygenerování pravidel. U algoritmu byl nastaven jako výstup množina pravidel a v módu expert bylo nastaveno prořezávání na hodnotu 95. Po spuštění algoritmu jsme získali 16 pravidel pro případy, kdy pacienti přežili a 339 pravidel pro případy, kdy pacienti zemřeli. Pro případ přežití jsme získali například pravidlo, které nám říkalo, že pokud pacient byl mladší než 82 let, narodil se v Egyptě a případ byl aktivně sledován programem SEER, tak přežil. Oproti tomu jsme pro případ smrti získali například pravidla, která nám říkala, že pokud byl pacient starší 72 let, pochází z Kentucky, narodil se v Kentucky a rok diagnózy byl 2001, tak pacient zemřel. Nebo například pravidlo, které nám říká, že pokud byl pacient starší 72 let, narodil se ve Wisconsinu, nemoc mu byla diagnostikována v roce 2003, nebyla provedena radiace ani žádná operace a nebyly potvrzeny žádné metastázy regionálních lymfatických uzlin, tak pacient zemřel.

Jelikož nám algoritmus neposkytoval důležitost u jednotlivých proměnných větší než 50%, tak bylo potřeba ověřit, zda poskytuje pravdivou predikci či nikoliv. Na výstup algoritmu byl tedy připojen modul "Analysis". Po ukončení procesu jsme získali informace, které jsou zobrazeny na obrázku 20. Na tomto obrázku je vidět, že algoritmus poskytoval predikci s 86,09% správnou přesností na testovaných datech

a predikci s 13,91% špatnou přesností na testovaných datech. Tímto bylo ukázáno, že algoritmus C5 poskytuje správné informace a že může dosahovat velmi dobrých výsledků nad těmito daty

Results for output field VitalstatusRecode

Comparing $C-VitalstatusRecode with VitalstatusRecode

'Partition'	1_Training		2_Testing	
Correct	166 483	86,79%	165 327	86,09%
Wrong	25 341	13,21%	26 722	13,91%
Total	191 824		192 049	

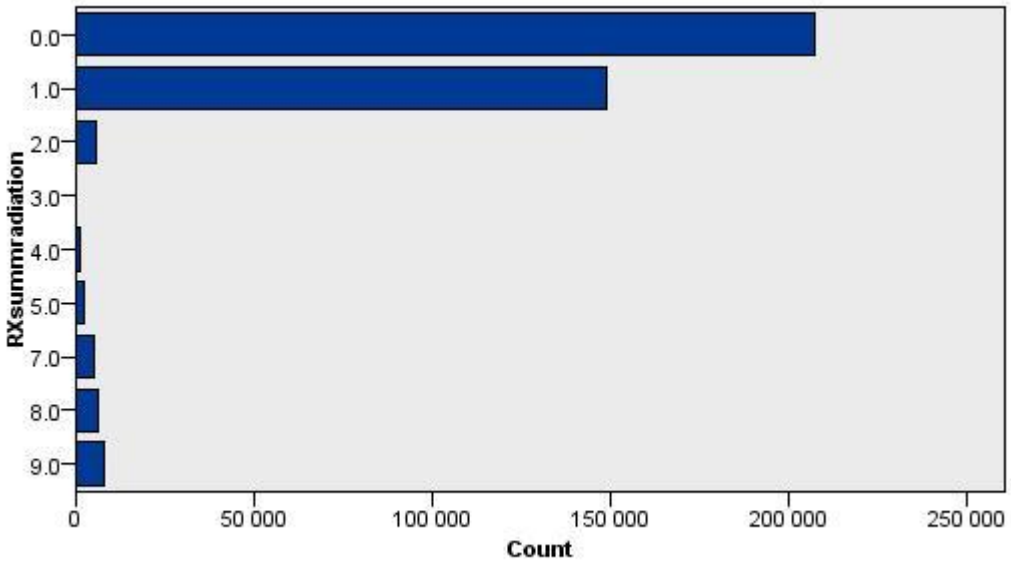
Obr. 20 Vital status recode – C5

4.4.2 RX Summ-radiation

Dalším atributem, který byl blíže zkoumán je RX Summ-radiation. Jedná se o atribut, který v sobě obsahuje informace ohledně provedené léčby radiací v prvním stádiu léčby. V tomto bodě si ukážeme kolik a jaké hodnoty atribut obsahoval. Obsah proměnné je znázorněn na obrázku číslo 21. Jak je na vygenerovaném grafu vidět, tak nejvíce případů mělo hodnotu 0 (nebyla provedena žádná radiace), druhý největší počet záznamů obsahovalo hodnotu 1 (tzv. radiace ozářením), následovanou hodnotou 9 (není známo, zda byla radiace zaznamenána) a hodnotou 2 (radioaktivní implantáty). Ostatní hodnoty, byly u pacientů použity v menším měřítku. Z tohoto grafu lze tedy vyčíst, že nejčastějším typem léčení radiací je metoda externího ozařování a druhou nejpoužívanější je implantace radioaktivních implantátů přímo k tumoru tak, aby byla léčba co možná nejúčinnější.

Po tomto kroku bylo potřeba přejít k definování vstupních a výstupních proměnných, které budou vstupovat do použitého algoritmu. Jako výstupní proměnnou byl tedy nastaven zmíněný atribut a jako vstupní proměnné byly ponechávány všechny ostatní atributy s výjimkou atributu RX Summ-surg/rad seq, neboť tento atribut je přímo spjatý s typem radiace a výsledky nám ukazovaly téměř 100% závislost mezi proměnnými a atributu RX Summ-surg prim site, který popisuje chirurgický zákrok, který odstraňuje anebo ničí tkáň nálezu a byl vykonán v prvním cyklu léčby. Druhý atribut byl vyloučen z důvodu velké závislosti na výstupnímu atributu, kde byly ovlivněny ostatní atributy takovým způsobem, že neměly téměř žádný vliv na výstupní atribut.

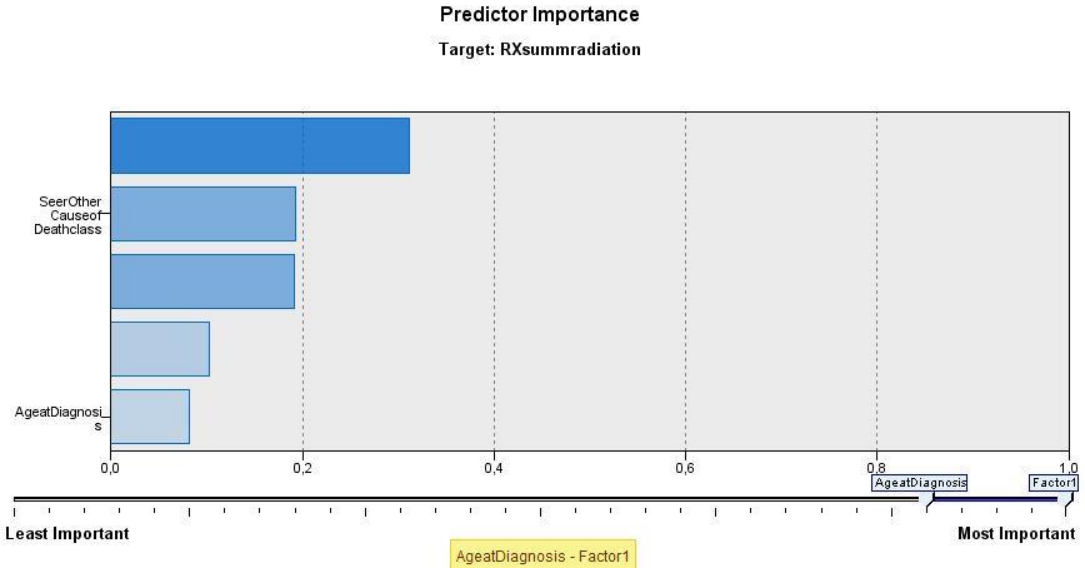
Po nastavení vstupních a výstupních proměnných, byla data nastavena na vstup modulu "Partition". Výstup modulu "Partition" poté sloužil jako vstup do jednotlivých algoritmů. Nastavení v tomto modulu bylo ponecháno, stejně jako u případu atributu Vital status recode, v defaultním stavu, který nám říká, že trénovací i testovací data budou tvořeny 50% vstupních dat a k jejich rozdělení dojde náhodným výběrem.



Obr. 21 RX Summ-radiation

Na vstup modulu "Partition" byl připojen algoritmus "Auto classifier", který nám měl doporučit, jaké algoritmy budou nad těmito daty poskytovat nejlepší výsledky. V tomto případě nám byl doporučen jen algoritmus rozhodovací strom C5. Nicméně jsme i přes toto doporučení použili rozhodovací strom CHAID a C&RT.

V prvním kroku si popíšeme nastavení a výsledky rozhodovacího stromu C&RT. Stejně jako v případě atributu Vitalstatus recode zůstala všechna nastavení v defaultním režimu. Bylo požadováno zobrazení důležitosti u predikovaných proměnných a také zobrazení modelu stromu. Na obrázku 22 je vidět výsledek algoritmu z hlediska důležitosti jednotlivých predikovaných proměnných. Na obrázku je vidět 5 nejvýznamnějších proměnných, které ovlivňují výstup algoritmu. Jako nejvýznamnější atribut byl určen CS site-specific Factor 1, který popisuje fázi, ve které se nacházel nález pacienta. Druhou nejvýznamnější proměnnou je SEER other cause of death classification, která nám říká, zda pacient přežil nebo umřel v důsledku rakoviny nebo zemřel na základě jiné příčiny. Jako třetí nejvýznamnější proměnnou, z hlediska predikce, je uvedena proměnná Derived AJCC 6 Stage GRP, která se stejně jako proměnná CS site-specific Factor 1 zabývá fází onemocnění. Čtvrtou nejvýznamnější proměnnou je Reason for no surgery, která blíže popisuje, zda-li byla provedena či zamítnuta operace a z jakého důvodu se tak stalo. Pátou nejvýznamnější proměnnou byla určena Age at diagnosis, která nám udává věk pacienta, kdy mu byla diagnostikována rakovina prsu.



Obr. 22 RX Summ-radiation – C&RT

Dalším výstupem algoritmu byl model stromu. Jelikož model stromu byl příliš rozvětvený a jeho největší výška byla rovna 5, tak je přiložen v příloze. Nicméně nám ukázal, že pokud měli pacienti nález, který byl pozitivní, negativní v běžných limitech nebo ohraničený, ale není uvedeno zda se jedná o pozitivní nebo negativní vzorek, přežili nebo zemřeli v důsledku rakoviny, fáze byla buď první (nádor byl mamograficky nebo xerograficky diagnostikován, ale není určena přesná velikost, nádor nebyl klinicky hmatný, nebyly potvrzeny žádné uzliny, žádné metastázy) a nebo třetí (na primárním místě, nebyl nalezen žádný nádor, ale nádor metastázuje a klinicky byly potvrzeny uzliny ve vnitřní části prsu), byli mladší 80-ti let a jejich druh nálezu odpovídal onemocnění bradavky, středové části prsu nebo obecně prsu, které nebylo blíže specifikováno, tak u 50,6 % (3686 pacientů) všech pacientů, kteří splňovali tyto kritéra, nebylo provedeno žádné ozařování, u 43,5 % (3169 pacientů) bylo provedeno externí ozařování, u 1,3 % (98 pacientů) všech pacientů byly zavedeny radioaktivní implantáty, u 1,7 % všech pacientů (120 pacientů) bylo ozařování doporučeno, ale není zaznamenáno, zda-li se provedlo, u 1,3 % (95 pacientů) všech pacientů nebylo známo, jestli bylo ozařování doporučeno, u 0,8% (55 pacientů) všech pacientů odmítlo ozařování, u 0,5 % (38 pecientů) všech pacientů byla provedena radiace, ale není blíže specifkována metoda, u 0,3 % (23 pacientů) všech pacientů bylo provedeno jak externí, interní tak i ozařování radioisotopy a u zbylých pacientů 0,08 % (6 pacientů) bylo provedeno ozařování radioisotopy Všechna tato kritéria splňovalo 5,43 % (7290 pacientů) všech pacientů, zanesených v testovacích datech.

Na výstup algoritmu C&RT byl připojen modul "Analysis", neboť bylo potřeba zjistit, s jakou celkovou přesností predikce algoritmus pracuje. Po ukončení procesu algoritmu, jsme získali výsledek, který je ukázán na obrázku 23. Jak je na výstupu

vidět, tak algoritmus poskytoval predikci s 59,01% správnou přesností na testovaných datech a predikci s 40,99% špatnou přesností na testovaných datech. Jelikož nám algoritmus předložil nízkou důležitost u jednotlivých proměnných sloužících k predikci, došlo k obavám, zda bude poskytovat správné výsledky. Tento předsudek jsme se snažili vyvrátit použitím modulu "Analysis", nicméně tyto výsledky nám zcela nepotvrdily správné použití algoritmu, ale ukázaly, že má stále lepší správnou než špatnou schopnost predikce.

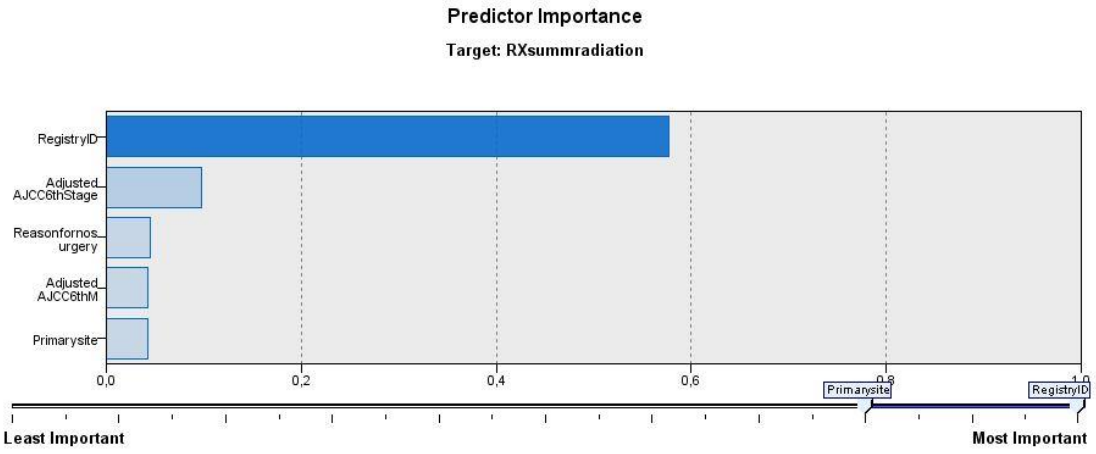
Results for output field RXsummradiation

Comparing $R-RXsummradiation with RXsummradiation

'Partition'	1_Training		2_Testing	
Correct	113 194	59,01%	113 326	59,01%
Wrong	78 630	40,99%	78 723	40,99%
Total	191 824		192 049	

Obr. 23 RX Summ-radiation – C&RT Analysis

Dalším použitým algoritmem u tohoto atributu byl algoritmus rozhodovací strom CHAID. Algoritmus byl použit na základě získaného výstupu algoritmu "Auto classifier". Na obrázku 24 je vidět výstup algoritmu z hlediska požadované důležitosti proměnných. Je zde vidět, že jako nejdůležitější atribut byl určen stát, ve kterém se pacient nachází, jako druhá nejdůležitější byla určena fáze onemocnění, další důležitou proměnnou je důvod pro neprovedení operace, jako čtvrtou nejdůležitější proměnnou je určeno rozšíření do vzdálených metastáz a jako čtvrtá významná proměnná je místo výskytu onemocnění.



Obr. 24 RX Summ-radiation – CHAID

Dalším požadovaným výstupem algoritmu byl model stromu. Model stromu je příliš rozvětvený a jeho největší výška byla rovna 5, a proto je ukázán v příloze práce. Model stromu nám ukázal, že pokud byli pacienti ze státu Greater California, neměli klinicky potvrzené metastázy, celková fáze onemocnění byla třetího stupně

(na primárním místě, nebyl nalezen žádný nádor, ale nádor metastázuje a klinicky byly potvrzeny uzliny ve vnitřní části prsu), přežili sledovaný den a byli buď pojištěni, ale nebylo blíže specifikováno jak, měli pojištění Medicaid a nebo nebyli pojištěni, tak ze všech pacientů, kteří tyto předpoklady splňovali, nebylo u 38,2 % (976 pacientů) provedeno ozařování, u 50,9 % (1300 pacientů) bylo provedeno externí ozařování, u 7,9 % (202) bylo ozařování doporučeno, ale není zaznamenáno, zda bylo provedeno, u 2,1 % (54 pacientů) pacient ozařování odmítl a u 0,9 % (24 pacientů) bylo ozařování provedeno, ale nebylo blíže specifikováno jakého typu. Všechna tato kritéria splňovalo 1,33 % (2556 pacientů) všech pacientů, zanesených v testovacích datech.

Jelikož nám algoritmus poskytoval důležitost větší než 50 % jen u atributu zabývajícího se státem, ze kterého pacient pocházel, tak bylo potřeba ověřit celkovou predikci poskytovanou algoritmem. Na výstup algoritmu byl tedy připojen modul "Analysis". Po ukončení procesu jsme získali informace na obrázku 25. Jak je na tomto obrázku vidět, tak algoritmus poskytoval predikci s 56,44% správnou přesností na testovaných datech a predikci s 43,56% špatnou přesností na testovaných datech. Modul "Analysis" ukázal, že jeho výsledky nám zcela nepotvrdily správné použití algoritmu, ale ukázaly, že má stále lepší správnou než špatnou schopnost poskytovat správnou predikci.

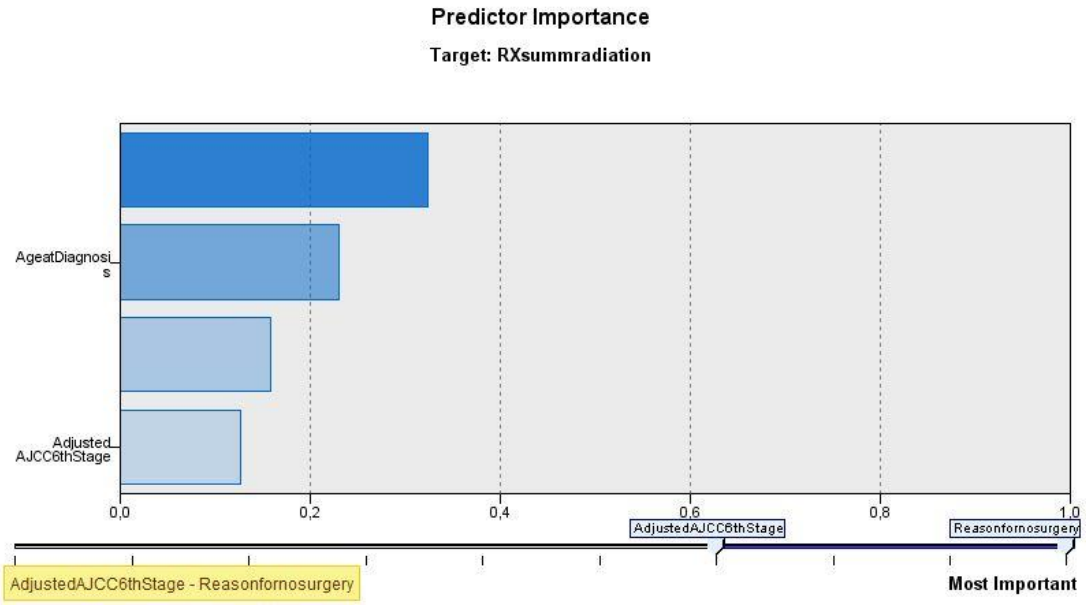
Results for output field RXsummradiation

Comparing $R-RXsummradiation with RXsummradiation

'Partition'	1_Training		2_Testing	
Correct	108 537	56,58%	108 400	56,44%
Wrong	83 287	43,42%	83 649	43,56%
Total	191 824		192 049	

Obr. 25 RX Summ-radiation – CHAID Analysis

Posledním algoritmem, který byl použit u daného atributu, byl rozhodovací strom C5. Algoritmus byl použit v modelu dvakrát, a to z důvodu potřebného vygenerování důležitosti proměnných z hlediska predikce, vybudování modelu stromu a k vygenerování rozhodovacích pravidel. Na obrázku 26 je vidět výstup algoritmu z hlediska požadované důležitosti proměnných. Je zde vidět, že jako nejdůležitější atribut byl určen důvod pro neprovedení operace, jako druhý nejdůležitější atribut byl věk v době diagnózy, jako třetí nejdůležitější atribut byl sledování pacienta programem SEER a jako čtvrtý nejdůležitější atribut byla fáze onemocnění.



Obr. 26 RX Summ-radiation – C5

Dalším požadovaným výstupem algoritmu byl model stromu. Model stromu je příliš rozvětvený a jeho největší výška byla rovna 10, a proto je ukázán v příloze práce. Model stromu nám ukázal, že pokud byly případy pacientů aktivně sledovány, u pacientů byla provedena operace, pacienti byli ženy mladší 77 let, pacienti měli první záznam v registru a nález ER byl pozitivní, fáze onemocnění byla 0 (nádor nenarušuje okolí a je omezen jen na primární místo nálezu, uzliny nebyly potvrzeny, žádné vzdálené metastázy), nebyly zkoumány žádné uzliny a onemocnění prsu bylo typu in-situ, ale nebylo blíže specifikováno, tak ze všech pacientů, kteří tyto předpoklady splňovali, nebylo u 34,0 % (2369 pacientů) provedeno ozařování, u 56,97 % (3969 pacientů) bylo provedeno externí ozařování, u 3,7 % (256 pacientů) byly implementovány radioaktivní implantáty, u 3 pacientů bylo léčení pomocí radioisotopů, u 0,3 % (24 pacientů) bylo léčení kombinací externího, interního a radioisotopního typu. U 1 % (67 pacientů) byla radiace provedena, ale nebyl blíže specifikován typ ozařování. U 2,1% (144 pacientů) bylo ozařování odmítnuto pacienty, u 1,3 % (92 pacientů) byla radiace doporučena, ale není zaneseno, zda byla provedena a u 43 případů není známo, zda bylo ozařování provedeno. Všechna tato kritéria splňovalo 3,63 % (6967 pacientů) všech pacientů, zanesených v testovacích datech

Druhé použití algoritmu C5 sloužilo pro vygenerování pravidel. U algoritmu byl nastaven jako výstup množina pravidel a v módu expert bylo nastaveno prořezávání na hodnotu 100. Po spuštění algoritmu jsme získali dohromady 16 pravidel. 7 pravidel pro případy, kdy nebyla provedena radiace, 7 pravidel pro případy, kdy bylo provedeno externí ozařování a 2 pravidla pro případy, kdy není známo, jestli byla radiace doporučena. Pro případ, kdy u pacientů nebyla provedena radiace, jsme získali například pravidlo, které nám říká, že nebyly zkoumány žádné uzliny a fáze

onemocnění je 1 (nádor byl mamograficky nebo xerograficky diagnostikován, ale není určena přesná velikost, nádor nebyl klinicky hmatný, nebyly potvrzeny žádné uzliny, žádné metastázy), tak radiace nebyla provedena. Oproti tomu jsme pro případ externího ozařování získali například pravidlo, které nám říká, že pokud pacient byl žena mladší 77 let, byla u pacienta provedena operace, jednalo se o první záznam pacienta v programu SEER, nález ER u pacienta byl pozitivní a onemocnění bylo ve fázi 3 (na primárním místě, nebyl nalezen žádný nádor, ale nádor metastázuje a klinicky byly potvrzeny uzliny ve vnitřní části prsu), tak bylo u pacienta provedeno externí ozařování.

Jelikož nám algoritmus neposkytoval důležitost větší než 50 % u žádné proměnné sloužící k predikci, tak bylo potřeba ověřit celkovou predikci poskytovanou algoritmem. Na výstup algoritmu byl tedy připojen modul "Analysis". Po ukončení procesu jsme získali informace na obrázku 27. Jak je na tomto obrázku vidět, tak algoritmus poskytoval predikci s 59,42% správnou přesností na testovaných datech a predikci s 40,58% špatnou přesností na testovaných datech. Modul "Analysis" ukázal, že jeho výsledky nám zcela nepotvrdily správné použití algoritmu, ale ukázaly, že má stále lepší správnou než špatnou schopnost poskytnuté predikce. Jak je dále na obrázku vidět, tak na testovaných datech jsme získali pouze o 0,9 % přesnější výsledky, než na trénovacích datech.

Results for output field RXsummradiation

Comparing $C-RXsummradiation with RXsummradiation

'Partition'	1_Training		2_Testing	
Correct	113 818	59,33%	114 120	59,42%
Wrong	78 006	40,67%	77 929	40,58%
Total	191 824		192 049	

Obr. 27 RX Summ-radiation – C5 Analysis

4.4.3 RX Summ-surg/rad seq

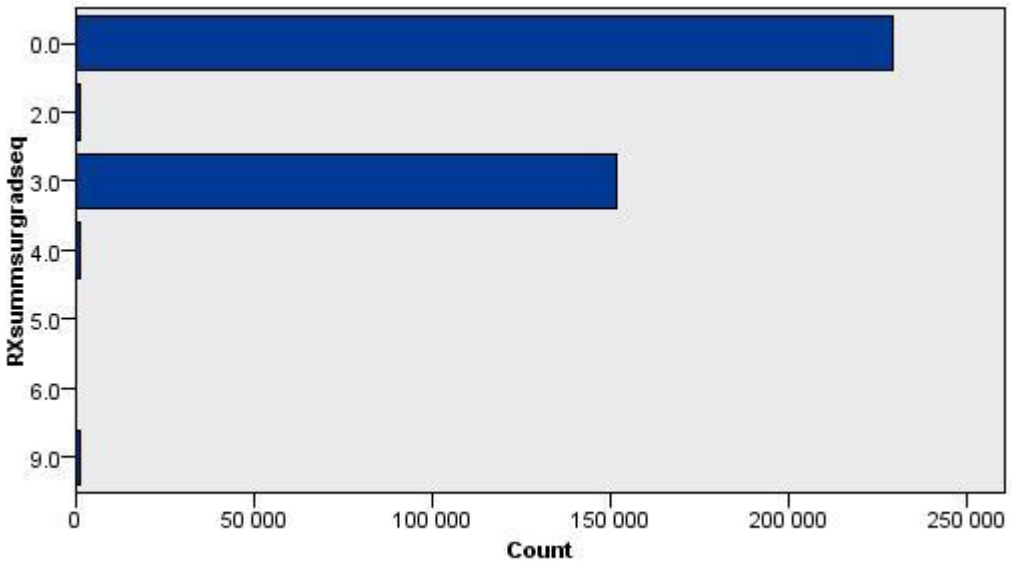
Posledním atributem, který byl blíže zkoumán je RX Summ-surg/rad seq. Jedná se o atribut, který v sobě obsahuje informace ohledně sekvence léčby radiací anebo chirurgickým zákrokem. Stejně jako dva cílové atributy výše, je tento atribut blíže popsán v kapitole věnující se zařazeným atributům do finálních dat. V tomto bodě si ukážeme kolik a jaké hodnoty atribut obsahoval. Obsah proměnné je znázorněn na obrázku číslo 29. Jak je na vygenerovaném grafu vidět, tak nejvíce případů mělo hodnotu 0 (nebyla provedena operace nebo radiace), druhý největší počet záznamů obsahovalo hodnotu 3 (radice byla provedena až po chirurgickém zákroku), následovanou hodnotou 2 (radiace před operací), hodnotou 4 (radiace byla provedena před i po operaci) a hodnotou 9 (pořadí je neznámé, ale bylo provedeno jako ozařování, tak i chirurgický zákrok). Z tohoto grafu lze tedy vyčíst, že nejpočetnější zastoupení v proměnné měly případy, kdy nebyl proveden jeden typ zmíněné léčby, druhé nejpočetnější zastoupení měly případy, kdy po chirurgickém

zákroku musel pacient podstoupit i léčbu formou ozařováním. Ostatním případy obsahující velmi malý počet všech případů (1000 záznamů/hodnota).

Po tomto kroku bylo potřeba nastavit vstupní a výstupní proměnné, které budou vstupovat do použitého algoritmu. Jako výstupní proměnnou byl tedy nastaven zmíněný atribut a jako vstupní proměnné byly ponechávány všechny ostatní atributy s výjimkou atributu RX Summ-radiation, neboť tento atribut je přímo spjatý s tímto atributem a výsledky nám ukazovaly téměř 100% závislost mezi proměnnými a atributu RX Summ-surg prim site, který popisuje chirurgický zákrok, který odstraňuje anebo ničí tkáň nálezu a byl vykonán v prvním cyklu léčby. Druhý atribut byl vyloučen z důvodu velké závislosti na výstupním atributu, kde byly ovlivněny ostatní atributy takovým způsobem, že neměly téměř žádný vliv na výstupní atribut.

Po nastavení vstupních a výstupních proměnných, byly data připojena na vstup modulu "Partition". Výstup modulu "Partition" poté sloužil jako vstup do jednotlivých algoritmů. Nastavení v tomto modulu bylo ponecháno opět v defaultním stavu.

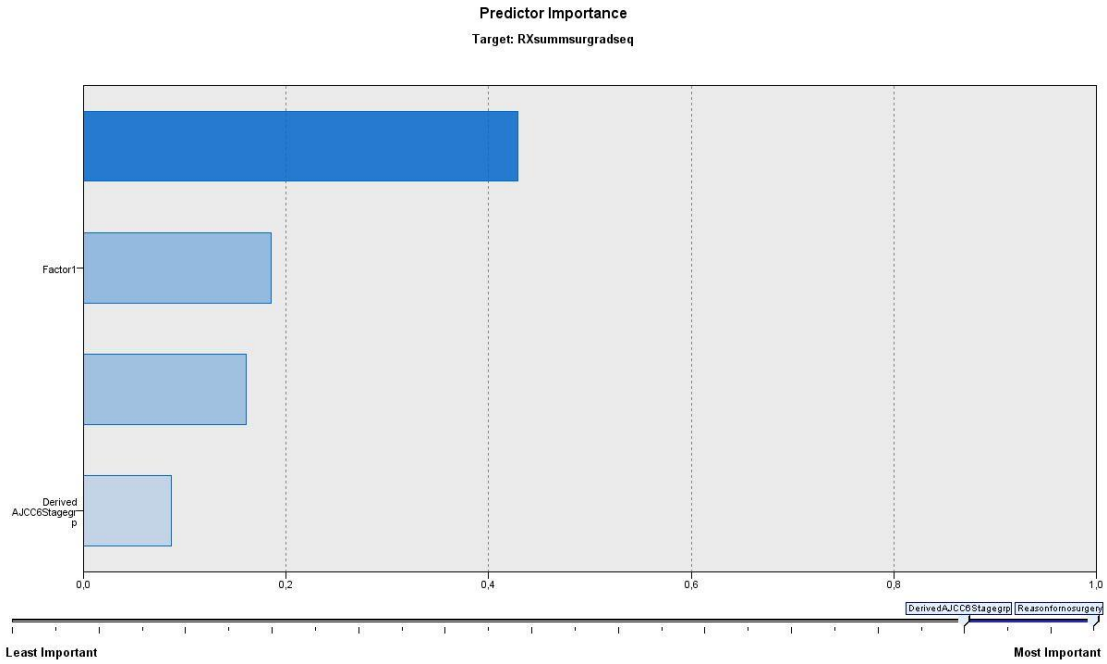
Na výstup modulu "Partition" byl připojen algoritmus "Auto classifier", který sloužil pro vygenerování algoritmů, které se nejlépe hodí na testovaná data. V našem případě byly doporučeny rozhodovací stromy CHAID, C&RT a C5, jejichž výsledky si popíšeme níže.



Obr. 28 RX Summ-surg/rad seq

V prvním kroku si popíšeme nastavení a výsledky rozhodovacího stromu C&RT. Všechna nastavení stromu zůstala v defaultním režimu. Bylo požadováno zobrazení důležitosti u predikovaných proměnných a také zobrazení modelu stromu.

Na obrázku 29 je vidět výsledek algoritmu z hlediska důležitosti jednotlivých predikovaných proměnných. Na obrázku jsou vidět 4 nejvýznamnější proměnné, které ovlivňují výstup algoritmu. Jako nejvýznamnější atribut je určen Reason for no surgery, který popisuje důvod pro neprovedení operace. Druhou nejvýznamnější proměnnou je CS site-specific Factor 1, který popisuje fázi, ve které se nacházel nález pacienta. Třetí nejvýznamnější proměnnou je SEER other cause of death classification, která nám říká, zda pacient přežil nebo umřel v důsledku rakoviny nebo zemřel na základě jiné nemoci. Jako čtvrtá nejvýznamnější proměnná je poté uvedena proměnná Derived AJCC 6 Stage GRP, která se stejně jako proměnná CS site-specific Factor 1 zabývá fází onemocnění.



Obr. 29 RX Summ-surg/rad seq – C&RT

Kromě důležitosti proměnných jsme získali také model stromu. Model stromu je přiložen v příloze, neboť byla jeho největší výška rovna 10. Model stromu nám ukázal, že pokud u pacienta byla provedena operace, měl nález pozitivní, v normálním rozšíření nebo ohraničený, ale nedefinovaný, zda-li se jedná o pozitivní nebo negativní nález), přežili nebo zemřeli v důsledku rakoviny, jejich fáze onemocnění byla 1 (nádor byl mamograficky nebo xerograficky diagnostikován, ale není určena přesná velikost, nádor nebyl klinicky hmatný, nebyly potvrzeny žádné uzliny, žádné metastázy) nebo 3 (na primárním místě, nebyl nalezen žádný nádor, ale nádor metastázuje a klinicky byly potvrzeny uzliny ve vnitřní části prsu) a bylo postiženo místo na horní vnitřní straně prsu, na dolní vnitřní straně prsu, na horní vnější straně prsu, na dolní vnější straně prsu, u místa konce prsu nebo obecně bylo postiženo prso, ale nebylo blíže specifikováno kde, tak ze všech pacientů, kteří tyto předpoklady splňovali, nebylo u 41,2 % (13738 pacientů) provedeno ozařování

nebo operace, u 57,85 % (19280 pacientů) bylo provedeno ozařování až po provedení operace, u 0,4 % (132 pacientů) bylo provedeno ozařování po i před operací, u 0,3 % (93 pacientů) bylo provedeno ozařování před operací, u 0,2 % (70 pacientů) bylo pořadí léčení neznámé, ale bylo provedeno jak ozařování, tak operace, u 8 pacientů bylo provedeno intraoperativní ozařování s kombinací s dalšími typy ozařování a u 9 pacientů bylo provedeno intraoperativní ozařování. Všechna tato kritéria splňovalo 24,83 % (33330 pacientů) všech pacientů, zanesených v testovacích datech.

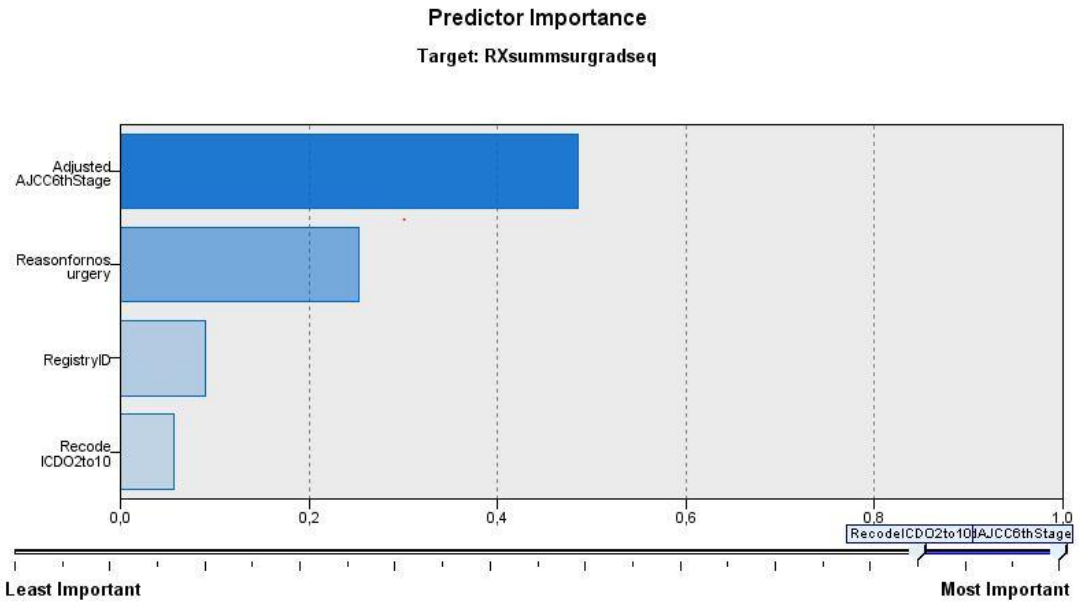
U algoritmu bylo potřeba ověřit celkovou predikci poskytovanou algoritmem. Na výstup algoritmu byl tedy připojen modul "Analysis". Po ukončení procesu jsme získali informace zobrazené na obrázku 30. Jak je na tomto obrázku vidět, tak algoritmus poskytoval predikci s 64,19% správnou přesností na testovaných datech a predikci s 35,81% špatnou přesností na testovaných datech. Modul "Analysis" ukázal, že jeho výsledky nám potvrdily správné použití algoritmu.

Comparing $R-RXsummsurgradseq with RXsummsurgradseq

'Partition'	1_Training		2_Testing	
Correct	123 147	64,2%	123 270	64,19%
Wrong	68 677	35,8%	68 779	35,81%
Total	191 824		192 049	

Obr. 30 RX Summ-rad/surg seq – C&RT Analysis

Dalším použitým algoritmem, na základě algoritmu "Auto classifier", u tohoto atributu byl algoritmus rozhodovací strom CHAID. Na obrázku 31 je vidět výstup algoritmu z hlediska požadované důležitosti proměnných. Je zde vidět, že jako nejdůležitější proměnná byla určena fáze popisující nález pacienta, jako druhý nejdůležitější atribut byl důvod pro neprovedení operace, další důležitý atribut byl stát, ve kterém se pacient nacházel a jako čtvrtou nejdůležitější proměnnou je určeno postižené místo nálezem.



Obr. 31 RX Summ-rad/surg seq – CHAID

Kromě důležitosti proměnných jsme získali také model stromu. Model stromu je přiložen v příloze, neboť byl velmi rozvětvený a jeho největší výška byla rovna 5. Model stromu nám ukázal, že pokud u pacienta byla provedena operace, měli nález ve fázi 1 (nádor byl mamograficky nebo xerograficky diagnostikován, ale není určena přesná velikost, nádor nebyl klinicky hmatný, nebyly potvrzeny žádné uzliny, žádné metastázy), přežili nebo zemřeli v důsledku rakoviny, pochází ze státu Greater California a jejich nález se nacházel na horní vnitřní straně prsu, horní vnější straně prsu, dolní vnější straně prsu nebo na konci prsu, tak ze všech pacientů, kteří tyto předpoklady splňovali, nebylo u 41,7 % (5318 pacientů) provedeno ozařování nebo operace, u 57,97 % (7389 pacientů) bylo provedeno ozařování až po provedení operace, u 10 pacientů bylo provedeno ozařování před operací, u 9 pacientů bylo provedeno ozařování před i po operaci, u 10 pacientů bylo pořadí léčení neznámé, ale bylo provedeno jak ozařování, tak operace, u 6 pacientů bylo provedeno intraoperativní ozařování s kombinací s dalšími typy ozařování a u 6 pacientů bylo provedeno intraoperativní ozařování. Všechna tato kritéria splňovalo 6,65 % (12748 pacientů) všech pacientů, zanesených v testovacích datech.

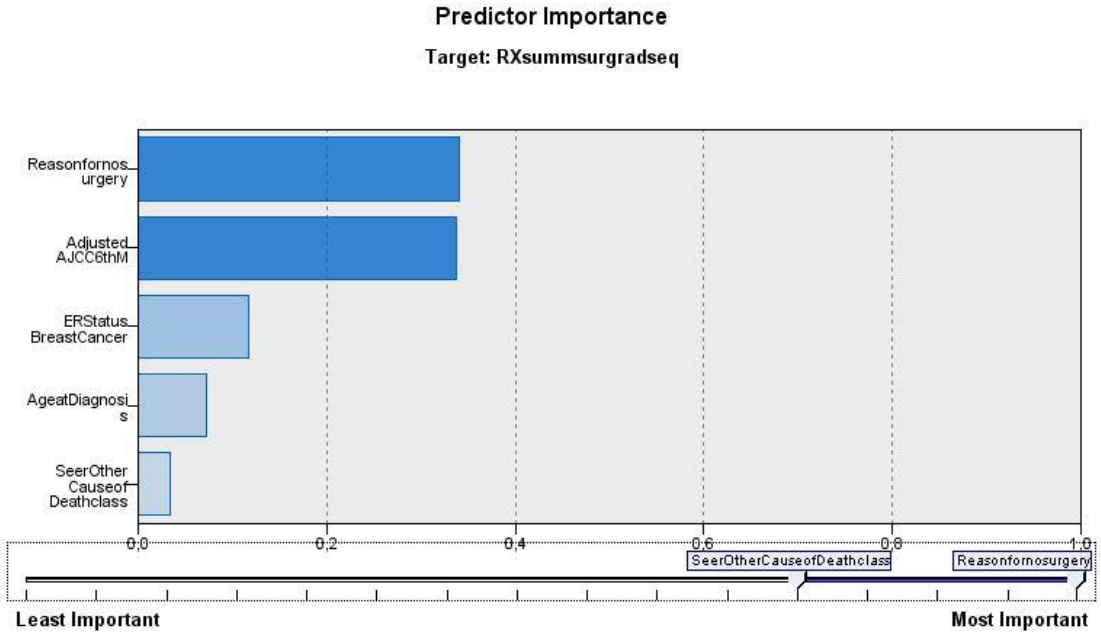
U algoritmu bylo potřeba ověřit celkovou predikci poskytovanou algoritmem. Na výstup algoritmu byl tedy připojen modul "Analysis". Po ukončení procesu jsme získali informace zobrazené na obrázku 32. Jak je na tomto obrázku vidět, tak algoritmus poskytoval predikci s 64,06% správnou přesností na testovaných datech a predikci s 35,94% špatnou přesností na testovaných datech. Modul "Analysis" ukázal, že jeho výsledky nám potvrdily správné použití algoritmu.

Comparing $R-RXsummsurgradseq with RXsummsurgradseq

'Partition'	1_Training		2_Testing	
Correct	122 958	64,1%	123 032	64,06%
Wrong	68 866	35,9%	69 017	35,94%
Total	191 824		192 049	

Obr. 32 RX Summ-rad/surg seq – CHAID analysis

Posledním algoritmem, který byl použit u daného atributu, byl rozhodovací strom C5. Algoritmus byl použit v modelu dvakrát, a to z důvodu potřebného vygenerování důležitosti proměnných z hlediska predikce, vybudování modelu stromu a k vygenerování rozhodovacích pravidel. Na obrázku 34 je vidět výstup algoritmu z hlediska požadované důležitosti proměnných. Je zde vidět, že jako nejdůležitější atribut byl určen důvod pro neprovedení operace, jako druhý nejdůležitější atribut byl stupeň metastáz, jako třetí nejdůležitější atribut byl nález ER, jako čtrvtý nejdůležitější atribut byl věk pacienta v době diagnózy a jako pátý atribut byl záznam, zda pacient přežil nebo zemřel na onemocnění rakoviny nebo na nějakou jinou nemoc.



Obr. 33 RX Summ-surg/rad seq – C5

Dalším požadovaným výstupem algoritmu byl model stromu. Model stromu je příliš rozvětvený a jeho největší výška byla rovna 12, a proto je ukázán v příloze práce. Model stromu nám ukázal, že pokud byla u pacienta provedena operace, pacient byl mladší než 79 let, pacient byla žena, stupeň metastáz byl roven 0 (nebyly potvrzeny žádné metastázy), nález ER byl u pacienta pozitivní, pacient přežil nebo zemřel díky rakovině, fáze onemocnění byla 1 (nádor byl mamograficky nebo xerograficky diagnostikován, ale není určena přesná velikost, nádor nebyl klinicky hmatný,

nebyly potvrzeny žádné uzliny, žádné metastázy), nález byl mikroskopicky potvrzen (pozitivní histologie), všechny zkoumané uzliny byly negativní a místo nálezu bylo na prsu obecně, které nebylo blíže specifikováno, tak ze všech pacientů, kteří tyto předpoklady splňovali, nebylo u 60,6 % (2164 pacientů) provedeno ozařování nebo operace, u 38,75 % (1383 pacientů) bylo provedeno ozařování až po operaci, u 6 pacientů bylo ozařování provedeno před operací, u 7 pacientů bylo provedeno ozařování jak před, tak i po operaci, u 2 pacientů bylo provedeno intraoperativní ozařování, u 1 pacienta bylo provedeno intraoperativní ozařování v kombinací s jinými typy ozařování a u 6 pacientů je pořadí léčení neznámé, ale bylo provedeno jak ozařování, tak i operace. Všechna tato kritéria splňovalo 1,86 % (3569 pacientů) všech pacientů, zanesených v testovacích datech.

Algoritmus C5 byl použit v modelu dvakrát a jeho druhé použití sloužilo k vygenerování konkrétních pravidel. Na výstupu algoritmu jsme získali 33 pravidel. 11 pravidel sloužilo k popisu stavu, kdy se pacient podrobil jen jednomu typu léčení (buď operaci nebo ozařování). Pro případ, kdy u pacientů nebyla provedena radiace, jsme získali například pravidlo, které nám říká, že pokud byl pacient starší 79 let, tak u něj nebylo doporučeno ozařování i operace, ale ve většině případů pacient podstoupil jen chirurgický zákrok. Oproti tomu 22 pravidel nám popisovalo stav, kdy pacient podstoupil ozařování až po provedení chirurgického zákroku. Jako příklad si můžeme uvést pravidlo, které nám říká, že pokud byl pacient žena, byl mladší 79 let, místo nálezu je na horní vnitřní straně prsu, nebyly potvrzeny žádné uzliny, pacient podstoupil operaci, přežil nebo zemřel na základě rakoviny, nález ER byl pozitivní a jeho fáze nálezu byla 1 (nádor byl mamograficky nebo xerograficky diagnostikován, ale není určena přesná velikost, nádor nebyl klinicky hmatný, nebyly potvrzeny žádné uzliny, žádné metastázy), tak pacient podstoupil mimo chirurgický zákrok i léčbu ozařováním.

U algoritmu bylo potřeba ověřit celkovou predikci poskytovanou algoritmem. Na výstup algoritmu byl tedy připojen modul "Analysis". Po ukončení procesu jsme získali informace zobrazené na obrázku 35. Jak je na tomto obrázku vidět, tak algoritmus poskytoval predikci s 65,19% správnou přesností na testovaných datech a predikci s 34,81% špatnou přesností na testovaných datech. Výstup modulu "Analysis" nám potvrdil správné použití algoritmu na testovaných datech.

Comparing $C-RXsummsurgradseq with RXsummsurgradseq

'Partition'	1_Training		2_Testing	
Correct	125 405	65,38%	125 200	65,19%
Wrong	66 419	34,62%	66 849	34,81%
Total	191 824		192 049	

Obr. 34 RX Summ-surg/rad seq – C5 Analysis

4.5 Diskuze

V předchozí podkapitole jsme si ukázali výstupy jednotlivých cílových atributů. V této kapitole si všechny získané poznatky sjednotíme a pokusíme se vyvodit znalosti, které budou přímo vycházet ze získaných informací.

U cílového atributu Vitalstatus recode jsme získali prediktivní proměnné, které nám nejvíce ovlivňují výstupní hodnotu. Těmito proměnnými byly: rozšíření nádoru a jeho rozpínavost, věk pacienta, informace ohledně metastáz, fáze onemocnění, kde tato proměnná v sobě zahrnovala informace o velikosti tumoru, o lymfatických uzlinách a o metastázách. Dále to byly proměnné, které se také zabývaly věkem pacienta, a to datum narození a rozdělení pacienta do věkových skupin. Můžeme si zde uvést, že největší pravděpodobnost úmrtí měli pacienti, u kterých byly diagnostikovány vzdálené metastázy (karcinomotáze[38]), nebo pacienti měli onemocnění rozšířeno na další sousedící místa (podpaží, druhé prso, prsní kost, horní část břicha, kosti, nadledvinové žlázy, plíce, vaječník). Dále měli největší pravděpodobnost úmrtí pacienti, kteří měli věk starší než 70 let, přitom největší poměr úmrtí ze všech pacientů, kteří se léčili na rakovinu prsu, měli pacienti, kteří byli starší než 85 let. Pokud se zaměříme na úmrtí u pacientů z hlediska fáze onemocnění, tak jsme na základě výstupů zjistili, že největší poměr úmrtí ze všech pacientů, byli ti, kteří měli fázi stádia 3 (na primárním místě nebyl nalezen žádný nádor, ale nádor metastázuje a klinicky byly potvrzeny uzliny ve vnitřní části prsu) nebo stádium 4 (tumor měl velikost 51 až 989 milimetrů, lymfatické uzliny byly klinicky znatelné, vzdálené metastázy s rozšířením na ostatní místa jako kůže, druhé prso, prsní kost, žaludek, plíce, podpaží). Všechny tyto důkazy jsou znázorněny na obrázcích přiložených v příloze práce.

Cílové atributy RX Summ-radiation a RX Summ-surg/rad seq, byly nejvíce ovlivněny proměnnými popisující chování nálezu v případě, pokud byl u pacienta potvrzen estrogen receptor, stavem popisující zda pacient zemřel v důsledku rakoviny nebo přežil a léčil se na rakovinu. Dále proměnnou popísující důvod pro neprovedení operace nebo zda operace byla provedena, fázi stádia rakoviny, státem, ve kterém se pacient nacházel, věkem pacienta v době diagnózy, informaci stahující se k výskytu metastáz a promněnou popisující primární místo výskytu rakoviny. Pokud se zaměříme na stav, kdy byla u pacienta provedena léčba externím ozařováním, z hlediska chování nálezu, tak jsme na základě výstupů zjistili, že největší poměr použití externí radiace ze všech pacientů je u případů, kdy chování

[38] Karcinomotáze – nádorový rozsev

estrogen receptoru bylo pozitivní nebo negativní v normálních mezích. Dále měli největší pravděpodobnost použití externího ozařování pacienti, u kterých byla již provedena operace a u pacientů, kteří přežili nebo zemřeli v důsledku rakoviny. Pokud se zaměříme na ovlivnění výstupů z hlediska fáze, ve které se nacházel nález pacienta, tak můžeme říct, že poměr použití externí radiace je u všech fází stádia přibližně stejný. Dalšími zkoumanými proměnnými byly věk pacienta a stát, ze kterého pocházel. Poměr použití externího ozařování u pacienta vycházel přibližně stejně a nezáleželo na tom, zda byl pacient mladší 30-ti let nebo starší 85-ti letům. To stejné můžeme říci o státu, ze kterého pocházel pacient, nicméně u této proměnné je potřeba říci, že nejvíce pacientů obsažených v datech pocházelo ze státu Greater California. Z hlediska metastáze můžeme říci, že největší poměr použití externí radiace ze všech pacientů je u případů, kdy stav metastáze byl M1 (došlo k rozšíření metastáz, rozšíření do lymfatických uzlin, byla napadena kůže, podpaží či další orgány a blízká místa). Pokud se zaměříme na stav, kdy byla u pacienta provedena léčba ozařováním po provedení chirurgického zákroku, tak z hlediska metastáze můžeme říci, že největší poměr ze všech pacientů, byl u případů, kdy nebyly potvrzeny žádné metastázy, ale byly zjištěny nádorové buňky v krvi, kostní dřeni nebo jiných ulových tkání.

Pokud si shrneme všechny získáné informace, tak můžeme říci, že pacienti zemřou s větší pravděpodobností, pokud jejich stádium rakoviny bude v pokročilé fázi (3, 4), budou v pokročilém věku (70+), budou potvrzeny vzdálené metastázy a u pacientů došlo k rozšíření nálezu na okolní místa. Z pohledu použití externího ozařování jsme dokázali, že tento druh léčby je nejpoužívanější léčebnou metodou z hlediska ozařování a že k jeho používání dochází i v nižších stádiích onemocnění bez rozdílu věku pacienta. Stejné znalosti platí i z hlediska léčby ozařování po provedení operace s výjimkou stavu, kdy nebyly potvrzeny vzdálené metastázy, ale byly zjištěny nádorové buňky v krvi nebo v kostní dřeni.

5 Závěr

Cílem práce bylo navrhnout a realizovat přípravu dat do formy vhodné pro aplikaci algoritmů a metod data-mining. Dále návrh a realizace experimentů vhodnými algoritmy a interpretace výsledků spolu s vyvozením patřičných závěrů o možnostech a vhodnosti vyzkoušených postupů.

V teoretické části jsme se seznámili s hlavními aspekty oboru Data mining. Popsali jsme si historii a hlavní metody tohoto oboru. Na základě těchto znalostí jsme přešli k řešení praktické části, ve které jsme se zaměřili na samostatnou práci s daty a programem IBM SPSS Modeler.

V průběhu praktické části se postupovalo dle metodologie CRISP-DM. V prvním kroku praktické části bylo potřeba stanovit, jakého cíle chceme dosáhnout (cíl práce). Následoval krok, ve kterém bylo potřeba porozumět datům (analýza dat). V dalším kroku bylo potřeba data upravit tak, aby do námi zvoleného nástroje vstupovala pouze validní data (úprava dat). Došlo k vyřazením nepotřebných atributů a zařazením potřebných atributů do finální množiny dat. Následoval výběr vhodných algoritmů, které budou hledat znalosti v datech. (použité algoritmy). V našem případě byly použity rozhodovací stromy C&RT, CHAID a C5. Nad testovanými daty byly použity i Neuronové sítě, ale jelikož nám neposkytovaly dostatečně velkou přesnost predikce, tak nebyly jejich výstupy použity k vlastní analýze cílových atributů. Dále bylo potřeba zhodnotit poskytnuté výstupy těchto technik (zhodnocení). Jednalo se o ověření poskytované přesnosti predikce modulem "Analysis". Všechny tyto kroky byly v naší práci použity a splněny a jsou podrobně popsány v jednotlivých kapitolách. Bohužel se však nepodařilo použití jiných algoritmů takovým způsobem, aby dosahovaly úspěšné predikce alespoň 50 %. Posledním krokem použité metodologie je nasazení modelu do praktického využití k zákazníkovi. Tento krok nebyl v naší práci zcela popsán a splněn, neboť by si vyžadoval další úpravu dat a možné nasazení dalších algoritmů.

Na závěr je potřeba říci, že poskytnutá data programem SEER jsou velmi rozsáhlá a kódování obsažené v těchto datech se během let měnilo, a proto zabralo dlouhý čas, než byla data správně pochopena a interpretována.

Nicméně na základě získaných výstupů můžeme říci, že znalosti, které jsme získali pomocí použitých nástrojů, se shodují s lékařskými znalostmi a logickými predikcemi a že veškeré uvedené cíle práce byly splněny.

6 Literatura

6.1 Tištěné knihy

BUREŠ, Vladimír. *Znalostní management a proces jeho zavádění: průvodce pro praxi*. 1. vyd. Praha: Grada, 2007, 212 s. Management v informační společnosti. ISBN 978-80-247-1978-8.

CLAUS WEIHS .. ED. Classification - the ubiquitous challenge proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Dortmund, March 9-11, 2004. [Online-Ausg.]. Berlin [u.a.]: Springer, 2005. ISBN 9783540280842.

GILLENSON, Mark L. *Fundamentals of database management systems*. 2nd ed. Hoboken, NJ: Wiley, 2012, xvi, 395 p. ISBN 9780470624708.

GÁLA, Libor, Jan POUR a Prokop TOMAN. Podniková informatika: počítačové aplikace v podnikové a mezipodnikové praxi, technologie informačních systémů, řízení a rozvoj podnikové informatiky. 1. vyd. Praha: Grada, 2006, 482 s. Management v informační společnosti. ISBN 80-247-1278-4

H.WITTEN, Ian a Eibe FRANK. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. 2.vydání. Burlington: Morgan Kaufmann, 2005. ISBN 0-12-088407-0.

JAMES, Gareth, Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRAMI. *An Introduction to Statistical Learning: with Applications in R*. 2.vydání. New York: Springer-Verlag New York, 2013. ISBN 978-1-4614-7138-7.

NISBET, Robert, Gary MINER a John ELDER. *Handbook of statistical analysis and data mining applications*. Boston: Elsevier Academic Press, 2009. ISBN 978-0-12-374765-5.

REFAAT, Mamdouh. *Data preparation for data mining using SAS*. Boston: Morgan Kaufmann Publishers, 2007, xxi, 399 p. Morgan Kaufmann series in data management systems.

SCHMUELI, Galit, Nitin R.PATEL, Peter C.BRUCE a . *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. 2.vydání. Hoboken, New Jersey, United States: John Wiley & Sons, Inc., 2010. ISBN 978-0-470-52682-8.

SUTTON, Richard S a Andrew G BARTO. *Reinforcement learning: an introduction*. Cambridge: Bradford Book, 1998, xviii, 322 s. Adaptive computation and machine learning. ISBN 0262193981.

Understanding metadata. Bethesda, MD: NISO, 2004. ISBN 1880124629.

WITTEN, Eibe FRANK a Mark A. HALL. *Data mining : practical machine learning tools and techniques*. Amsterdam: Morgan Kaufman, 2005. ISBN 0-12-088407-0.

WU, Xindong, Vipin KUMAR, J. ROSS QUINLAN, Joydeep GHOSH, Qiang YANG, Hiroshi MOTODA, Geoffrey J. MCLACHLAN, Angus NG, Bing LIU, et al. *Knowledge and*

Information Systems. 2008, (1): 1-37. DOI: 10.1007/s10115-007-0114-2. ISSN 0219-1377. Dostupné také z: http://link.springer.com/10.1007/s10115-007-0114-2

XIAOJIN ZHU AND ANDREW B. GOLDBERG. *Introduction to semi-supervised learning*. San Rafael, Calif.: Morgan & Claypool, 2009. ISBN 9781598295474.

6.2 Elektronické citace

Aboutdm: History of machine learning. *Aboutdm* [online]. 2013, 26.4.2013 [cit. 2015-12-14]. Dostupné z: http://www.aboutdm.com/2013/04/history-of-machine-learning.html

Cancer: seer. *Seer.cancer.gov* [online]. 2014 [cit. 2014-10-25]. Dostupné z: http://seer.cancer.gov/

Classification and Regression Trees. *Www.statsoft.com* [online]. 2015 [cit. 2015-10-09]. Dostupné z: http://www.statsoft.com/Textbook/Classification-and-Regression-Trees ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/ModelingNodes.pdf

Clustering. *Oracle* [online]. 2015 [cit. 2015-10-09]. Dostupné z: http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/clustering.htm

CHAID-Analysis. *Dell*:software [online]. 2015 [cit. 2015-12-14]. Dostupné z: http://documents.software.dell.com/Statistics/Textbook/CHAID-Analysis

IBM: developerworks. *IBM* [online]. 2014, 10.6.2014 [cit. 2015-12-14]. Dostupné z:http://www.ibm.com/developerworks/bpm/library/techarticles/1407_chandran/index.html

Inflow: znalostni informacni management. *Inflow* [online]. 2012, 7.11.2012 [cit. 2015-12-14]. Dostupné z: http://www.inflow.cz/znalostni-informacni-management

NATIONAL CANCER INSTITUTE. *Surveillance, Epidemiology, and End Results Program: Turning Cancer Data Into Discovery* [online]. 2015 [cit. 2015-10-09]. Dostupné z: http://seer.cancer.gov/archive/manuals/2010/SPCSM_2010_AppendixB.pdf

Regression. *Oracle* [online]. 2015 [cit. 2015-10-09]. Dostupné z: http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm

SEER: *About the SEER Program* [online]. 2015 [cit. 2015-12-15]. Dostupné z: http://seer.cancer.gov/about/overview.html

SEER*: Registry Groupings for Analyses* [online]. 2015 [cit. 2015-12-15]. Dostupné z: http://seer.cancer.gov/registries/terms.html

SEER*: SEER Data Management System* [online]. 2015 [cit. 2015-12-15]. Dostupné z: http://seer.cancer.gov/seerdms/

SEER*: SEER Registries* [online]. 2015 [cit. 2015-12-15]. Dostupné z: http://seer.cancer.gov/registries/

SPSS software: Predictive analytics software and solutions. IBM. *Www.ibm.com: software* [online]. 2015 [cit. 2015-10-09]. Dostupné z: http://www-01.ibm.com/software/analytics/spss/

Přílohy

A Nezařazená data

- Behavior code ICD-O-2 – atribut popisuje chování nálezu a jeho kódy jsou také zařazeny v ICD-O-2. V našem případě byly hodnoty odvozeny z novější verze ICD-O-3. Obsahovaly tedy stejné hodnoty, jako atribut Behavior code ICD-O-3.
- Patient ID – tento parametr je použit v registru SEER k jednoznačné identifikaci osoby. Atribut má délku 8 numerických znaků. Každá osoba může mít více výskytů různé rakoviny, ale bude mít vždy jen jedno a to stejné ID. Jednou použité ID pacienta nemůže být už nikdy použito v žádném dalším registru. V našem případě jsme toto pole nezařadili do výzkumu, neboť nedokazovalo žádnou důležitost u jednotlivých algoritmů.
- EOD-Extension Prost Path – tento atribut slouží pouze pro rakovinu prostaty a pouze za účelem odrážení informací z radikální prostatektomie a pro diagnózy od roku 1995. Pro ostatní druhy rakoviny je tento atribut prázdný, a proto nebyl zařazen do finálních dat.
- EOD—Lymph node involv – položka zaznamenává nejvyšší specifický řetězec lymfatických uzlů, který je zapojen v nádoru. Dostupné hodnoty pro atribut jsou 0 až 9. Pro naše účely však nebyl atribut zapojen do finálních dat z důvodu bsahování velmi mnoho prázdných záznamů, neboť v něm jsou obsaženy informace pouze do roku 2004. Tyto informace jsou ve finálních datech reprezentovány atributem CS Lymph nodes.
- EOD—Old 13 digit – atribut sloužil pouze pro případy do roku 1982.
- EOD—Old 2 digit – atribut sloužil pouze pro případy do roku 1982 a jen pro případy, které neměly kódování EOD13.
- EOD—Old 4 digit – atribut sloužil pouze pro případy v letech 1983 až 1987.
- Coding System for EOD – atribut zahrnuje SEER EOD kód aplikovaný do tumoru. V našem případě byly v atributu obsaženy záznamy, které obsahovaly dva druhy informace. První byla 4, která odpovídala 10-ti místnému číslu, které určovalo rozsah nemoci anebo zde byly prázdné záznamy, které značily, že EOD kódovací schéma nebylo aplikováno pro diagnostikované případy od roku 2004.
- Tumor marker 2 – atribut nám neposkytoval žádné bližší informace, neboť popisoval jen případy do roku 2003 a nesloužil pro případy rakoviny prsů. Ve finálních datech je nahrazen tento atribut atributem CS Site-specific Factor 2, který slouží pro popis rakoviny prsů.
- Tumor marker 3 – atribut slouží pouze pro popis onemocnění varlat.
- CS site-specific factor 25 – atribut obsahoval informaci o CS faktor site-specific (SSF). Každý faktor je závislý na daném schématu. Tento faktor může poskytnout informace potřebné k fázi případu, klinicky relevantní informace nebo prognostické informace. Je k dispozici pro různé období a schémata v závislosti na standardním nastavení požadavků. V našem případě obsahovaly záznamy hodnoty 988 nebo nebyla hodnota obsažena. Hodnota 988 odpovídá případům, kdy nebyl aplikován daný faktor.

- Derived AJCC-Flag – atribut slouží jen k informaci, zdali AJJC fáze byla odvozena z kódování CS, EOD nebo nebyla odvozena. Obsahoval jen dvě hodnoty, a to 1, která odpovídala tomu, že fáze byla odvozena z CS a prázdnou hodnotu, která odpovídala tomu, že fáze nebyla odvozena.
- Derived SS1997-Flag – atribut slouží jen k informaci, zda fáze shrnutí 1997 byla odvozena z kódování CS, EOD nebo nebyla odvozena. Obsahoval jen dvě hodnoty, a to 1, která odpovídala tomu, že fáze byla odvozena z CS a prázdnou hodnotu, která odpovídala tomu, že fáze nebyla odvozena.
- Derived SS2000-Flag – atribut slouží jen k informaci, zda fáze shrnutí 2000 byla odvozena z kódování CS, EOD nebo nebyla odvozena. Obsahoval jen dvě hodnoty, a to 1, která odpovídala tomu, že fáze byla odvozena z CS a prázdnou hodnotu, která odpovídala tomu, že fáze nebyla odvozena.
- CS Version derived – položka udává číslo verze CS používané v poslední době k odvození výstupního pole CS. První dvě číslice představují hlavní číslo verze, další dvě číslice představují drobné změny verze a poslední dvě číslice znamenají i méně významné změny, například opravy typografických chyb, které neovlivňují kódování nebo odvození výsledků. Atribut však nebyl zařazen do finálních dat, neboť neposkytoval žádnou bližší informaci.
- CS Version input current – položka udává číslo verze CS po tom, co byly aktualizovány nebo zaznamenány vstupní pole. První dvě číslice představují hlavní číslo verze, další dvě číslice představují drobné změny verze a poslední dvě číslice znamenají i méně významné změny, například opravy typografických chyb, které neovlivňují kódování nebo odvození výsledků. Atribut však nebyl zařazen do finálních dat, neboť neposkytoval žádnou bližší informaci.
- RX Summ-Reg Ln Examined – datová položka zaznamenává počet regionálních lymfatických uzlin zkoumaných v kombinaci s chirurgicky provedenými zákroky, v rámci prvního cyklu léčby na všech zařízeních. Tato položka je k dispozici pouze pro případy, které byly diagnostikovány mezi léty 1998 a 2002.
- RX Summ—Surg site 98-02 – v atributu jsou informace o typu operace na primárním místě nálezu v rámci prvního cyklu léčby, která byla provedena na všech možných zařízeních pro diagnostikované případy v letech 1998-2002. Atribut není zařazen ve finálních datech z důvodu velkého množství chybějících záznamů.
- RX Summ-Scope Reg 98-02 – tento atribut popisuje odebrání, biopsii nebo aspiraci regionálních lymfatických uzlin v době operace primárního výskytu nebo při samostatném chirurgickém zákroku na všech zařízeních. Atribut obsahuje informace pro diagnostikované případy v letech 1998-2002. V našem případě jsou v datech dva typy hodnot 9 a prázdné hodnoty. Prázdné hodnoty jsou pro případy, které byly zaznamenány v jiných letech a 9 odpovídá neznámým případům nebo případům, na které nebyl aplikován daný záznam. Atribut není zařazen ve finálních datech z důvodu velkého množství chybějících záznamů.

- RX Summ—surg oth 98-02 – v této položce je zaznamenáno odstranění vzdálených lymfatických uzlin, jiných tkání nebo orgánů za primárním místem nálezu v rámci prvního cyklu léčby pro diagnostikované případy v letech 1998-2002. Atribut není zařazen ve finálních datech z důvodu velkého množství chybějících záznamů.
- Over-ride age/site/morph – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Over-ride seqno/dxconf – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Over-ride sit/lat/seqno – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Over-ride surg/dxconf – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Over-ride site/type – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Over-ride histology – atribut obsahoval prázdné hodnoty a hodnotu 1, 2 nebo 3. Hodnota 1 odpovídala neobvyklé kombinaci histologie chování, 2 neobvyklé kombinaci diagnostického potvrzení a chování, hodnota 3 značila přezkoumání obou případů (1,2). Prázná hodnota pak odpovídala případům, kdy tento atribut nebyl hodnocen. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Over-ride report source – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Over-ride ill-define – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.

- Over-ride Leuk, Lymphoma – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Over-ride site/behavior – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Over-ride site/eod/dxdt – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Over-ride site/lat/eod – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Over-ride site/lat/morph – atribut obsahoval prázdné hodnoty nebo hodnotu 1, která odpovídala případům, kdy byl proveden chirurgický zákrok, ale mikroskopické potvrzení ukázalo, že tkáň nebyla dostatečně odstraněna. Do finálních dat nebyl atribut zařazen z důvodu velkého množství chybějících záznamů.
- Site Recode ICD-O-3/WHO 2008 – atribut obsahuje informace o primárním výskytu a ICD-O-3 histologii za účelem umožnit jednodušší analýzu výskytu nebo histologii skupin. Pro naše případy obsahovala data jednu hodnotu 26000, která odpovídala výskytu rakoviny prsu. Kvůli tomuto byl atribut vyřazen z finálních dat.
- Recode ICD-O-2 to 9 – v atributu jsou hodnoty, které byly získány převodem z ICD-O verze 2 do ICD-O verze 9. Ve výsledných finálních datech není tento atribut zahrnut, neboť je v datech atribut Recode ICD-O-10, který obsahuje novější verzi těchto dat.
- Histology recode—brain groupings – položka slouží pro identifikaci druhu histologie v části mozku. V tomto případě atribut obsahoval jen hodnotu 98, která odpovídala tomu, že se nejedná o onemocnění mozku.
- CS Schema v0204 – atribut obsahuje CS informaci, shromážděnou podle specifikace konkrétního schématu založeném na místě a dané histologii. Tento záznam by měl být použit v jakékoliv analýze AJCC 7th ed fáze a T, N, M. V daných datech byla u tohoto atributu uvedena jen jedna hodnota 13 odpovídající rakovině prsu.
- SEER modified AJCC stage 3rd ED (1988-2003) – atribut byl odvozený algoritmem z rozsahu onemocnění (EOD). Atribut není k dispozici pro všechny

roky, nebo pro všechny onemocnění. Tato upravená verze zahrnuje případy, které by nebyly zařazené podle přísných pravidel AJCC. V našem případě nebyl atribut zahrnut do finálních dat, neboť obsahoval spoustu prázdných záznamů.

- SEER summary stage 1977 – jedná se o atribut, který byl zaznamenáván do roku 2000 a který popisoval shrnutí fáze 1997. V našem případě bylo v datech zahrnuto velmi malé množství záznamů a tedy nebyl atribut zahrnut do finálních dat.
- SEER summary stage 2000 – jedná se o atribut, který byl zaznamenáván do roku 2000 a který popisoval shrnutí fáze 2000. V našem případě bylo v datech zahrnuto velmi malé množství záznamů a tedy nebyl atribut zahrnut do finálních dat.
- State-county recode – atribut je stvořen kombinací kraje a státu, kde první dva znaky reprezentují stát pomocí FIPS kódu a poslední tři číslice reprezentují kraj pomocí FIPS kódu. Po bližším prozkoumání nám atribut neposkytoval bližší informace, a tak nebyl zařazen do finálních dat.
- COD To Site REC KM – jedná se o atribut, který obsahuje záznamy založených na základní příčině smrti. Na základě této příčiny dochází k rozdělení do skupin podobných výskytu onemocnění záznamů s KS a mezoteliom. Atribut obsahoval stejné hodnoty jako atribut Cause of death to SEER site recode, a proto nebyl zařazen do finálních dat.
- Summary stage 2000 (1998+) – atribut obsahuje informace o fáze shrnutí 2000, která je odvozena od Collaborative Stage (CS) pro případy od roku 2004 a rozsahu onemocnění (EOD), v letech 1998-2003. Jedná se o zjednodušenou verzi, která v sobě obsahuje informace o jednotlivé fázi in situ, lokalizovaná, regionální, vzdálená anebo neznámá. V našem případě obsahovala data velké množství prázdných záznamů, a proto nebyl atribut zařazen do finálních dat.
- Lymphoma subtype recode/WHO 2008 – atribut obsahuje informace o místě a histologii a používá se hlavně pro analýzu údajů o dospívajících a mladých dospělých. Záznam byl aplikován na všechny případy bez ohledu na věk, aby bylo možné provést vzájemné porovnání mezi jednotlivými skupinami. V našem případě atribut obsahoval jen hodnotu 99, která odpovídala neklasifikovanému případu, proto také nebyl atribut zařazen do finálních dat.
- CS schema—AJCC 6th ED (previously called V1) – CS informace shromážděné podle specifikace konkrétního schéma založené na místě a histologie. Tento záznam by měl být použit v jakékoliv analýze AJCC 6th ed fázi a T, N, M. Založená na CS verze 1, neměl by být použit pro SSF[39]s shromážděné nebo upravené podle CS v02. V našem případě atribut obsahoval malé množství vyplněných záznamů, a tak nebyl zařazen do finálních dat.
- CS site-specific factor 8, 10, 11, 13, 15 a 16 – jedná se o atributy CS faktor site-specific (SSF), které jsou závislé na daném schématu. Tato schémata mohou poskytnout informace potřebné k fázi případu, klinicky relevantní informace, nebo prognostické informace k dispozici pro různé roky a schémata v závislosti

[39] SSFs – Site specific factors

na standardních nastavených požadavcích. V našem případě atributy obsahovaly velké množství prázdných záznamů, a tak nebyly zařazeny do finálních dat.

- Survival months flag – atribut obsahuje informaci o kompletních datech, včetně dnů, a proto se může lišit od doby přežití, která je vypočítaná pouze z roku a měsíce. Atribut nebyl zařazen do finálních dat, protože po bližším prozkoumání neposkytoval žádné další informace.
- Survival months – presumed alive – atribut obsahuje informaci o kompletních datech, včetně dnů, a proto se může lišit od doby přežití, která je vypočítaná pouze z roku a měsíce. Pokud poslední známý stav je naživu, potom databáze studie mezního dne je použita jako datum posledního kontaktu. Atribut nebyl zařazen do finálních dat, protože po bližším prozkoumání neposkytoval žádné další informace.
- Survival months – presumed alive flag – atribut vytvořen pomocí kompletních dat, včetně dnů, a proto se může lišit od doby přežití, vypočítán pouze z roku a pouze měsíce. Pokud je poslední zásadní stav naživu, tak se databázová studie mezního data používá jako datum posledního kontaktu. Atribut nebyl zařazen do finálních dat, protože po bližším prozkoumání neposkytoval žádné další informace.
- Derived AJCC-7 T – jedná se o položku, ve které je obsažena AJCC "T" složka, která je odvozena pomocí algoritmu CS, z CS kódovaných polí. Tato položka je efektivní pro případy diagnostikovaných od roku 2010. V našem případě nebyla položka zařazena do finálních dat z důvodu velkého množství prázdných záznamů.
- Derived AJCC-7 N – jedná se o položku, ve které je obsažena AJCC "N" složka, která je odvozena pomocí algoritmu CS, z CS kódovaných polí. Tato položka je efektivní pro případy diagnostikovaných od roku 2010. V našem případě nebyla položka zařazena do finálních dat z důvodu velkého množství prázdných záznamů.
- Derived AJCC-7 M – jedná se o položku, ve které je obsažena AJCC "M" složka, která je odvozena pomocí algoritmu CS, z CS kódovaných polí. Tato položka je efektivní pro případy diagnostikovaných od roku 2010. V našem případě nebyla položka zařazena do finálních dat z důvodu velkého množství prázdných záznamů.
- RX Summ—RAD to CNS – proměnná byla zaznamenávána pouze pro roky 1988-1997 a pouze pro případy rakoviny plic a leukémie.
- AJCC stage 3rd edition (1988-2003) – atribut je odvozený algoritmem z rozsahu onemocnění (EOD). Není k dispozici pro všechny roky nebo pro všechny druhy onemocnění. Jelikož byl atribut zaznamenáván pouze do roku 2003, tak obsahoval velmi malé množství záznamů a tedy nebyl zařazen do finálních vstupních dat.
- Lymph vascular invasion – atribut sloužil pouze pro onemocnění rakoviny penisu a varlat. Z tohoto důvodu nebyl zařazen do finálních dat.

B Zařazená data

- Spanish/Hispanic origin – datová položka, která je složena z jednoho numerického znaku, reprezentuje osoby, které měly španělské nebo hispanské příjmení nebo španělský původ. Osoby se španělským nebo s hispánským příjmením nemusí být však stejného původu. Pro ukázku si můžeme uvést, že osoby, které jsou původem z Dominikánské republiky, mají hodnotu 8.
- Grade – atribut pro ICD-O-2 má hodnotu 1 až 4 nebo 9 a znázorňuje nám stupeň a diferenciaci onemocnění. Pro roky před 1977 může být tento parametr neúplný a pro případy v brzkých 90 letech specifikují kódy, zda se jednalo o T-buňky, B-buňky nebo nulové buňky u lymfomů a u onemocnění leukémií. Tento parametr se tedy během let měnil a je tedy složité ho analyzovat.
- Regional nodes examined – dvounumerický atribut, který byl použit od roku 1988 a zaznamenává celkový počet lokálních lymfatických uzlin, které byly odstraněny nebo vyšetřeny patologem. Hodnoty jsou buď 00 = žádné nebyly odstraněny, 01 – 89 = exaktní počet vyšetřených uzlin, 90 = 90 nebo více uzlin nebo například 98 = není známo, zda byly vyšetřeny nějaké uzliny.
- RX Summ – Oth reg/dis – Atribut byl použit u záznamů od roku 2003 a je tvořen jedním numerickým znakem, který nám popisuje chirurgické odstranění vzdálených lymfatických uzlin, jiné tkáně nebo orgánů mimo primárního výskytu. Hodnota 0 nám například vyjadřuje, že neproběhlo žádné odstranění a další výskyty byly diagnostikovány až při pitvě. Hodnota 1 vyjadřuje, že nebyla vykonána žádná chirurgická operace kromě odstranění primárního výskytu nebo například 9, která reprezentuje možnost, že atribut nebyl vyplněn nebo byl potvrzen až úmrtním listem.
- Derived AJCC-6 T – tento atribut popisuje AJCC "T" složku, která je odvozena z CS kódovaných polí, pomocí algoritmu CS. Atribut je efektivní u případů, které byly diagnostikovány od roku 2004+.
- Derived AJCC-6 N – atribut popisuje AJCC "N" složku, která je odvozena z CS kódovaných polí, pomocí algoritmu CS. Atribut je efektivní u případů, které byly diagnostikovány od roku 2004+.
- Derived AJCC-6 M – atribut popisuje AJCC "M" složku, která je odvozena z CS kódovaných polí, pomocí algoritmu CS. Atribut je efektivní u případů, které byly diagnostikovány od roku 2004+.
- Derived SS1977 – datová položka je odvozená ze "SEER Summary Stage 1977" z algoritmu CS. Atribut je efektivní u případů diagnostikovaných od roku 2004.
- SEER record number – Záznamové číslo je jedinečné pořadové číslo. Jedná se o nejvyšší číslo pro každého pacienta, které určuje počet záznamů, které byly předloženy k léčení pro daného konkrétního pacienta. Čísla záznamů jsou sekvenční a začínají číslem 01. V našem případě je maximum záznámů u jednoho pacienta číslo 7.
- ICCC site rec extended ICD-O-3/WHO 2008 – jedná se o záznam, který nám udává informace o místě/hostologii a který se používá hlavně pro analýzu dat

u dětí. Záznam byl aplikován na všechny případy bez ohledu na věk, aby mohlo dojít k věkovému srovnání s těmito uskupeními (ICD-O-3 a WHO 2008). Jedná se o rozšíření atributu výše.

- Bahavior for record analysis – tento atribut byl vytvořen za účelem, aby data mohla eliminovat hlavní skupiny histologie nebo chování, které nebyly v průběhu času konzistentně vybrány. Atribut byl vytvořen z ICD-O-3 chování a histologie.
- CS tumor size ext/eval – atribut blíže popisuje nález tumoru a tento záznam byl zaznamenáván u případů od roku 2004.
- CS lymph nodes eval – atribut nám prezentuje přesný počet lymfatických uzlin v zaznamenaných případech od roku 2004.
- CS mets eval – atribut blíže popisuje nález tumoru a tento záznam byl zaznamenáván u případů od roku 2004.
- PR Status recode breast cancer (1990+) – atribut vytvořený kombinací informací z atributu Tumor marker 2 (1990-2003), s informacemi z CS site-specific faktoru 2 (2004+). S tím, že toto pole je prázdné pro případy, kdy nebyla zjištěna rakovina prsu a u případů diagnostikovaných před rokem 1990. Hodnoty jsou 1 – pozitivní, 2 – negativní, 3 – hraniční, 4 – neznámý, 9 – ne 1990+ prsa.
- Race recode (white, black, other) – Záznam týkající se původu daného pacienta, je založen na rasových proměnných a proměnné American Indian/Native American IHS. Tento záznam by měl být použit k propojení s populací pro bílé, černé a další a je nezávislý na hispánské etnice. Hodnoty jsou 1 = běloch, 2 = černoch, 3 = jiné, 9 = neznámé.
- Race recode (W, B, AI, API) – Při použití této proměnné je zapotřebí pečlivosti. Atribut nám specifikuje, stejně jako předchozí, původ pacienta. Hodnoty jsou 1 = běloch, 2 = černoch, 3 = americký indián/původní obyvatel Aljašky, 4 = Asiat nebo obyvatel pacifického ostrova, 9 = neznámé.
- Origin recode NHIA (Hispanic, Non-hisp) – u tohoto atributu platí to stejné jako u předchozího, a to, že je potřeba být u použití této proměnné velmi opatrní. Atribut nám říká, zda daný pacient byl hispánského původu (1) či ne (0).
- SEER histologic stage A – Atribut vychází z Collaborative Stage (CS) pro 2004+ a rozsahu onemocnění (EOD), od 1973-2003. Jedná se o zjednodušenou verzi stádia onemocnění: in situ, lokalizované, lokální, vzdálené a neznámé. Postupem času bylo použito několik různých schémat EOD. Kódování v datech je následující.

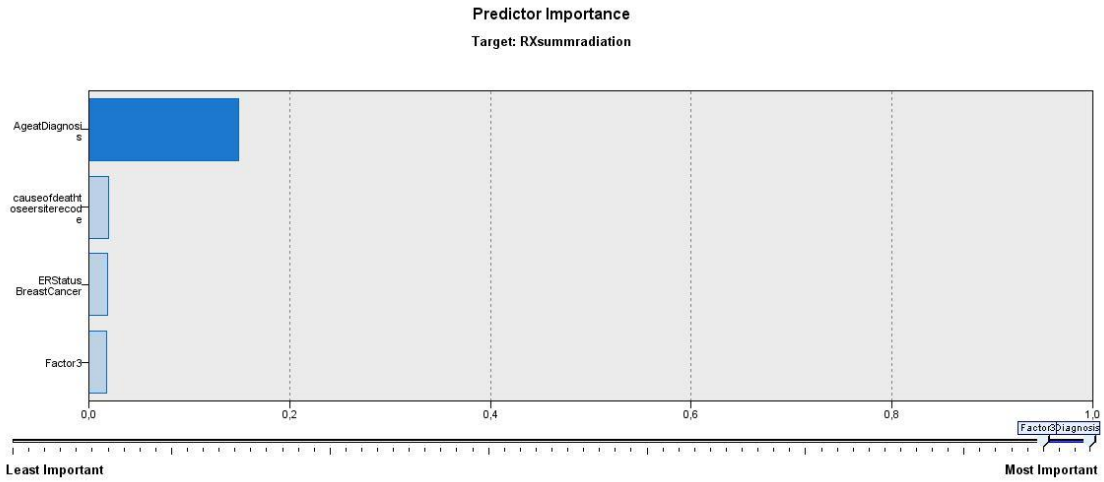
Tab. 6 Histologické stádia A

Kód	**Popis**
0	In situ. Jedná se o neinvazivní nádor, který neprorazil na bazální membránu, ani nebyl rozšířen přes epiteliální tkáně.
1	Lokalizovaný – Invazivní nádor omezený výhradně na originální původ na daném orgánu.
2	Lokální – Nádor rozšířený za hranice orgánu původu přímo do okolních orgánů nebo tkáně nebo do lokálních lymfatických uzlin prostřednictvím lymfatického systému nebo kombinací rozšíření a lokálních lymfatických uzlin.
4	Vzdálený – Nádor, který se rozšířil do části těla, odvrácené od primárního místa buď přímým rozšířením nebo diskontuálními metastázemi do vzdálených orgánů.
8	Lokalizované/regionální – používá se pouze v případě onemocnění prostaty.
9	Nezařazené – informace není dostačují pro přiřazení do stádia.

C Výstupy algoritmu Neuronová síť

Use?	Graph	Model	Build Time (mins)	Overall Accuracy (%)	No. Fields Used
☑		C5 1	157	63,16	63

Obr. 35 RX summ-radiation – Auto classifier

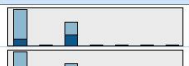
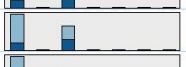


Obr. 36 RX summ-radiation – Neuronová síť

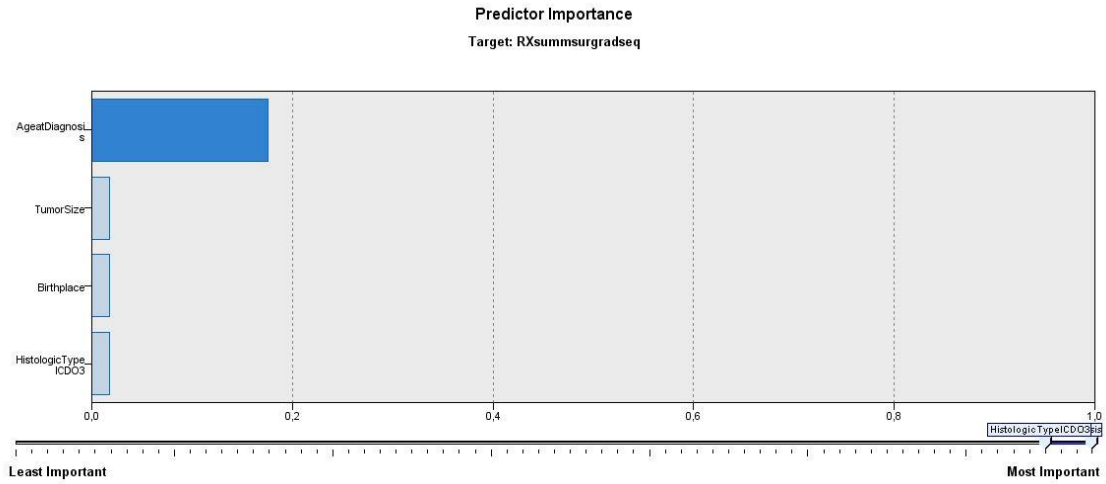
Comparing $N-RXsummradiation with RXsummradiation

'Partition'	1_Training		2_Testing	
Correct	43 590	22,72%	43 533	22,67%
Wrong	148 234	77,28%	148 516	77,33%
Total	191 824		192 049	

Obr. 37 RX summ-radiation, Neuronová síť – Analysis

Use?	Graph	Model	Build Time (mins)	Overall Accuracy (%)	No. Fields Used
☑		C5 1	68	66,98	53
☑		C&R Tree 1	1	64,2	26
☑		Quest 1	1	64,2	69
☑		CHAID 1	1	64,2	23

Obr. 38 RX summ-surg/rad seq – Auto classifier



Obr. 39 RX summ surg/rad seq – Neuronová síť

Comparing $N-RXsummsurgradseq with RXsummsurgradseq

'Partition'	1_Training		2_Testing	
Correct	46 558	24,27%	46 745	24,34%
Wrong	145 266	75,73%	145 304	75,66%
Total	191 824		192 049	

Obr. 40 RX summ-surg%rad seq, Neuronová síť – Analysis

D Diskuze – důkazy

V této části budou zobrazeny obrázky tabulek, které byly vygenerovány v programu IBM SPSS Modeler, a které slouží jako podklad pro diskuzi. Obrázky 41 až 44 se zabývají atributem Vitalstatus recode a obrázky 45 až 49 atributy RX summ – radiation a RX summ – surg/rad seq.

DerivedSS2000		1.0	4.0
$null$	Count	92101	41907
	Row %	68.728	31.272
0.0	Count	46305	1895
	Row %	96.068	3.932
1.0	Count	112465	10595
	Row %	91.390	8.610
2.0	Count	3559	1542
	Row %	69.771	30.229
3.0	Count	42567	5967
	Row %	87.706	12.294
4.0	Count	6023	2748
	Row %	68.669	31.331
7.0	Count	4512	6604
	Row %	40.590	59.410
9.0	Count	2256	2827
	Row %	44.383	55.617

Obr. 41 Rozšíření nádoru – vitalstatus recode

AdjustedAJCC6thStage		1.0	4.0
0.0	Count	66611	5354
	Row %	92.560	7.440
10.0	Count	119893	16696
	Row %	87.776	12.224
32.0	Count	57308	11573
	Row %	83.199	16.801
33.0	Count	23071	6158
	Row %	78.932	21.068
51.0	Count	459	635
	Row %	41.956	58.044
52.0	Count	14351	5241
	Row %	73.249	26.751
53.0	Count	3931	3468
	Row %	53.129	46.871
54.0	Count	5990	4208
	Row %	58.737	41.263
70.0	Count	4317	10605
	Row %	28.930	71.070
88.0	Count	278	218
	Row %	56.048	43.952
99.0	Count	13579	9929
	Row %	57.763	42.237

Obr. 42 Fáze onemocnění – vitalstatus recode

CSMetsatDX		1.0	4.0
$null$	Count	92101	41907
	Row %	68.728	31.272
0.0	Count	209683	22149
	Row %	90.446	9.554
10.0	Count	386	235
	Row %	62.158	37.842
40.0	Count	1519	2546
	Row %	37.368	62.632
42.0	Count	56	74
	Row %	43.077	56.923
44.0	Count	1691	2567
	Row %	39.713	60.287
5.0	Count	8	1
	Row %	88.889	11.111
50.0	Count	494	978
	Row %	33.560	66.440
60.0	Count	135	66
	Row %	67.164	32.836
99.0	Count	3715	3562
	Row %	51.051	48.949

Obr. 43 Fáze metastáz – vitalstatus recode

agerecode1yearolds		1.0	4.0
	Row %	90.216	9.784
11.0	Count	41363	5151
	Row %	88.926	11.074
12.0	Count	41806	5890
	Row %	87.651	12.349
13.0	Count	39801	6209
	Row %	86.505	13.495
14.0	Count	35553	7084
	Row %	83.385	16.615
15.0	Count	28932	8467
	Row %	77.360	22.640
16.0	Count	23298	10502
	Row %	68.929	31.071
17.0	Count	14381	10153
	Row %	58.617	41.383
18.0	Count	7674	11053
	Row %	40.978	59.022
3.0	Count	4	0
	Row %	100.000	0.000
4.0	Count	16	6
	Row %	72.727	27.273
5.0	Count	210	49
	Row %	81.081	18.919
6.0	Count	1171	286
	Row %	80.371	19.629
7.0	Count	3758	732
	Row %	83.697	16.303
8.0	Count	10060	1606
	Row %	86.233	13.767

Obr. 44 Věk pacienta do skupin – vitalstatus recode

AdjustedAJCC6thM		0.0	1.0	2.0	3.0	4.0	5.0	7.0	8.0	9.0
0.0	Count	189617	143146	5567	133	891	2006	4346	5878	5187
	Row %	53.148	40.123	1.560	0.037	0.250	0.562	1.218	1.648	1.454
10.0	Count	9727	4420	10	6	22	79	247	158	253
	Row %	65.186	29.621	0.067	0.040	0.147	0.529	1.655	1.059	1.695
88.0	Count	337	133	1	0	0	3	6	5	11
	Row %	67.944	26.815	0.202	0.000	0.000	0.605	1.210	1.008	2.218
99.0	Count	7567	1136	27	0	1	70	252	93	2538
	Row %	64.764	9.723	0.231	0.000	0.009	0.599	2.157	0.796	21.722

Obr. 45 Fáze metastáz – rx summ radiation

AdjustedAJCC6thM		0.0	2.0	3.0	4.0	5.0	6.0	9.0
0.0	Count	205861	816	148120	921	129	73	851
	Row %	57.701	0.229	41.517	0.258	0.036	0.020	0.239
10.0	Count	12440	252	2124	67	2	1	36
	Row %	83.367	1.689	14.234	0.449	0.013	0.007	0.241
88.0	Count	364	6	126	0	0	0	0
	Row %	73.387	1.210	25.403	0.000	0.000	0.000	0.000
99.0	Count	10505	23	1006	5	2	0	143
	Row %	89.909	0.197	8.610	0.043	0.017	0.000	1.224

Obr. 46 Fáze metastáz – rx summ surg/rad seq

AdjustedAJCC6thStage		0.0	1.0	2.0	3.0	4.0	5.0	7.0	8.0	9.0
0.0	Count	44653	23237	1121	22	144	296	904	647	941
	Row %	62.048	32.289	1.558	0.031	0.200	0.411	1.256	0.899	1.308
10.0	Count	65290	60876	3814	85	440	845	1582	1807	1850
	Row %	47.800	44.569	2.792	0.062	0.322	0.619	1.158	1.323	1.354
32.0	Count	39025	25623	483	14	145	368	785	1344	1094
	Row %	56.656	37.199	0.701	0.020	0.211	0.534	1.140	1.951	1.588
33.0	Count	16316	11153	70	3	54	146	369	699	419
	Row %	55.821	38.157	0.239	0.010	0.185	0.500	1.262	2.391	1.434
51.0	Count	629	367	0	0	3	7	31	27	30
	Row %	57.495	33.547	0.000	0.000	0.274	0.640	2.834	2.468	2.742
52.0	Count	7734	10536	7	4	35	156	204	668	248
	Row %	39.475	53.777	0.036	0.020	0.179	0.796	1.041	3.410	1.266
53.0	Count	3549	3320	4	0	12	52	153	189	120
	Row %	47.966	44.871	0.054	0.000	0.162	0.703	2.068	2.554	1.622
54.0	Count	4093	5411	3	0	23	82	92	353	141
	Row %	40.135	53.059	0.029	0.000	0.226	0.804	0.902	3.461	1.383
70.0	Count	9727	4420	10	6	22	79	247	158	253
	Row %	65.186	29.621	0.067	0.040	0.147	0.529	1.655	1.059	1.695
88.0	Count	337	133	1	0	0	3	6	5	11
	Row %	67.944	26.815	0.202	0.000	0.000	0.605	1.210	1.008	2.218
99.0	Count	15895	3759	92	5	36	124	478	237	2882
	Row %	67.615	15.990	0.391	0.021	0.153	0.527	2.033	1.008	12.260

Obr. 47 Fáze onemocnění – rx summ radiation

Factor1		0.0	1.0	2.0	3.0	4.0	5.0	7.0	8.0	9.0
$null$	Count	72406	52590	351	44	650	523	1459	1768	4217
	Row %	54.031	39.244	0.262	0.033	0.485	0.390	1.089	1.319	3.147
10.0	Count	84974	71761	4308	70	207	1215	2458	3103	1452
	Row %	50.118	42.325	2.541	0.041	0.122	0.717	1.450	1.830	0.856
20.0	Count	23589	17653	603	11	24	260	494	946	422
	Row %	53.609	40.119	1.370	0.025	0.055	0.591	1.123	2.150	0.959
30.0	Count	285	158	11	0	0	6	4	3	6
	Row %	60.254	33.404	2.326	0.000	0.000	1.268	0.846	0.634	1.268
996.0	Count	29	6	1	0	0	0	2	0	0
	Row %	76.316	15.789	2.632	0.000	0.000	0.000	5.263	0.000	0.000
997.0	Count	2952	1075	64	7	12	21	38	69	92
	Row %	68.176	24.827	1.478	0.162	0.277	0.485	0.878	1.594	2.125
998.0	Count	12654	3275	136	5	9	33	258	138	220
	Row %	75.646	19.578	0.813	0.030	0.054	0.197	1.542	0.825	1.315
999.0	Count	10359	2317	131	2	12	100	138	107	1580
	Row %	70.250	15.713	0.888	0.014	0.081	0.678	0.936	0.726	10.715

Obr. 48 Faktor ER – rx summ radiation

RegistryID		0.0	1.0	2.0	3.0	4.0	5.0	7.0	8.0	9.0
1541.0	Count	87092	65570	2230	67	73	1459	3685	4335	795
	Row %	52.685	39.666	1.349	0.041	0.044	0.883	2.229	2.622	0.481
1542.0	Count	17161	14257	543	10	39	343	622	444	5329
	Row %	44.289	36.794	1.401	0.026	0.101	0.885	1.605	1.146	13.753
1543.0	Count	21114	14354	619	32	203	88	15	20	332
	Row %	57.411	39.030	1.683	0.087	0.552	0.239	0.041	0.054	0.903
1544.0	Count	54894	36872	1028	20	517	190	48	153	891
	Row %	58.020	38.971	1.087	0.021	0.546	0.201	0.051	0.162	0.942
1547.0	Count	26987	17782	1185	10	82	78	481	1182	642
	Row %	55.725	36.718	2.447	0.021	0.169	0.161	0.993	2.441	1.326

Obr. 49 Stát – rx summ radiation