UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

DIPLOMOVÁ PRÁCE

POUŽITÍ KLASIFIKAČNÍCH STROMŮ PRO DIAGNOSTIKU RAKOVINY PROSTATY

Vedoucí diplomové práce:
Mgr. Ondřej Vencálek, Ph.D.

Vypracovala:
Bc. Andrea Luterová

Olomouc 2014

Bibliografický záznam

Autor:	Bc. Andrea Luterová Přírodovědecká fakulta, Univerzita Palackého v Olomouci

Název práce:	Použití klasifikačních stromů pro diagnostiku rakoviny prostaty
Studijní program:	N1101 Matematika
Studijní obory:	Učitelství biologie pro střední školy Učitelství matematiky pro střední školy
Vedoucí práce:	Mgr. Ondřej Vencálek, Ph.D.
Akademický rok:	2013/2014
Počet stran:	93
Klíčová slova:	Klasifikační a regresní stromy, rakovina prostaty, Giniho koeficient, CART

Bibliographic Entry

Author	Bc. Andrea Luterová Faculty of Science, Palcky University in Olomouc

Title of Thesis:	Use of classification trees for prostate cancer diagnosis
Degree programme:	N1101 Mathematics
Fields of Study:	Teaching Biology for High Schools Teaching Mathematics for High Schools
Supervisor:	Mgr. Ondřej Vencálek, Ph.D.
Academic Year:	2013/2014
Number of Pages:	93
Keyword:	Classification and regression trees, prostate cancer, Gini coefficient, CART

Abstrakt

V této diplomové práci se věnujeme použití klasifikačních a regresních stromů na data týkající se karcinomu prostaty. Práce nás nejprve seznamuje s problematikou karcinomu prostaty a potřeby zpracování těchto dat pro lékařské účely. Dále se zabývá přípravou dat pro účely zpracování, opravou možných chyb a výběrem použitelných záznamů, aby výsledky práce byly co nejvíce pravděpodobné. Na tomto souboru chceme prozkoumat závislosti mezi proměnnými a zařazením do tříd a naučit se je co nejpřesněji odhadovat.

Diplomová práce vychází z české a cizojazyčné literatury, která je uvedena v textu a v seznamu na konci práce, a také z vlastních zkušeností získaných při jejím vzniku. K neparametrickým odhadům a popisné statistice byl využit program STATISTICA.

Abstract

In this thesis we study the use of classification and regression trees to data related to prostate cancer. This work first introduces us to the issue of prostate cancer and the need of data processing for medical purposes. It also deals with the preparation of data for processing, repairs on possible errors and selecting the applicable records, that the work will be most likely. In this file we want to explore dependencies between variables and included in the classes and learn how to estimate as accurately as possible.

The thesis is based on Czech and foreign literature, which is mentioned in the text and in the list at the end of the work, and also from my own experience gained during its creation. The non-parametric estimates and descriptive statistics were used STATISTICA program.

Poděkování

Na tomto místě bych chtěla poděkovat vedoucímu mé bakalářské práce panu Mgr. Ondřeji Vencálkovi Ph.D. za jeho trpělivost, čas i odbornou pomoc a za jeho další cenné rady při zpracování této diplomové práce.

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana Mgr. Ondřeje Vencálka, Ph.D. s použitím uvedené literatury.

V Olomouci dne 31. března 2014

.....

Bc. Andrea Luterová

Obsah

ÚVOD **7**

1 KARCINOM PROSTATY **8**

1.1 Příznaky a léčba karcinomu prostaty 11
1.2 Varianty karcinomu prostaty 12

2 DATA A JEJICH POPIS **14**

2.1. Typy dat 14
2.2 Datový soubor 15
2.3 Prostatický specifický antigen PSA 17
2.4 Čištění datového souboru 18
2.5 Testové statistiky 19
2.6 Popisná statistika souboru 23

3 KLASIFIKAČNÍ A REGRESNÍ STROMY **32**

3.1 Mnohonásobná regrese *33*
3.2 Klasifikační stromy *35*
3. 2. 1 Rozhodovací pravidla větvení *38*
3.2.2 Možnosti ukončení větvení *52*
3.2.3 Ověření velikosti stromu *54*
3.3 CART *54*
3.4 Regresní stromy *56*
3.4.1 Regresní metody *58*

4 TVRZENÍ A HYPOTÉZY **59**

4.1 Klasifikace prvních případů *60*
4.2 Klasifikace rebiopsií *64*

ZÁVĚR **68**

SEZNAM POUŽITÉ LITERATURY **71**

SEZNAM GRAFŮ A OBRÁZKŮ **74**

SEZNAM VZORCŮ **76**

SEZNAM TABULEK **77**

PŘÍLOHY **78**

Klasifikační stromy pro první pozorování *78*
Data2 ..... *78*
Data3 ..... *80*
Data 4 ..... *82*

Klasifikační stromy pro rebiopsie 84
Data2.. 84
Data 3 86
Data 4 88
Krabicové grafy 90

Úvod

Karcinom prostaty patří spolu s karcinomem plic k jednomu z nejčastějších nádorových onemocnění u mužů. Zpracování dat týkajících se tohoto onemocnění je tedy pro lékařské účely velice důležité. Ke zpracování mohou být použité odlišné statistické metody, kdy každá z nich může být přínosná pro lepší odhad budoucích výskytů onemocnění případně pro odhad jeho průběhu.

V této práci je použita pro odhad výsledků druhých biopsií (první rebiopsie) metoda klasifikačních stromů. Tuto metodu aplikujeme na data z Fakultní nemocnice v Olomouci, která byla nasbírána v letech 2006 až 2012. Soubor obsahuje pacienty, kteří byli v nemocnici na preventivním vyšetření nebo přišli již s nějakými obtížemi. Najdeme vztahy, které existují mezi hodnotami jednotlivých vyšetření u pacientů s diagnostikovaným a nediagnostikovaným karcinomem prostaty, a ty můžeme použít ke klasifikaci budoucích případů.

Budeme zde tedy srovnávat výsledky jednotlivých vyšetření u pacientů a porovnávat hladiny těchto prediktorů, které jsou důležitými ukazateli pro výskyt karcinomu prostaty. Na základě prediktorů (vysvětlujících proměnných), které budou nejvhodnější pro co nejpřesnější zařazení, pak budeme moci odhadnout, zda by při biopsii (rebiopsii) byl nález pozitivní nebo by byl karcinom prostaty nediagnostikován.

V práci jsou popsány jednotlivé metody užívané při tvorbě klasifikačních a regresních stromů. Metoda aplikovaná na náš datový soubor bude obsahovat i ilustrační příklad klasifikace jednotlivých případů do skupin.

1 Karcinom prostaty

Nádorová onemocnění jsou jednou z nejběžnějších typů chorob moderní populace a bohužel jsou i častou příčinou úmrtí. Jsou specifické nekontrolovatelným dělením zmutovaných tělních buněk. Zmutovaná buňka se vymkne kontrole organismu, neproběhne apoptóza buňky (programovaná smrt poškozené buňky) a ani její oprava a tak buňka nekontrolovatelně roste a množí se. Dochází k napadání okolní tkáně a zároveň se tak oslabuje obranyschopnost daného jedince. Česká republika stojí na prvních příčkách v celosvětových statistikách incidence nádorových onemocnění.

Karcinom prostaty se vedle kolorektálního karcinomu a karcinomu plic řadí mezi nejčastější nádorové onemocnění postihující muže. Uvádí se, že karcinom prostaty postihne asi 60 z 100 000 mužů. Navíc v posledních letech výskyt tohoto onemocnění stále stoupá Obr. 1.2, a předpokládá se, že stále stoupat bude Obr. 1.3, a proto je velice důležitá prevence. V národním onkologickém registru je karcinom prostaty označován kódem C61.

Výskyt karcinomu prostaty je stejně jako i další nádorová onemocnění závislý na věku jedince, na genetických faktorech, ale kromě toho také na barvě kůže. Obecně negroidní rasa je více náchylná na onemocnění karcinomem prostaty než europoidní a u mužů starších čtyřiceti-pěti let je riziko onemocnění vyšší než u mužů mladšího věku. Mladší věk muže, ale možnost onemocnění nevylučuje. Spekuluje se i o dalších možných faktorech, které by mohly ovlivňovat výskyt této nemoci, ale dosud jejich vliv nebyl přímo prokázán. Patří sem obezita, kouření, ale také prodělání určitých infekčních onemocnění. Z genetického hlediska příbuzenství prvního stupně s nemocným (tzn. přímý vztah otec-syn) má vliv na vyšší rizikovost výskytu karcinomu prostaty.

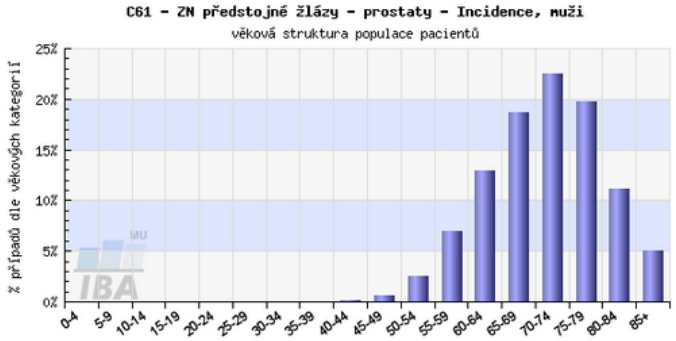
V České republice jsou muži posíláni na preventivní vyšetření, pokud jsou starší padesáti let. Toto vyšetření ale není součástí státem organizovaného screeningu. Nejvíce jsou karcinomem prostaty postiženi muži ve věku šedesáti-pěti až osmdesáti let Obr. 1.1, ale nejlepší by bylo navštěvovat preventivní vyšetření již od čtyřicátého věku života. Ve sledovaném výskytu tohoto onemocnění není zahrnuta a ani nelze zahrnout latentní (skrytou)

formu karcinomu, což znamená, že se příznaky onemocnění ještě na jedinci neprojevily. Výskyt karcinomu prostaty je tedy pravděpodobně ještě vyšší.

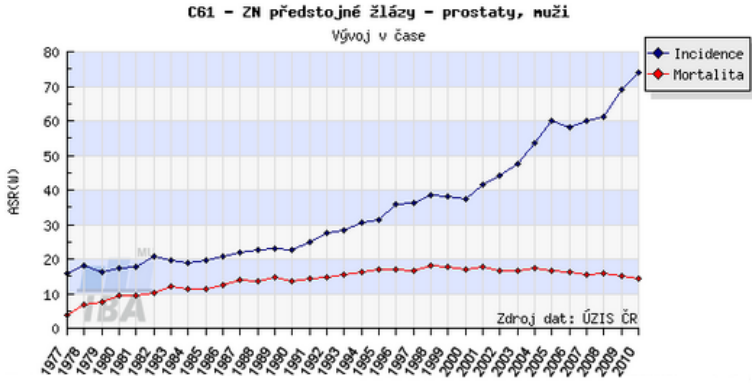
Mezi základní úkoly lékařství dnes patří prevence. Správná životospráva může snížit riziko vzniku karcinomu. Riziko výskytu můžeme snížit skladbou jídelníčku, kde snížíme příjem tuků a naopak zvýšíme příjem vitamínu E a D, selenu, izoflavonoidů (obsažené například v sóji) a podobně.

Mezi doporučovaná preventivní vyšetření patří vyšetření na prostatický specifický antigen (PSA), který vylučují nádorové buňky prostaty do krve, a vyšetření per rectum. Dalším krokem je správná diagnostika a registrace pacienta do seznamu nemocných - do Národního onkologického registru. Účelem registrace je lepší povědomí o rozšíření karcinomů v populaci a také budoucí úspěšnější odhalování nemoci i odhad jejího následujícího průběhu. V neposlední řadě je důležitá léčba, která je nejúspěšnější většinou tehdy, když je nemoc zachycena v raném stádiu, kdy nádorové buňky ještě neopustily prostatu a nedostaly se do jiných částí těla. Většinu těchto pacientů jde zcela vyléčit.

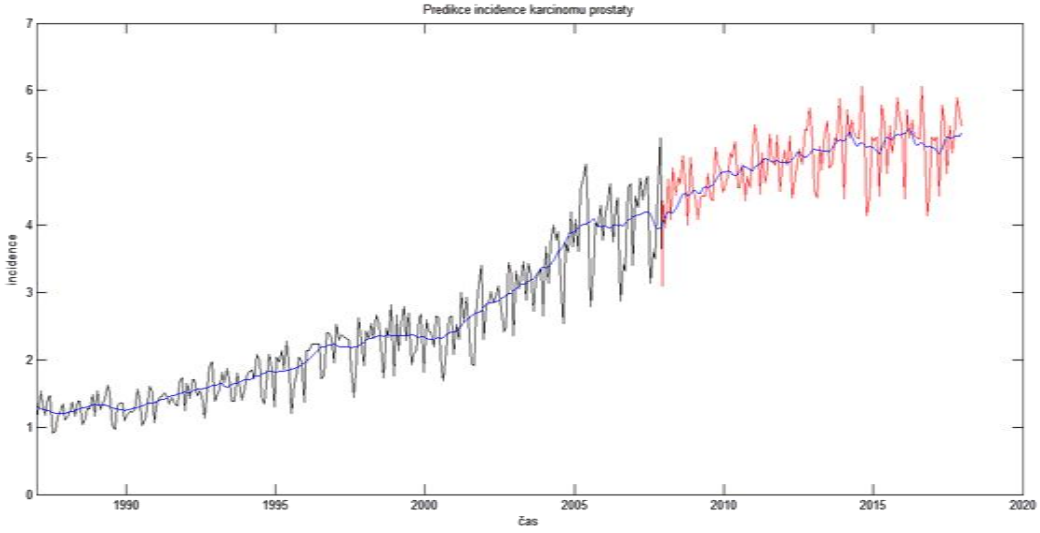
Na druhé straně většina mužů umírajících ve vyšším věku na nejrůznější onemocnění má menší nebo větší ložisko karcinomu prostaty, které jim nedělá žádné obtíže, ačkoli nejsou léčeni. Jinak řečeno, většina mužů umírá s karcinomem prostaty, nikoli na karcinom prostaty. V současné době však většinou není možno včas rozhodnout, který karcinom poroste a bude dělat pacientovi obtíže, a který karcinom bude bezvýznamný. Není proto zejména u starších mužů možno říci, zda časná detekce karcinomu prostaty zvýší konkrétnímu pacientovi šanci na prodloužení života odstraněním karcinomu prostaty. Je třeba konstatovat, že karcinom prostaty je druhou nejčastější příčinou smrti na nádory u mužů a že u všech těchto mužů byl někdy jejich karcinom malý, omezený jen na prostatu a tudíž vyléčitelný. (Jarolím L., 2012)



Obr. 1.1 Ilustrační obrázek, Věková struktura populace pacientů s karcinomem prostaty v ČR v letech 1977-2010; (http://www.swod.cz, 2012)



Obr. 1.2 Ilustrační obrázek, Incidence karcinomu prostaty a mortalita v ČR v letech 1977-2010; (http://www.swod.cz, 2012)



Obr.1.3 Ilustrační obrázek, Predikce incidence karcinomu prostaty v ČR modelováno s využitím inverzních filtrů, pomocí časových řad (Luterová A., 2012)

1.1 Příznaky a léčba karcinomu prostaty

Nádorová onemocnění jsou obecně známá tím, že nejsou zpočátku bolestivá a za nález nádoru pak často může náhoda. V některých případech za objev karcinomu může vyšetření, které pacient podstupuje kvůli jiným obtížím, u karcinomu prostaty to může být například vyšetření související s operací či problémy s močovou trubicí, anebo je to právě preventivní vyšetření, které zachraňuje i životy.

U pokročilejšího stádia karcinomu prostaty patří mezi symptomy onemocnění potíže při močení, krev v moči nebo jiné potíže s ledvinami a močovým ústrojím. Jestliže karcinom pokročil do fáze, kdy metastazoval, pak se objevuje často bolestivost zad, kyčlí a končetin. Prostatický karcinom obvykle metastazuje právě do kostí, ale také do plic a jater. Metastáze v kostech se objevují u 5 % mužů, kterým byl nově diagnostikován karcinom prostaty a u 80 – 85 % mužů, kteří nádoru prostaty podlehli. K vyšetření kostních metastáz se používá scintigrafie skeletu. Metastáze se mohou objevit i v ledvinách a nadledvinách, ty zjišťujeme pomocí transabdominální ultrasonografie. Další potíže, které se mohou vyskytovat, jsou podobné jako u většiny nádorových onemocnění, patří sem nechutenství, únava a celková slabost jedince.

Nejběžnějším prvním krokem při vyšetření u lékaře, je vyšetření per rectum. Lékař tak zjistí, zda je prostata zvětšená. Následně je dobré udělat krevní testy a stanovit hladinu PSA (prostatického specifického antigenu) v séru. Zvýšena hladina se může vyskytovat i u zánětů prostaty či benigních nádorů. Prostata je pouze zvětšená a nejedná se o zhoubný karcinom prostaty. Pro vyloučení karcinomu při vysokých hodnotách PSA se používá magnetická rezonance, či odběr tkáně, která se pošle k histologii (biopsie prostaty). Biopsie prostaty se používá jako doplňující vyšetření při podezření na karcinom prostaty při vyšetření per rectum a vyšších hodnotách PSA či PSA velocitě. Pokud je třeba, provadí se rebiopsie a to v odstupu třech až šesti měsíců.

Léčba karcinomu prostaty závisí na celkovém zdravotním stavu jedince. Často bývá používána radioterapie. Jestliže se karcinom rozšířil, pak se nejprve operativně vyjmou nádory z uzlin a míst, kde byly nádory lokalizovány a poté komise lékařů určí další léčebný postup.

Na některých pracovištích se dělá i brachyradioterapie (ozařování „zevnitř“, kdy se pomocí speciálních instrumentů zavede zářič do blízkosti prostaty, paprsky tedy nejdou přes kůži, nemají tolik vedlejších účinků a jejich léčebný efekt je vyšší). Výsledky ozařování nádoru jsou srovnatelné s chirurgickou léčbou. (URL2)

U karcinomu prostaty lze podobně jako u nádorového onemocnění prsu u žen, použít hormonální léčbu. Radikálnějším řešením je vyjmutí varlat (orchiektomie). Tato operace není náročná a výhodou je, že se v mužském těle přestane tvořit testosteron, který ovlivňuje růst karcinomu. Vliv testosteronu se může potlačit i pomocí inhibujících léků. Poslední možností je chemoterapie, která se užívá při léčbě různých nádorových onemocnění a je pro organismus vysilující. Současně je dobré používat při léčbě doplňky stravy posilující imunitu, která je u nemocného oslabena.

1.2 Varianty karcinomu prostaty

Prostata nebo také předstojná žláza se řadí k mužským pohlavním žlázám. Vylučuje sekret, který je odpovědný za pohyblivost spermatu. Normální prostata má hmotnost 15 - 20g, velikost a tvar jako kaštan či mandarinka (délka asi 3,3 cm, výška 2,4 cm a šířka 3,9 - 5,3 cm). Objem prostaty se pak nachází v rozmezí 12 - 27 ml. Muži středního věku jsou častými pacienty se změnou prostaty. Nemusí se ale nutně jednat o postižení karcinomem, změnu může vyvolat i zánět nebo se zde může nacházet cysta.

Nádory často vznikají ve více ohniskách. V periferní (okrajové) zóně se vyskytuje okolo 70 – 80 % všech karcinomů prostaty. Adenokarcinom patří mezi nádory, jejichž růst je závislý na hormonech a prostata je jedním z cílových orgánů androgenů (mužských pohlavních hormonů). Prostata roste celý život a její zvětšování je závislé právě na mužských pohlavních hormonech. Testosteron se naváže na receptory buňky a mění se v účinnější formu. Nejvíce receptorů se nachází v epiteliálních buňkách prostaty, a proto je zde i nejčastější nález karcinomu.

Dobře diferencované karcinomy (označení G1) mají volnější progresi oproti málo diferencovaným (označení v systému je až G5), které častěji metastazují.

Nejčastějším histologickým nálezem je adenokarcinom s různým stupněm diferenciace buněk. Adenokarcinom tvoří více než 95% maligních nádorů prostaty. Adenokarcinom prostaty vzniká z epiteliálních buněk prostatických acinů (acinární karcinom) nebo vzácněji ve velkých periuretrálních prostatických vývodech (duktální karcinom). Mezi další vzácné varianty karcinomu prostaty patří např. acinózní karcinom, malobuněčný karcinom, karcinom z prsténčitých buněk, adenoidně bazocelulární karcinom, sarkomatoidní karcinom či karcinom z přechodného epitelu. (Lukeš M., 2013)

2 Data a jejich popis

Pro konkrétní výsledky jakékoliv analýzy používáme různé datové soubory. Data definujeme jako číselný nebo slovní záznam studovaného objektu, který musí být smysluplný a musí souviset s problematikou, kterou chceme dále popisovat či studovat.

V každém datovém souboru se nacházejí nepřesnosti a chyby v měření. Tyto chyby mohou být snadno odstranitelné nebo nepodstatné pro danou studii. Některé chyby jsou ale neodstranitelné a mohly by studii zkreslovat, před analýzou je tedy třeba ze souboru tyto chyby odstranit či minimalizovat.

2.1. Typy dat

Data můžeme dělit podle jejich vlastností na kvalitativní a kvantitativní.

Kvalitativní neboli kategoriální data můžeme řadit do kategorií, ale nemůžeme jim přiřadit konkrétní číselnou hodnotu. Dále je můžeme dělit na data binární, nominální, ordinální. Binární data nabývají pouze dvou hodnot, často jsou to data obsahující odpověď ano a ne (např. karcinom diagnostikován / nediagnostikován). Nominální data můžeme roztřídit do více kategorií, ale tyto kategorie dále nelze seřadit (krevní skupina A, B, AB, 0 – nemůžeme říci, která je lepší, větší či menší a podobně). Ordinální data se od nominálních liší tím, že je můžeme seřadit podle nějakého kritéria.

Kvantitativní data můžeme vyjádřit konkrétní číselnou hodnotou. Tato data dále dělíme na spojitá a diskrétní. Spojitá data mohou nabývat jakýchkoliv hodnot v určitém intervalu (např. výška, hmotnost apod.). Data diskrétní mohou naopak nabývat pouze spočetně mnoha hodnot (např. počet dětí v rodině).

2.2 Datový soubor

Datový soubor zahrnuje záznamy o karcinomu prostaty z Fakultní nemocnice Olomouc z let 2006 -2012. Velikost toho souboru čítá 2570 případů na 2024 pacientů. To znamená, že jeden pacient může být zanesen v souboru vícekrát, a to z důvodu, že se u něj karcinom opět vyskytl nebo byl poslán na preventivní rebiopsii či krevní vyšetření.

Každý z pacientů je zde pod identifikačním číslem. V záznamu je uvedeno datum, kdy byly jednotlivé hodnoty vyšetření zapsány. Dále jsou zde informace o věku pacienta, rodinné anamnéze, hladině PSA, fPSA, index hodnotě, vyšetření per rectum, objemu prostaty (volum), objemu tranzitorní zóny (volumTZ) a výsledcích biopsie, pokud byla u pacienta provedena.

Seznam jednotlivých proměnných:

RA – anamnéza v rodokmenu: 0 – žádný příbuzný postižený karcinomem
1 – vzdálený příbuzný postižen (např. dědeček, babička)
2 – postižen blízký příbuzný (např. otec, bratr)
– Jestliže má jedinec blízkého příbuzného postiženého karcinomem prostaty, riziko výskytu nemoci u něj se minimálně dvakrát zvýší, při onemocnění dvou a více příbuzných se riziko zvyšuje dokonce pětkrát až jedenáctkrát.

PSA – krevní hodnoty prostatického specifického antigenu; hodnoty nad 4 ng/ml jsou podezřelé, ale záleží i na věku pacienta. Hodnoty zaznamenané v souboru jsou v rozmezí 0,01 – 5857 ng/ml.

fPSA – složka PSA – volný glykoprotein, nevázaný na sérový protein; hodnoty v souboru jsou v rozmezí 0 - 22,74

index – poměr volného a celkového PSA
indikátor k provedení biopsie; pokud je index vyšší než 0,20 biopsie se neprovádí

pr – vyšetření per rectum: 0 – norma, čím více, tím vyšší podezření, nejvyšší hodnota v datovém souboru - 8

volum – objem prostaty; normální objem prostaty se nachází v rozmezí 12-37 ml, hodnoty zaznamenané v souboru: 8 – 278 ml

volum TZ – tranzitorní zóna prostaty; měla by tvořit 2 - 5 % předstojné žlázy, hodnoty zaznamenané v souboru: 0 – 160 ml, tvoří až 94 % prostaty

y – 0 – biopsií nediagnostikovaný karcinom
1– pozitivní nález

Z jednotlivých údajů jde pomocí jednoduchého vzorce určit PSA denzita (PSAD), která je u pacientů s karcinomem prostaty vyšší a hraniční hodnota je 0,15. Výpočet je definován jako poměr celkové hladiny PSA v séru a celkového objemu prostaty:

$$PSAD = \frac{PSA\ (\frac{ng}{ml})}{volum\ (cm^3)}\ . \tag{2.1}$$

Podobně můžeme určit i PSAD_TZ denzitu přechodné zóny, kde je za hraniční hodnotu doporučována hodnota 0,35. Jde o poměr celkové hladiny PSA v séru a objemu tranzitorní zóny prostaty:

$$PSAD_{TZ} = \frac{PSA\ \left(\frac{ng}{ml}\right)}{volumTZ\ (cm^3 = ml)}\ . \tag{2.2}$$

PSAD a PSAD_TZ mohou pomoci při rozlišování benigního onemocnění prostaty a karcinomu. Nejdůležitější jsou pak tyto hodnoty pro tzv. šedou zónu, kdy hodnoty PSA v séru se nachází v rozmezí 4 - 10 ng/ml. Datový soubor byl o tyto parametry doplněn.

Další podobnou pomůckou při rozlišování karcinomů může být PSA velocita (PSAV), která udává vzestup hladiny PSA v séru za určitý čas (maximální hodnota je 0,75 ng/ml za rok):

$$PSAV = \frac{PSA_1 - PSA_2\ (\frac{ng}{ml})}{čas\ (rok)}\ . \tag{2.3}$$

Často se používá i hodnota PSA doubling time (PSADT), který udává čas, za který se hladina PSA v séru zdvojnásobí.

2.3 Prostatický specifický antigen PSA

Prostatický specifický antigen byl objeven v roce 1979 (Wang a kol. – gelová elektroforéza). Ve své molekulární podstatě jde o jedno-řetězový glykoprotein s 237 aminokyselinami, který produkují epiteliální buňky prostaty (jak zdravé tkáně tak i postižené karcinomem). Vysoká hladina antigenu není pro tělo nijak nebezpečná, ale signalizuje problém, který se týká prostaty. Hodnoty hladiny antigenu se udávají v nanogramech na mililitr a získávají se ze vzorku krve pacienta.

Zvýšenou hladinu celkového PSA v séru můžeme pozorovat u karcinomu prostaty, avšak i u jiných onemocnění, např. benigní hyperplazie prostaty (BHP), zánětu prostaty, při akutní retenci moče, po některých urologických manipulacích, ale též po pohlavním styku. Po biopsii prostaty je nutné počkat na objektivní výsledek přibližně 6 týdnů. Vyšší přítomnost PSA zřejmě souvisí s porušením bazální membrány epitelu prostatických buněk a kontaktem obsahu prostatických tubulů s krevním řečištěm. (Lukeš M., 2013)

Benigní hyperplazie prostaty je běžné nenádorové zvětšení prostaty, které se vyskytuje většinou u mužů po padesáti letech. Třetina mužů s hyperplazií má hladinu celkového PSA v séru až do 10ng/ml.Vysoké hodnoty PSA, které nesignalizují karcinom prostaty, poukazují na tzv. falešnou pozitivitu. Naopak falešná negativita znamená, že u pacienta s karcinomem prostaty jsou hodnoty PSA v optimu.

Hladina PSA je závislá na věku jedince. Pro muže věku 40 - 49 let je za normální považována hladina nižší než 2,5 ng/ml. Ve věku 50 - 59 let se tato hraniční hladina zvyšuje na 3,5 ng/ml a pro věk 70 - 79 let dokonce na 6,5 ng/ml. Vyšetření krve je doprovázeno vyšetřením prostaty konečníkem a výsledky jsou pak posuzovány lékařem společně.

vyšetření prostaty konečníkem	**hladina PSA**		
	0-2,5	*2,5-10*	*10 a více*
normální	nízké	střední	vysoké
nenormální	střední	vysoké	vysoké

Tab. 1 Riziko karcinomu prostaty závislé na vyšetření konečníkem a hladiny celkového PSA v séru (Jarolím L., 2012)

2.4 Čištění datového souboru

Z původního počtu 2024 pacientů je do výzkumu zahrnuto maximálně 1986 pacientů (data1) a to z důvodu chybějících klíčových proměnných či jejich chybného zadání. Zbytek souboru byl zkontrolován a případně opraven.

Mezi klíčové proměnné byly zahrnuty hodnoty PSA a věk pacienta a datum vyšetření. Mezi některými záznamy data vyšetření a věkem pacienta byly nalezeny nesrovnalosti, konkrétně u 43 záznamů. Tyto nesrovnalosti jsme se snažili minimalizovat tak, že jsme u daného jedince zprůměrovali věk a následně jej dopočetli podle data vyšetření (např. v roce 2010 měl pacient 46 let a v roce 2011 57 let, nevíme který z údajů je špatně, proto byl věk pomocí průměru opraven na 51 a 52 let). Nález biopsie nebyl zaznamenán u 478 případů. Biopsie buď nebyla vůbec provedena a nebo nebyl zaznamenán výsledek do datového souboru. Navíc u více než poloviny pacientů neexistuje žádný druhý záznam o jejich následující kontrole. Index a hodnota fPSA nebyly vyplněny a nešly doplnit z jiných údajů u 1104 případů. Objem prostaty nebyl zaznamenán u 57 případů. Z důvodu velkého počtu chyb do klíčových proměnných není zahrnuta ani rodinná anamnéza. V rodinné anamnéze byly zaneseny chybné údaje o blízkých příbuzných postižených karcinomem. Chybně zanesených je 402 případů z celkového počtu 2570. Pro příklad jedinec uvedl již v roce 2010, že má blízkého příbuzného postiženého karcinomem, ale v roce 2011, že nemá žádného takového příbuzného. Takové informace nemůžeme vyhodnotit, protože nevíme, která z nich je správná. Vyšetření per rectum není příliš spolehlivé, může být zaneseno značnou chybou (subjektivní hodnocení lékaře, který jej prováděl) a navíc nebylo zaznamenáno u 777 případů, a proto jej také nepočítáme mezi klíčové proměnné. Tranzitorní objem prostaty nebyl zanesen u 539 případů.

V datovém souboru byly vypočteny hodnoty fPSA nebo index, podle vzájemného vztahu hodnot PSA, fPSA a indexu:

$$index = \frac{fPSA}{PSA} \ . \tag{2.4}$$

Lze vynásobit 100 a získat tak procentuelně vyjádřený index. U jednoho pacienta byla opravena hodnota indexu 226,2570 tak, že byla přepočítána s platnou fPSA, byla nalezena chyba v desetinné čárce. Hodna fPSA byla vyšší než hodnota celkové hladina PSA v séru a index uvedený v procentech nemůže nabývat hodnoty kolem 226 %. Nová hodnota indexu je tedy 22,6 %. Podobně byla opravena hodnota fPSA ještě u dvou pacientů, kteří ji měli také vyšší než celkovou hladinu PSA, bylo tedy zřejmé, že šlo o chybu v desetinné čárce. Hodnoty indexů u všech případů byly přepočítány podle vzorce (2.4) a případné chyby byly opraveny.

Vysoké hodnoty PSA (udáváno i 5857 ng/ml) jsou brány jako vysoce nepravděpodobné a tedy chybně zadané.

2.5 Testové statistiky

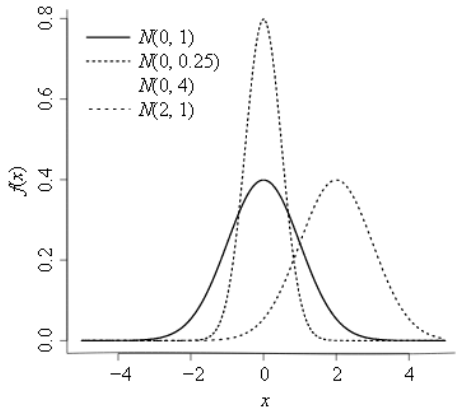
Pro popis datového souboru budeme využívat různých testových statistik. Budeme testovat především normalitu dat a korelaci jednotlivých proměnných. Normální rozdělení bývá častým předpokladem základních testů a modelů.

Normální rozdělení je spojité rozdělení pravděpodobnosti, které popisuje celou řadu veličin, jejichž hodnoty se symetricky shlukují kolem střední hodnoty a vytvářejí tak charakteristický tvar hustoty pravděpodobnosti, která je známá také pod pojmem Gaussova křivka. (Pavlík T., Dušek L., 2012)

Normální rozdělení pravděpodobnosti je zcela popsáno dvěma parametry, které jsou standardně označovány jako μ a σ^2, kdy první z nich představuje střední hodnotu normálního rozdělení a druhý představuje rozptyl normálního rozdělení. Fakt, že náhodná veličina X má normální rozdělení pravděpodobnosti se střední hodnotou μ a rozptylem σ^2, zapisujeme jako $X \sim N(\mu, \sigma^2)$. Hustota náhodné veličiny X pak má následující tvar:

$$f(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad . \tag{2.5}$$

Ukázky hustot náhodných veličin s normálním rozdělením pro různé hodnoty parametrů μ a σ^2 jsou uvedeny na Obr. 2.1. (Pavlík T., Dušek L.,2012)



Obr.2.1 Ukázky hustot náhodných veličin s normálním rozdělením. (Pavlík T., Dušek L.,2012)

Pro testování normality se používá i Kolmogorovův-Smirnonův test, který srovnává výběrové distribuční funkce s teoretickou distribuční funkcí odpovídající normálnímu rozdělení (ale může testovat i shodu s jiným rozdělením). Hodnotí maximální vzdálenost mezi dvěma funkcemi. Test může být jedno-výběrový nebo dvou-výběrový. Modifikací tohoto testu je Lillieforse test, který je určený přímo k hodnocení shody s normálním rozdělením.

Výsledky testování bývají často vyjádřeny pomocí p-hodnoty. Ta vyjadřuje pravděpodobnost za platnosti nulové hypotézy, s níž bychom získali stejnou nebo méně pravděpodobnou (extrémnější) hodnotu testové statistiky. Zde nulová hypotéza odpovídá tvrzení, že rozdělení je shodné s normálním rozdělením pravděpodobnosti. Čím nižší je p-hodnota, tím menší je pravděpodobnost, že platí nulová hypotéza. Za hladinu významnosti, při níž zamítáme nulovou hypotézu, bývá často považována hranice 5 % nebo 1 %.

Korelační analýzou zjišťujeme vztahy mezi jednotlivými proměnnými (náhodnými veličinami). Pro popis těchto vztahů používáme Pearsonův korelační koeficient:

$$R(X,Y) = \frac{E((X-EX)(Y-EY))}{\sqrt{DX}\,\sqrt{DY}}\,, \tag{2.6}$$

kde X a Y jsou náhodné veličiny, EX vyjadřuje střední hodnotu a DX rozptyl (variance). Podobně jsou definovány EY a DY.

Pearsonův korelační koeficient nabývá hodnot z intervalu $\langle -1,1 \rangle$, kde hodnoty blízké 0 značí nekorelovanost (či velice nízkou korelaci) proměnných a naopak hodnoty blízké 1 a -1 vysokou korelaci. Kladnou hodnotu koeficientu získáme tehdy, když vyšší hodnoty proměnné X souvisí s vyššími hodnotami proměnné Y. Naopak záporný koeficient získáme, když nižší hodnoty proměnné X nějak souvisí s vyššími hodnotami proměnné Y. Pearsonův korelační koeficient odraží pouze lineární závislost.

Pro nelineární závislost používáme k hodnocení Spearmanův korlační koeficient. Jde o neparametrickou metodu, která je odolná vůči odlehlým hodnotám a odchylkám od normality. Spearmanův korelační koeficient vypočítáme podle vzorce:

$$r_s = \frac{\sum_{i=1}^{n} x_{ri} y_{ri} - n \bar{x}_r \bar{y}_r}{(n-1) s_{x_r} s_{y_r}} \ , \tag{2.7}$$

kde x_{ri} je pořadí hodnoty x_i v rámci vzestupně uspořádaných hodnot x_1, ..., x_n, podobně vyjadřuje pořadí y_{ri}. Čísla $\bar{x}_r$ a $\bar{y}_r$ jsou průměry hodnot x_{ri} a y_{ri} (vyjadřují průměrná pořadí) a s_{x_r} a s_{y_r} představují směrodatné odchylky od $\bar{x}_r$ a $\bar{y}_r$.

Podobně jako Pearsonův korelační koeficient muže i r_s nabývat hodnot z intervalu $\langle -1,1 \rangle$. Pokud koeficient nabývá hodnot blízkých nebo rovných nule, pak mezi sledovanými proměnnými není žádný monotónní vztah nebo je minimální. Jestliže koeficient nabývá hodnot -1 a 1, pak mezi nimi existuje monotónní vztah. Spearmanův korelační koeficient je možné používat i pro diskrétní veličiny s ordinálními hodnotami.

Výpočetní alternativnou ke vzorci (2.7) je výpočet založený na diferencích pořadí pozorovaných hodnot, které definujeme následovně:

$$d_i = x_{ri} - y_{ri} \ . \tag{2.8}$$

Hodnotu Spearmanova korelačního koeficientu pak odhadneme pomocí vztahu

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i}{n(n^2-1)} \ . \tag{2.9}$$

Tento výpočet r_s platí přesně pouze pro neopakovaná pozorování, což znamená, že je citlivý na opakující se hodnoty, které vedou k průměrování pořadí. Vyskytuje-li se mezi hodnotami $x_1, ..., x_n$ respektive $y_1, ..., y_n$, množství shodných hodnot, je vhodnější použít k výpočtu Spearmanova korelačního koeficientu definiční vztah (2.7). (Pavlík T., Dušek L.,2012)

V našem datovém souboru nemají proměnné normální rozdělení, proto budeme používat k výpočtu korelací Spearmanův korelační koeficient. Hypotézu o nekorelovanosti proměnných testujeme pomocí Fischerovy z-transformace (2.10) a výpočtu z-skóre (2.11) pro Spearmanův korelační koeficient r_s. Z-skóre vyjadřuje číselné hodnoty pro standardní normální rozdělení, tedy aby výsledná čísla po transformaci měla průměr 0 a směrodatnou odchylku 1. Na základě těchto hodnot jsme schopni vypočítat p-hodnotu. Pomocí p-hodnoty (2.12) určíme na zvolené hladině významnosti ($\alpha = 0{,}05$), zda zamítáme ($p < \alpha$) nebo nezamítáme ($p > \alpha$) nulovou hypotézu, která odpovídá tvrzení, že proměnné nejsou korelované.

Výpočet Fischerovy z-transformace, kde r_s představuje Spearmanův korelační koeficient:

$$F(r) = \frac{1}{2} ln \frac{1+r_s}{1-r_s}. \tag{2.10}$$

Z-skóre pro Fischerovu transformaci:

$$z = \sqrt{\frac{n-3}{1{,}06}} F(r), \tag{2.11}$$

kde F(r) představuje odpovídající Fischerovu transformaci a n představuje počet vzorků v datovém souboru, pro něž byla korelace počítána.

P- hodnotu p zjistíme pomocí z-skóre a $P(X \leq z)$ označující hodnotu distribuční funkce standardizovaného normálního rozdělení v bodě z, kde $X \sim N(0,1)$.

Vzorec pro výpočet p-hodnoty:

$$p = 2 \times \left(1 - P(X \leq z)\right) \ . \qquad (2.12)$$

V následující sekci 2.6 z ilustračních důvodů provedeme výpočet z-transformace a následně p-hodnoty pro vybrané proměnné. Ostatní proměnné budou realizovány pomocí statistického softwaru.

2.6 Popisná statistika souboru

Byly připraveny celkem 4 datové soubory, na kterých bude použita metoda klasifikačních stromů. Jednotlivé soubory obsahují různě „ořezané“ klíčové proměnné o odlehlé hodnoty. Uvidíme tak, jak ovlivňují odlehlé hodnoty některých vysvětlujících proměnných výsledný strom.

V prvním ze souborů (data1) byly odstraněny pouze záznamy pacientů s chybějícími klíčovými proměnnými PSA a neopravitelným a nezjistitelným datem vyšetření. Pouze u dvou z těchto pacientů tento chybějící záznam neměl vliv na jejich úplné odstranění ze souboru, jelikož se jednalo o třetí či čtvrtou rebiopsii, u nichž nebyly klíčové proměnné zaznamenány a ty zbylé byly do studie započteny. U jednoho ze třech pacientů se špatně zadaným datem vyšetření šlo toto datum doplnit ze zbylých údajů (věk a rebiopsie). Dále byl odstraněn záznam pacienta, u něhož byla zaznamenána rebiopsie tentýž den s odlišnými hodnotami pro všechny proměnné. Tímto promazáním nám v souboru zůstalo 1986 z 2024 pacientů, což je asi 98 % původního datového souboru. Nás budou zajímat především první případy a první rebiopsie. Datový soubor pak zahrnuje 92, 87 % původního souboru. Tabulka Tab.9 na konci sekce ukazuje zastoupení prvních případů a rebiopsií v připravených souborech.

Bylo zjištěno, že soubor obsahuje pacienty s průměrným věkem 63,43 let, s průměrnou hodnotou PSA 15,4 ng/ml, objemem prostaty 49,5 cm^3 a průměrným objemem tranzitorní zóny 28,3 cm^3. Z grafů na Obr.2.2, Obr. 2.3, Obr. 2.4 a z tabulky Tab. 2 můžeme vidět, že

data nejsou normálně rozdělené. Na krabicových grafech Obr. 2.5, Obr. 2.6 můžeme pozorovat průměrné i odlehlé hodnoty proměnných a nenormální rozdělení souboru. Krabicové grafy zvlášť vytvořené pro první případy a první rebiopsie, které poukazují na nenormalitu rozdělení, jsou k vidění v příloze.

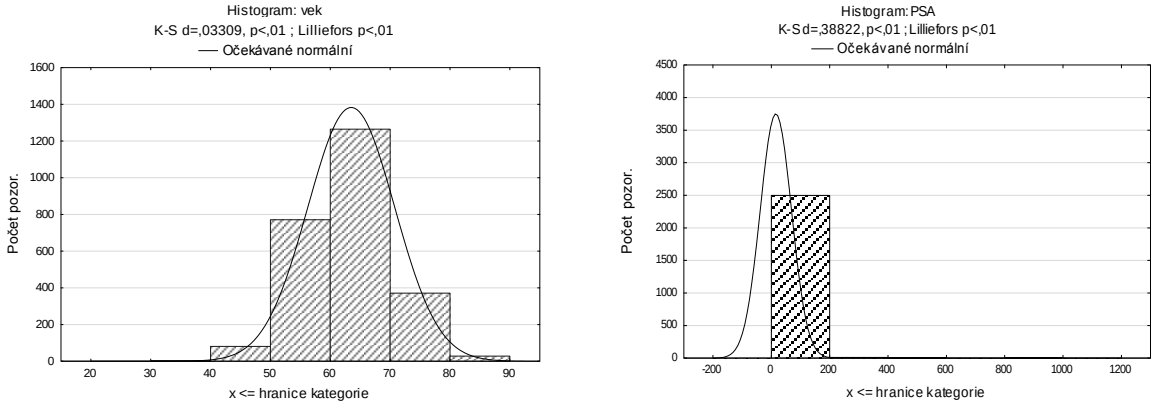
Průměrný věk pacientů 63,41 (zahrnuje pouze pacienty, kteří přišli do nemocnice s prvními problémy s prostatou) potvrzuje studie, které uvádí rozmezí postižených mužů 65 - 80 let. Dokonce o něco nižší průměrný věk pacientů může poukazovat na možnost včasného záchytu možné nemoci.

Průměrné hodnoty v souboru:

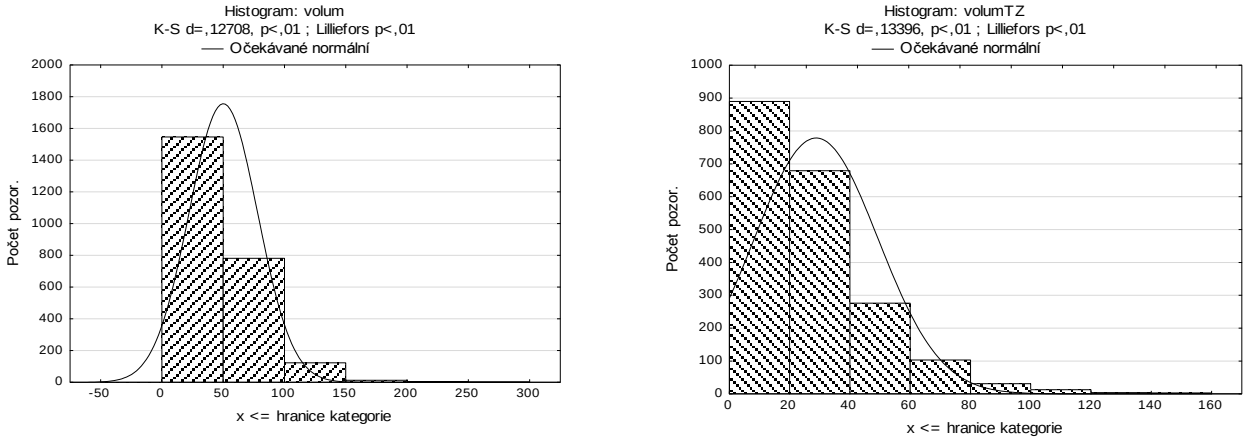
Proměnná	jednotka	hodnota	Směrodatná odchylka	Průměrná hodnota, pouze první záznam pacientů
Věk	rok	63,4	7,3	63,41
index	%	16,7	8,7	16,89
Volum_TZ	cm^3	28,3	20,1	27,50
Volum	cm^3	49,5	27,8	48,06
PSA	ng/ml	15,4	54,3	16,16
fPSA	ng/ml	1,13	1,05	1,12
PSA_V	ng/ml*rok	3,19	22,1	-
PSAD	ng/ml^2	0,35	1,5	0,37
PSAD_TZ	ng/ml^2	0,74	7,7	0,81

Tab.2 Průměrné hodnoty v datovém souboru data1 a jejich směrodatné odchylky

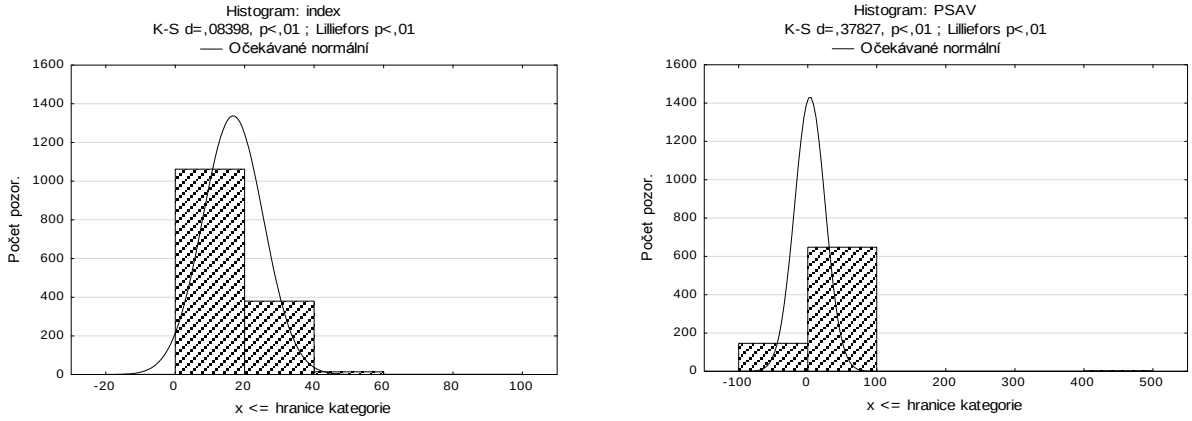
Z grafů na Obr.2.2, Obr. 2.3, Obr. 2.4 a z tabulky Tab. 2 můžeme vidět, že data nejsou normálně rozdělená. Kdyby rozdělení datového souboru bylo normální, pak by jednotlivé sloupce v histogramech kopírovaly křivku, která znázorňuje právě očekávanou normalitu v souboru. Výsledky potvrzují i Kolmogorovův-Smirnovův test normality a Lilliefors test normality. Na krabicových grafech Obr. 2.5, Obr. 2.6 můžeme také pozorovat průměrné i odlehlé hodnoty proměnných a nenormální rozdělení souboru.



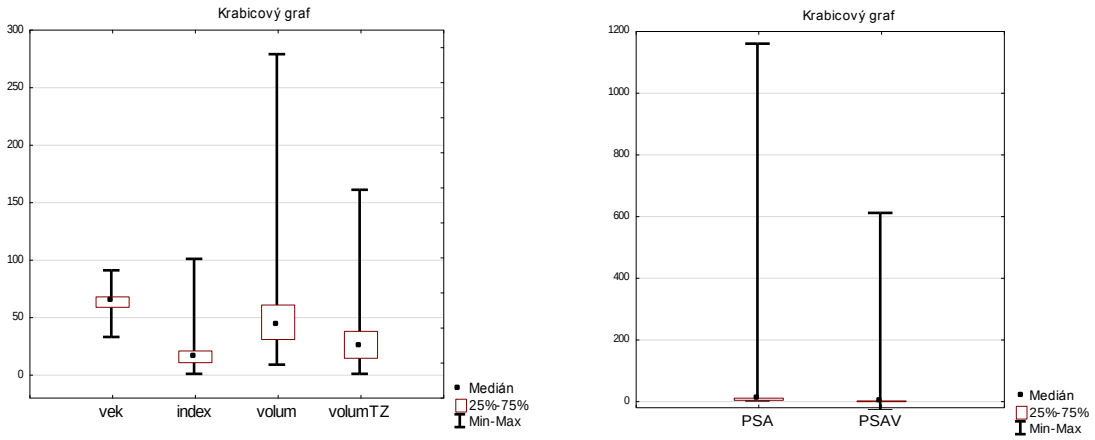
Obr.2.2 Histogramy rozdělení datového souboru postupně podle věku a hodnoty PSA v séru a jejich očekávané normální rozdělení.



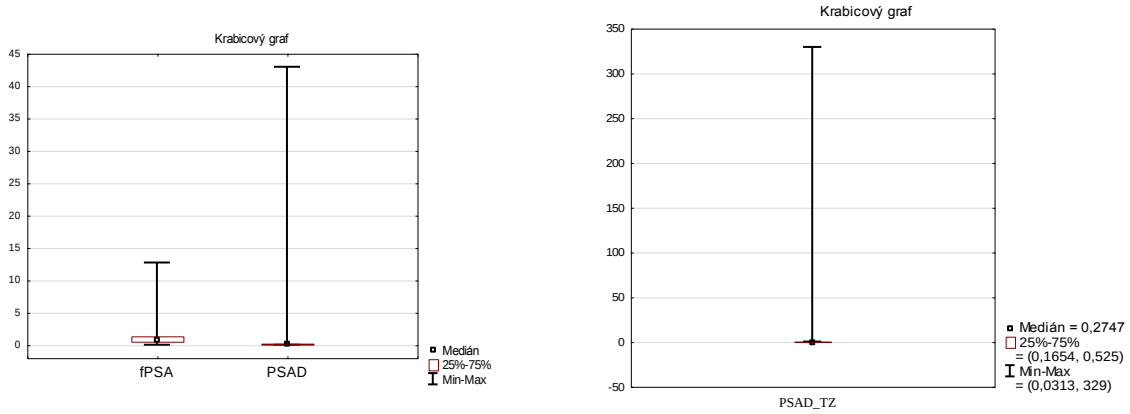
Obr. 2.3 Histogramy rozdělení datového souboru postupně podle objemu prostaty a objemu tranzitorní zóny a jejich očekávané normální rozdělení.



Obr. 2.4 Histogramy rozdělení datového souboru postupně podle indexu a PSA_V a jejich očekávané normální rozdělení.



Obr. 2.5 Boxploty jednotlivých vysvětlujících proměnných



Obr. 2.6 Boxploty jednotlivých vysvětlujících proměnných

Proměnná	Spearmanovy korelace (data1_1a2pripady) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000								
	vek	PSA	fPSA	PSAV	index	volum	volumTZ	PSAD	PSAD-TZ
vek	1,000	0,338	0,371	0,083	0,194	0,237	0,238	0,157	0,075
PSA	0,338	1,000	0,675	0,361	-0,131	0,260	0,245	0,745	0,576
fPSA	0,371	0,675	1,000	0,174	0,590	0,558	0,522	0,148	0,003
PSAV	0,083	0,361	0,174	1,000	-0,069	0,059	0,029	0,307	0,253
index	0,194	-0,131	0,590	-0,069	1,000	0,443	0,415	-0,469	-0,478
volum	0,237	0,260	0,558	0,059	0,443	1,000	0,935	-0,386	-0,532
volumTZ	0,238	0,245	0,522	0,029	0,415	0,935	1,000	-0,350	-0,596
PSAD	0,157	0,745	0,148	0,307	-0,469	-0,386	-0,350	1,000	0,923
PSAD-TZ	0,075	0,576	0,003	0,253	-0,478	-0,532	-0,596	0,923	1,000

Tab. 3 Spearmanův korelační koeficient pouze pro první případy a první rebiopsie

Z tabulky Tab.3 jsme schopni určit závislosti proměnných, kdy většina z vysvětlujících proměnných je vzájemně korelovaná (značeno červenou barvou). To znamená, že jedna z nich zahrnuje informaci druhé. Označené korelace jsou významné na hladině významnosti

α = 0,05. Nejsilnější korelaci pak můžeme pozorovat u hodnot objemu prostaty a objemu tranzitorní zóny, dále u hodnot PSAD tranzitorní zóny a PSAD nebo také PSA a PSA denzity. Zajímavá je korelace mezi PSA velocitou a věkem, kdy pravděpodobně s vyšším věkem roste i hodnota PSA velocity. Mezi nekorelované proměnné patří PSA velocita s objemem, objemem tranzitorní zóny a s hodnotou indexu. Dále PSAD tranzitorní zóny není korelované s hodnotou volného PSA. K výpočtu korelací je použitý Spearmanův korelační koeficient a je proveden na souboru pacientů, u kterých známe všechny z proměnných. V tabulce Tab.4 můžeme vidět, že v případě kdy nás zajímají pouze první případy, není v korelacích výrazná změna. V tabulce nemůžeme vidět PSA velocitu, která u prvních záznamů nejde vypočítat.

Hypotézu o nekorelovanosti proměnných testujeme pomocí Fischerovy z-transformace, jak bylo popsáno v kapitole 2.5. Pro ilustraci vybereme například korelační koeficient r_s = 0,194 pro index a věk, kde počet záznamů je roven n = 1423. Fischerovu transformaci $F(r) = \frac{1}{2} ln \frac{1,194}{0,806} = 0,196$ a z-skóre $z = \sqrt{\frac{1420}{1,06}} \times 0,196 = 7,17$ jsme vypočítali postupně podle vzorců 2.10 a 2.11. Pokud bude p-hodnota vyšší než hladina významnosti α = 0,05, pak nezamítáme nulovou hypotézu, která říká, že proměnné nejsou korelované.

$p = 2 \times (1 - P(X < 7,17) = 2 \times P(X < -7,17)$, kde X ~ N(0,1)

$p = 7,5 \times 10^{-13} \Rightarrow$ p < 0,05.

P-hodnota je menší než hladina významnosti α, proto zamítáme nulovou hypotézu (proměnné jsou nekorelované). Shodujeme se tedy s tabulkou Spearmanových korelací Tab.3, kde je označena vzájemná korelace těchto dvou proměnných červenou barvou (proměnné jsou korelované).

	Spearmanovy korelace (prvni_pripady_data1) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000							
Proměnná	**vek**	PSA	fPSA	index	volum	PSAD	volumTZ	PSAD-TZ
vek	1,000	0,338	0,378	0,196	0,249	0,150	0,251	0,068
PSA	0,338	1,000	0,671	-0,130	0,249	0,753	0,237	0,576
fPSA	0,378	0,671	1,000	0,593	0,551	0,141	0,519	-0,007
index	0,196	-0,130	0,593	1,000	0,438	-0,467	0,407	-0,471
volum	0,249	0,249	0,551	0,438	1,000	-0,384	0,936	-0,539
PSAD	0,150	0,753	0,141	-0,467	-0,384	1,000	-0,354	0,923
volumTZ	0,251	0,237	0,519	0,407	0,936	-0,354	1,000	-0,600
PSAD-TZ	0,068	0,576	-0,007	-0,471	-0,539	0,923	-0,600	1,000

Tab.4 Spearmanovy korelace pouze pro první případy

V druhém datovém souboru (data2) byly odstraněny ze souboru záznamy jedinců s hodnotami PSA v séru nad 300 ng/ml. V takto upraveném souboru nám zbylo 1973 pacientů z 2024, což je 97 % původního souboru. Ponechali jsme v souboru první a druhé záznamy pacientů a dostali jsme se na velikost 92,3 % původního souboru.

Rozdělení datového souboru není normální. Histogramy ověřující normalitu byly obdobné jako v předchozím případě Obr. 2.2. V tabulce Tab.5 můžeme vidět, že záznam PSA velocity je citlivý na odlehlá pozorování, kdy odstraněním 1 % souboru vznikl rozdíl v průměru o více než 1 jednotku. Korelace byly obdobné jako v prvním případě.

Průměrné hodnoty v souboru:

Proměnná	jednotka	Hodnota	Směrodatná odchylka
Věk	rok	63,4	7,3
Index	%	16,7	8,7
Volum_TZ	cm^3	28,3	20,1
Volum	cm^3	49,5	27,8
PSA	ng/ml	11,9	21,8
fPSA	ng/ml	1,13	1,05
PSA_V	ng/ml*rok	2,14	6,9
PSAD	ng/ml^2	0,27	0,5
PSAD_TZ	ng/ml^2	0,50	0,8

Tab.5 Průměrné hodnoty v datovém souboru data2 a jejich směrodatné odchylky

V souboru data3 byli navíc odstraněni ze souboru jedinci s hodnotami PSA nad 100 ng/ml. Nebyli úplně odstraněni pouze dva jedinci, u nichž se jednalo o třetí či čtvrtou rebiopsii. Dostali jsme 95 % souboru, což je 1934 pacientů z 2024. První záznamy a první rebiopsie tvoří 90,8 % původního datového souboru.

Rozdělení datového souboru není normální, histogramy rozdělení podle všech proměnných nevykazovaly normalitu a byly obdobné jako v prvním souboru. Podobně vypadala i tabulka korelací.

Průměrné hodnoty v souboru:

Proměnná	jednotka	Hodnota	Směrodatná odchylka
Věk	rok	63,3	7,2
Index	%	16,7	8,7
Volum_TZ	cm^3	28,2	20,1
Volum	cm^3	49,2	27,7
PSA	ng/ml	9,49	10,5
fPSA	ng/ml	1,13	1,05
PSA_V	ng/ml*rok	1,95	6,2
PSAD	ng/ml^2	0,23	0,3
PSAD_TZ	ng/ml^2	0,47	0,6

Tab.6 Průměrné hodnoty v datovém souboru data3 a jejich směrodatné odchylky

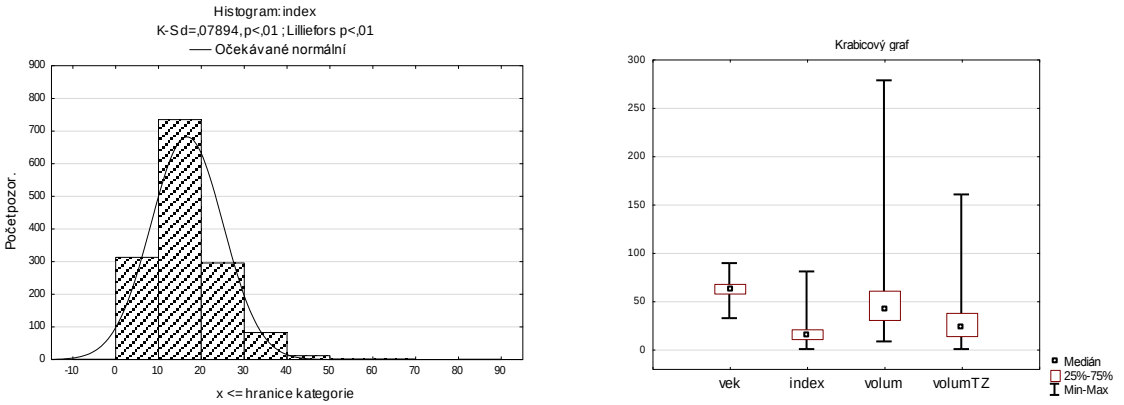
V posledním ze souborů (data4) byli odstraněni ze souboru dat pacienti s hodnotami PSA v séru nad 30 ng/ml, kromě 3 pacientů u kterých bychom ztratili důležitou informaci pro rebiopsii (rozdíly v hodnotách zde byly možné – hodnoty si byly blízké). Dále byli ze souboru smazáni ti, kterým chyběly hodnoty objemu prostaty a zároveň většina proměnných v souboru. Pacienti, kteří měli záznam PSA, chybějící objem prostaty, ale známé výsledky biopsie, zde byli ponecháni.

V takto upraveném souboru zbylo 1836 pacientů, což je necelých 91 % původního souboru. První a druhé případy tvoří pouze 86,4 % původního souboru.

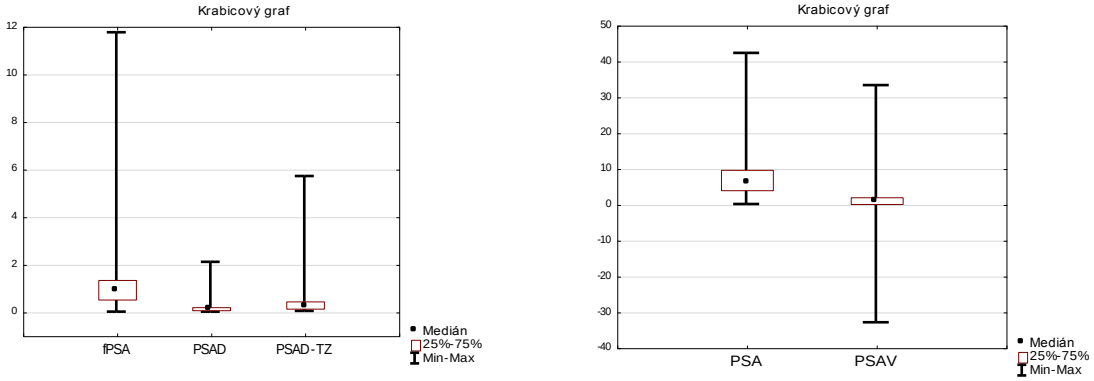
Průměrné hodnoty v souboru:

Proměnná	jednotka	Hodnota	Směrodatná odchylka
Věk	rok	63,1	7,04
Index	%	16,7	8,5
Volum_TZ	cm^3	28,2	20,1
Volum	cm^3	49,1	27,5
PSA	ng/ml	7,71	5,4
fPSA	ng/ml	1,09	0,9
PSA_V	ng/ml*rok	1,55	3,6
PSAD	ng/ml^2	0,19	0,2
PSAD_TZ	ng/ml^2	0,39	0,4

Tab.7 Průměrné hodnoty v datovém souboru data4 a jejich směrodatné odchylky



Obr. 2.7 Histogram rozdělení datového souboru podle indexu a jejich očekávané normální rozdělení, vpravo boxplot vysvětlujících proměnných (věk, index, volum, volumTZ)



Obr. 2.8 Krabicové grafy dalších vysvětlujících proměnných (PSA, PSAV, fPSA, PSAD a PSAD_TZ)

Na Obr. 2.7 můžeme vidět změnu v rozdělení souboru podle indexu v porovnání s Obr.2.4. Na krabicových grafech Obr. 2.7 a 2.8 můžeme pozorovat změny v souboru. Rozdělení datového souboru není normální, pro tento druh souboru použijeme neparametrické odhady. V tomto souboru nám jako v jediném vyšly navíc nekorelované proměnné věk a PSAD tranzitorní zóny.

	Spearmanovy korelace (data4_1a2pripady) ChD vynechány párově Označ. korelace jsou významné na hl. p <,05000								
Proměnná	**vek**	PSA	fPSA	PSAV	index	volum	PSAD	volumTZ	PSAD-TZ
vek	1,000	0,307	0,373	0,062	0,199	0,230	0,103	0,236	0,031
PSA	0,307	1,000	0,665	0,331	-0,129	0,268	0,691	0,255	0,513
fPSA	0,373	0,665	1,000	0,166	0,607	0,565	0,128	0,521	-0,014
PSAV	0,062	0,331	0,166	1,000	-0,076	0,021	0,276	0,001	0,218
index	0,199	-0,129	0,607	-0,076	1,000	0,450	-0,468	0,420	-0,474
volum	0,230	0,268	0,565	0,021	0,450	1,000	-0,458	0,934	-0,589
PSAD	0,103	0,691	0,128	0,276	-0,468	-0,458	1,000	-0,394	0,912
volumTZ	0,236	0,255	0,521	0,001	0,420	0,934	-0,394	1,000	-0,655
PSAD-TZ	0,031	0,513	-0,014	0,218	-0,474	-0,589	0,912	-0,655	1,000

Tab.8 Spearmanovy korelace pro datový soubor data4

Počet případů	První vyšetření	Rebiopsie	Celkem případů	Cekem pacientů
Data 1	1986	400	2386	1986
Data 2	1973	399	2372	1973
Data 3	1934	398	2332	1934
Data 4	1836	384	2220	1836

Tab.9 Zastoupení prvních vyšetření a prvních rebiopsií v rámci připravených souborů

Tabulka Tab.9 byla připravena pro přehlednost změn v počtech pacientů i případů v jednotlivých souborech z původních 2570 případů a 2024 pacientů.

3 Klasifikační a regresní stromy

Klasifikační a regresní stromy spolu úzce souvisí. Tato metoda je obdobou mnohonásobné regrese, kdy máme jednu vysvětlovanou proměnnou a několik vysvětlujících proměnných (prediktorů), které se ji snaží vysvětlit. Pomocí metody klasifikačních a regresních stromů najdeme klasifikátor, podle kterého můžeme v budoucnu předpovídat, do jaké třídy objekt budeme moci zařadit. V našem případě zjišťujeme, podle kterých symptomů můžeme pacienta zařadit do kategorie, kdy bude při biopsii (rebiopsii) potvrzen či nebude potvrzen nález karcinomu prostaty. Stromy využívají těchto klasifikátorů a s jejich pomocí odhadují budoucí nález. Pokud se jedná o kvantitativní (spojitou) vysvětlovanou proměnnou, počítá se regresní strom, pokud o kvalitativní (kategoriální závislost), jde o klasifikační strom. Naše vysvětlovaná proměnná (biopsie / rebiopsie - je nebo není pozitivní) je kvalitativní a budeme tedy používat klasifikační strom.

Výhodou klasifikačních a regresních stromů je, že nejsou kladeny vysoké nároky na tvar vysvětlujících proměnných a přitom dosahují přibližně stejné přesnosti, jako parametrické metody. Parametrické metody jsou ve statistice častěji užívané, jelikož jejich výsledky jsou lépe interpretovatelné. Jsou u nich kladeny vysoké požadavky na rozdělení dat v souboru, jedná se většinou o standardizovaný datový soubor, kde jsou jednotlivé proměnné nekorelované, aby o souboru vypovídaly co nejvíce. Korelované proměnné se snažíme ze souboru odstranit tím, že použijeme jen jednu z proměnných, která daný problém vysvětluje co nejlépe a nejpřesněji a zahrnuje ostatní korelované proměnné. V metodě klasifikačních a regresních stromů mohou být vysvětlující proměnné navzájem korelované, nejsou kladeny žádné podmínky na typ rozdělení, prediktory mohou být všech typů a algoritmy, podle kterých jsou stromy vytvořeny, nejsou náchylné na odlehlé hodnoty. Navíc poskytují přehledný a názorný model pro interpretaci výsledků. Naopak nevýhodu představuje jejich nestabilita, kdy stačí mírně pozměnit vstupní data nebo parametry a dostaneme odlišný strom. Nastala by změněna v klasifikaci, a proto musíme být opatrní, jak výsledky interpretujeme. Musíme tedy vybírat takové proměnné, které nám data rozdělí na co nejhomogennější skupiny. Často se

používá k přesnější interpretaci výsledků kombinace většího množství stromů, to minimalizuje jejich nestabilitu a variabilitu.

Velikost stromu nemá přímou souvislost s jeho kvalitou. Jestliže máme obsáhlý strom, může odpovídat pouze datům, na kterých byl strom sestaven, ale nepopisuje již všeobecně platné závislosti. Pokud bychom model použili na jiná data, nemusel by data dobře popsat. Naopak málo obsáhlý strom nemusí postihnout celou strukturu dat a nevypovídá příliš o závislostech mezi proměnnými.

3.1 Mnohonásobná regrese

Klasifikační stromy jsou obdobou mnohonásobné lineární regrese, která je účinnou metodou pro analýzu vztahů mezi závislou proměnnou a vysvětlujícími proměnnými. Touto metodou vysvětlujeme hodnoty závislé proměnné pomocí lineární kombinace několika vysvětlujících proměnných (dvou a víc):

$$EY = \ a + \ b_1 x_1 + \ b_2 x_2 + \cdots b_p x_p \ \ , \tag{3.1}$$

kde a je konstanta, b_1, b_2, ... ,b_p představují regresní koeficienty, které vysvětlují vliv jednotlivých vysvětlujících proměnných a x_1, x_2, ..., x_p jsou hodnoty vysvětlujících proměnných (prediktorů).

Mnohonásobná regrese má za úkol vysvětit co největší část variability. Využívá se koeficient determinace R^2, který popisuje, jak velkou část variability vysvětlované proměnné model pokrývá. Vysvětluje tedy rozptyl v závislé proměnné. Odhad regresních koeficientů spočívá v tom, že je kontrolováno působení ostatních proměnných vstupujících do modelu. Standardizované regresní koeficienty (β) vyjadřují sílu vlivu jednotlivých proměnných na vysvětlovanou proměnnou. Můžeme tak určit, které proměnné mají velký nebo malý vliv na rozptyl vysvětlované proměnné. Takto sestavenou regresní rovnicí můžeme odhadnout hodnoty vysvětlované proměnné pro jednotlivé případy.

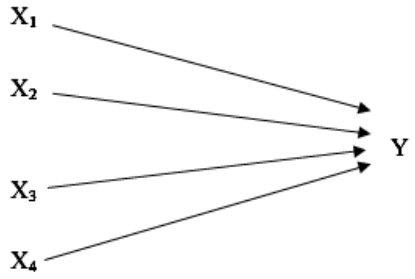
Pro použití mnohonásobné lineární regrese musí datový soubor splňovat určité předpoklady:

- Vysvětlovaná proměnná musí být metrická, jinak se používá logistická regrese.
- Vysvětlující proměnné jsou měřeny, stejně jako závislá proměnná, na intervalové úrovni (jsou metrické) nebo mohou být dichotomické. Jestliže není splněn tento předpoklad, existuje zde způsob, jak tuto podmínku „obejít".
- Vysvětlující proměnné by neměly být vzájemně korelované. Výsledky regrese jsou pak nespolehlivé. Vlivem korelace se může stát, že některý významný prediktor je vyřazen z modelu.
- Proměnné tedy musejí být i v lineárním vztahu a vzájemné korelace popisujeme Pearsonovým korelačním koeficientem. To znamená, že musejí být splněny i požadavky na normalitu dat. Normalita nemusí být splněna pouze v případě, kdy máme k dispozici velký výběrový soubor.
- Mezi proměnnými existuje homogenita rozptylu.
- Mnohonásobná regrese je citlivá na odlehlá pozorování.

(zkráceno od Rabušic L., 2004)

Pokud jsou splněny požadavky na datový soubor, můžeme použít mnohonásobnou regresi. Mnohonásobná regrese může mít různé formy:

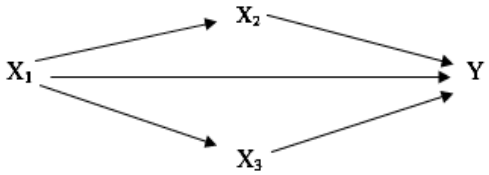
1.) Deskriptivní model mnohonásobné regrese



Mezi vysvětlujícími proměnnými nepředpokládáme žádnou strukturu vztahů. Sděluje nám pouze sílu vlivu jednotlivých proměnných na vysvětlovanou proměnnou, a jak velký podíl rozptylu závisle proměnné je jimi vysvětlen.

Obr. 3.1 Deskriptivní model mnohonásobné regrese (Rabušic L., 2004)

2.) Kauzální model mnohonásobné regrese



Model popisuje vliv vysvětlujících proměnných mezi sebou i vliv na vysvětlovanou (závislou) proměnnou.

Obr. 3.2 Kauzální model mnohonásobné regrese (Rabušic L., 2004)

Sestavení modelu, ať již jednoduchého deskriptivního nebo složitějšího kauzálního, vyžaduje vždy rozvahu o počtu proměnných, které necháme vstoupit do mnohonásobné regrese. Samotné adjektivum „mnohonásobná“ by mohlo nezkušeného analytika svádět k tomu, aby pracoval s co možná největším počtem proměnných – s vírou, že čím více proměnných do regrese zahrne, tím vyšší podíl rozptylu vysvětlí. To je samozřejmě špatný přístup. Ve vědě, stejně jako v životě, platí princip efektivity, tedy snaha dosáhnout s minimálními výstupy maximálně možného efektu. Do rovnice zahrnujeme pouze takové proměnné, o nichž víme z teorie nebo empirických zobecnění vyplívajících z analýzy jiných autorů, že jsou pro daný problém relevantní. (Rabušic L., 2004)

Metoda klasifikačních a regresních stromů využívá podobných principů k vysvětlení závislé proměnné proměnnými vysvětlujícími. Každý z prediktorů má jinou důležitost a podle ní jsou dále stanovovány podmínky větvení. Výhodou oproti mnohonásobné regresi je fakt, že nejsou přísné podmínky pro datový soubor, na němž můžeme pracovat. Právě kvůli těmto podmínkám není vhodné na náš datový soubor používat přímo mnohonásobnou regresi.

3.2 Klasifikační stromy

Klasifikační strom představuje model pro data, kde každé pozorování patří do některé z tříd $T_1,...,T_k$, $k \geq 2$. Současně je pozorování charakterizováno vektorem $x = (x_1,...,x_p)$ hodnot vysvětlujících proměnných (prediktorů) $X_1,...,X_p$, kde prediktory mohou být jak kvalitativní, tak kvantitativní. (Klashka J., Kotrč E., 2004)

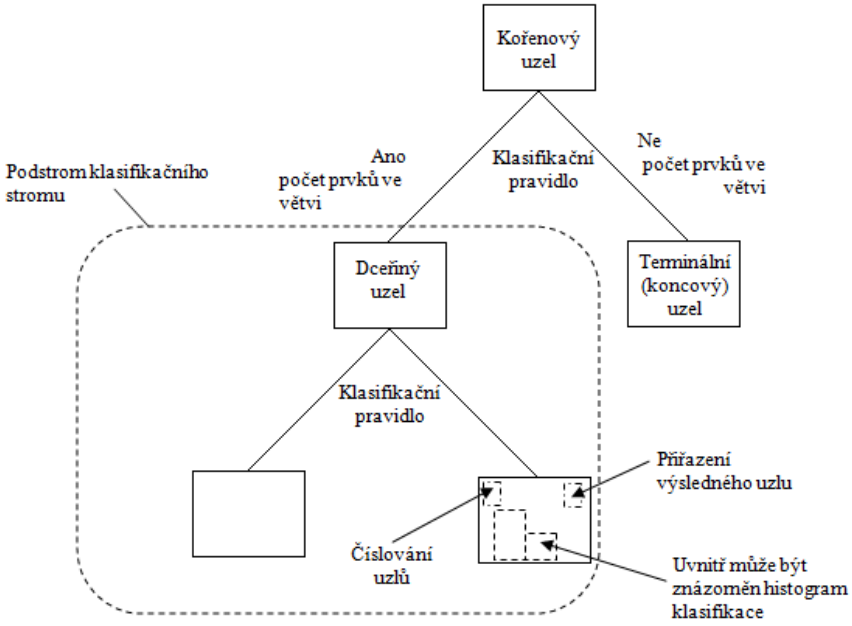
Máme tedy datový soubor v němž jednotlivá pozorování mají podobu $(X,Y) = (X_1, X_2, ..., X_p, Y)$. Model je znázorněn stromovým grafem, který je složen z uzlů a orientovaných hran (orientace nebývá vyznačena, hrana vede shora dolů). Uzly se dělí na kořenový uzel, ze kterého strom vychází, neterminální, dceřiné a terminální uzly.

Kořenový uzel je nejobsáhlejší, obsahuje všechna cvičná data. Z kořenového uzlu se může strom větvit do neterminálních či terminálních uzlů. V neterminálních uzlech se strom dále větví a hrany z něj vedou do uzlů dceřiných. Větvení závisí právě na prediktorech. Podle kriteriální statistiky probíhá výběr všech možností větvení. Kriteriální statistika zkoumá stejnorodost (homogennost) vzorků uvnitř možných dceřiných uzlů a zároveň nakolik jsou uzly odlišné. Nejlepší možné větvení (s maximální hodnotou kriteriální statistiky) pak vytvoří nový terminální uzel. Data z kořenového uzlu se tedy podle hodnot prediktorů rozdělí mezi nové dceřiné uzly a celý proces se opakuje, hledáme další větvení stromu. Nové větvení stromu se hledá tak dlouho, dokud je to přínosné (viz metody ukončení), poté proces končí. Další dělení by nepřineslo významné zlepšení odhadu. Můžeme také sestavit co největší strom T_{max}, který se následně „prořezává“ a odstraňují se nevýznamné uzly, podle předem zvolených mezí a odhadů skutečných chyb.

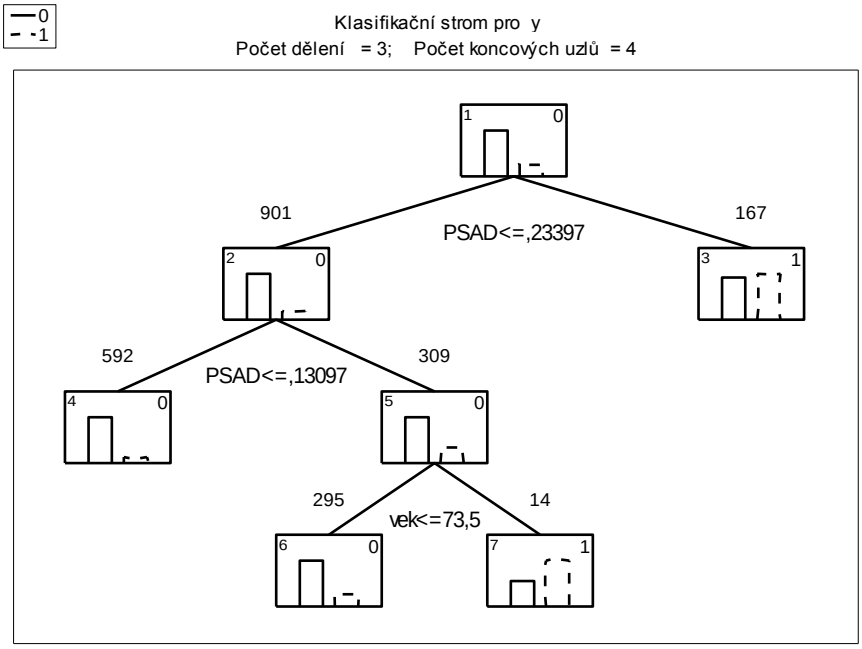
Většinou bývá na stromech binární větvení (z neterminálního uzlu vychází dvě orientované hrany), ale může se větvit (klasifikovat) i do více dceřiných uzlů (nebinární stromy). Pokud se strom dále nevětví, pak uzel, který nemá žádné dceřiné uzly, se nazývá terminální. Pro terminální uzly se v literatuře používá také název listy.

Množina všech listů určuje disjunktní rozklad prostoru hodnot prediktorů X. Terminálnímu uzlu a zároveň pozorováním, která do něj patří, je přiřazena některá z tříd $T_1,...,T_k$. Strom T tak určuje klasifikační funkci d_T definovanou na X s hodnotami v množině $\{T_1,...,T_k\}$. (Klashka J., Kotrč E., 2004)

Na Obr. 3.3 je pro přehlednost schéma stromu a Obr. 3.4 znázorňuje konkrétní příklad takového stromu.



Obr. 3.3 Diagram klasifikačního stromu s binárním větvením



Obr. 3.4 Konkrétní příklad diagramu klasifikačního stromu

Na Obr. 3.4 vidíme, že klasifikační strom je binární – tedy pokud uzel není koncový, z každého vycházejí dvě větve. Kořenový uzel je dělen do dvou větví, z nichž jeden uzel je neterminální a větví se dále podle vybraného nejsilnějšího prediktoru PSAD a druhý uzel je koncový. Tento konkrétní příklad klasifikačního stromu bychom pak mohli interpretovat následovně: Klasifikační strom Obr. 3.4 určil jako významného ukazatele pro diagnostikování karcinomu prostaty hladinu PSAD vyšší než 0,23 pro pacienty všech věkových kategorií a pro pacienty s věkem nad 73 let dokonce i hladinu v rozmezí 0,13 až 0,23.

Při konstrukci klasifikačního stromu se snažíme dosáhnout co nejmenší skutečné klasifikační chyby

$$R_P(T) = P(d_T(X) \neq Y) \ , \tag{3.2}$$

kde P představuje sdružené rozdělení vektoru prediktorů X a závisle proměnné Y s hodnotami v $\{T_1,...,T_k\}$, $d_T(x)$ je zařazení vektoru X do jedné z tříd $\{T_1,...,T_k\}$podle klasifikačního stromu T. Podobně se v regresních úlohách používá (skutečná) střední kvadratická chyba

$$R_P(T) = E_P\left(Y - d_T(X)\right)^2 \ . \tag{3.3}$$

Pokud velikost stromu roste, tak chyba na cvičných datech stále klesá (nebo alespoň neroste), ale skutečná chyba v mnoha typických situacích klesá jen do určité velikosti, pak s dalším zvětšováním stromu opět roste.

Konstrukce klasifikačního stromu se skládá ze tří kroků (podle Keprta S., 1994):

1) výběr štěpícího pravidla v každém uzlu,
2) rozhodnutí, kdy je uzel koncový,
3) přiřazení třídy vysvětlované proměnné každému koncovému uzlu.

3. 2. 1 Rozhodovací pravidla větvení

Algoritmy pro konstrukci klasifikačních stromů obvykle pracují shora dolů. V každém kroku je vybírána proměnná, která co nejlépe rozděluje soubor do jednotlivých uzlů. Různé algoritmy používají různé metriky pro měření "nejvhodnější" vysvětlující proměnné.

STATISTICA nabízí diskriminační jednorozměrné dělení pro kategoriální a spojité proměnné, diskriminační dělení (lineární kombinace) pro spojité proměnné a nebo metodu CART, která je využívána i při konstrukci regresních stromů.

Diskriminační dělení

Diskriminační dělení jsou založená na kvadratické diskriminační analýze, kdy je cílem diskriminovat objekty na základě kvantitativních proměnných do jednotlivých skupin. Tato

analýza se zabývá závislostí jedné kvalitativní proměnné (v našem případě biopsie je negativní či pozitivní) na několika kvantitativních proměnných (např. hodnota PSA, věk apod.). Používá se k sestavení binárního stromu.

Vstupem diskriminační analýzy je tedy datový soubor obsahující několik kvantitativních proměnných a jednu vysvětlovanou (kvalitativní) proměnnou. Výstupem analýzy je pak diskriminační funkce, klasifikační funkce či ordinační diagram (nemáme předem definovanou závislou proměnnou, klasifikujeme podle podobnosti jednotlivých skupin).

Diskriminační funkcí zjišťujeme relativní příspěvek jednotlivých vysvětlujících proměnných k celkové diskriminaci skupin. Hledáme tedy proměnné, které jsou pro diskriminaci významné. Počet diskriminačních funkcí d je roven počtu skupin, do kterých jsou objekty děleny snížených o jednu. V případě dvou skupin (tedy i v našem případě) je diskriminační funkce d rovna mnohonásobné regresi (3.1), přičemž d = EY a b_1, ..., b_p jsou koeficienty diskriminační funkce.

Klasifikační diskriminační analýza slouží k identifikaci objektů. Výsledkem jsou klasifikační funkce, které mohou být použity k určení pravděpodobnosti příslušnosti objektů do skupin. V tomto případě máme skupinu objektů se známým zařazením do skupin (trénovací soubor, informativní výběr) a skupinu objektů, které musíme zařadit do jedné ze skupin. Na základě trénovacího souboru sestavíme klasifikační funkce, pomocí kterých odhadneme pravděpodobnost zařazení neznámých objektů do skupin. (Jarkovský J. a kol., 2012)

Jednou z možností odvození klasifikačního pravidla je výpočet lineární klasifikační funkce pro každou skupinu. Počet klasifikačních funkcí je tedy roven počtu skupin. Každá funkce umožní vypočítat klasifikační skóre pro každý objekt pro každou skupinu při použití vzorce:

$$s_i = c_1 + w_{i1}x_1 + w_{i2}x_2 + \cdots + w_{ip}x_p \ , \qquad (3.4)$$

kde i určuje skupinu, 1, 2, ..., p označují p proměnných, c_i je konstanta pro i-tou skupinu, w_{ip} je váha p-té proměnné ve výpočtu klasifikačního skóre pro i-tou skupinu; x_p je pozorovaná hodnota pro příslušný objekt a p-tou proměnnou, s_i je výsledné klasifikační skóre. (Jarkovský J. a kol., 2012)

Zařazení do skupiny je závislé právě na klasifikačním skóre. Objekt je řazen do skupiny, pro kterou je skóre nejvyšší. Ověření klasifikačního kritéria se provádí pomocí resubstituce nebo pomocí křížové validace (3.2.3).

Diskriminační analýza se dále dělí na lineární a kvadratickou. Lineární diskriminační analýza se používá pro normální rozdělení, které se liší pouze středními hodnotami (značené jako EX či μ) jednotlivých proměnných. Jestliže se navíc liší i kovariančními maticemi proměnných, používá se kvadratická diskriminační analýza.

Kovarianční matice má na hlavní diagonále variance σ^2 (DX) jednotivých proměnných a mimo diagonálu leží jednotlivé kovariance cov(x_{ki}, x_{kj}), kde i≠j a i,j =1, 2, ..., n. Kovarianční matice je symetrická. Kovariance vypočítáme ze vztahu:

$$cov(x_{ki}, x_{kj}) = E[(x_{ki} - Ex_{ki})(x_{kj} - Ex_{kj})] = E(x_{ki}, x_{kj}) - E(x_{ki})E(x_{kj}) \ , \qquad (3.5)$$

kde E(x_{ki}), E(x_{kj}) představují střední hodnoty dvou proměnných mezi kterými hledáme závislosti.

Pokud nejsou kovarianční matice stejné (obecně $C_A \neq C_B$, dále jen pro dvě $C_1 \neq C_2$), vede pravidlo pro zařazení do první skupiny $f_1(x)\pi_1 > f_2(x)\pi_2$ (π_1 a π_2 představují apriorní pravděpodobnosti zařazení do jednotlivých skupin a $f_1(x)$, $f_2(x)$ jsou hustoty pravděpodobností proměnných) ke kvadratické nerovnosti

$$x^T G x + h^T x + C > 0 \ , \qquad (3.6)$$

kde matice

$$G = 0{,}5(\ C_2^{-1} - C_1^{-1}) \ , \qquad (3.7)$$

vektor

$$h^T = \mu_1\ C_1^{-1} - \mu_2 C_2^{-1} \ , \qquad (3.8)$$

a konstanta C

$$C = 0{,}5ln\frac{\det(C_2)}{\det(C_1)} - 0{,}5(\mu_1{}^T\, C_1^{-1}\mu_1 - \mu_2{}^T C_2^{-1}\mu_2) - \left(\ln\frac{\pi_2}{\pi_1}\right). \tag{3.9}$$

Platí-li pro nové x_0 tato kvadratická nerovnost, zařazuje se objekt do první skupiny a v opačném případě do druhé skupiny. Lze také definovat kvadratické diskriminační kritérium

$$QK_j(x) = -0{,}5ln\, detC_j - 0{,}5(x-\mu_1)^T C_1^{-1}\left(\mathrm{x} - \mu_j\right) + ln\pi_j \ , \tag{3.10}$$

Objekt x_0 se pak zařazuje do třídy, které odpovídá maximální hodnota $QK_j(x_0)$. Při kvadratické diskriminační analýze se objekty zařazují do tříd podle minima Mahalanobisových vzdáleností od středů tříd μ_j. (Meloun M., 2011)
Výpočet minima Mahalanobisových vzdáleností:

$$M = \ arg\min_{j=1\ldots n}\left(x-\mu_j\right)^T \sum_j^{-1}\left(x-\mu_j\right) \ . \tag{3.11}$$

Kvadratická diskriminační analýza může využívat k diskriminaci také Bayesovo kritérium. Předpokládáme normální rozdělení vysvětlujících proměnných. Bayesovo kritérium je zobecněním kritéria maximální věrohodnosti, které zohledňuje apriorní pravděpodobnosti skupin:

$$B = \ arg\max_{j=1\ldots n} \pi_j f_j(x)\,, \tag{3.12}$$

kde f_j je hustota pravděpodobnosti a π_j označuje apriorní pravděpodobnosti. Apriorní pravděpodobnost je relativní četnost určité hodnoty proměnné pro všechny případy.
Nevýhodou diskriminační analýzy jsou opět požadavky na vstupní data, jde totiž o parametrickou metodu. Potřebujeme, aby vysvětlující proměnné měly normální rozdělení a aby nebyly příliš vzájemně korelované. Analýza je také citlivá na odlehlé hodnoty. Z těchto důvodů budeme používat metodu CART a tedy Giniho koeficient.

Giniho koeficient

Giniho koeficient G pro daný uzel vyjadřuje, jak často by byl náhodně vybraný prvek z tohoto uzlu nesprávně zařazen, kdyby pravděpodobnost zařazení do jednotlivých skupin byla rovna relativní četnosti zástupců těchto skupin v daném uzlu. V metodě CART je vypočítán Giniho index pro každý uzel. Lze vypočítat vynásobením pravděpodobnosti, že náhodně vybraný prvek z uzlu je z i-té skupiny s pravděpodobností chyby v kategorizaci případu.

$$G = \sum_{i=1}^{K} \hat{p}_i(1 - \hat{p}_i) = \sum_{i=1}^{K} (\hat{p}_i - \hat{p}_i^{\,2}) = \sum_{i=1}^{K} \hat{p}_i - \sum_{i=1}^{K} \hat{p}_i^{\,2} = 1 - \sum_{i=1}^{K} \hat{p}_i^{\,2}\,, \tag{3.13}$$

kde i ϵ {1, 2, ..., K}, K je počet skupin a $\hat{p}_i$ je relativní četnost prvků v i-té skupině v daném uzlu. Giniho index tedy můžeme vyjádřit jako sumu pravděpodobností:

$$G = \sum_{i=1}^{K} P(A_i)\,, \tag{3.14}$$

kde

$$P(A_i) = P(A_i) \times P(B/A_i)\,, \tag{3.15}$$

$P(A_i)$ je pravděpodobnost, že náhodně vybraný prvek je z i-té skupiny, $P(B/A_i)$ je potom podmíněná pravděpodobnost toho, že tento prvek chybně kategorizujeme za předpokladu, že je z i-té skupiny.

Pro každé větvení je spočítán celkový Giniho index, který je roven váženému součtu Giniho indexů všech dceřiných uzlů (3.16).

$$G_{celk} = \sum_{i=1}^{U} \frac{n_i}{n_t} G(i)\,, \tag{3.16}$$

kde U značí počet dceřiných uzlů, do kterých se mateřský uzel větví (pro binární stromy U = 2), n_i je počet prvků v dceřiných uzlech a n_t počet prvků v mateřském uzlu.

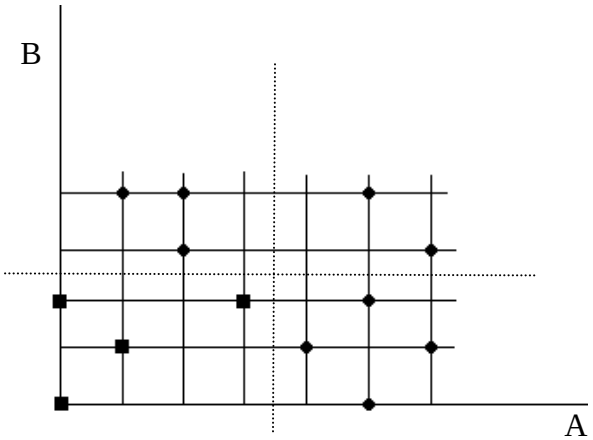
Koeficient může nabývat hodnot z intervalu $\langle 0, 1)$. Obecně pro K skupin, kde $\hat{p}_1 = \hat{p}_2 = \cdots = \hat{p}_K = \frac{1}{K}$ je relativní četnost prvků v jednotlivých skupinách platí, že Giniho koeficient je roven $G = \sum_{i=1}^{K} \hat{p}_i \times (1 - \hat{p}_i) = k \times \frac{1}{K} \times \left(1 - \frac{1}{K}\right) = 1 - \frac{1}{K}$. Tedy pro K→∞ platí, že G konverguje k 1, ale pro konečná K platí $1 - \frac{1}{K} < 1$. Jestliže koeficient nabývá nízkých hodnot (až nula), pak všechny případy spadají do jedné cílové skupiny (uzlu), v opačném případě mohou být případy řazeny do více skupin a je tedy málo pravděpodobné, že bude případ zařazen správně.

Jako nejlepší možné větvení mateřského uzlu je vybráno to dělení, pro které je Giniho koeficient minimální.

Ilustrační příklad

Mějme datový soubor, který obsahuje jednu kategoriální vysvětlovanou proměnnou Y a dvě spojité vysvětlující proměnné, tedy vstupní matice X = (A, B, Y), kde A může nabývat hodnot z množiny {1,2,3,4,5,6,7} a B z množiny {1, 2, 3, 4, 5} viz Obr. 3.5.

	A	B	Y
2	1	1	1
3	1	3	1
4	2	2	1
5	2	5	0
6	3	4	0
7	3	5	0
8	4	3	1
9	5	2	0
10	6	1	0
11	6	3	0
12	6	5	0
13	7	2	0
14	7	4	0



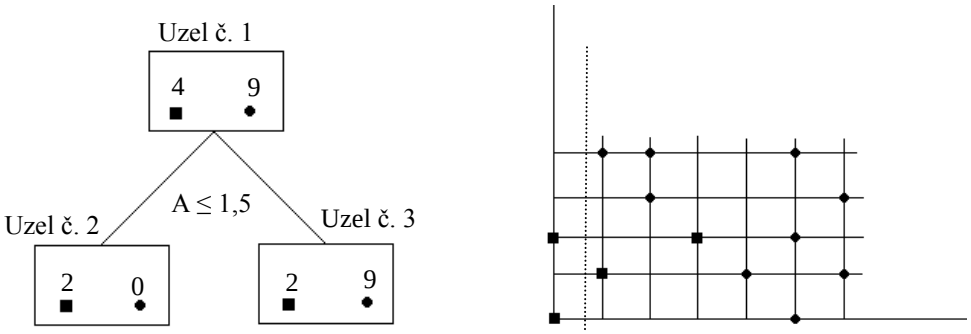
Obr. 3.5. Datový soubor a rozložení jednotlivých prvků, kde čtverec odpovídá skupině 1 a kolečko 0, přerušované čáry pak symbolizují odhad pro nejlepší možné rozdělení souboru

Nyní pomocí Giniho koeficientu nalezneme klasifikační pravidla pro rozdělení souboru. Vypočtené hodnoty srovnáme s výsledky softwaru STATISTICA.

1.) Možnosti pro první větvení souboru

Abychom zahrnuli všechna možná větvení, musíme vypočítat Giniho index postupně pro případy A ≤ 1,5; A ≤ 2,5; A ≤ 3,5; A ≤ 4,5; A ≤ 5, 5; A ≤ 6,5 (A = 7 nemá smysl, jelikož bychom zahrnuli celý soubor) a podobně B ≤ 1,5; B ≤ 2,5; B ≤ 3,5; B ≤ 4,5.

a) A ≤ 1,5



Obr 3.6 Ilustrační obrázek větvení, za podmínky A ≤ 1,5, vpravo ilustrace rozdělení (přerušovaná čára)

V mateřském (kořenovém) uzlu (č. 1) máme prvků celkem $n_t = 13$.

Pro dceřiný uzel č. 2 platí, že do skupiny 1 spadají dva prvky a do skupiny 0 nepatří žádný prvek. Nyní si spočítejme jednotlivé pravděpodobnosti $\hat{p}_0$, že náhodně vybraný prvek z uzlu je ze skupiny 0 a $\hat{p}_1$, že náhodně vybraný prvek z uzlu je ze skupiny 1.

$$\hat{p}_i = \frac{n_m}{n_i} \,, \tag{3.17}$$

kde n_i je počet prvků v dceřiném uzlu a n_m počet prvků zastoupených v dané skupině (0,1).

Tedy: $\hat{p}_1 = \frac{2}{2} = 1 \qquad \hat{p}_0 = \frac{0}{2} = 0$

$G_{uzel\ č.2} = \sum_{i=1}^{K} \hat{p}_i(1 - \hat{p}_i) = \hat{p}_1(1 - \hat{p}_1) + \hat{p}_0(1 - \hat{p}_0) = 1 \times 0 + 0 \times 1 = 0$

Pro dceřiný uzel č. 3 podobně vypočítáme:

$\hat{p}_1 = \frac{2}{11} \qquad \hat{p}_0 = \frac{9}{11}$

$G_{uzel\ č.3} = \frac{2}{11}\left(1 - \frac{2}{11}\right) + \frac{9}{11}\left(1 - \frac{9}{11}\right) = \frac{2}{11}\frac{9}{11} + \frac{9}{11}\frac{2}{11} = \frac{36}{121}$

Celkový Giniho koeficient pro toto větvení je tedy podle vzorce (3.16) roven:

$$G_{celk} = \sum_{i=1}^{U} \frac{n_i}{n_t} G(i) = \frac{2}{13} \times 0 + \frac{11}{13} \times \frac{36}{121} = \frac{36}{141} \cong 0{,}2517$$

Podobně provedeme i výpočty pro ostatní možnosti větvení

b) A ≤ 2,5

Uzel č.2 obsahuje 3 prvky ve skupině 1 a 1 prvek ve skupině 0, tedy

$\hat{p}_1 = \frac{3}{4}$ a $\hat{p}_0 = \frac{1}{4}$, pro tento uzel je Giniho index: $G_{uzel\ č.2} = \frac{3}{8}$

Uzel č.3 obsahuje 1 prvek ve skupině 1 a 8 prvků ve skupině 0, tedy

$\hat{p}_1 = \frac{1}{9}$ a $\hat{p}_0 = \frac{8}{9}$, pro tento uzel je Giniho index: $G_{uzel\ č.3} = \frac{16}{81}$

Celkový Giniho index pro větvení: $G_{celk} = \frac{59}{234} \cong 0{,}2521$

c) A ≤ 3,5

Uzel č.2 obsahuje 3 prvky ve skupině 1 a 3 prvky ve skupině 0, tedy

$\hat{p}_1 = \frac{1}{2}$ a $\hat{p}_0 = \frac{1}{2}$, pro tento uzel je Giniho index: $G_{uzel\ č.2} = \frac{1}{2}$

Uzel č.3 obsahuje 1 prvek ve skupině 1 a 6 prvků ve skupině 0, tedy

$\hat{p}_1 = \frac{1}{7}$ a $\hat{p}_0 = \frac{6}{7}$, pro tento uzel je Giniho index: $G_{uzel\ č.3} = \frac{12}{49}$

Celkový Giniho index pro větvení: $G_{celk} = \frac{139}{364} \cong 0{,}3819$

d) A ≤ 4,5

Uzel č.2 obsahuje 4 prvky ve skupině 1 a 3 prvky ve skupině 0, tedy

$\hat{p}_1 = \frac{4}{7}$ a $\hat{p}_0 = \frac{3}{7}$, pro tento uzel je Giniho index: $G_{uzel\ č.2} = \frac{24}{49}$

Uzel č.3 obsahuje 0 prvků ve skupině 1 a 6 prvků ve skupině 0, tedy

$\hat{p}_1 = 0$ a $\hat{p}_0 = 1$, pro tento uzel je Giniho index: $G_{uzel\ č.3} = 0$

Celkový Giniho index pro větvení: $G_{celk} = \frac{24}{91} \cong 0{,}2637$

e) A ≤ 5, 5

Uzel č.2 obsahuje 4 prvky ve skupině 1 a 4 prvky ve skupině 0, tedy

$\hat{p}_1 = \frac{1}{2}$ a $\hat{p}_0 = \frac{1}{2}$, pro tento uzel je Giniho index: $G_{uzel\ č.2} = \frac{1}{2}$

Uzel č.3 obsahuje 0 prvků ve skupině 1 a 4 prvky ve skupině 0, tedy

$\hat{p}_1 = 0$ a $\hat{p}_0 = 1$, pro tento uzel je Giniho index: $G_{uzel\ č.3} = 0$

Celkový Giniho index pro větvení: $G_{celk} = \frac{4}{13} \cong 0{,}3077$

f) $A \leq 6{,}5$

Uzel č.2 obsahuje 4 prvky ve skupině 1 a 7 prvků ve skupině 0, tedy

$\hat{p}_1 = \frac{4}{11}$ a $\hat{p}_0 = \frac{7}{11}$, pro tento uzel je Giniho index: $G_{uzel\ \text{č.2}} = \frac{56}{121}$

Uzel č.3 obsahuje 0 prvků ve skupině 1 a 2 prvky ve skupině 0, tedy

$\hat{p}_1 = 0$ a $\hat{p}_0 = 1$, pro tento uzel je Giniho index: $G_{uzel\ \text{č.3}} = 0$

Celkový Giniho index pro větvení: $G_{celk} = \frac{56}{143} \cong 0{,}3916$

g) $B \leq 1{,}5$

Uzel č.2 obsahuje 1 prvek ve skupině 1 a 1 prvek ve skupině 0, tedy

$\hat{p}_1 = \frac{1}{2}$ a $\hat{p}_0 = \frac{1}{2}$, pro tento uzel je Giniho index: $G_{uzel\ \text{č.2}} = \frac{1}{2}$

Uzel č.3 obsahuje 3 prvky ve skupině 1 a 8 prvků ve skupině 0, tedy

$\hat{p}_1 = \frac{3}{11}$ a $\hat{p}_0 = \frac{8}{11}$, pro tento uzel je Giniho index: $G_{uzel\ \text{č.3}} = \frac{48}{121}$

Celkový Giniho index pro větvení: $G_{celk} = \frac{59}{143} \cong 0{,}4126$

h) $B \leq 2{,}5$

Uzel č.2 obsahuje 2 prvky ve skupině 1 a 3 prvky ve skupině 0, tedy

$\hat{p}_1 = \frac{2}{5}$ a $\hat{p}_0 = \frac{3}{5}$, pro tento uzel je Giniho index: $G_{uzel\ \text{č.2}} = \frac{12}{25}$

Uzel č.3 obsahuje 2 prvky ve skupině 1 a 6 prvků ve skupině 0, tedy

$\hat{p}_1 = \frac{1}{4}$ a $\hat{p}_0 = \frac{3}{4}$, pro tento uzel je Giniho index: $G_{uzel\ \text{č.3}} = \frac{24}{64}$

Celkový Giniho index pro větvení: $G_{celk} \cong 0{,}4154$

i) $B \leq 3{,}5$

Uzel č.2 obsahuje 4 prvky ve skupině 1 a 4 prvky ve skupině 0, tedy

$\hat{p}_1 = \frac{1}{2}$ a $\hat{p}_0 = \frac{1}{2}$, pro tento uzel je Giniho index: $G_{uzel\ \text{č.2}} = \frac{1}{2}$

Uzel č.3 obsahuje 0 prvků ve skupině 1 a 5 prvků ve skupině 0, tedy

$\hat{p}_1 = 0$ a $\hat{p}_0 = 1$, pro tento uzel je Giniho index: $G_{uzel\ \text{č.3}} = 0$

Celkový Giniho index pro větvení: $G_{celk} = \frac{4}{13} \cong 0{,}3077$

j) $B \leq 4{,}5$

Uzel č.2 obsahuje 4 prvky ve skupině 1 a 6 prvků ve skupině 0, tedy

$\hat{p}_1 = \frac{2}{5}$ a $\hat{p}_0 = \frac{3}{5}$, pro tento uzel je Giniho index: $G_{uzel\ \text{č.2}} = \frac{24}{50}$

Uzel č.3 obsahuje 0 prvků ve skupině 1 a 3 prvky ve skupině 0, tedy

$\hat{p}_1 = 0$ a $\hat{p}_0 = 1$, pro tento uzel je Giniho index: $G_{uzel\ č.3} = 0$

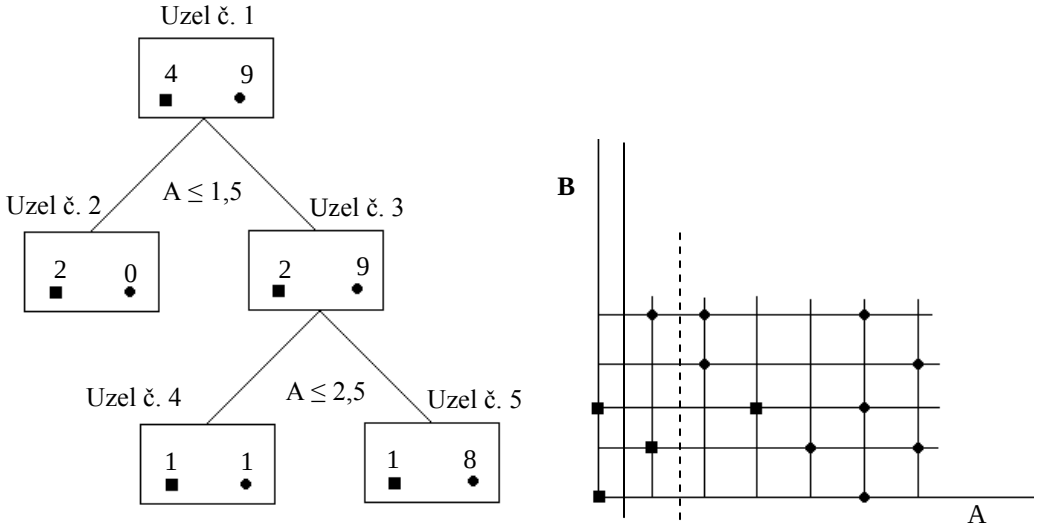
Celkový Giniho index pro větvení: $G_{celk} = \frac{24}{65} \cong 0{,}3692$

Ze všech možností větvení najdeme nejlepší možnost tak, že vybere to, pro které je Celkový Giniho index nejnižší, tedy A ≤ 1,5. Jelikož v uzlu č.2 jsou všechny prvky řazeny do jedné skupiny, nemá jej dále smysl dělit. Budeme tedy hledat další větvení pro uzel č.3.

2.) Možnosti pro druhé větvení

Nyní se budeme snažit rozdělit všechny prvky z uzlu č.3, tedy máme $n_t = 11$ a možnosti větvení A ≤ 2,5; A ≤ 3,5; A ≤ 4,5; A≤ 5, 5; A ≤ 6,5 a podobně B ≤ 1,5; B ≤ 2,5; B ≤ 3,5; B ≤ 4,5.

a) A ≤ 2,5



Obr 3.7 Ilustrační obrázek větvení, za podmínky A ≤ 2,5, vpravo ilustrace rozdělení (přerušovaná čára)

Uzel č.4 obsahuje 1 prvek ve skupině 1 a 1 prvek ve skupině 0, tedy

$\hat{p}_1 = \frac{1}{2}$ a $\hat{p}_0 = \frac{1}{2}$, pro tento uzel je Giniho index: $G_{uzel\ č.2} = \frac{1}{2}$

Uzel č.5 obsahuje 1 prvek ve skupině 1 a 8 prvků ve skupině 0, tedy

$\hat{p}_1 = \frac{1}{9}$ a $\hat{p}_0 = \frac{8}{9}$, pro tento uzel je Giniho index: $G_{uzel\ č.3} = \frac{16}{81}$

Celkový Giniho index pro větvení: $G_{celk} \cong 0{,}2525$

Pro ostatní případy už budeme psát přímo celkový Giniho koeficient.

b) $A \leq 3{,}5: \ G_{celk} \cong 0{,}2922$

c) $A \leq 4{,}5: \ \boxed{G_{celk} \cong 0{,}2182}$

d) $A \leq 5{,}5: G_{celk} \cong 0{,}2424$

e) $A \leq 6{,}5: G_{celk} \cong 0{,}2828$

f) $B \leq 1{,}5: G_{celk} \cong 0{,}2909$

g) $B \leq 2{,}5: G_{celk} \cong 0{,}2922$

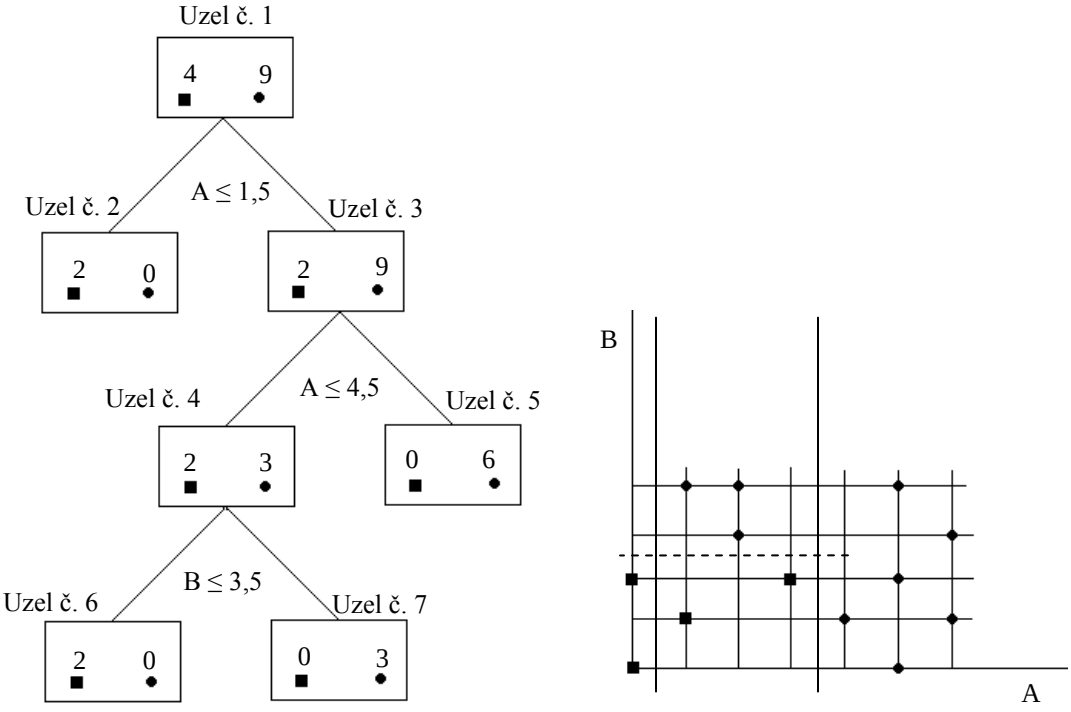
h) $B \leq 3{,}5: G_{celk} \cong 0{,}2424$

i) $B \leq 4{,}5: G_{celk} \cong 0{,}2727$

Nejnižší hodnotou pro druhé větvení je Giniho koeficient pro $A < 4{,}5$($G_{celk} \cong 0{,}2182$)

3.) Možnosti pro třetí větvení

V případě, že soubor máme rozdělen hranicemi $A \leq 1{,}5$ a $A \leq 4{,}5$, máme již omezené možnosti dělení souboru Obr. 3.8. Soubor můžeme dělit rozhraními $A \leq 2{,}5$; $A \leq 3{,}5$; $B \leq 2{,}5$; $B \leq 3{,}5$ a $B \leq 4{,}5$. V uzlu č.4, který má smysl dále dělit, máme $n_t = 5$.



Obr 3.8 Ilustrační obrázek větvení, za podmínky d) $B \leq 3{,}5$, vpravo ilustrace rozdělení (přerušovaná čára – možné dělení, plná čára – dříve provedené dělení)

a) $A \leq 2{,}5: \ G_{celk} \cong 0{,}4666$

b) $A \leq 3{,}5: G_{celk} \cong 0{,}3$

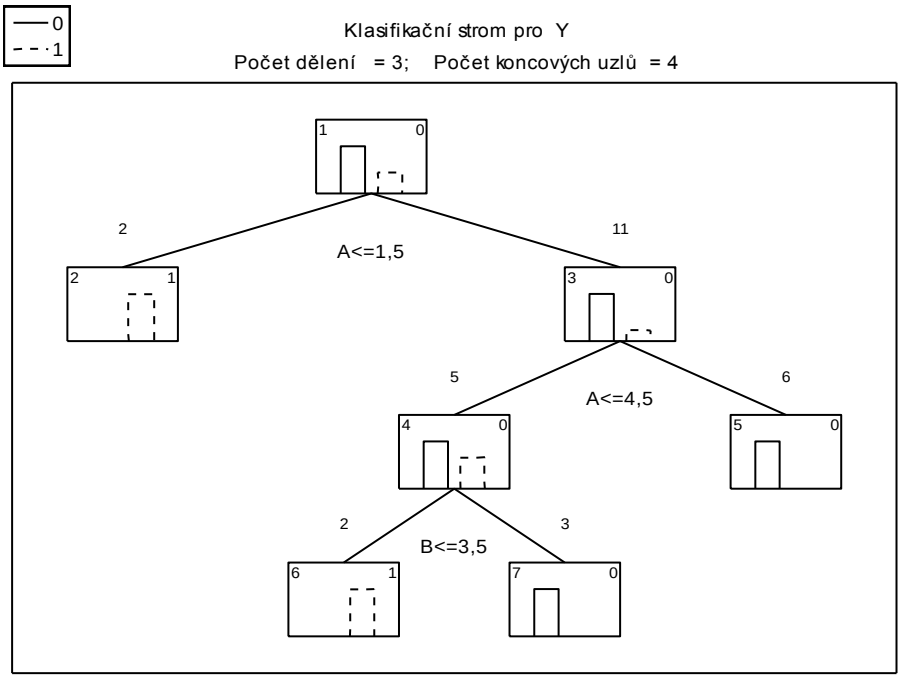
c) $B \leq 2{,}5: G_{celk} = 0{,}3$

d) $B \leq 3{,}5: \boxed{G_{celk} = 0}$

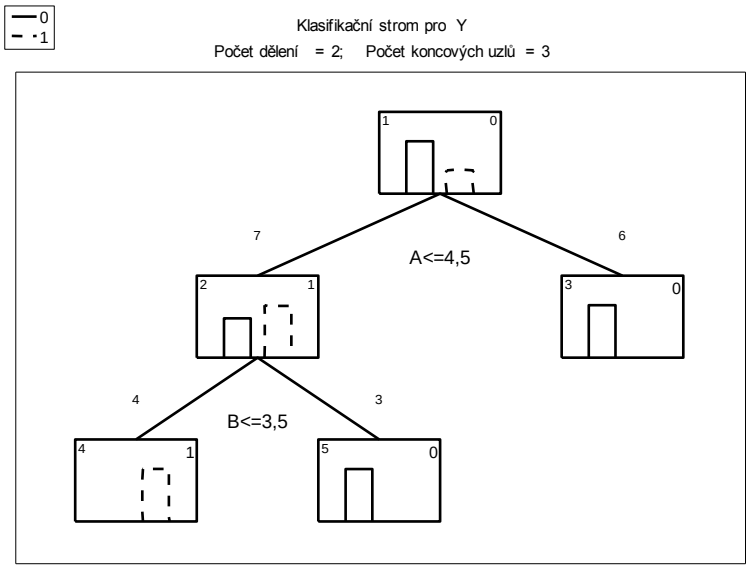
e) $B \leq 4{,}5: G_{celk} \cong 0{,}2666$

Nejnižší hodnotou pro třetí větvení je Giniho koeficient pro $B \leq 3{,}5$. Můžeme vidět Obr.3.8, že všechny prvky v uzlu č. 6 i v uzlu č. 7 jsou zařazeny do jedné ze skupin, nemá tedy již smysl žádné další dělení. Nyní provedeme klasifikaci pomocí softwaru STATISTICA. Měli bychom sestrojit shodný klasifikační strom jako na Obr. 3.8.

Obr. 3.9 potvrzuje, že klasifikace prvků se ve všech uzlech všech pater shoduje s našimi výpočty. Námi očekávané dělení souboru Obr. 3.5 bylo provedeno podle výpočtu Giniho koeficientů až v druhém větvení klasifikačního stromu. Pro porovnání jsme v softwaru vyzkoušeli i chí-kvadrát test a G-kvadrát, přičemž chí-testem vyšla klasifikace stejná jako v případě použití Giniho koeficientu a až maximálně věrohodný chí-kvadrát (G-kvadrát) našel námi předpokládanou klasifikaci souboru (Obr. 3.10).



Obr 3.9 Diagram klasifikačního stromu ilustračního příkladu



Obr 3.10 Diagram klasifikačního stromu ilustračního příkladu za použití G-kvadrát klasifikace

Chí-kvadrát

Chí-kvadrát test (test dobré shody) je jedním z nejpoužívanějších statistických testů. Srovnává pozorované četnosti a očekávané četnosti jednotlivých kombinací proměnných.

Nulová hypotéza je stanovovaná tak, že proměnné X_1 a X_2 jsou nezávislé. Jestliže proměnná X_1 je rovna hodnotě i a proměnná X_2 je rovna hodnotě j, pak n_{ij} označuje počet všech případů, kdy tato situace nastala. Marginální četnosti příslušné i-té variantě proměnné X_1, respektive j-té variantě proměnné X_2, vypočteme

$$n_{i.} = \sum_{j=1}^{c} n_{ij} \quad , \qquad n_{.j} = \sum_{i=1}^{r} n_{ij} \quad , \tag{3.18}$$

kde i = 1,..., r a j=1, ...,c.

Za platnosti nulové hypotézy lze očekávané četnosti jednotlivých kombinací, kdy X_1 = i a X_2 = j, které budeme značit e_{ij}, vypočítat pomocí výrazu

$$e_{ij} = n\frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}. \tag{3.19}$$

Karl Pearson již v roce 1904 odvodil, že statistika

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij} - e_{ij})^2}{e_{ij}}\ , \tag{3.20}$$

má za platnosti nulové hypotézy o nezávislosti asymptoticky chí-kvadrát rozdělení pravděpodobnosti s parametrem (r – 1)(c – 1), tedy že platí $X^2 \sim \chi^2_{(r-1)(c-1)}$. V případě chí-kvadrát testu proti nulové hypotéze hovoří pouze extrémně velké hodnoty testové statistiky, neboť ty indikují významnou neshodu mezi pozorovanými a očekávanými četnostmi. Naopak velmi malé hodnoty testové statistiky hovoří pro nulovou hypotézu, proto nulovou hypotézu o nezávislosti X a Y zamítáme na hladině významnosti α, když hodnota testové statistiky X^2 přesáhne příslušný 100(1 – α)% kvantil rozdělení χ^2, tedy když

$$X^2 \geq \chi^2_{(r-1)(c-1)}(1-\alpha)\ . \tag{3.21}$$

(Pavlík T., Dušek L.,2012)

Ověřování míry shody pomocí chí-kvadrátu využívá metoda CHAID. Strom typu CHAID se řadí mezi klasifikační stromy. Používá se pouze pro kvalitativní proměnné (kategoriální), jelikož tato metoda využívá k nalezení nejlepšího možného větvení právě testování pomocí χ^2 (chí-kvadrátu). Pro každou závisle proměnou (vysvětlovanou proměnou) a vysvětlující proměnné (prediktory) se vytváří kontingenční tabulka a pro všechny dvojice hodnot vysvětlujících proměnných se spočítá χ^2 test a najde se tak nejvhodnější kombinace prediktorů. Chí-testem dostaneme statistickou významnost kombinací prediktorů, která je dána p-hodnotou. Nejvhodnější kombinací prediktorů, je potom ta, kde je p-hodnota nejnižší. Strom typu CHAID vytváří nebinární větvení, takže se může stát, pokud nemáme dostatečné množství dat, že nevzniknou další patra stromu.

Další možnosti

G-kvadrát je maximálně věrohodný chí-kvadrát test. Jeho užití by mělo zvyšovat přesnost klasifikace datového souboru. Mezi další kriteriální statistiky pro větvení se řadí informační

zisk a entropie. Entropie udává míru neuspořádanosti daného systému či neurčitost procesu větvení. Informační zisk pak využívá výpočtu entropie. Je definován jako rozdíl entropie pro celý datový soubor a soubor, o kterém se rozhoduje.

3.2.2 Možnosti ukončení větvení

Klasifikační strom nemůže tvořit neustále nové větve. Klasifikace je omezena velikostí souboru. Jestliže uzel obsahuje pouze jeden případ nebo všechna pozorování v uzlu mají stejné hodnoty všech prediktorů, nemůže se strom dále větvit. Další možností ukončení větvení je, že všechny případy v daném uzlu mají stejnou hodnotu vysvětlované proměnné – v tomto případě říkáme, že uzel je „čistý“. Omezit větvení můžeme ale i nastavením některých parametrů pro ukončení.

Software STATISTICA nabízí možnost přímého ukončení FACT (frakce objektů), pokles chyb špatné klasifikace a pokles odchylky, kdy nastavujeme parametry pro ukončení, jako minimální počet n (za list se prohlásí uzel, do kterého patří méně než n pozorování, např. n = 5 až 10 pozorování, poté se strom „prořezává“) nebo pravidlo směrodatné chyby. Tedy uzel se nerozdělí, pokud střední kvadratická chyba (MSE) nebo procento nesprávně klasifikovaných vzorků v důsledku rozdělení překročí určitou hranici.

Pokles odchylky vybíráme v případě, že tvoříme regresní strom. Jestliže naše vysvětlovaná proměnná je kategoriální, pak použijeme možnost přímého ukončení FACT (frakce objektů). Přímé ukončení FACT pokračuje v růstu stromu, dokud všechny terminální uzly nejsou čisté. Nastavením parametru α určíme prořezání stromu na základě počtu prvků ve skupinách v možném terminálním uzlu, který (pokud uzel není čistý) nesmí být v majoritně zastoupené skupině nižší, než je stanovené minimum objektů třídy. Minimum objektů třídy (h_1; h_0), se kterými jsou počty v dceřiných uzlech srovnávány, vypočteme pomocí zastoupení prvků v i-té skupině mateřského uzlu a parametru $\alpha \in (0;1)$, který nastavujeme. Když $n_1 < n_0$, hodnota h_i pak představuje dolní celou část:

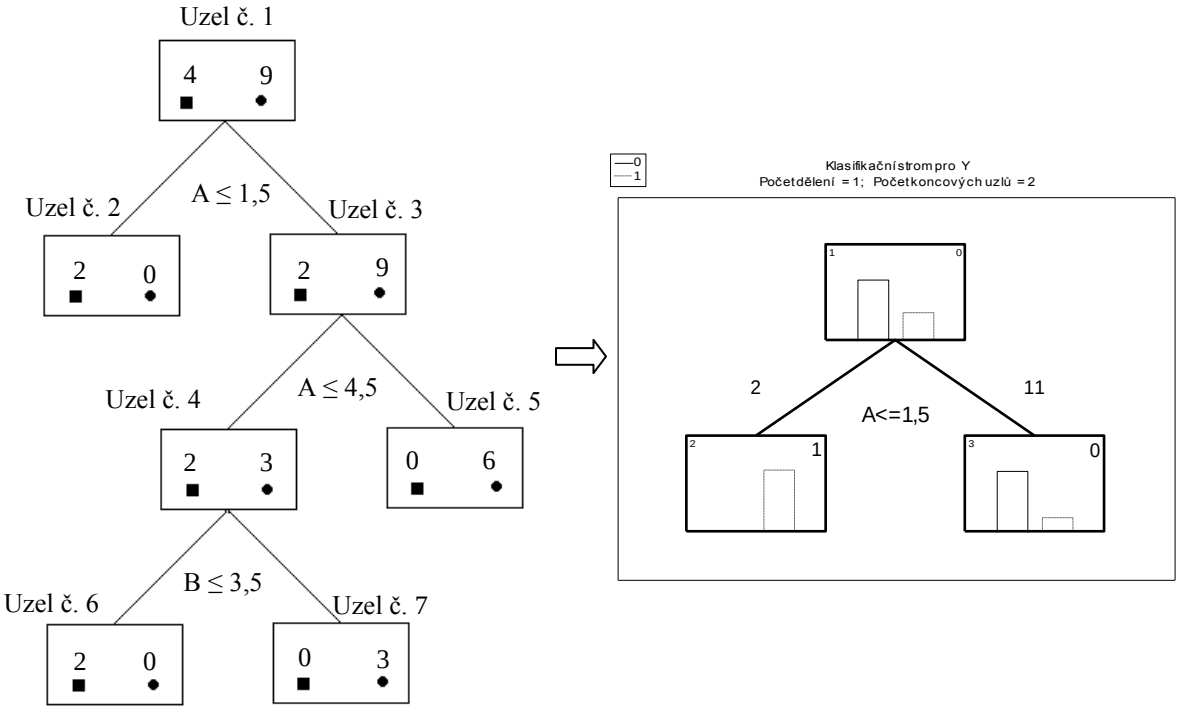
$$(h_1; h_0) = \left(\lfloor n_1 \alpha \rfloor; \left\lfloor n_0 \alpha \frac{n_0}{n_1} \right\rfloor\right), \tag{3.22}$$

kde n_i je počet prvků v i-té skupině, $i \in \{0,1\}$.

Převedeme – li toto pravidlo na náš ukázkový ilustrační příklad, pak nastavením parametru α dosáhneme různých změn ve velikosti klasifikačního stromu.

Pro α = 0,2 platí: $(h_1; h_0) = \left(\lfloor 4 \times 0{,}2 \rfloor; \left\lfloor 9 \times 0{,}2 \times \frac{9}{4} \right\rfloor\right) = (\lfloor 0{,}8 \rfloor; \lfloor 4{,}05 \rfloor) = (0; 4)$.

Nyní postupujme po jednotlivých uzlech. Uzel č.1 se zastoupením prvků $n_1 = 4$, $n_0 = 9$ splňuje podmínku větvení – pro majoritně zastoupenou skupinu platí $n_0 \geq h_0$, tedy $9 \geq 4$. Uzel č. 2 je čistý a tedy podmínky větvení splňuje. Pro uzel č. 3 máme majoritně zastoupenou skupinu 0, srovnáme tedy počet prvků s minimem objektů v třídě h_0, $9 \geq 4$ a tedy i tento uzel splňuje podmínku větvení. Pro uzel č. 4 máme počty $n_1 = 2$, $n_0 = 3$, majoritně zastoupenou skupinou je zde skupina 0 a tedy srovnáváme hodnoty n_0 a h_0. Jestliže $n_0 < h_0$ ($3 < 4$), není splněna podmínka pro větvení. Uzel č. 4 a tedy i uzly č.5, č.6 a č.7 budou „odřezány“. Původní strom tedy bude prořezán na strom o velikosti 3 uzlů (Obr. 3.11).



Obr. 3.11 Prořezávání klasifikačního stromu možností ukončení FACT

3.2.3 Ověření velikosti stromu

Ověření klasifikačního kritéria a tedy i k ověření vhodné velikosti stromu se používá resubstituce nebo křížová validace.

Resubstituce používá k ověřování stejný datový soubor, z něhož bylo počítáno klasifikační kritérium. Využívá se v případě, že nemáme dostatečně velký datový soubor. Jestliže máme dostatečné množství dat, je vhodnější rozdělení souboru, kdy vytvoříme z datového souboru dvě skupiny, jednu použijeme k vytvoření odhadu a na druhé testujeme kvalitu tohoto odhadu. (podrobně viz Komprdová K., 2012)

Křížová validace vybírá z m případů (datový soubor rozdělí na m-částí) m – 1 a ty použije jako trénovací soubor, ze kterého odvodí klasifikační kritérium. Toto kritérium aplikuje na zbývající část datového souboru a postup opakuje m-krát. Výsledky testování se vyjadřují procentuelně tabulkou.

V případě, že sestrojujeme strom T_{max}, který následně prořezáváme, lze použít k určení optimální velikosti stromu kritérium cost of komplexity. Hledáme strom, který pro každý parametr $\alpha \geq 0$, který představuje kompromis mezi přesností a velikostí stromu, minimalizuje $C_\alpha T$.

Kritérium cost-komplexity:

$$C_\alpha(T_1) = DT_1 + \alpha|T_1| \ , \tag{3.23}$$

kde $|T_1|$ značí počet terminálních uzlů, DT_1 je deviance, čili chyba stromu T_1. K odhadu parametru α se používá křížová validace.

3.3 CART

Metoda CART (C&RT) se používá k vytvoření klasifikačních i regresních stromů. Tyto stromy využívají sestavení stromu T_{max}, který se následně „prořezává“ a odstraňují se nevýznamné uzly, podle předem zvolených mezí a odhadů skutečných chyb. Strom sestavený

metodou CART obsahuje pouze binární větvení. Klasifikační stromy CART využívají jako klasifikační kritérium Giniho koeficient.

Algoritmus růstu stromů CART (podle Komprdové K., 2012):

1.) Rozdělení souboru na trénovací a testovací. Tento poměr se určuje na základě počtu pozorování a účelu studie.

2.) Nalezení nejlepšího rozdělení každého z prediktorů:

a) Pro spojité vysvětlující proměnné – seřadí hodnoty každého z prediktorů od nejmenší po největší. Projde všechny hodnoty prediktoru X a spočítá kriteriální statistiku všech možných rozdělení proměnné Y na dva možné potenciální dceřinné uzly. Pokud je dělící hodnota „a“ prediktoru X větší nebo rovna hodnotě x_i, pozorování y_i náleží do levého uzlu, jinak do pravého (popřípadě naopak). Hodnota „a“ pro kterou je kriteriální statistika minimální je vybrána jako nejlepší možné dělení závislé proměnné Y pomocí daného prediktoru. Pro každý prediktor získáme jednu hodnotu (nejlepší potenciální rozdělení) kriteriální statistiky. Následně je vybrán prediktor s nejnižší hodnotou kriteriální statistiky a hodnota a je použita k rozdělení souboru (hodnoty y_i) do dvou dceřiných uzlů.

b) Pro kategoriální prediktor se za účelem nalezení nejlepšího rozdělení projdou všechny možné kombinace tvořené jednotlivými kategoriemi prediktoru a hodnot nebo kategorií závislé proměnné. Opět se použije dělení s nejnižší hodnotou kriteriální statistiky.

3.) Rozdělení souboru na dva dceřinné uzly t_1 a t_2 podle hodnoty prediktoru vybrané v kroku 2

4.) Opakování kroku 2 a 3, dokud se dělení nezastaví na předem definované hodnotě (dokud není dosaženo některého z pravidel pro zastavení růstu stromu). Stejný prediktor může být použitý vícekrát, protože vybíráme vždy z celé množiny vysvětlujících proměnných.

5.) Použití testovacího souboru k ověření vhodné velikosti stromu. Pokud je strom příliš velký, strom se následně „prořezává“.

Srovnání metody CART a dalších algoritmů pro tvorbu stromů můžeme nalézt v článku (Savický P. a kol., 2000), kde bylo provedeno testování algoritmů na různých typech experimentálních dat. V tabulce Tab.10 můžeme vidět rozdíly mezi jednotlivými klasifikačními metodami.

Metoda	Rozhodovací kritérium	Větvení	Typ proměnných
Kvadratická diskriminační analýza (QUEST)	Kvadratická diskriminační analýza	Binární	Spojité, kategoriální
CART	Gini index	Binární	Spojité
CHAID	χ^2 – kvadrát	Nebinární	kategoriální

Tab.10 Srovnání jednotlivých metod tvorby klasifikačních stromů

3.4 Regresní stromy

Regresní stromy můžeme zařadit mezi neparametrické odhady, přesněji mezi odhady po částech konstantní. Při vzniku stromu se postupuje stejně jako u klasifikačních stromů, je složen z orientovaných hran a jednotlivých uzlů, kde je v každém uzlu ukryt princip větvení stromu.

Regresní strom se od klasifikačního stromu liší tím, že každému terminálnímu uzlu je přiřazena reálná konstanta – odhad kvantitativní závisle proměnné Y. Regresní strom T definuje reálnou regresní funkci d_T, která je uvnitř množin odpovídajících terminálním uzlům konstantní. (Klashka J., Kotrč E., 2004)

V regresi se využívá různých metod, mezi nejpoužívanější patří metoda CART, která se využívá i u klasifikačních stromů, dále metoda PRIM a MARS.

Řekněme, že všechny možné hodnoty vysvětlujících proměnných X padnou do vektorového prostoru $X = X_1 * ... * X_M$.

V prvním kroku je prostor $\mathbb{X}$ rozložen posloupností rekurzivních dělení na velice mnoho co možná nejmenších podmnožin. Máme-li k dispozici velkou paměť počítače, každá podmnožina výsledného rozkladu T_{max} bude obsahovat pouze jedno pozorování. V druhém kroku je aplikován algoritmus kolapsující (rekombinující) tento počáteční rozklad až do $\mathbb{X}$. Při kolapsování se používá téže míry kvality odhadů, jako při konstrukci, tj. střední čtvercové (absolutní) chyby, modifikované však o člen penalizující nás za příliš rozsáhlé rozklady. Výsledkem kolapsování je posloupnost do sebe vnořených rozkladů prostoru $\mathbb{X}$, počínající T_{max} a končící samotným prostorem $\mathbb{X}$. Z této množiny je třeba vybrat řešení optimální. (Antoch J., 1988)

K výpočtu regrese se využívají dvě klasické varianty – metoda nejmenších čtverců a nejmenších absolutních odchylek.

V případě varianty nejmenších čtverců má statistika φ tvar

$$\varphi = -\sum_{j=1}^{2} \sum_{p \in P_j(s)} (Y(p) - \bar{Y}_j)^2 \ , \tag{3.24}$$

kde předpokládáme, že rozklad s prostoru $\mathbb{X}$ indukuje rozklad množiny n případů P na množiny $P_1(s)$ a $P_2(s)$ o velikosti $n_1(s)$ a $n_2(s)$ a kde

$$\bar{Y}_j = \frac{1}{n_j(s)} \sum_{p \in P_j(s)} y(p) \ , \ \text{j} = 1,2. \tag{3.25}$$

Varianta nejmenších absolutních odchylek má statistiku φ tvaru

$$\varphi = -\sum_{j=1}^{2} \sum_{p \in P_j(s)} \left| Y(p) - \tilde{Y}_j \right| \ , \tag{3.26}$$

kde $\tilde{Y}_j$ je medián hodnot Y(p) v množině Pj(s). (Klaschka J., Antoch J., 1996)

K určení optimální velikosti stromu se zde běžně užívá kritérium „cost of komplexity“.

3.4.1 Regresní metody

Tak jako se ke klasifikaci používají metody CHAID či CART, máme i metody pro regresi.

U metody MARS se používá namísto reálné konstanty, která je přiřazena terminálnímu uzlu, lineární aproximace. Je vhodná v případech, kdy máme velké množství vysvětlujících proměnných, zahrnuje jejich interakci. Chybí zde typický stromový model, který je vhodnější pro interpretaci, výstupem je zde regresní rovnice.

PRIM je metoda primárně určena pro regresi. Rozděluje prostor na pravoúhelníky – vyhledávají se takové, ve kterých je odpovídající průměr hodnot závisle proměnné nejvyšší. Pravoúhelník se postupně zmenšuje - na začátku algoritmus vybere nejvýhodnější osu podle pozorování, mající nejvyšší nebo nejnižší hodnoty prediktoru. Vybere se takové „zmenšení“, které má nejvyšší průměr hodnot závisle proměnné ve zbývajícím pravoúhelníku. To se opakuje do předem definované hodnoty minimálního počtu pozorování v pravoúhelníku. Oproti CART je výhodou, že se probere větší škála pravidel a můžeme najít optimální řešení. Nevýhoda je, že není k dispozici stromová struktura, pouze pravidla. (Kubošová K., 2013)

4 Tvrzení a hypotézy

Hypotéza je tvrzení, které můžeme statisticky vyhodnotit na základě datového souboru (skutečně naměřených či pozorovaných hodnot proměnných). Jako první stanovujeme hypotézu nulovou H_0, kterou můžeme na určité hladině významnosti zamítnout a potvrdit hypotézu alternativní H_A, která je jejím opakem. Případně můžeme zjistit, že nulovou hypotézu zamítnout nemůžeme. Klasifikační stromy bohužel nejsou sestaveny k ověřování klasických statistických hypotéz. Smíme tedy pouze srovnat výsledky odhadů klasifikačních stromů a předpokládané kritické hodnoty proměnných, jež stanovili specialisti v daném oboru.

Z výsledných klasifikačních stromů se pokusíme ověřit platnost jednotlivých tvrzení, která jsou v medicíně všeobecně známá a považují se za platná. Budeme sledovat klasifikaci do jednotlivých uzlů podle vysvětlujících proměnných, které vypovídají o tom, zda při následné biopsii byl nebo nebyl diagnostikován karcinom prostaty.

Tvrzení určená k ověření:

1.) Uvádí se, že PSA denzita vyjadřující poměr celkové hladiny PSA v séru a celkového objemu prostaty je u pacientů s karcinomem prostaty vyšší než 0,15 (URL5).

2.) Za rizikovou skupinu mužů, u kterých se ve větší míře objevuje karcinom prostaty, je považována věková kategorie 65 až 80 let (URL1).

3.) Lékaři pokládají hladinu prostatického specifického antigenu za důležitý ukazatel při vyšetření prostaty. Obecně hladina nad 10 ng/ml v séru (URL2) je považována za podezřelou.

4.) V případě, že se hodnota hladiny PSA v séru nachází kolem hraniční hodnoty, může index poukazovat právě na riziko karcinomu. Index nižší než 0,2 (uváděno i 0,25) (URL5, URL7) značí možnost nálezu karcinomu prostaty při následné biopsii.

5.) Objem tranzitorní zóny prostaty by měl u zdravého muže tvořit 2 – 5 % celkového objemu prostaty (URL6).

6.) Podobně jako PSAD i Psa denzita tranzitorní zóny prostaty může poukazovat na karcinom prostaty. Je známo, že hodnoty nad 0,35 jsou podezřelé (URL5).

7.) Vzestup hladiny PSA v séru za rok nesmí překračovat hodnotu 0,75 ng/ml (URL4). Pokud je překročena tato hladina, je zde podezření, že muži bude při biopsii diagnostikován karcinom prostaty.

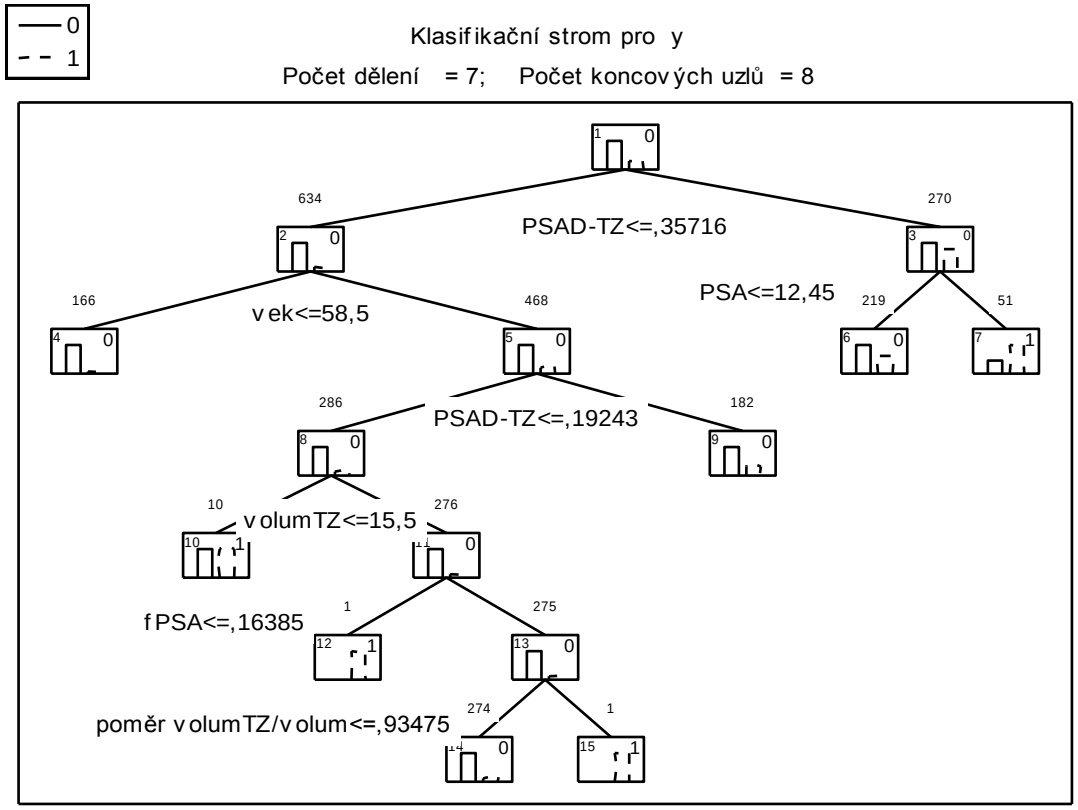
4.1 Klasifikace prvních případů

Ke konstrukci klasifikačních stromů budeme používat metodu CART a výsledky srovnáme s jednotlivými tvrzeními. Naším úkolem není zahrnout do analýzy co nejvíce vysvětlujících proměnných, protože to není podmínkou toho, aby výsledný strom zahrnoval co nejvíce informací. Může se stát, že zahrnutím příliš mnoha proměnných do analýzy se kategorizace do skupin spíše znepřesní až znemožní.

Jestliže zahrneme do analýzy všechny proměnné, metoda CART i metody založené na kvadratické diskriminační analýze mohou vytvořit pouze triviální klasifikaci, která nebude mít žádnou výpovědní hodnotu. Nebo naopak můžeme vytvořit příliš velký strom s mnoha větveními, které již nemají velký smysl.

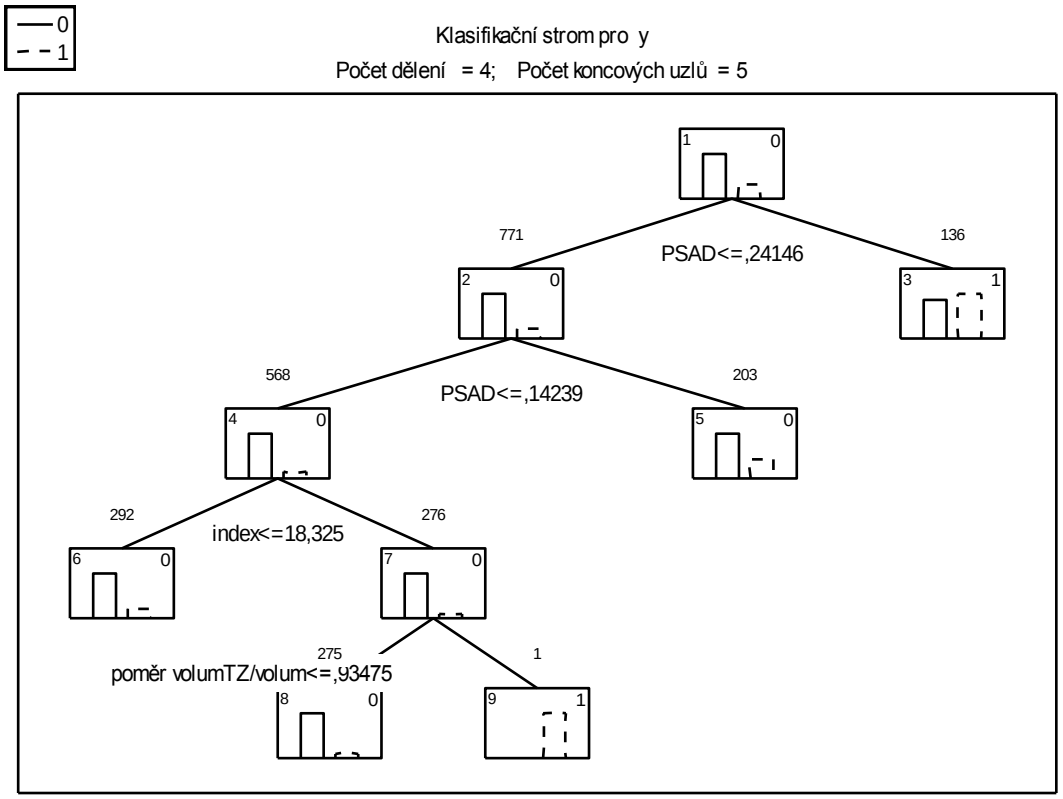
V případě, že v souboru zanecháme všechny vysvětlující proměnné, dostaneme sedm větví stromu (Obr. 4.1). Nejzajímavější je hned první a druhé větvení, kdy klasifikační strom poukazuje na důležitost prediktoru PSAD tranzitorní zóny. Je patrné, že pokud je u pacienta PSAD tranzitorní zóny vyšší než 0,357 a zároveň hodnota PSA v séru vyšší než 12,45 ng/ml, je více pravděpodobné, že pacient bude mít diagnostikován karcinom prostaty. Dostáváme se tedy k tvrzením specialistů, kteří poukazují na zvýšené riziko karcinomu prostaty, pokud hladina PSA v séru překročí hodnotu 10 ng/ml a pro bližší specifikaci v této tzv. šedé zóně (hodnoty kolem

10 ng/ml) využívají právě některé z hodnot PSA denzity (i denzity tranzitorní zóny) či v případě rebiopsií PSA velocity, přičemž právě hodnota PSAD tranzitorní zóny nesmí přesáhnout hladinu 0,35. Náš klasifikační strom se shoduje s těmito předpokládanými prediktory. V dalších větveních zmíníme ještě věk, kdy věk vyšší 58,5 let byl identifikátorem spolu s dalšími hodnotami prediktorů k diagnostice karcinomu prostaty.



Obr. 4.1 Klasifikační strom pro data1 pouze s prvními záznamy, metoda CART přímé ukončení – zlomek objektů 0,065, minimu objektů tříd (13;152)

Abychom dostali i další zákonitosti ukryté v klasifikaci, odebereme nejdůležitější prediktor PSAD tranzitorní zóny a podíváme se na klasifikaci takového stromu (Obr. 4.2).

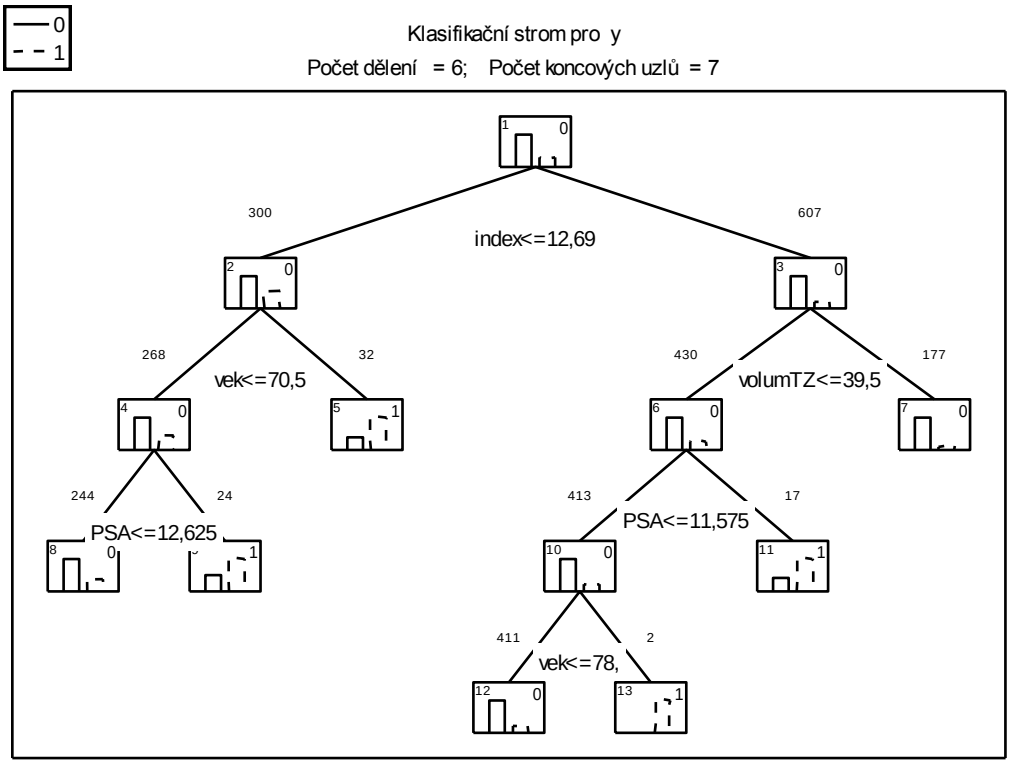


Obr. 4.2 Klasifikační strom pro data1 pouze s prvními záznamy bez PSAD-TZ, metoda CART přímé ukončení – zlomek objektů 0,1, minimu objektů tříd (20;234)

Klasifikační strom Obr. 4.2 určil jako významného ukazatele pro diagnostikování karcinomu prostaty hladinu PSAD vyšší než 0,24 pro pacienty všech věkových kategorií. Pro pacienty s nižší hodnotou PSAD je potom rozhodující hodnota indexu a poměr objemů prostaty, ovšem v tomto klasifikačním stromu tato hodnota není směrodatná – do diagnostikovaných zde v koncovém uzlu č. 9 spadá pouze jeden pacient. Pro zjištění přesnějších hodnot indexu vypustíme z vysvětlujících proměnných i hodnotu PSAD.

Jako podezřelé hodnoty indexu se uvádí hodnoty pod 0,2, tedy podíl volného PSA tvoří 20 % celkového PSA v séru. Klasifikační strom Obr. 4.3 klasifikoval případy podle hraniční hodnoty 12,7 %, kdy pro případy s nižší hodnotou indexu jsou stěžejní dále věk a hodnota PSA. Jestliže má pacient zároveň index nižší 12,7 % a věk nad 70 let, je pravděpodobné, že mu bude diagnostikován karcinom prostaty. Pokud pacient patří do kategorie s indexem nižším než 12,7 % a věkové kategorie pod 70 let, stále nemá vyhráno, pokud má zároveň i hladinu PSA v séru vyšší než 12, 7 ng/ml. I v tomto případě mu pravděpodobně bude diagnostikován karcinom prostaty. Pravá větev stromu nám nedává žádné nové informace o

kategorizaci případů. Udává pouze jako podezřelou hodnotu PSA nad 11,6 ng/ml v případě, že je objem tranzitorní zóny nižší než 39,5 cm^3 a index vyšší než 12,69 %.

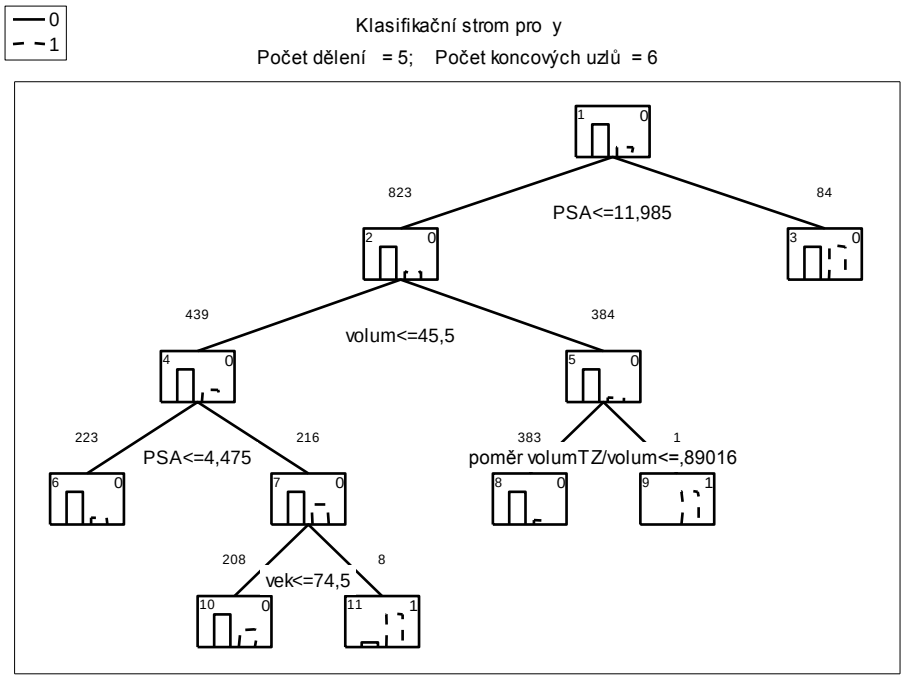


Obr. 4.3 Klasifikační strom pro data1 pouze s prvními záznamy bez PSAD a PSAD-TZ, metoda CART přímé ukončení – zlomek objektů 0,07, minimu objektů tříd (14;164)

Pro postihnutí i dalších zákonitostí (především poměru objemů, ke kterému jsme ještě neměli možnost se vyjádřit) nyní vynechejme hodnotu indexu. Můžeme vidět, že klasifikační strom (Obr. 4.4) již ztratil schopnost kategorizovat případy s hladinou PSA nad 11,9 ng/ml. Histogram případů v obou skupinách je poměrně vyrovnaný a tedy bez dalších znalostí o případech nejsme schopni určit správně jejich třídu zařazení.

Vynechání tolika vysvětlujících proměnných nemělo žádný vliv ani na kategorizaci podle objemů prostaty. Jeden případ, kterému by byl diagnostikován karcinom prostaty v případě vyššího objemu prostaty než 45cm^3 a zároveň s poměrem objemů nad 0,89, nemá v rámci celého souboru vyšší význam.

Při vynechání i dalších proměnných jsme nezískali žádné zlepšení v klasifikaci souboru, nenašli jsme tedy žádnou spojitost s diagnózou karcinomu prostaty a objemem prostaty, objemem tranzitorní zóny či s jejich poměrem.



Obr. 4.4 Klasifikační strom pro data1 pouze s prvními záznamy bez PSAD, PSAD-TZ a indexu, metoda CART přímé ukončení – zlomek objektů 0,05, minimu objektů tříd (10;117)

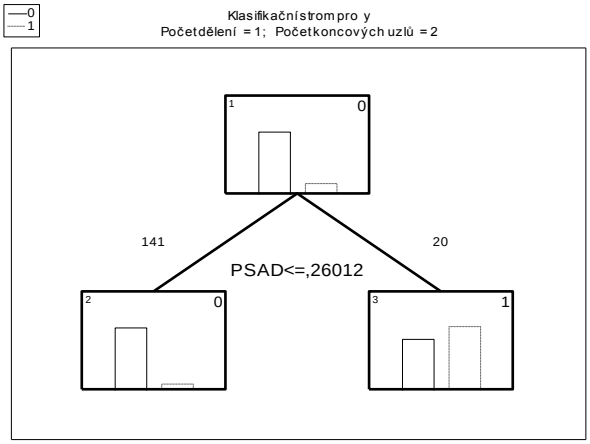
Pro další připravené datové soubory vycházejí klasifikační stromy stejně nebo velice podobně, odstranění některých případů ze souboru tedy nemělo vliv pro klasifikaci prvních případů. (viz Příloha)

4.2 Klasifikace rebiopsií

Pro diagnostikování karcinomu prostaty nám vyšly významné některé hodnoty vysvětlujících proměnných, jako PSA vyšší než 12 ng/ml, PSA denzita tranzitorní zóny nad 0,35 apod. Uvidíme, jak tomu bude v případě rebiopsií, kdy pacient byl na vyšetřeních s určitými obtížemi již podruhé. Může se stát, že nám vyjdou pro kategorizaci významné jiné vysvětlující proměnné či jiné hodnoty těchto proměnných.

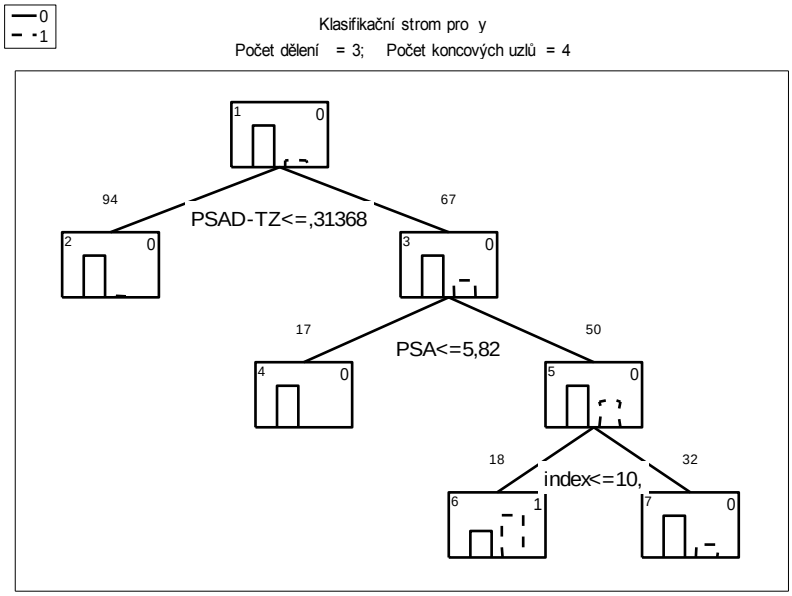
V prvním kroku ponecháme ke klasifikaci všechny vysvětlující proměnné, tentokrát máme k dispozici i hodnotu PSA velocity. Na Obr. 4.5 vidíme, že ke klasifikaci postačila pouze jediná hodnota vysvětlující proměnné a to hodnota PSA denzity. V případě, že je PSAD

u pacienta vyšší než 0,26, je pravděpodobné, že mu bude diagnostikován karcinom prostaty a měl by být poslán na biopsii.

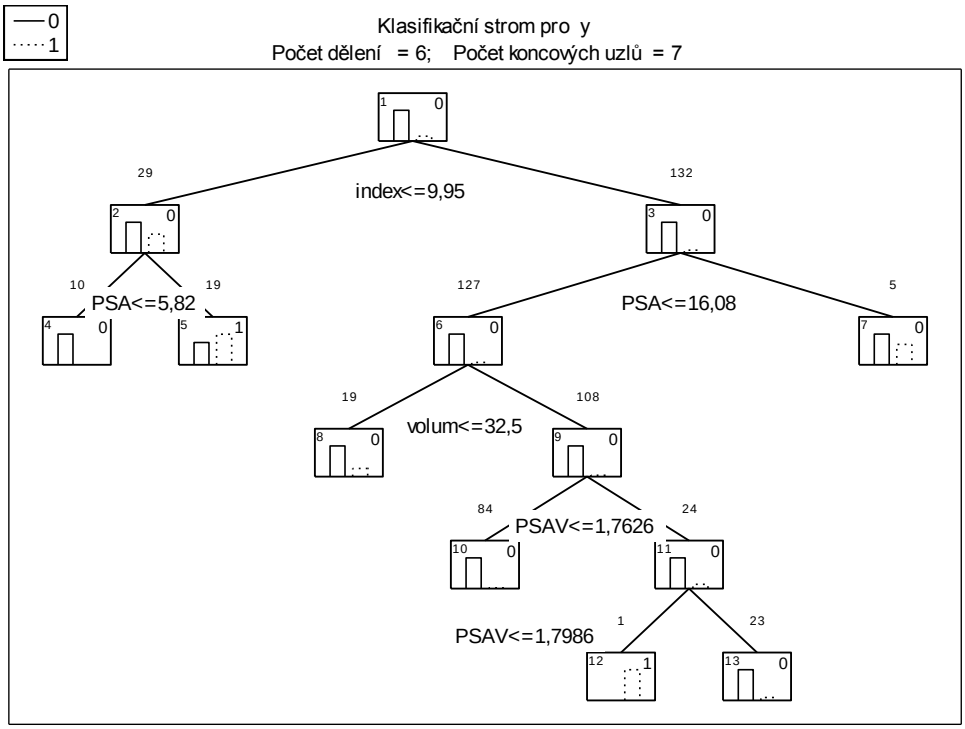


Obr. 4.5 Klasifikační strom pro data1 pouze s rebiopsiemi pro všechny proměnné, metoda CART přímé ukončení – zlomek objektů 0,05, minimu objektů tříd (1;43)

Po vypuštění z klasifikace pouze prediktoru PSA denzity jsme získali klasifikaci s významnou vysvětlující proměnnou PSAD tranzitorní zóny (PSAD_TZ). Pro pacienty s PSAD_TZ nad 0,31 a zároveň s PSA nad 5,8 ng/ml a hodnotou indexu pod 10 % je pravděpodobné, že jim bude diagnostikován karcinom prostaty.



Obr. 4.6 Klasifikační strom pro data1 pouze s rebiopsiemi,bez PSAD metoda CART přímé ukončení – zlomek objektů 0,03, minimu objektů tříd (0;26)

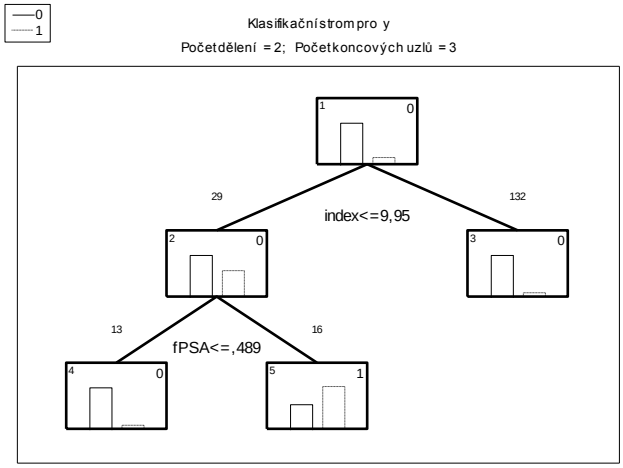


Obr. 4.7 Klasifikační strom pro data1 pouze s rebiopsiemi,bez PSAD a PSAD-TZ metoda CART přímé ukončení – zlomek objektů 0,02, minimu objektů tříd (0;17)

Odebrání další vysvětlující proměnné nám v klasifikaci případů významně nepomohlo. Mírně se snížila pouze hraniční hodnota indexu, přičemž pacienti s hodnotou indexu pod 9,95 % a zároveň PSA vyšším než 5,8 ng/ml jsou zařazeni do třídy, kdy jim je diagnostikován karcinom prostaty. Další větvení nemá smysl uvádět z hlediska zastoupení prvků v jednotlivých třídách vzhledem k celému datovému souboru.

V případě, že ještě odebereme prediktor PSA, dostaneme poměrně jednoduchý klasifikační strom (Obr. 4.8). Karcinom prostaty je diagnostikován pacientům s hodnotami indexu pod 9,95 % a zároveň s hodnotami volného PSA v séru nad 0,49 ng/ml.

Další změny v počtu vysvětlujících proměnných nepřinesly zajímavé klasifikace. Nezjistili jsme žádné přímé souvislosti diagnostikování karcinomu prostaty s hodnotami PSA velocity či s hodnotami objemu prostaty, objemu tranzitorní zóny či jejich poměru. Jediný případ uvedený u Obr. 4.7 v rámci celého souboru není významný.



Obr. 4.8 Klasifikační strom pro data1 pouze s rebiopsiemi,bez PSA, PSAD a PSAD-TZ metoda CART přímé ukončení – zlomek objektů 0,02, minimu objektů tříd (0;17)

V případě rebiopsií jsme dostávali ve všech připravených souborech stejné či podobné výsledky klasifikací, pouze u datového souboru č.4 (kdy jsme provedli největší „ořezání“ souboru) nám vyšly dva diagramy odlišně.

Odebereme-li z datového souboru PSAD, dostaneme klasifikační strom (příloha P21), který určil za rozhodující prediktor hodnotu PSAD tranzitorní zóny prostaty. Jestliže pacient má hodnoty PSAD_TZ pod 0,31, pak je kategorizován do skupiny pacientů, kterým nebyl diagnostikován karcinom prostaty. Pokud má jedinec hodnoty PSAD_TZ vyšší než udaná hranice a zároveň hladina PSA v séru u něj překračuje hodnotu 15 ng/ml, pak patří do rizikové skupiny s karcinomem prostaty. U pacientů s hladinou PSA v rozmezí 5,8 – 15 ng/ml a s PSAD_TZ nad 0,31 je rozhodující hodnota indexu. V případě že je index nižší než 10,29 %, bude jim pravděpodobně diagnostikován karcinom prostaty, v opačném případě pacienti spadají do kategorie bez diagnózy karcinomu prostaty.

V případě odebrání z vysvětlujících proměnných PSAD a PSAD_TZ (příloha P22) metoda CART vybrala za stěžejní hodnotu PSA. Pacienti s hladinou PSA v séru nad 16 ng/ml byli zařazeni do kategorie s diagnostikovaným karcinomem prostaty. V případě nižší hladiny PSA než je udaná hranice, nemá smysl kategorizaci podle dalších prediktorů dále rozebírat. V tomto případě spadá do skupiny s diagnostikovaným karcinomem pouze jeden případ a ostatní jsou řazeni do kategorie bez pozitivní diagnózy.

Závěr

Klasifikační a regresní stromy jsou účinnou metodou k odhadu budoucích hodnot závislé proměnné na základě známých hodnot vysvětlujících (predikujících) proměnných.

U klasifikačních stromů vytvořených např. metodou CART nebo CHAID je listu přiřazena určitá hodnota klasifikační funkce. Regresní strom vypočtený metodou CART definuje regresní funkci, která je uvnitř každé množiny odpovídající listu konstantní. (Klaschka J, Antoch J., 1996)

Klasifikační a regresní stromy jsou vhodné k nalezení souvislostí mezi prediktory a vysvětlovanou proměnou v případě, že naše data nesplňují podmínky pro parametrické metody. Výhodou klasifikačních stromů jsou malé nároky na podobu vstupních dat. Jediné, co musíme dodržet, aby výsledné klasifikace případů do skupin byly věrohodné, je vyšší počet případů v datovém souboru. Datový soubor, který by obsahoval málo případů, by sice popisoval určité vztahy uvnitř tohoto souboru, ale nebylo by možné tyto zákonitosti vztáhnout na celou populaci. Další výhodou této metody je malá citlivost vůči odlehlým hodnotám, o čemž jsme se přesvědčili užitím metody CART na různě „ořezané“ datové soubory. Výsledky klasifikačních stromů se shodovaly ve většině případů. Ve chvíli, kdy jsme dostali mírně odlišný klasifikační strom, měli jsme podobné hraniční hodnoty pro vysvětlující proměnné.

Užili jsme metodu CART ke klasifikaci prvních případů, kdy pacienti došli na vyšetření do fakultní nemocnice v Olomouci s určitými problémy či v rámci preventivního opatření, ale také pro klasifikaci prvních rebiopsií. Rebiopsie se prováděly u mužů, kteří již jednou byli v nemocnici s obtížemi a jsou sledováni nebo jim již jednou byl zjištěn nález na prostatě. Tato data byla nasbírána v letech 2006 až 2012.

Zjistili jsme, že velmi důležitou proměnou v rámci klasifikace je PSA denzita tranzitorní zóny prostaty (PSAD_TZ), která vyjadřuje poměr celkové hladiny PSA v séru vůči objemu tranzitorní zóny prostaty. Tranzitorní neboli přechodná zóna prostaty je místem, kde se tvoří největší množství prostatického specifického antigenu. V naučných článcích je udávanou hranicí hodnota 0,35 ng/ml^2. S vyššími hodnotami je i vyšší riziko karcinomu prostaty. Metoda CART tuto hraniční hodnotu potvrdila, přičemž důležitá byla zároveň i vyšší

hodnota prostatického specifického antigenu. Z našich výsledků vyplývá, že pokud má jedinec hodnotu PSA denzity tranzitorní zóny prostaty vyšší než 0,357 a zároveň množství PSA v séru přesahuje hodnotu 12,45 ng/ml, spadá do rizikové skupiny pacientů, u nichž je více pravděpodobné, že při následné biopsii bude zjištěn nález karcinomu prostaty.

V případě, že nemáme k dispozici hodnotu objemu tranzitorní zóny prostaty k výpočtu PSAD_TZ, ale známe celkový objem prostaty, ke kategorizaci případů do skupin využijeme PSA denzitu (PSAD). Zjistili jsme, že pokud hodnota PSAD přesáhne hranici 0,24, řadíme případy do rizikové skupiny pacientů s možným nálezem karcinomu při následné biopsii. Jestliže chceme zjistit rizikovost nálezu karcinomu při rebiopsiích, hraniční hodnota byla stanovena na 0,26, přičemž tento prediktor je v tomto případě považován za nejdůležitější klasifikátor případů do jednotlivých skupin. Udávanou hranici pro PSAD 0,15 jsme tedy mírně překročili.

Může se stát, že u pacientů nemáme k dispozici hodnoty objemu prostaty či tranzitorní zóny prostaty, ale pravděpodobně známe hodnoty PSA, volného PSA a věk pacienta. Pro klasifikaci využijeme indexu, který vyjadřuje poměr volného a celkového PSA. Pacient spadá do rizikové skupiny s možným nálezem karcinomu prostaty v případě, že hodnota indexu je nižší než 12,7 % a věk pacienta přesahuje 70 let. Jestliže muž spadá do nižší věkové kategorie, pak závisí na hodnotě PSA, která nesmí přesáhnout hodnotu 12,6 ng/ml, aby pacient nepatřil do rizikové skupiny.

U rebiopsií jsme zaznamenali mírné snížení rizikových hodnot prediktorů. Do rizikové skupiny pacientů, kterým by mohl být při rebiopsii nalezen karcinom prostaty, spadají všichni s hodnotami PSAD tranzitorní zóny prostaty nad 0,31 ng/ml^2 a zároveň s PSA nad 5,8 ng/ml a indexem pod 10 %.

Nenalezli jsme žádné souvislosti v klasifikaci pacientů a prediktory jako je PSA velocita, poměr objemů prostaty a podobně. Doporučovali bychom lékařům, aby za signifikantní prediktory považovali hodnoty PSA denzity a PSA denzity tranzitorní zóny prostaty, index a hladinu PSA v séru. V případě, že budou u pacientů překročeny limitní hodnoty prediktorů, měli by být posláni na biopsii pro vyloučení karcinomu prostaty.

Při použití logistické regrese Fačevicová K. zjistila, že pravděpodobnost výskytu karcinomu prostaty je tím vyšší, čím vyšší je věk pacienta, či hladina PSA v krvi a čím nižší je objem prostaty (Fačevicová K., 2012). My jsme nenalezli žádnou souvislost mezi objemem

prostaty a klasifikací případů do skupiny s rizikem nálezu karcinomu prostaty při následné biospii, ale souvislost s věkem či s PSA ano.

Souhrn hodnot prediktorů pro zařazení pacientů do rizikové skupiny v porovnání s udávanými hodnotami nalezneme v tabulce Tab.11.

Prediktor	**jednotka**	**Hodnota: udávaná**	**pro 1.případy**	**pro rebiopsie**
PSAD_TZ	ng/ml^2	$\geq 0{,}35$	$\geq 0{,}36$	$\geq 0{,}31$
PSAD	ng/ml^2	$\geq 0{,}15$	$\geq 0,\ 24$	$\geq 0,\ 26$
PSA	ng/ml	≥ 10	$\geq 12{,}6$	$\geq 5{,}8$
Věk	rok	65 -80	≥ 70	nehraje roli
Index	%	≤ 20	$\leq 12{,}7$	≤ 10

Tab. 11 Hodnoty nejdůležitějších prediktorů pro klasifikaci do rizikové skupiny

Seznam použité literatury

[1] ANTOCH J., Klasifikace a regresní stromy. [online]. Robust, 1988
Dostupné z www:<http:// www.statspol.cz/robust/1988_antoch88.pdf >

[2] FAČEVICOVÁ K., Použití logistické regrese pro diagnostiku výskytu rakoviny prostaty. Olomouc, 2012, diplomová práce (Mgr.). UNIVERZITA PALACKÉHO V OLOMOUCI. Přírodovědecká fakulta

[3] HOLČÍK J., Analýza a klasifikace dat. [online]. Brno: Akademické nakladatelství CERM, s.r.o., [2012], ISBN 978-80-7204-793-2, první vydání.
Dostupné z www: <http://www.iba.muni.cz/res/file/ucebnice/holcik-analyza-klasifikace-dat .pdf>

[4] JARKOVSKÝ J., LITTNEROVÁ S., DUŠEK L., HARUŠTIAKOVÁ D., Brno: Vícerozměrné statistické metody v biologii. Akademické nakladatelství CERM, s.r.o., [2012], ISBN 978-80-7204-791-8, první vydání.
Dostupné z www: <https://www.iba.muni.cz/res/file/ucebnice/jarkovsky-vicerozmerne-statisti cke-metody.pdf >

[5] JAROLÍM L., Stanovení diagnózy karcinomu prostaty a příslušná vyšetření. [online]. 2012
Dostupné z www: <*http://*www.gentlemanplus.cz/karcinom-prostaty~stanoveni-diagnozy-karcinomu-prostaty-a-prislusna-vysetreni/>

[6] KEPRTA S., Nebinární klasifikační stromy. [online]. Robust, 1994
Dostupné z www:< http://www.statspol.cz/robust/1994_keprta94.pdf >

[7] KLASCHKA J., KOTRČ E., Klasifikační a regresní lesy. [online]. Robust, 2004
Dostupné z www:<http:// www.statspol.cz/robust/robust2004/klaschka.pdf >

[8] KLASCHKA J., ANTOCH J., Jak rychle pěstovat stromy. [online]. Robust, 1996
Dostupné z www:< http://www.statspol.cz/robust/1996_klasch96.pdf >

[9] KOMPRDOVÁ K., Rozhodovací stormy a lesy. [online], Brno: Akademické nakladatelství CERM, s.r.o., [2012], ISBN 978-80-7204-785-7, první vydání.
Dostupné z www:< http://www.iba.muni.cz/res/file/ucebnice/komprdova-rozhodovaci-stromy-lesy.pdf>

[10] LUKEŠ M., Karcinom prostaty. [online]. Androgeos, [2013], ISBN 978-80-254-1859-8
Dostupné z www: <http://www.urologieprostudenty.cz>

[11] LUTEROVÁ A., Modely dynamiky nádorových onemocnění. [online]. Brno, 2012
Dostupné z www: <http://is.muni.cz/th/356907/prif_b/Luterova.pdf>

[12] MELOUN M., Počítačová analýza víerozměrných dat v oborech přírodních, technických a společenských věd. [online]. 2011
Dostupné z www:< http://www.crr.vutbr.cz/system/files/prezentace_05_1106_07a.pd>

[13] PAVLÍK T., DUŠEK L., Biostatistika. Brno: Akademické nakladatelství CERM, s.r.o., [2012], ISBN 978-80-7204-782-6, první vydání.
Dostupné z www: <http:// www.iba.muni.cz/res/file/ucebnice/pavlik-biostatistika.pdf >

[14] RABUŠIC L., Mnohonásobná lineární regrese. [online]. 2004
Dostupné z www:
<http://www.is.muni.cz/el/1423/podzim2004/SOC418/multipl_**regres**_1.pdf >

[15] SAVICKÝ P., KLASCHKA J., ANTOCH J., Optimální klasifikační stromy. [online]. Robust, 2000
Dostupné z www:< http://www.statspol.cz/robust/2000_savick00.pdf >

[16] ŠAFARČÍK K., PSA a jeho izoformy pro časnou diagnostiku. [online]. 2009
Dostupné z www:<http:// www.europauomo.cz >

URL1: Epidemiologické údaje zhoubných nádorů v České republice. [online]. 2011
Dostupné z www: <http:// http://www.svod.cz/>

URL2: Rakovina prostaty. [online]. 2012
Dostupné z www: <http:// *www*.nemoci.vitalion.cz/rakovina-prostaty/ >

URL3: Program preventivních prohlídek. [online]. 2013
Dostupné z www: <http:// www.linkos.cz/prevence/program-preventivnich-prohlidek/ >

URL4: Státní zdravotní ústav. [online]. 2013, Dostupné z www: <http:// www.szu.cz >

URL5: Diagnostika. [online]. 2009
Dostupné z www: <http://www.urologieprostudenty.cz/diagnostika-3>

URL6: Směrnice pro diagnostiku nezhoubného zvětšení prostaty ve Švédsku. [online]. 2001
Dostupné z www: <http://www.urologiepropraxi.cz/pdfs/uro/2001/03/06.pdf>

URL7: Karcinom prostaty – molekulární podstata, diagnostika a ekonomika prevence. [online]. 2008
Dostupné z www: <http://www.europauomo.cz/pdf/kpmp.pdf>

Další použité zdroje informací:
Přednášky Kubošové K., Pokročilé neparametrické metody, cit. 28.10.2013
Dostupné z www: <http://www.iba.muni.cz/>

Seznam grafů a obrázků

Obr. 1.1 Ilustrační obrázek, Věková struktura populace pacientů s karcinomem prostaty v ČR v letech 1977-2010; (http://www.swod.cz, 2012).....10

Obr. 1.2 Ilustrační obrázek, Incidence karcinomu prostaty a mortalita v ČR v letech 1977-2010; (http://www.swod.cz, 2012).....10

Obr.1.3 Ilustrační obrázek, Predikce incidence karcinomu prostaty v ČR modelováno s využitím inverzních filtrů, pomocí časových řad (Luterová A., 2012).....10

Obr.2.1 Ukázky hustot náhodných veličin s normálním rozdělením. (Pavlík T., Dušek L.,2012).....20

Obr.2.2 Histogramy rozdělení datového souboru postupně podle věku a hodnoty PSA v séru a jejich očekávané normální rozdělení.....25

Obr. 2.3 Histogramy rozdělení datového souboru postupně podle objemu prostaty a objemu tranzitorní zóny a jejich očekávané normální rozdělení.....25

Obr. 2.4 Histogramy rozdělení datového souboru postupně podle indexu a PSA_V a jejich očekávané normální rozdělení.....25

Obr. 2.5 Boxploty jednotlivých vysvětlujících proměnných.....26

Obr. 2.6 Boxploty jednotlivých vysvětlujících proměnných.....26

Obr. 2.7 Histogram rozdělení datového souboru podle indexu a jejich očekávané normální rozdělení, vpravo boxplot vysvětlujících proměnných (věk, index, volum, volumTZ).....30

Obr. 2.8 Krabicové grafy dalších vysvětlujících proměnných (PSA, PSAV, fPSA, PSAD a PSAD_TZ).....30

Obr. 3.1 Deskriptivní model mnohonásobné regrese (Rabušic L., 2004).....34

Obr. 3.2 Kauzální model mnohonásobné regrese (Rabušic L., 2004).....35

Obr. 3.3 Diagram klasifikačního stromu s binárním větvením.....37

Obr. 3.4 Konkrétní příklad diagramu klasifikačního stromu.....37

Obr. 3.5. Datový soubor a rozložení jednotlivých prvků, kde čtverec odpovídá skupině 1 a kolečko 0, přerušované čáry pak symbolizují odhad pro nejlepší možné rozdělení souboru....43

Obr 3.6 Ilustrační obrázek větvení, za podmínky A < 1,5, vpravo ilustrace rozdělení (přerušovaná čára).....44
Obr 3.7 Ilustrační obrázek větvení, za podmínky A < 2,5, vpravo ilustrace rozdělení (přerušovaná čára).....47
Obr 3.8 Ilustrační obrázek větvení, za podmínky d) B < 3,5, vpravo ilustrace rozdělení (přerušovaná čára – možné dělení, plná čára – dříve provedené dělení).....48
Obr 3.9 Diagram klasifikačního stromu ilustračního příkladu.....49
Obr 3.10 Diagram klasifikačního stromu ilustračního příkladu za použití G-kvadrát klasifikace.....50
Obr. 3.11 Prořezávání klasifikačního stromu možností ukončení FACT.....53
Obr. 4.1 Klasifikační strom pro data1 pouze s prvními záznamy, metoda CART přímé ukončení – zlomek objektů 0,065, minimu objektů tříd (13;152).....61
Obr. 4.2 Klasifikační strom pro data1 pouze s prvními záznamy bez PSAD-TZ, metoda CART přímé ukončení – zlomek objektů 0,1, minimu objektů tříd (20;234).....62
Obr. 4.3 Klasifikační strom pro data1 pouze s prvními záznamy bez PSAD a PSAD-TZ, metoda CART přímé ukončení – zlomek objektů 0,07, minimu objektů tříd (14;164).....63
Obr. 4.4 Klasifikační strom pro data1 pouze s prvními záznamy bez PSAD, PSAD-TZ a indexu, metoda CART přímé ukončení – zlomek objektů 0,05, minimu objektů tříd (10;117).....64
Obr. 4.5 Klasifikační strom pro data1 pouze s rebiopsiemi pro všechny proměnné, metoda CART přímé ukončení – zlomek objektů 0,05, minimu objektů tříd (1;43).....65
Obr. 4.6 Klasifikační strom pro data1 pouze s rebiopsiemi,bez PSAD metoda CART přímé ukončení – zlomek objektů 0,03, minimu objektů tříd (0;26).....65
Obr. 4.7 Klasifikační strom pro data1 pouze s rebiopsiemi,bez PSAD a PSAD-TZ metoda CART přímé ukončení – zlomek objektů 0,02, minimu objektů tříd (0;17).....66
Obr. 4.8 Klasifikační strom pro data1 pouze s rebiopsiemi,bez PSA, PSAD a PSAD-TZ metoda CART přímé ukončení – zlomek objektů 0,02, minimu objektů tříd (0;17).....67

Seznam vzorců

(2.1) Výpočet PSAD

(2.2) Výpočet PSAD_TZ

(2.3) Výpočet pro prediktor PSA velocita

(2.4) Výpočetní vztah pro index

(2.5) Hustota náhodné veličiny X

(2.6) Pearsonův korelační koeficient

(2.7) Spearmanův korelační koeficient

(2.8) Výpočet diferencí pořadí pozorovaných hodnot d_i

(2.9) Výpočet Spearmanova korelačního koeficientu pomocí d_i

(2.10) Výpočet Fischerovy z-transformace

(2.11) Výpočet z-skóre

(2.12) Výpočet p-hodnoty

(3.1) Mnohonásobná lineární regrese

(3.2) Skutečná klasifikační chyba

(3.3) Střední kvadratická chyba

(3.4) Klasifikační skóre diskriminační analýzy

(3.5) Výpočet kovariancí

(3.6) Kvadratická nerovnost (kvadratické diskriminační analýzy)

(3.7) Sestavení matice G

(3.8) Vektor h^T *definován pro kvadratickou nerovnost*

(3.9) Konstanta C v kvadratické nerovnosti

(3.10) Kvadratické diskriminační kritérium

(3.11) Výpočet minima Mahalanobisových vzdáleností

(3.12) Bayesovo kritérium

(3.13) Giniho koeficient

(3.14) Giniho koeficient vyjadřený součtem pravděpodobností

(3.15) Výpočet pravděpodobností pro Giniho koeficient

(3.16) Celkový Giniho koeficient

(3.17) Relativní četnost prvků v i-té skupině

(3.18) Marginální četnosti

(3.19) Očekávané četnosti

(3.20) Výpočet chí-kvadrát statistiky

(3.21) Kritérium pro zamítnutí nulové hypotézy na hladině významnosti α

(3.22) Výpočet minima objektů třídy pro ukončení FACT

(3.23) Cost-komplexity kritérium

(3.24) Statistika φ v případě varianty metody nejmenších čtverců pro regresní stromy

(3.25) Výpočet $\bar{Y}_j$ pro statistiku φ v případě varianty metody nejmenších čtverců

(3.26) Statistika φ varianty nejmenších absolutních odchylek pro regresní stromy

Seznam tabulek

Tab. 1 Riziko karcinomu prostaty závislé na vyšetření konečníkem a hladiny celkového PSA v séru (Jarolím L., 2012)

Tab.2 Průměrné hodnoty v datovém souboru data1 a jejich směrodatné odchylky

Tab.3 Spearmanův korelační koeficient pouze pro první případy a první rebiopsie

Tab.4 Spearmanovy korelace pouze pro první případy

Tab.5 Průměrné hodnoty v datovém souboru data2 a jejich směrodatné odchylky

Tab.6 Průměrné hodnoty v datovém souboru data3 a jejich směrodatné odchylky

Tab.7 Průměrné hodnoty v datovém souboru data4 a jejich směrodatné odchylky

Tab.8 Spearmanovy korelace pro datový soubor data4

Tab.9 Zastoupení prvních vyšetření a prvních rebiopsií v rámci připravených souborů

Tab.10 Srovnání jednotlivých metod tvorby klasifikačních stromů

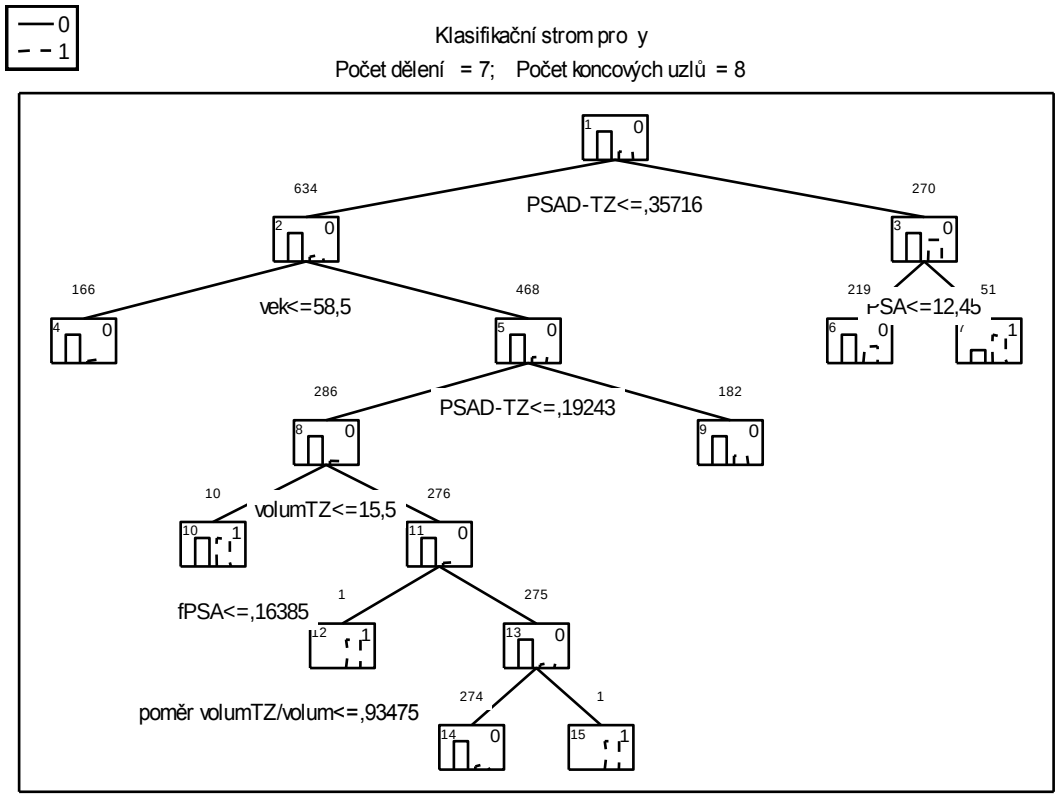
Tab. 11 Hodnoty nejdůležitějších prediktorů pro klasifikaci do rizikové skupiny

Přílohy

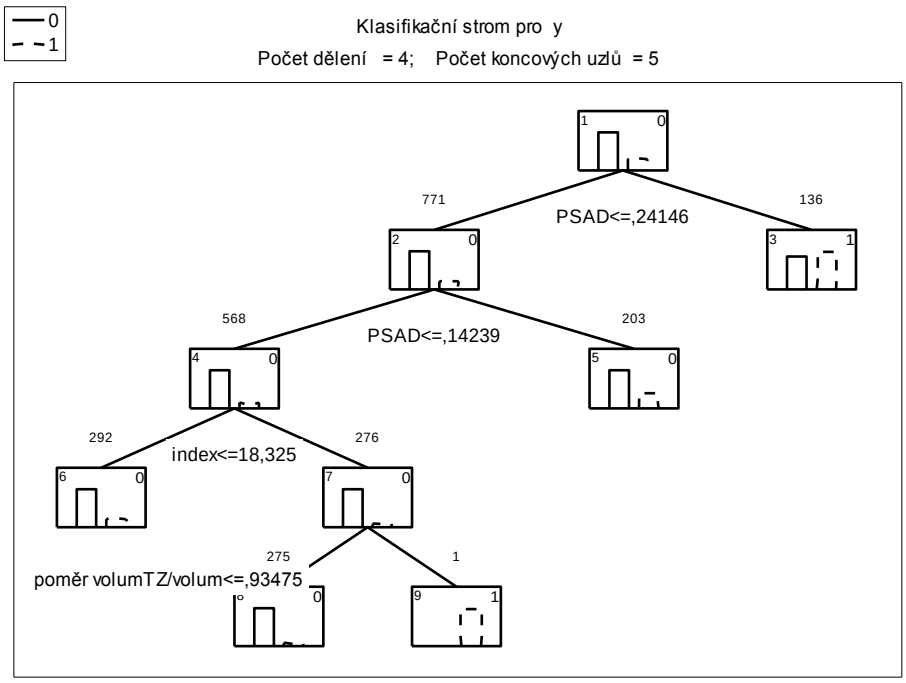
Klasifikační stromy pro první pozorování

Pro srovnání jsme sestavili klasifikační stromy pro různě „ořezané “ datové soubory. Na následujících stranách jsou k vidění jednotlivé diagramy.

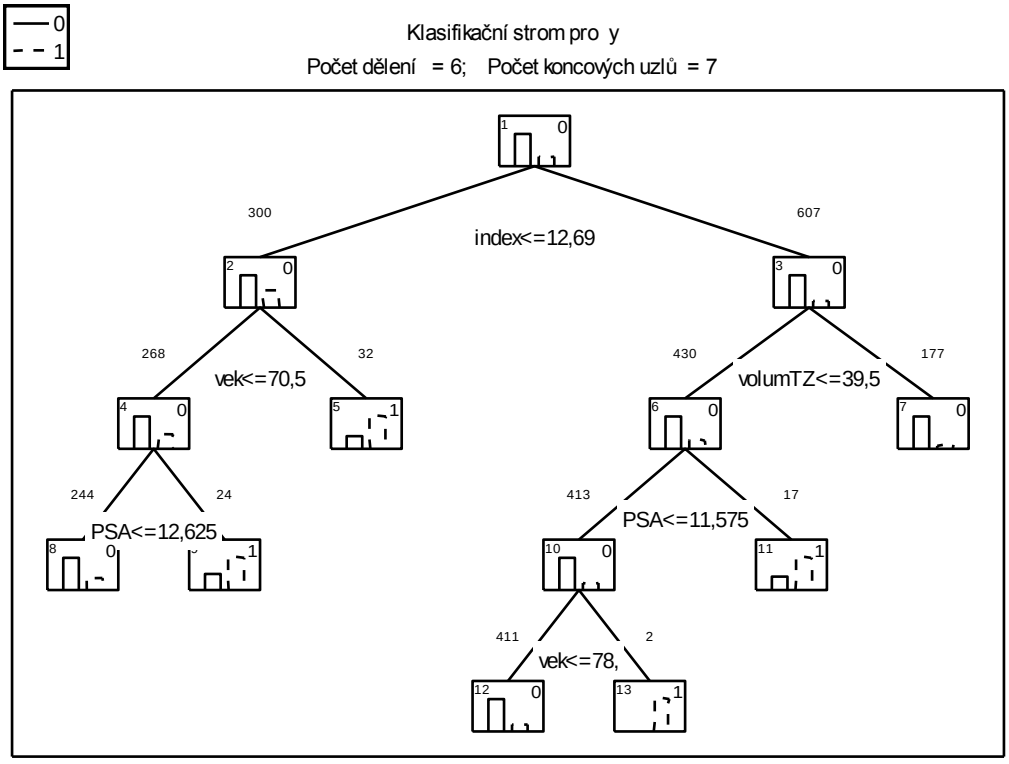
Data2



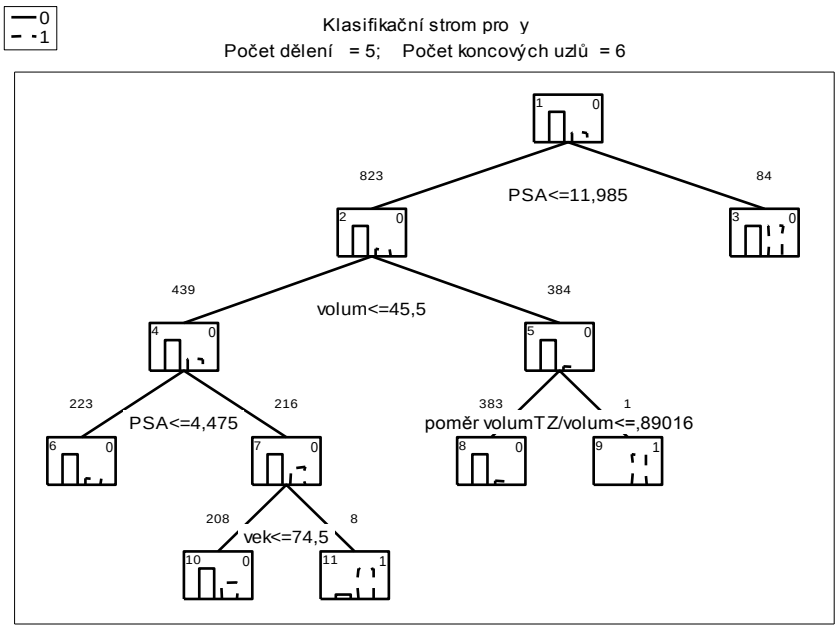
Obr. P1 Klasifikační strom pro data2 pouze s prvními záznamy, CART přímé ukončení – FACT = 0,065, minimu objektů tříd (13;152)



Obr. P2 Klasifikační strom pro data2 pouze s prvními záznamy bez PSAD-TZ, metoda CART přímé ukončení – zlomek objektů 0,1, minimu objektů tříd (20;234)

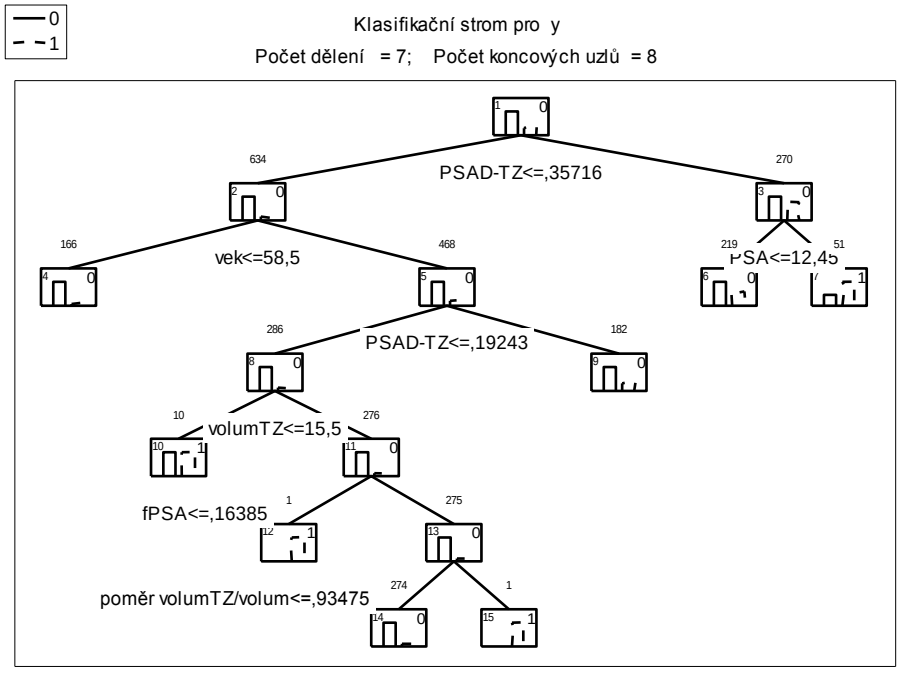


Obr. P3 Klasifikační strom pro data2 pouze s prvními záznamy bez PSAD a PSAD-TZ, metoda CART přímé ukončení – zlomek objektů 0,07, minimu objektů tříd (14;164)

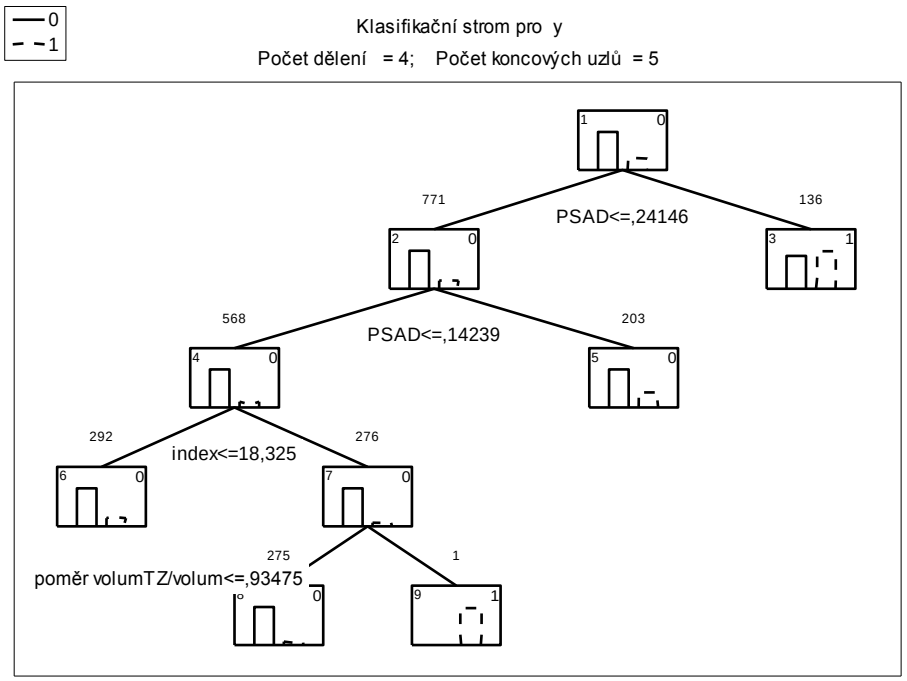


Obr. P4 Klasifikační strom pro data2 pouze s prvními záznamy bez PSAD, PSAD-TZ a indexu, metoda CART přímé ukončení – zlomek objektů 0,05, minimu objektů tříd (10;117)

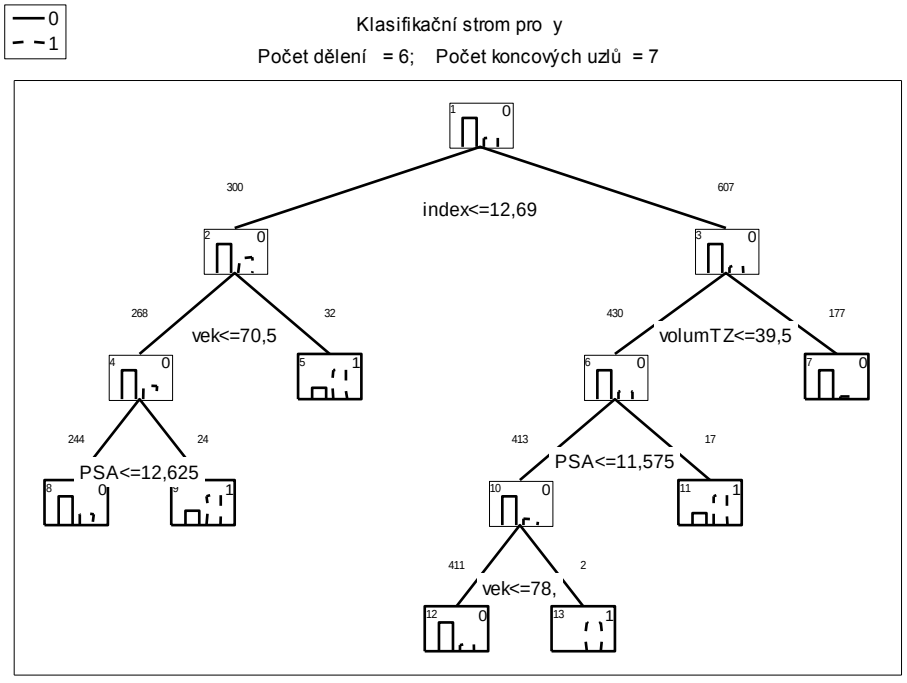
Data3



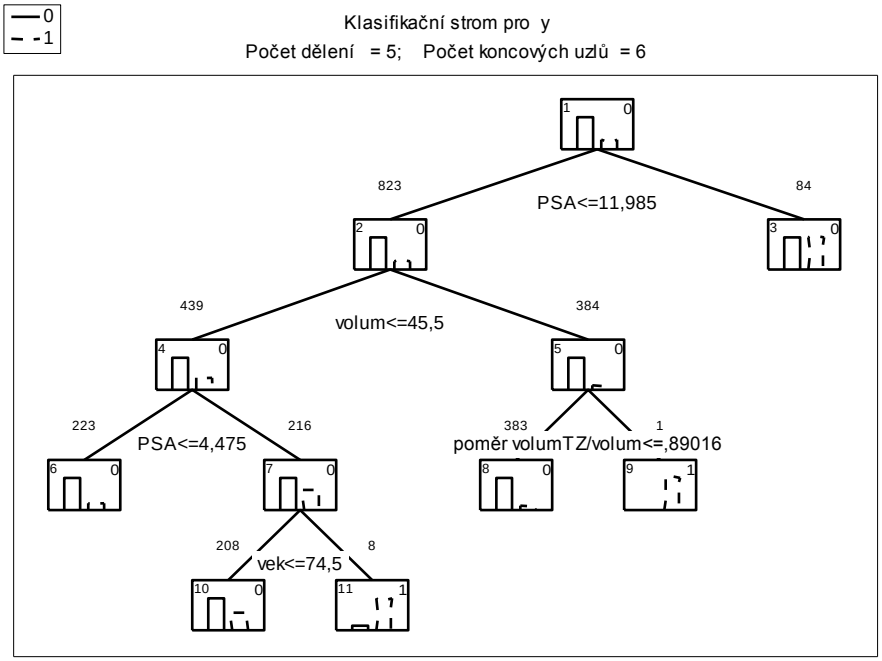
Obr. P5 Klasifikační strom pro data3 pouze s prvními záznamy, CART přímé ukončení – FACT = 0,065, minimu objektů tříd (13;152)



Obr. P6 Klasifikační strom pro data3 pouze s prvními záznamy bez PSAD-TZ, metoda CART přímé ukončení – zlomek objektů 0,1, minimu objektů tříd (20;234)

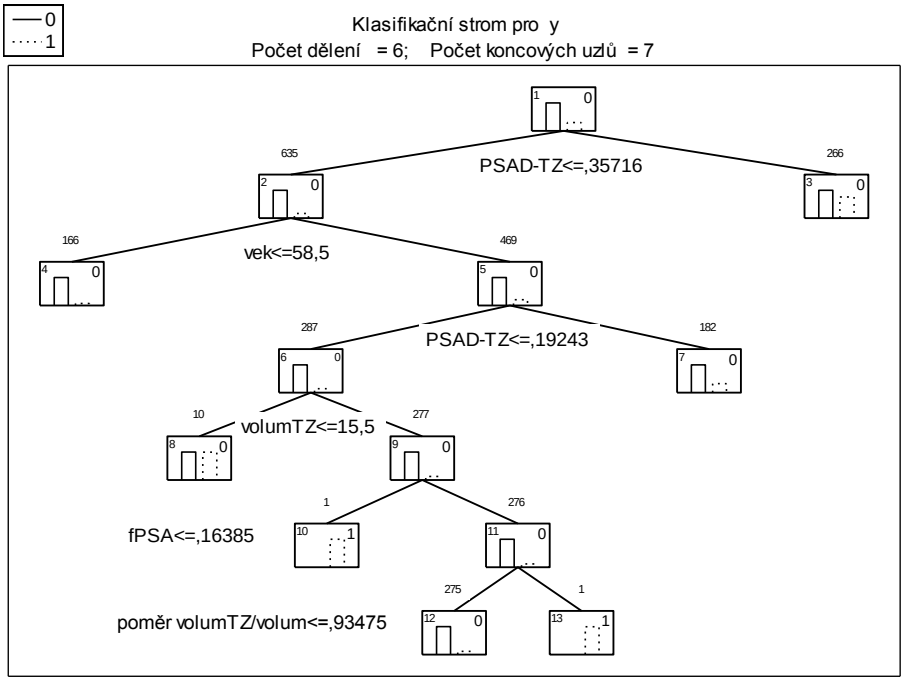


Obr. P7 Klasifikační strom pro data3 pouze s prvními záznamy bez PSAD a PSAD-TZ, metoda CART přímé ukončení – zlomek objektů 0,07, minimu objektů tříd (14;164)

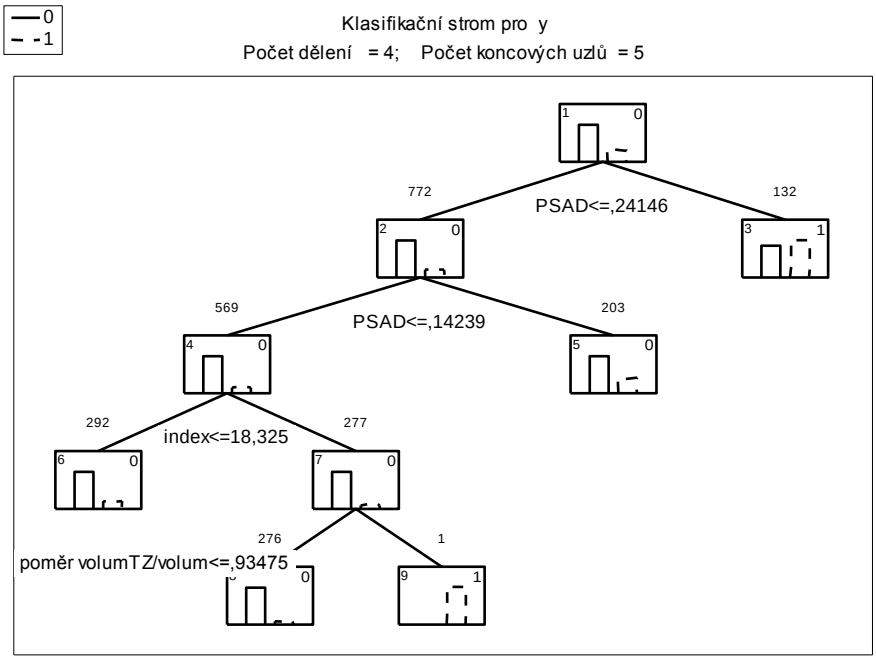


Obr. P8 Klasifikační strom pro data3 pouze s prvními záznamy bez PSAD, PSAD-TZ a indexu, metoda CART přímé ukončení – zlomek objektů 0,05, minimu objektů tříd (10;117)

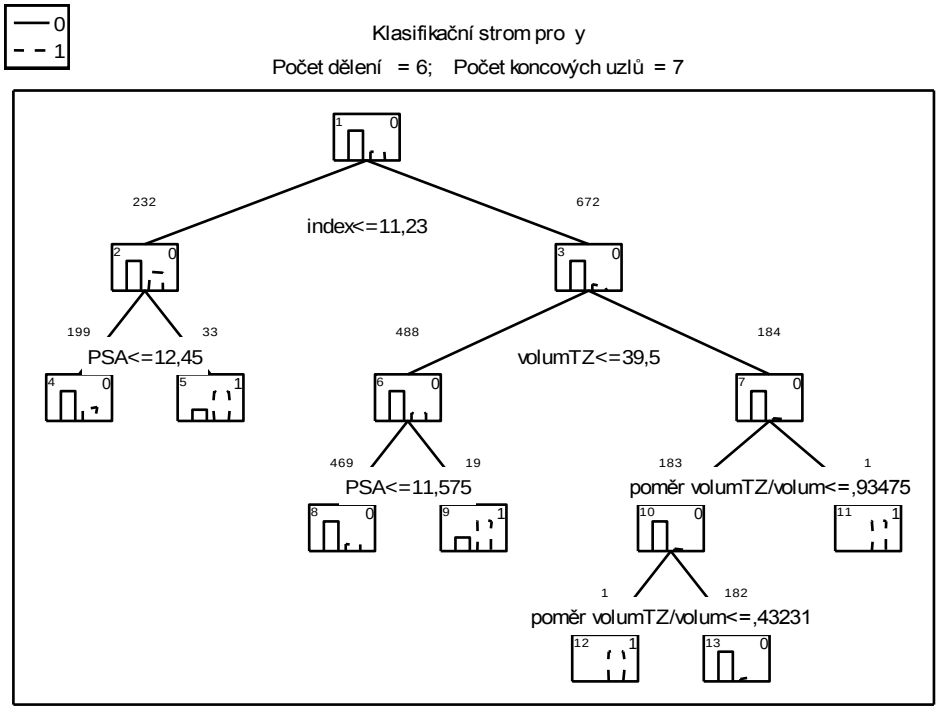
Data 4



Obr. P9 Klasifikační strom pro data4 pouze s prvními záznamy, CART přímé ukončení – FACT = 0,065, minimu objektů tříd (13;153)



Obr. P10 Klasifikační strom pro data4 pouze s prvními záznamy bez PSAD-TZ, metoda CART přímé ukončení – zlomek objektů 0,1, minimu objektů tříd (20;236)

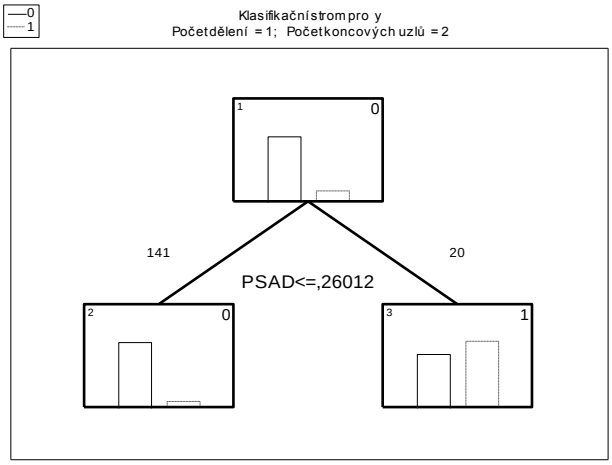


Obr. P11 Klasifikační strom pro data4 pouze s prvními záznamy bez PSAD a PSAD-TZ, metoda CART přímé ukončení – zlomek objektů 0,06, minimu objektů tříd (12;141)

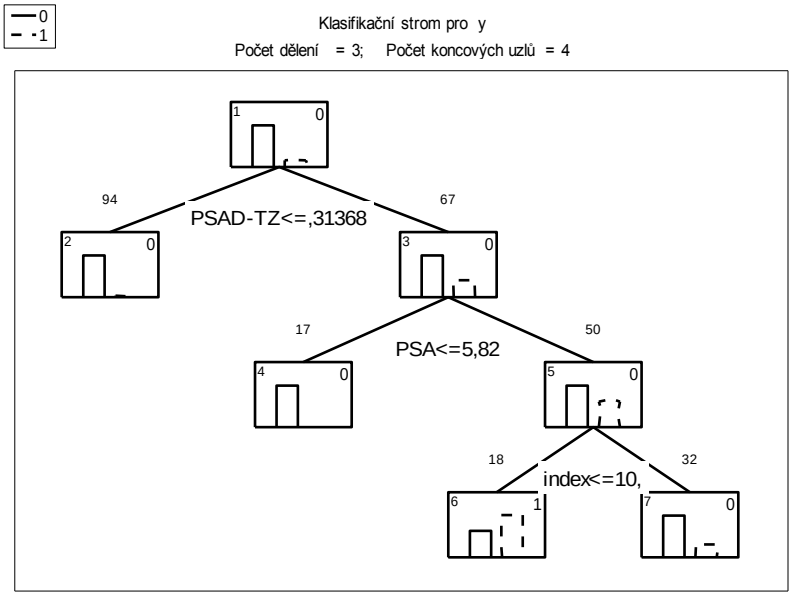
Vypuštěním další proměnné index jsme získali rozsáhlý strom bez zajímavých výsledků klasifikace.

Klasifikační stromy pro rebiopsie

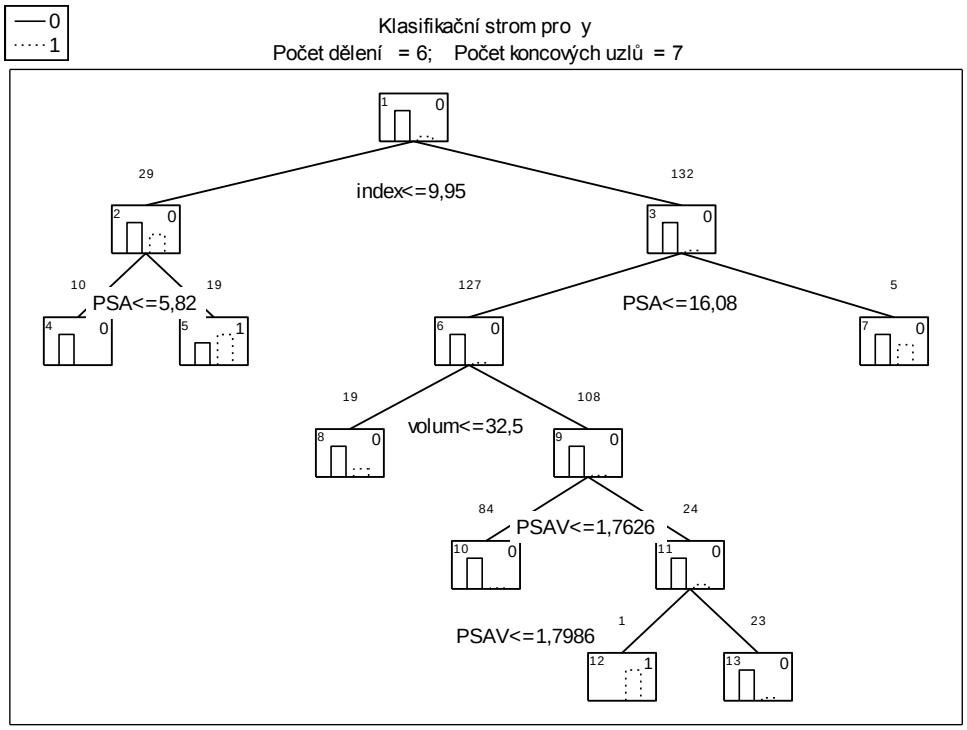
Data2



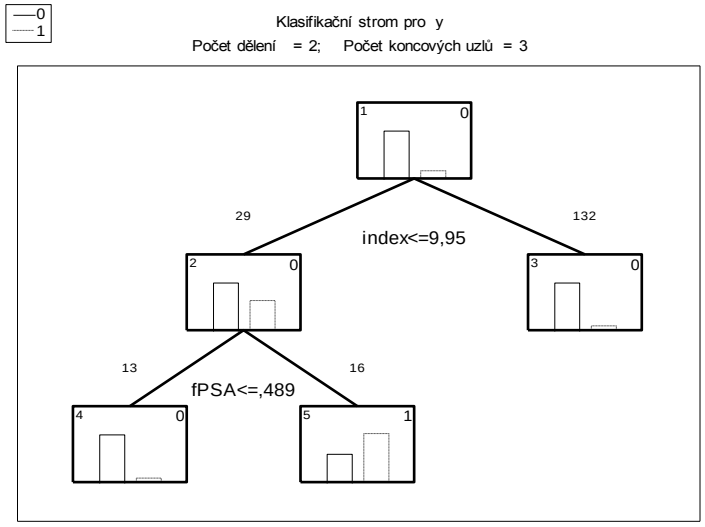
Obr. P12 Klasifikační strom pro data2 pouze s rebiopsiemi pro všechny proměnné, metoda CART přímé ukončení – zlomek objektů 0,05, minimu objektů tříd (1;43)



Obr. P 13 Klasifikační strom pro data2 pouze s rebiopsiemi,bez PSAD metoda CART přímé ukončení – zlomek objektů 0,03, minimu objektů tříd (0;26)

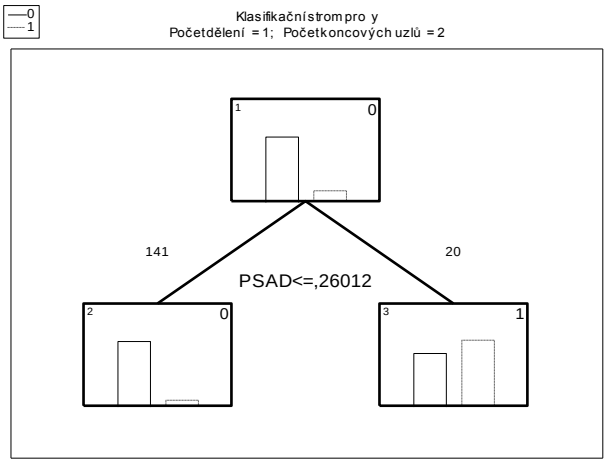


Obr. P14 Klasifikační strom pro data2 pouze s rebiopsiemi,bez PSAD a PSAD-TZ metoda CART přímé ukončení – zlomek objektů 0,02, minimu objektů tříd (0;17)

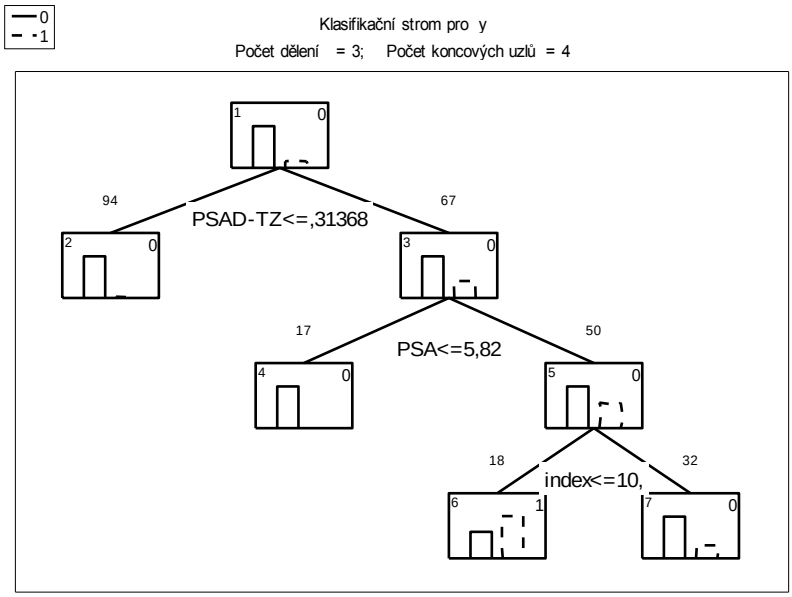


Obr. P15 Klasifikační strom pro data2 pouze s rebiopsiemi,bez PSA, PSAD a PSAD-TZ metoda CART přímé ukončení – zlomek objektů 0,02, minimu objektů tříd (0;17)

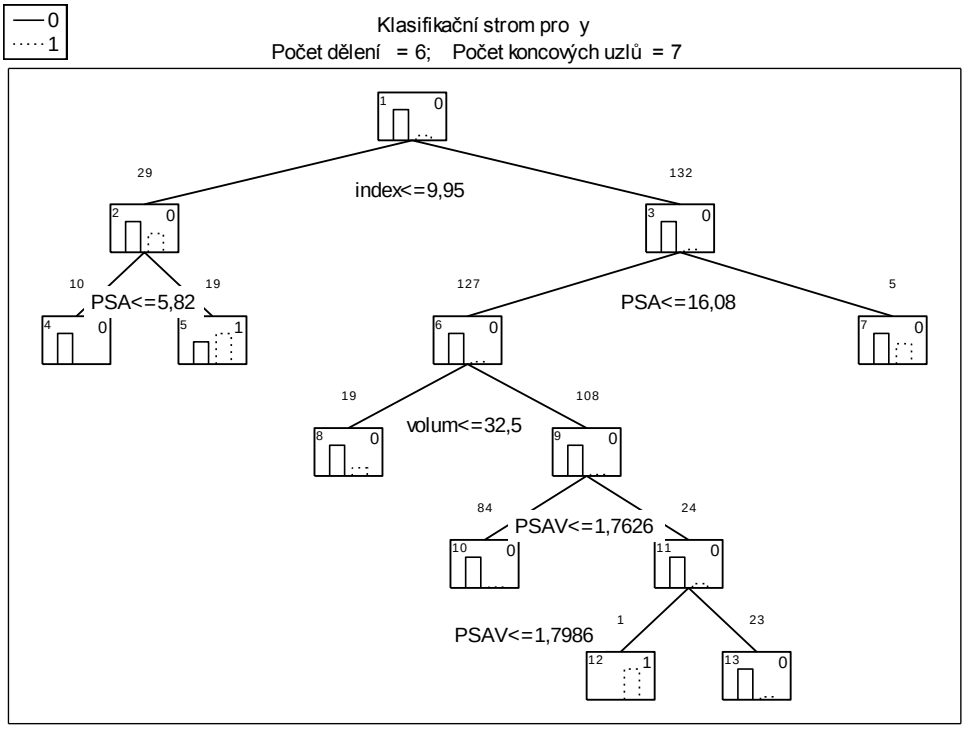
Data 3



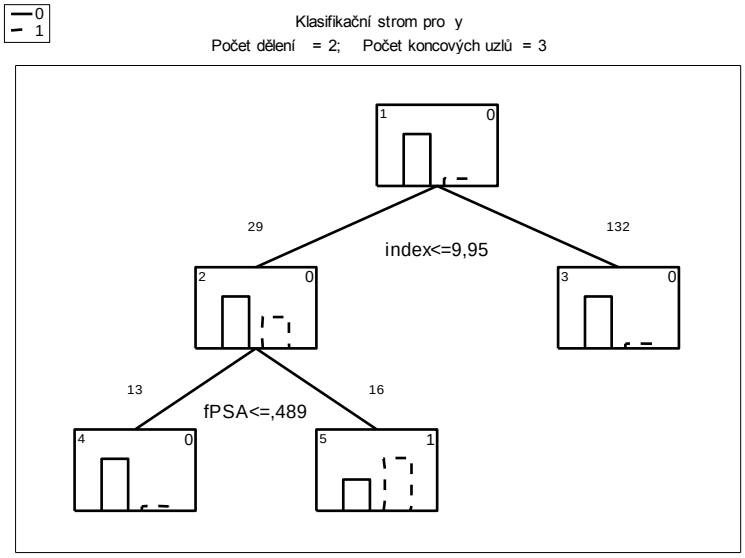
Obr. P16 Klasifikační strom pro data3 pouze s rebiopsiemi pro všechny proměnné, metoda CART přímé ukončení – zlomek objektů 0,05, minimu objektů tříd (1;43)



Obr. P 17 Klasifikační strom pro data3 pouze s rebiopsiemi,bez PSAD metoda CART přímé ukončení – zlomek objektů 0,03, minimu objektů tříd (0;26)

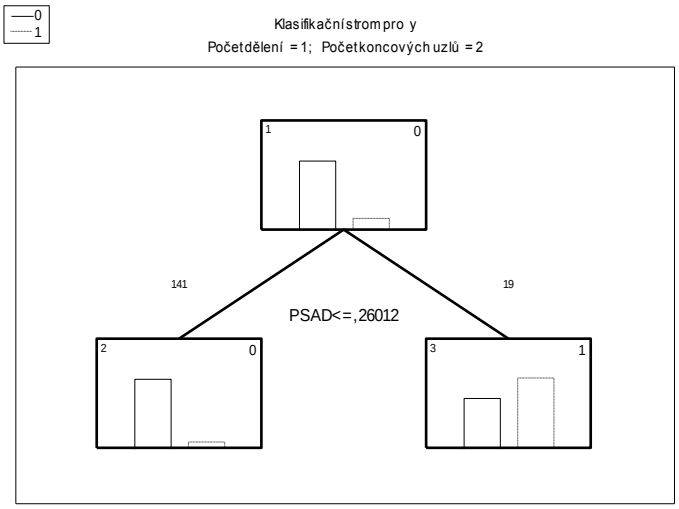


Obr. P18 Klasifikační strom pro data3 pouze s rebiopsiemi,bez PSAD a PSAD-TZ metoda CART přímé ukončení – zlomek objektů 0,02, minimu objektů tříd (0;17)

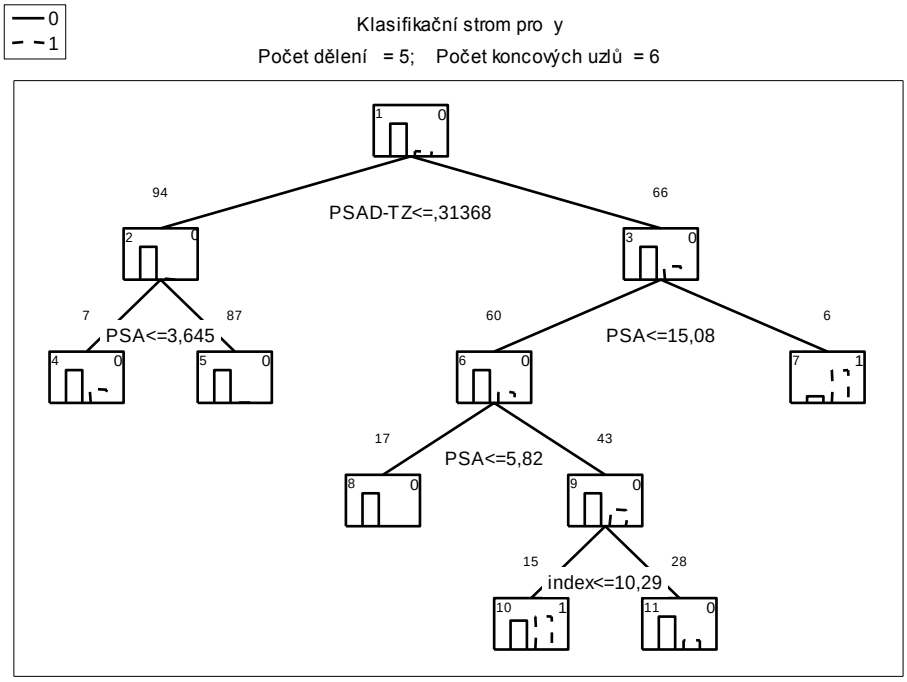


Obr. P19 Klasifikační strom pro data3 pouze s rebiopsiemi,bez PSA, PSAD a PSAD-TZ metoda CART přímé ukončení – zlomek objektů 0,02, minimu objektů tříd (0;17)

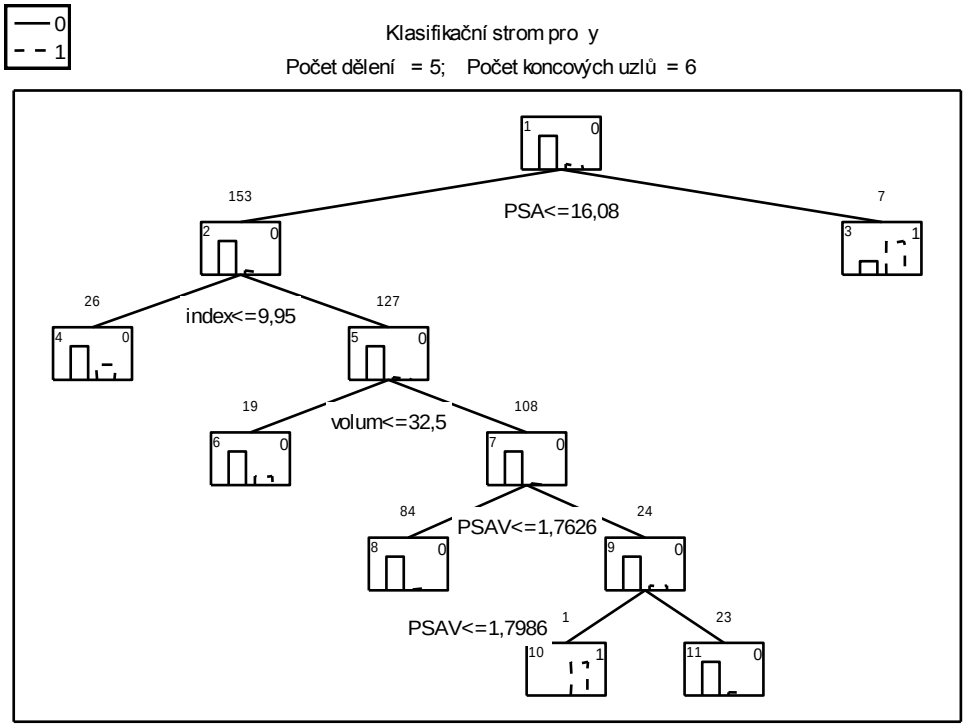
Data 4



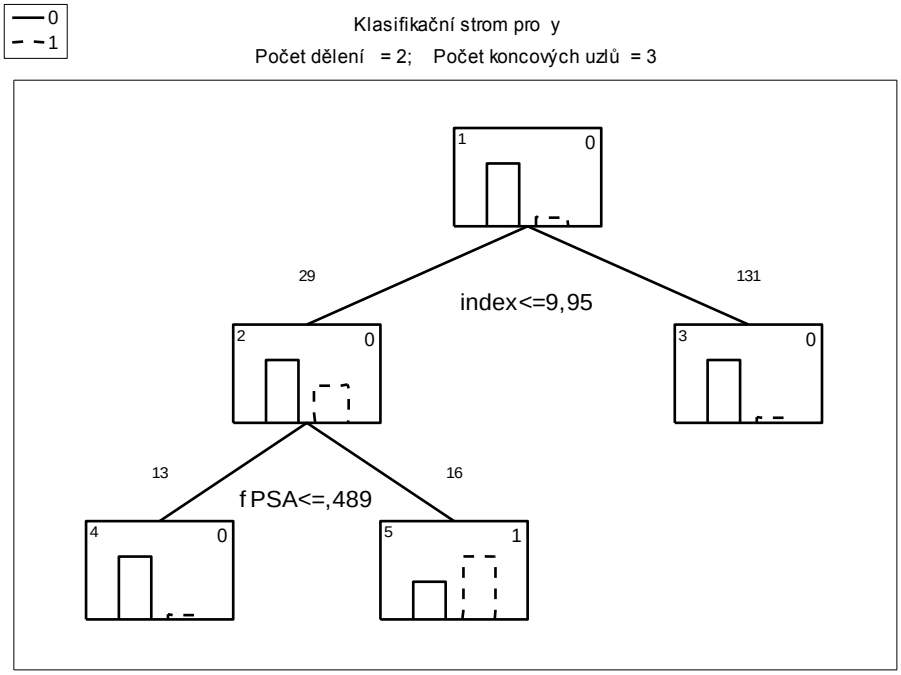
Obr. P20 Klasifikační strom pro data4 pouze s rebiopsiemi pro všechny proměnné, metoda CART přímé ukončení – zlomek objektů 0,05, minimu objektů tříd (1;43)



Obr. P 21 Klasifikační strom pro data4 pouze s rebiopsiemi,bez PSAD metoda CART přímé ukončení – zlomek objektů 0,03, minimu objektů tříd (0;25)



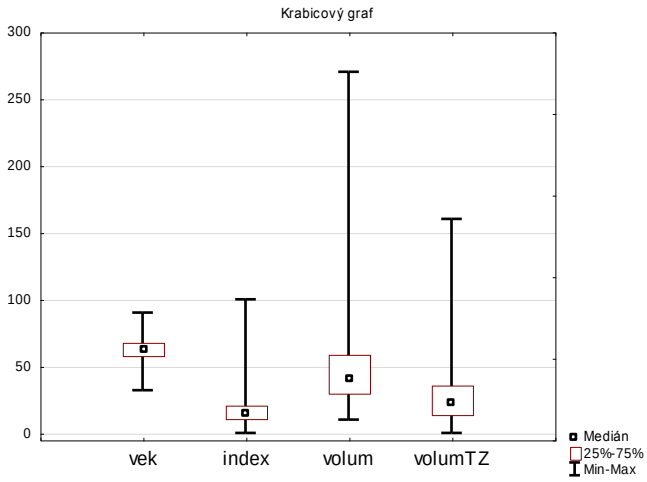
Obr. P22 Klasifikační strom pro data4 pouze s rebiopsiemi,bez PSAD a PSAD-TZ metoda CART přímé ukončení – zlomek objektů 0,02, minimu objektů tříd (0;17)

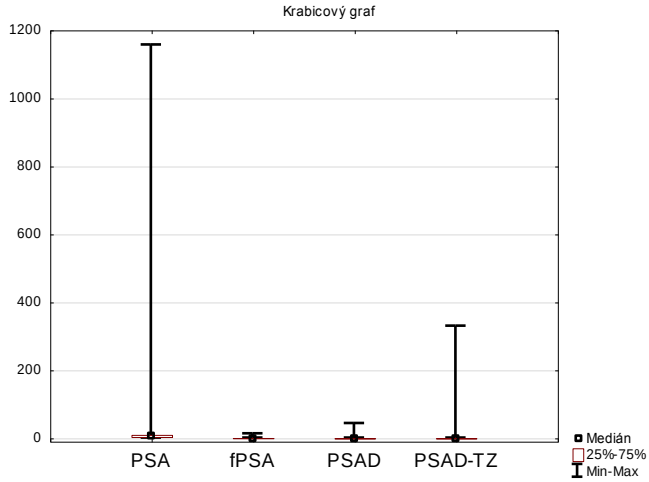


Obr. P23 Klasifikační strom pro data4 pouze s rebiopsiemi,bez PSA, PSAD a PSAD-TZ metoda CART přímé ukončení – zlomek objektů 0,02, minimu objektů tříd (0;17)

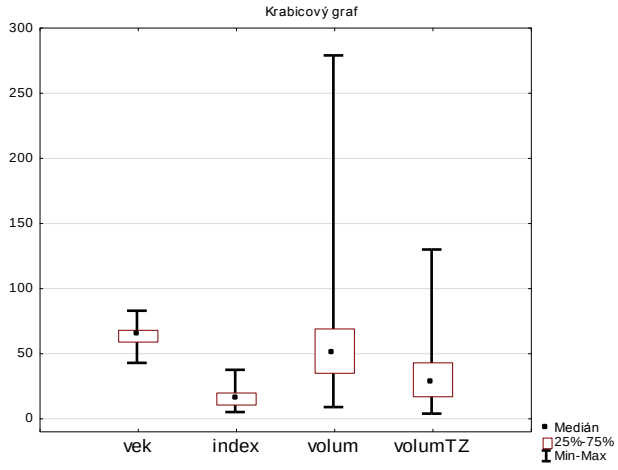
Krabicové grafy

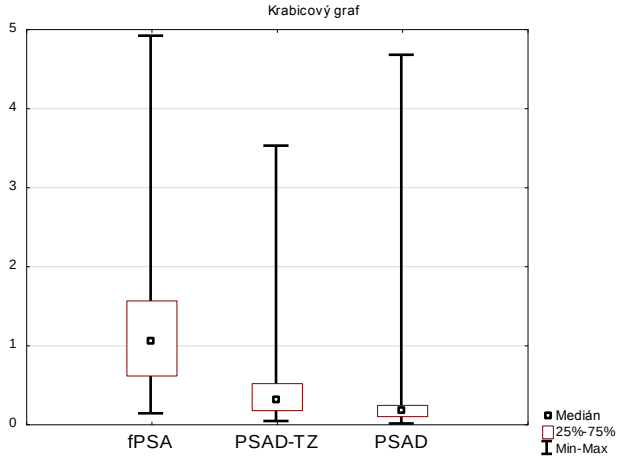
Data1 – první případy

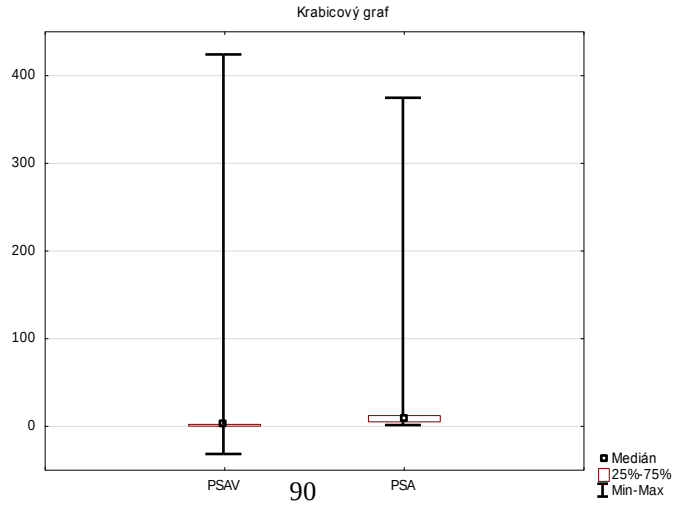




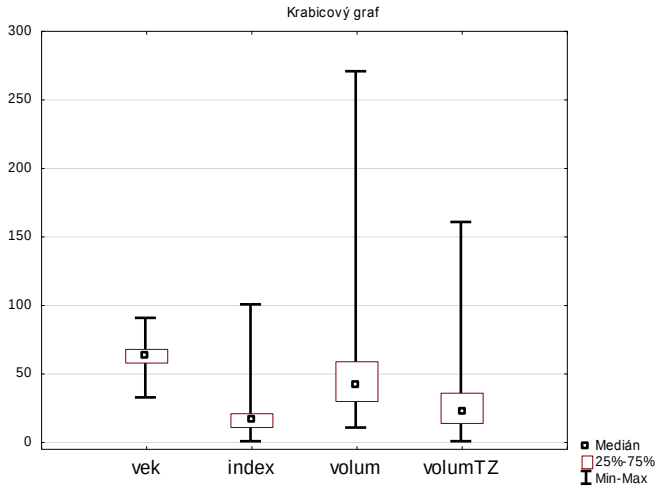
Data 1 – rebiopsie

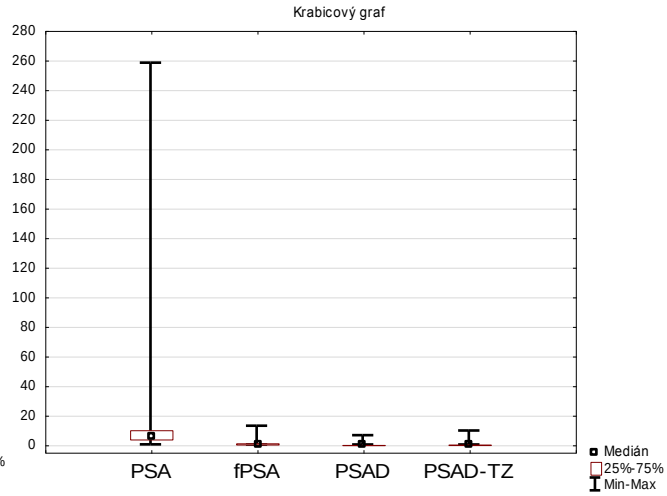




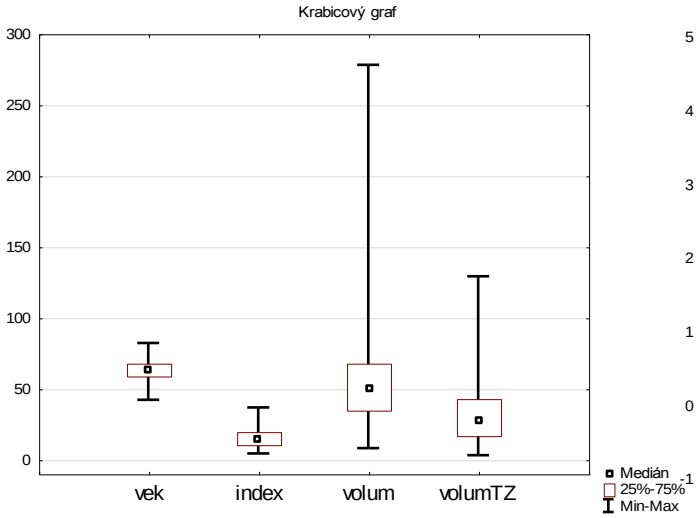


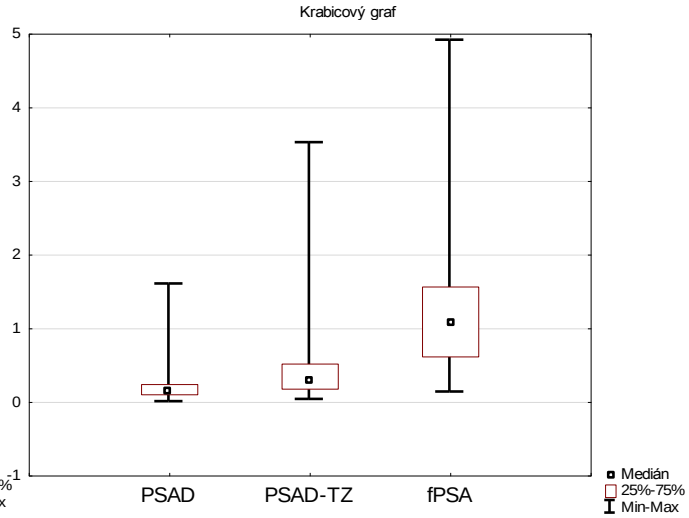
Data 2 – první případy

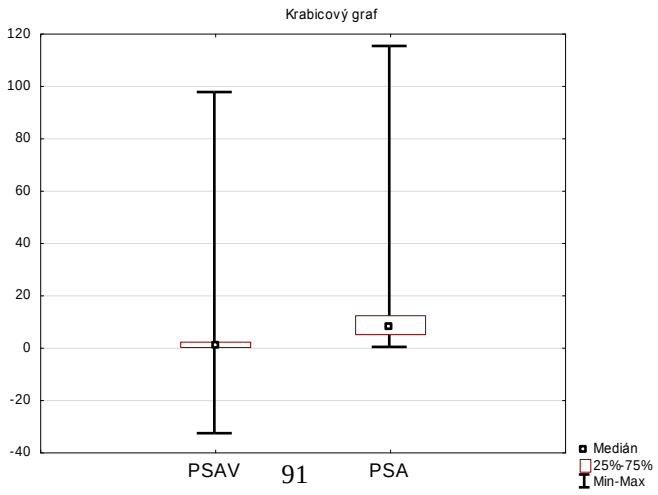




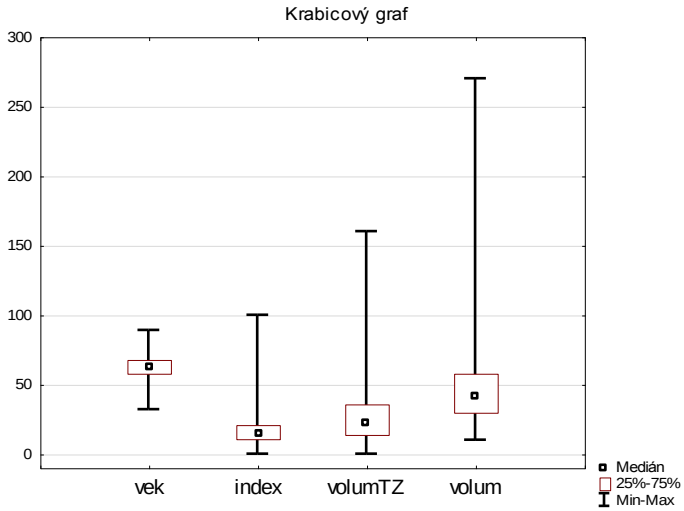
Data 2 – rebiopsie

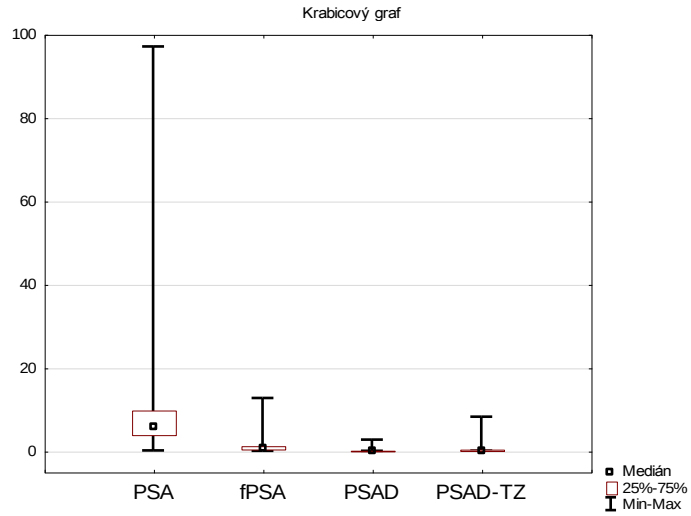




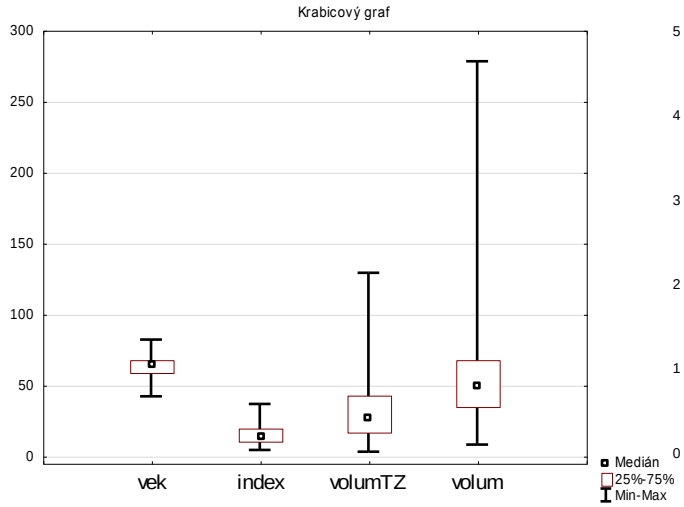


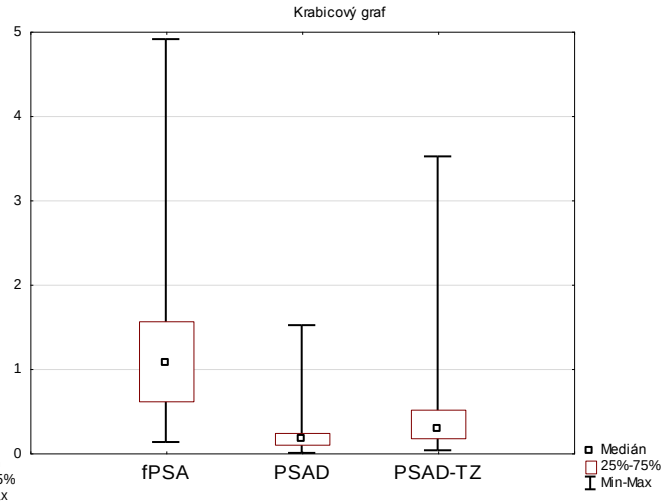
Data 3 – první případy

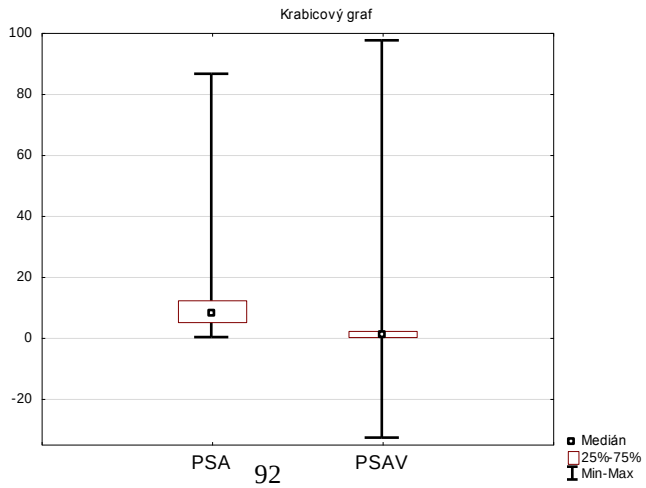




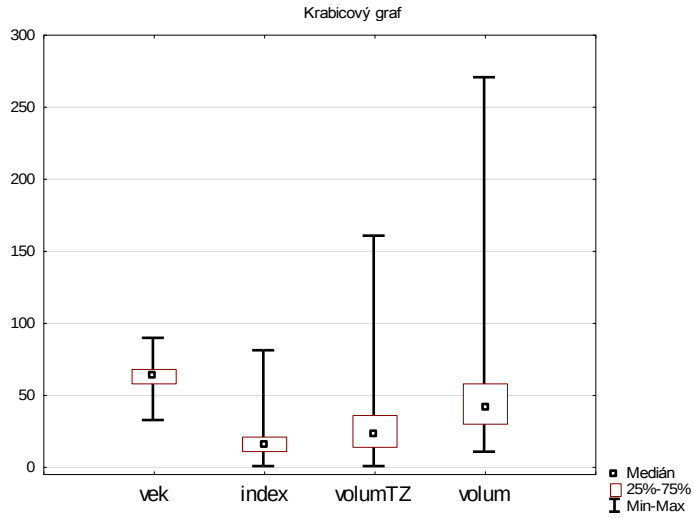
Data 3 – rebiopsie

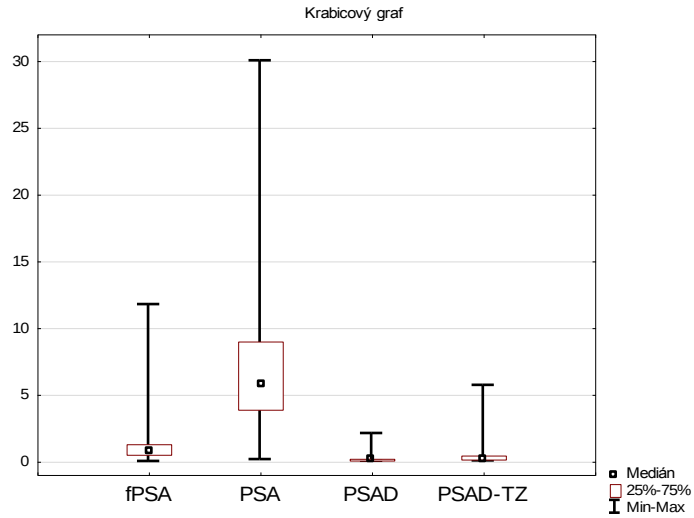






Data 4 – první případy





Data 4 – rebiopsie