Analýza datasetu ADAMEK z oblasti preventivní kardiologie v prostředí a jazyce R

4IZ450 Dobývání znalostí z databází

Lubomír Štěpánek

27. května 2017

Obsah

1 Zadání úlohy

Cílem úlohy je předzpracovat a analyzovat jeden ze dvou nabídnutých reálných datasetů, tj.

- dataset ADAMEK obsahující data z oblasti preventivní kardiologie – kardiovaskulární, resp. interní diagnózy a symptomy pacientů a jejich obou rodičů a další laboratorní či elektrokardiologické záznamy pacientů;

Vysázeno pomocí R-ka, YAML-u, TEX-u a BibTEX-u

- dataset TimeDeposit_10K obsahující sociodemografická a socioekonomická data klientů bankovních ústavů a vždy informace o tom, zda byl klientovi založen termínovaný účet;

a poté vhodně interpretovat získané výsledky. Smyslem úlohy je kvalitní předzpracování dat, „porozumění“ datům, formulace vhodné (klasifikační, asociační, segmentační či regresní) úlohy a výběr odpovídajícího algoritmu strojového učení, nakonec i vhodná interpretace a reportování výsledků. Metodika by neměla vybočovat z dobré praxe založené na principech CRISP-DM (Cross Industry Standard Process for Data Mining).

2 Řešení úlohy

Pro realizaci řešení úlohy byl zvolen dataset ADAMEK založen na datech z oboru preventivní kardiologie.

2.1 Doménový úvod

Kardiologie spolu s onkologií[1] patří mezi obory medicíny, kde je kvalita výzkumu na velmi vysoké úrovni. Designy studií jsou obvykle velmi komplexní, dobře precisované; studie probíhají obvykle multicentricky. Příkladem budiž i série světoznámých klinických studií PRAGUE české provenience [1]; Česká republika má na poli kardiologie dobré jméno, unikátnost plyne i ze zajímavého paradoxu špičkově dostupných metod „ischemické“ kardiologie v protikladu k relativně malé zeměpisné rozloze republiky – tato konstelace a zmíněné klinické studie vedly k tuzemskému objevení a ustanovení některých mezinárodně uznávaných doporučených postupů v kardiologii, např. preference včasné katetrizace před fibrinolýzou u akutního infarktu myokardu, což je v protikladu k rovněž vyspělým, ale rozlehlým zemím typu státy USA, které stále preferují technicky méně náročnou a dostupnější fibrinolýzu.

Preventivní kardiologie je rovněž „badatelsky“ relativně pokročilým oborem; řada výsledků plyne z velké obliby výzkumů hypolipidemických léčiv typu *statinů*, jež široce podporují farmakologické koncerny; proto jsou výsledky přesvědčivě významné.

V klinické praxi kardiologů se běžně uplatňují výsledky aktuálních epidemiologicky-kardiologických výzkumů, např. různé skórovací systémy pro rizikovost pacientů stran objevení se akutní myokardiální ischemie v následujícím roce života vzhledem k některým aktuálním klinickým hodnotám pacienta [2] apod. Kardiologicky a „civilizačně“ laděné studie totiž probíhají od druhé poloviny minulého století, některé dokonce kontinuálně dodnes, např. známá framinghamská studie začala již v roce 1948 a produkuje cenná longitudinální data o metabolických a kardiovaskulárních anamnézách pacientů [3].

Je tedy relativně velkou výzvou snažit se objevit zajímavé vztahy v datasetech obsahujících epidemiologicky-kardiologická data; zajímavé a silnou evidencí podložené vztahy budou pravděpodobně patřit mezi doménové znalosti, tedy mezi vztahy, které jsou doménovým expertům již dobře známé.

[1] Autor je lékařem, bývalým onkologem.

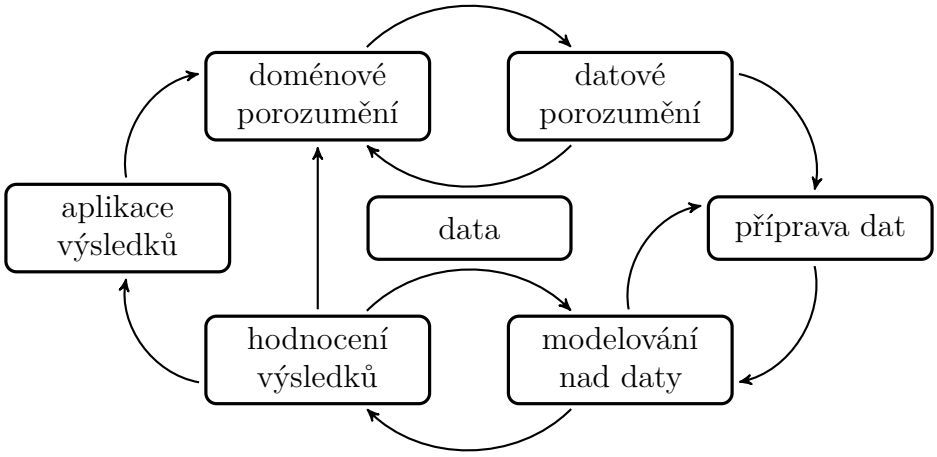
2.2 Metodika CRISP-DM

Metodika CRISP-DM (z anglického Cross Industry Standard Process for Data Mining) shrnuje principy dobré praxe v rámci dataminingového procesu tak, jak by měly být rámcově dodržovány pro zachování kvality celé analýzy [4].

Dataminingový proces by měl být zahájen porozuměním jak doménovým (alespoň částečně), tak datovým, tj. měla by být zřejmá paralela mezi doménovými výzkumnými otázkami a jejich datovým (rigidním, matematickým) zněním, resp. definováním – k tomu je obvykle nutná vzájemná komunikace doménových a datových expertů.

Poté by měla být data správně připravena, aby nad nimi mohlo následovat modelování výzkumných otázek a realizace samotné analýzy. Získané výsledky se mohou do jisté míry míjet se zamýšlenými výstupy, proto je někdy nutné původně zvolené modely analýzy upravit a její podmínky precisovat nebo naopak relaxovat a danou analýzu zopakovat, i mnohokrát po sobě.

Zhodnocené a interpretované výsledky mohou vést k jejich aplikaci v praxi, resp. ovlivnit doménové porozumění celému problému – i v této fázi je možné upravit výzkumné otázky a celý dataminingový proces, jak je naznačen na obrázku 1, opakovat.



Obrázek 1: Schéma fází dataminingového procesu podle metodiky CRISP-DM

2.3 Popis a kódování dat

Dataset ADAMEK o rozměrech 1122 × 200 obsahuje hodnoty pozorování o celkem 1122 pacientech pro jednotlivé proměnné, kterých je celkem 200. Proměnné se obecně týkají anamnestický, sociodemografických, vnitřně-lékařských a kardiologických dat pacienta, resp. jeho rodičů; proměnné vytváří některé tématicky související skupiny, jak ukazuje tabulka 1. Podrobnější informace o jednotlivých proměnných jsou dostupné v přiloženém souboru `Adamek_popis_pro_KIZI.doc`.

skupina proměnných	# proměnných ve skupině	popis proměnných ve skupině
identifikátory	2	unikátní identifikátory pacientů
osobní údaje	4	datum narození, věk, pohlaví, adresa
informace o otci	24	eventuální příčina úmrtí a známé diagnózy otce
informace o matce	24	eventuální příčina úmrtí a známé diagnózy matky
další osobní údaje	6	rodinný stav, žije pacient sám?, vzdělání, míra psychické zátěže, míra fyzické zátěže, tělesná aktivita
kouření	6	zda kuřák či {ne, ex}-kuřák, míra kouření, začátek a konec kouření, kolik cigaret denně
alkohol	3	míra spotřeby piva, vína, destilátů denně
alergie	2	přítomnost známých lékových či jiných alergií
ICHS	16	přítomnost některých jednotek ICHS (ischemická choroba srdeční) – infarktu myokardu, anginy pectoris, němé ischemie, arytmií; léčba obtíží
ICHPT	7	přítomnost některých jednotek ICHS (ischemická choroba periferních tepen), léčba obtíží
CMP	6	přítomnost, resp. datum CMP (cévní mozkové příhody), léčba
diabetes	6	přítomnost, eventuálně léčba diabetes mellitus
hypertenze	5	přítomnost, eventuálně léčba vysokého krevního tlaku
hyperlipoproteinémie	6	přítomnost, eventuálně léčba hyperlipoproteinémie
další rizika	4	přítomnost aneurysmatu aorty, menopauzy
jiné choroby	5	volnotextové údaje o dalších chorobách
obtíže	8	přítomnost typických symptomů (dušnost, bolest na hrudi, palpitace, otoky, synkopa, kašel, hemoptýza, klaudikace)
dieta	5	zda jsou uplatňovány některé dietetické zásady
léky	5	zda a kdy jsou aplikována perorální léčiva

Tabulka 1: Skupiny proměnných v datasetu ADAMEK

skupina proměnných	# proměnných ve skupině	popis proměnných ve skupině
míry	5	biometrické míry pacienta (hmotnost, výška, obvod pasu, obvod boků, dominantní)
krevní tlak	4	systolický a diastolický krevní tlak na levé a pravé končetině
fyzikální nález	4	tělesná teplota, tepová a dechová frekvence, je fyzikální nález „bpn“ (bez patologického nálezu)?
patologie	6	přítomnost a popis patologického nálezu na karotidách, tepnách, srdci, plicích či jinde
laboratoř	6	sérové hodnoty glykémie, kyseliny močové, celkového cholesterolu, HDL cholesterolu, LDL cholesterolu, triacylglycerolů
EKG	20	elektrofyziologické a deskriptivní parametry EKG křivky (rytmus, frekvence, intervaly, poruchy vedení, extrasystoly, hrubé patologie křivky)

Tabulka 1 (pokračování): Skupiny proměnných v datasetu ADAMEK

Matice dat je obecně hodně řídká (*sparse*), časté jsou chybějící hodnoty. Kódování datasetu bylo pravděpodobně jen základní ASCII, proto se ztratily některé znaky s diakritickými znaménky.

Dataset popisuje komplexní problematiku preventivní kardiologie; na první pohled se nenabízí právě jedna vhodná závisle proměnná.

2.4 Analýza dat

Dle popisu skupiny proměnných je zřejmé, že dataset se zabývá velmi komplexní oblastí, která kombinuje řadu již známých doménových znalostí, ale nabízí i prostor pro hypotetické rozklíčování znalostí nových.

Relativně velká složitost a bohatost datasetu otevírá možnost použití některých metod *učení bez učitele*, které nevyžadují explicitní určení závislé proměnné. To by zde bylo výhodou, neboť volba závislých proměnných je v dané problematice vzhledem k velké vzájemné provázanosti otazná (*chick-and-egg phenomenon* – mnohdy není zřejmé, co mohlo být příčinnou a co následkem).

Řada souvislostí v datasetu, byť by se primárně mohly zdát vhodné k analýze a např. k hledání asociací, je jistě doménovým odborníkům (lékařům) dobře známá a je podpořena rozsáhlejší evidencí, např. přítomnost některých diagnóz a k nim odpovídající léčba. Je-li

známá diagnóza, pak je apriorně zřejmé, že bude i přítomna i léčba, obě veličiny jsou na sobě téměř funkčně závislé – pacient s diagnózou, ale bez léčby by se neměl v datasetu vůbec objevit. Při volbě prediktorů je tedy nutné uplatnit alespoň nějaké (doménové) znalosti oboru, například že informace o přítomnosti léčby mohou kvůli pozitivní korelaci falešně zesílit váhu některých diagnóz.

Otázky, které jsme si kladly, souvisí tedy s nalezením asociací některých proměnných, resp. s nalezením hodnot některých proměnných typických pro zatřídění do hodnoty jiné, závislé proměnné (klasifikační úloha). Celá úloha byla řešena v prostředí R, které je určeno pro statistické výpočty a následné grafické náhledy [5].

2.4.1 Příprava dat

Data jsou uložena v souboru `Adamek06.csv` a byla do prostředí R nahrána jako text (s kódováním UTF-8), aby nedošlo k mylné koerci datových typů.

Pro účely další analýzy byly všechny proměnné programasticky upraveny tak, že pokud při automatické koerci na numerickou proměnnou nevznikla nová chybějící hodnota (což by např. pro textový řetězec „ano“ vznikla), byla daná proměnná kódována jako numerická (kvantitativní). V opačném případě byla kódována jako kategorická. Manuální kontrola však je nutná, některé proměnné byly pro jistotu kódovány ručně – např datum narození `Dat_nar` a datum vyšetření `DatVys` nejsou v jistém slova smyslu ani numerické, ani kategorické proměnné, je třeba je převést na formát datum pomocí patternu, kterým jsou uloženy v původním datasetu (tedy `den.měsíc.rok.`).

Dále byly vytypovány vhodné závislé proměnné, za ty považujeme jednak přítomnost dané diagnózy (ICHS, DM apod.), dále přítomnost symptomatologie (jakožto manifestace diagnóz, tedy bolest na hrudi, dušnost apod.) a léčbu léky či dietou.

Ve fázi seznamování s daty byl odhadnut počet případů bez chybějících hodnot. Zcela bez chybějících hodnot je pouze jedno pozorování ze všech 1122! Pomineme-li proměnné, které nebudou dozajista testovány (ty, které mají volnotextové odpovědi nebo jsou u nich nejsou známy hodnoty pro žádného pacienty), je počet pozorování bez chybějících hodnot několik málo desítek.

Kvůli rozhodovacím stromům typu CART, které byly v analýze použity, byly chybějící hodnoty ve všech smysluplných proměnných použitých pro další analýzu podle následujícího pravidla: buď $x_{i,j}$ hodnota i-tého pozorování j-té proměnné pro všechna $i \in \{1, 2, \ldots, 1122\}$ a $j \in \{1, 2, \ldots, 200\}$, pak

$$x_{i,j} = \begin{cases} \bar{X}_j, & \text{pokud je veličina } X_j \text{ numerická} \\ \hat{X}_j, & \text{pokud je veličina } X_j \text{ kvalitativní,} \end{cases}$$

kde $\bar{X}_j$ je průměr nechybějících hodnot j-té veličiny datasetu a $\hat{X}_j$ je modus nechybějících hodnot j-té veličiny datasetu. Tím bylo docíleno rozumných výsledků, které rozhodovací stromy vracely.

2.4.2 Použité algoritmy strojového učení

V rámci analýzy a modelování nad daty byly použity rozhodovací klasifikační stromy typu CART (Classification And Regression Trees). Výhodou této metody je, že je dobře čitelná a interpretovatelná i pro audienci mimo datovou vědu, zde např. pro doménové odborníky, lékaře. Klasifikační stromy řeší klasifikační úlohu, je tedy třeba určit závislou proměnnou, která je kategorická.

Klasifikační stromy jsou založeny na principu, že trénovací množina je postupně rozdělována na stále menší podmnožiny tak, aby v každé podmnožině převládaly prvky jedné třídy [6], jde tedy o princip „rozděl a panuj“ (“divide and conquer“); metoda je známa jako *top-down induction of decision tree* (TDIDT). V každé iteraci je vybrán některý z atributů a je určena jeho hodnota tak, že trénovací množina je pak hodnotou této proměnné „nejlépe“ rozdělena ve smyslu některého diskriminačního kritéria. Vzniká tak graf typu strom. Mezi kritéria, které jsou pro atributy maximalizovány, patří Giniho index

$$Giniho\ index_i = 1 - \sum_{j=1}^{k} p_{ij}^2,$$

informační zisk

$$informační\ zisk_i = -\sum_{j=1}^{k} p_{ij} \log_2 p_{ij}$$

a deviance

$$deviance_i = -2 \sum_{j=1}^{k} n_{ij} \ln p_{ij},$$

kde p_{ij} je pravděpodobnost existence j–té třídy v i–tém uzlu, n_{ij} je počet pozorování j–té třídy v podmnožině i–tého uzlu, k je počet tříd. *Pruning* je prořezání výsledného stromu (tj. neuvažování koncových větví stromu od určitého stupně větvení), děje se pomocí k-násobné křížové validace nebo porovnáním nevysvětlené variability vs. počtu uzlů stromu (v diagramu *elbow fenomén*) či jinak.

Jako závisle proměnné byly postupně voleny přítomnost dané diagnózy (ICHS, DM apod.), dále přítomnost symptomatologie (jakožto manifestace diagnóz, tedy bolest na hrudi, dušnost apod.) a přítomnost léčby léky či dietou. Většina z nich je pouze dvouhodnotová (binární).

Pro každou volbu závisle proměnné byl dataset náhodně rozdělen v poměru 70 : 30 na trénovací a testovací množinu, kdy na trénovací množině strom vyrostl a naučil se klasifikovat, na testovací množině byla vyhodnocena jeho přesnost. Přesnost hodnotíme na matici záměn (*konfuzní matici*), která má tvar dle tabulky 2, kde $n_{i,j}$ je počet pozorování z testovací

množiny, které byly zařazeny do třídy j a patří to třídy i, kde $i, j \in \{1, 2, \ldots, k\}$, má-li závislá proměnná právě k tříd.

		přiřazená hodnota			
		1	2	…	k
skutečná hodnota	1	n_{11}	n_{12}	$\cdots$	n_{1k}
	2	n_{21}	n_{22}	$\cdots$	n_{2k}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	k	n_{k1}	n_{k2}	$\cdots$	n_{kk}

Tabulka 2: Obecná matice záměn pro závisle proměnnou o k třídách

Přesnost (*accuracy*) pak vyčíslíme jako podíl stopy a součtu matice záměn, tedy

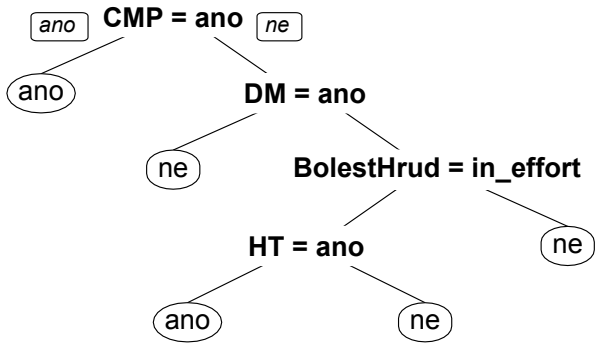
$$\text{přesnost} = \frac{\sum_{i=1}^{k} n_{ii}}{\sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij}}.$$

Snadno nahlédneme, že predikoval-li by klasifikační strom pouze jako náhodný mechanismus, pak pravděpodobnost správného přiřazení i-tého pozorování do své třídy je rovna $\frac{1}{k}$. Kdykoliv je tedy predikční přesnost $> \frac{1}{k}$, lze tvrdit, že klasifikační strom predikuje lépe, než náhodný mechanismus (tím je např. hod k-stěnnou nebiasovanou kostkou).

2.5 Výsledky

Klasifikační strom byl naučen postupně pro následující závisle proměnné.

V případě závisle proměnné ICHS (v datasetu jako `ICHS`), ischemická choroba srdeční – přítomna, nepřítomna, byl získán klasifikační strom podle obrázku 2. Matice záměn má podobu, jakou ukazuje tabulka 3. Predikce dosáhla přesnosti 93,8 %.

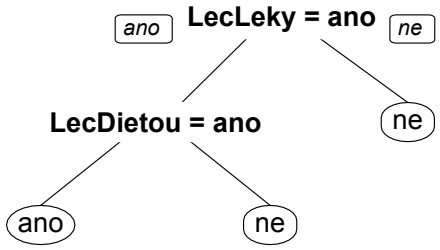


Obrázek 2: Klasifikační strom pro závisle proměnnou ICHS (**in_effort** = námahová)

		přiřazená hodnota	
		ano	ne
skutečná hodnota	ano	2	12
	ne	9	314

Tabulka 3: Matice záměn pro závisle proměnnou ICHS

V případě závisle proměnné HT (v datasetu jako `HT`), hypertenze – přítomna, nepřítomna, byl získán klasifikační strom podle obrázku 3. Matice záměn má podobu, jakou ukazuje tabulka 4. Predikce dosáhla přesnosti 76,6 %.

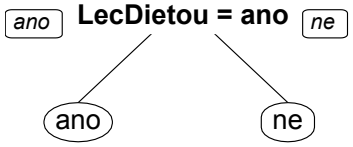


Obrázek 3: Klasifikační strom pro závisle proměnnou HT

		přiřazená hodnota	
		ano	ne
skutečná hodnota	ano	61	45
	ne	34	197

Tabulka 4: Matice záměn pro závisle proměnnou HT

V případě závisle proměnné HLP (v datasetu jako `HLP`), hyperlipoproteinémie – přítomna, nepřítomna, byl získán klasifikační strom podle obrázku 4. Matice záměn má podobu, jakou ukazuje tabulka 5. Predikce dosáhla přesnosti 83,7 %.

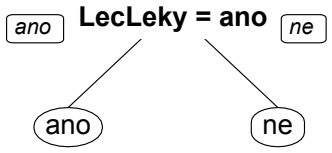


Obrázek 4: Klasifikační strom pro závisle proměnnou HLP

		přiřazená hodnota	
		ano	ne
skutečná hodnota	ano	100	17
	ne	38	182

Tabulka 5: Matice záměn pro závisle proměnnou HLP

V případě závisle proměnné JineChoroby (v datasetu jako `JineChoroby`), komorbidity – přítomny, nepřítomny, byl získán klasifikační strom podle obrázku 5. Matice záměn má podobu, jakou ukazuje tabulka 6. Predikce dosáhla přesnosti 61,1 %.

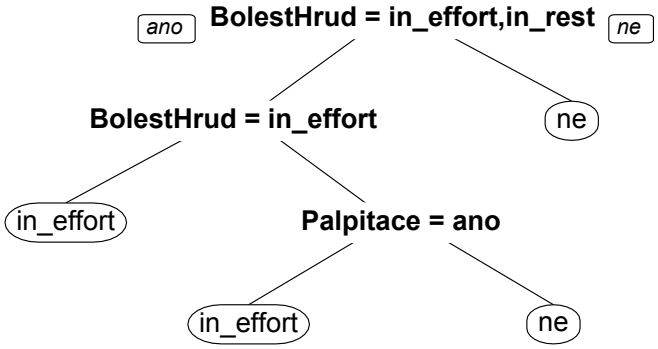


Obrázek 5: Klasifikační strom pro závisle proměnnou JineChoroby

		přiřazená hodnota	
		ano	ne
skutečná hodnota	ano	116	77
	ne	54	90

Tabulka 6: Matice záměn pro závisle proměnnou JineChoroby

V případě závisle proměnné Dusnost (v datasetu jako `Dusnost`), dušnost – námahová, klidová, vůbec, byl získán klasifikační strom podle obrázku 6. Matice záměn má podobu, jakou ukazuje tabulka 7. Predikce dosáhla přesnosti 88,1 %.

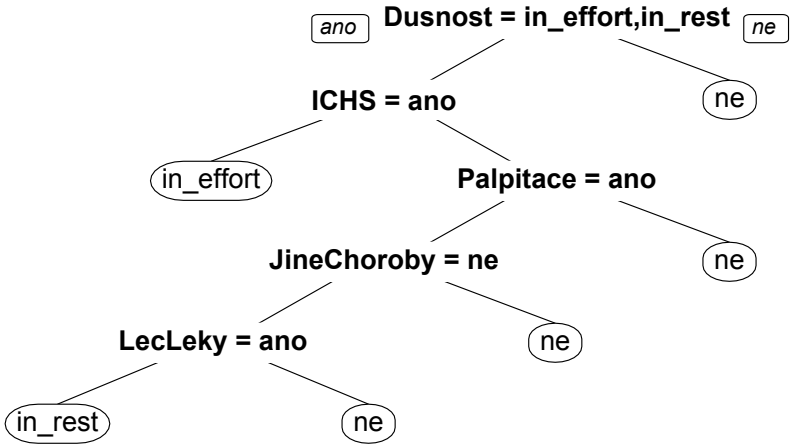


Obrázek 6: Klasifikační strom pro závisle proměnnou Dusnost (**in_effort** = námahová, **in_rest** = klidová)

		přiřazená hodnota		
		námahová	klidová	vůbec
skutečná hodnota	námahová	3	0	22
	klidová	2	0	3
	vůbec	13	0	294

Tabulka 7: Matice záměn pro závisle proměnnou Dusnost

V případě závisle proměnné BolestHrud (v datasetu jako `BolestHrud`), bolest na hrudi – námahová, klidová, vůbec, byl získán klasifikační strom podle obrázku 7. Matice záměn má podobu, jakou ukazuje tabulka 8. Predikce dosáhla přesnosti 91,7 %.

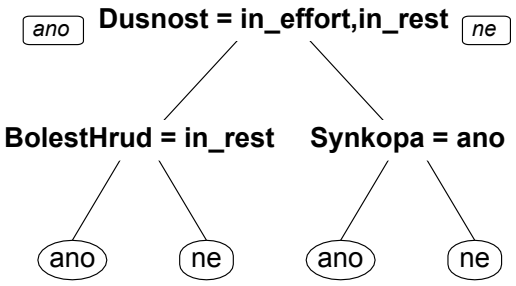


Obrázek 7: Klasifikační strom pro závisle proměnnou BolestHrud (**in_effort** = námahová, **in_rest** = klidová)

		přiřazená hodnota		
		námahová	klidová	vůbec
skutečná hodnota	námahová	0	0	13
	klidová	1	1	11
	vůbec	1	2	308

Tabulka 8: Matice záměn pro závisle proměnnou BolestHrud

V případě závisle proměnné Palpitace (v datasetu jako `Palpitace`), palpitace – přítomny, nepřítomny, byl získán klasifikační strom podle obrázku 8. Matice záměn má podobu, jakou ukazuje tabulka 9. Predikce dosáhla přesnosti 85,1 %.

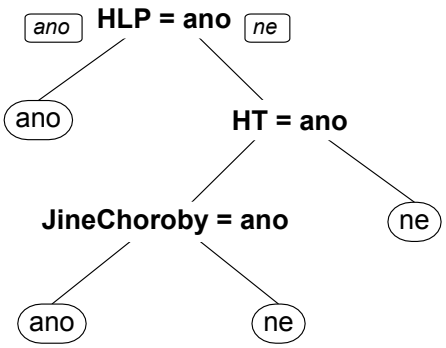


Obrázek 8: Klasifikační strom pro závisle proměnnou Palpitace (**in_effort** = námahová, **in_rest** = klidová)

		přiřazená hodnota	
		ano	ne
skutečná hodnota	ano	3	44
	ne	6	284

Tabulka 9: Matice záměn pro závisle proměnnou Palpitace

V případě závisle proměnné LecDietou (v datasetu jako `LecDietou`), léčba dietou – přítomna, nepřítomna, byl získán klasifikační strom podle obrázku 9. Matice záměn má podobu, jakou ukazuje tabulka 10. Predikce dosáhla přesnosti 84,0 %.

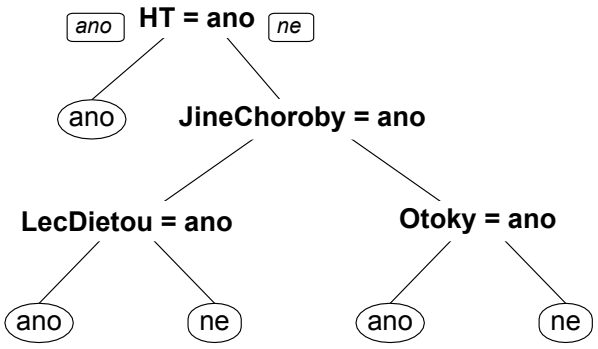


Obrázek 9: Klasifikační strom pro závisle proměnnou LecDietou

		přiřazená hodnota	
		ano	ne
skutečná hodnota	ano	115	23
	ne	31	168

Tabulka 10: Matice záměn pro závisle proměnnou LecDietou

V případě závisle proměnné LecLeky (v datasetu jako `LecLeky`), léčba léky – přítomna, nepřítomna, byl získán klasifikační strom podle obrázku 10. Matice záměn má podobu, jakou ukazuje tabulka 11. Predikce dosáhla přesnosti 73,0 %.



Obrázek 10: Klasifikační strom pro závisle proměnnou LecLeky

		přiřazená hodnota	
		ano	ne
skutečná hodnota	ano	123	47
	ne	44	123

Tabulka 11: Matice záměn pro závisle proměnnou LecLeky

Pro ostatní závisle proměnné, pro které byly rovněž vytvářeny klasifikační stromy, tj. `ICHPT`), CMP `CMP`, `DM`, `AneurysmaAorty`, `Otoky`, `Synkopa`, `Kase`, `Hemoptyza` a `Klaudikace`, vyšly tyto stromy jako jednouzlové grafy, které byly samy o sobě koncovým uzlem řadícím každé pozorování v testovací množině do třídy a větším zastoupením v trénovací množině.

2.6 Diskuze

Klasifikační stromy na vstupním dataset a zvoleném rozdělení do trénovací a testovací množiny předvedly relativně dobrou predikční přesnost.

V případě závisle proměnné ICHS má nejdůležitější pozici přítomnost CMP, tedy cévní mozkové příhody v anamnéze. To dává dobrý smysl, neboť lze čekat, že u pacientů budou přítomny obě diagnózy, mají totiž naprosto shodnou etiologii – aterosklerózu tepen, liší se jen lokalizací tepen v rámci organismu. Přítomnost diabetu mellitu zde svědčí proti ischemické chorobě srdeční. Naopak přítomnost námahové bolesti na hrudi je dokonce diagnostický znak srdeční ischemie, její přítomnost vede pak ještě v kombinaci s hypertenzí k pravděpodobné přítomnosti ischemie myokardu. Tento klasifikační strom je tedy dobře slučitelný s dosavadními znalostmi domény.

V případě hypertenze jako závisle proměnné svědčí pro její přítomnost to, že je pacient léčen léky i dieteticky, jinak nejspíš postižen hypertenzí není. Smysl to dává, hypertonikům se známou diagnózou se doporučuje oboje; naopak absence farmakologické léčby lze čekat skutečně jen u pacientů bez hypertenze.

Klasifikační strom pro hyperlipoproteinémii vyhodnotil jako postačující pro zařazení do diagnózy ordinovanou léčbu dietou – v praxi je velká část pacientů zajištěna skutečně jen tímto postupem.

Komorbidity se dle dat poznají podle přítomnosti farmakologické léčby, to dává relativně dobrý smysl.

Dendrogram pro dušnost vyšel téměř učebnicově. Jako nejdůležitější rozhodovací krok je zde přítomnost námahové či klidové bolesti na hrudi, což je v dobré klinické korelaci s představou o ischemii myokardu, která vyvolává bolest na hrudi a zároveň snižuje pacientovu fyzickou výkonnost – proto je dušný. Má-li pacient bolest námahovou či má palpitace, lze u něj očekávat námahovou dušnost, jak bychom i očekávali.

V případě bolesti na hrudi je situace obdobná – přítomnost jakékoliv formy dušnosti a známé diagnózy ICHS svědčí pro možnost námahové bolesti na hrudi. Když bude stižen místo ICHS palpitacemi, a léčenými komorbiditami, lze očekávat, že bude mít zkušenost s těžší formou dušnosti, klidovou. To lze klinicky vysvětlit tak, že je-li ušetřen myokard (není diagnóza ICHS), je etiologie pravděpodobně pulmonální, proto je dýchání zasaženo ještě více a dušnost je i klidová, námaha k jejímu vyvolání není nutná.

Palpitace se zdají být spojeny s dušností a bolestí na hrudi (nejspíše obraz anginy pectoris) či pouze se synkopami (obraz typický pro němou ischemii či arytmickou poruchu srdce). Obdobné souvislosti se vyučují i v rámci vnitřního lékařství.

U léčby dietou je zásadní přítomnost hyperlipoproteinémie nebo hypertenze, která je kombinovaná s dalšími nemocemi. U farmakologické léčby lze čekat, že pacient má buďto hypertenzi, případně jiné choroby, na které mu již byla doporučena dietetická léčba, anebo má samostatně otoky (které vyžadují léčbu, nejspíše diuretickou).

Je třeba kriticky dodat, že výsledky, byť vypadají na první pohled velmi povedeně, jsou zatíženy limitacemi CART stromů. Mezi ně například patří závislost konkrétního stromu na rozdělení dat do trénovací a testovací množiny. Dále je třeba uvážlivá volba vysvětlujících a vysvětlovaných proměnných – problémem je někdy *leak* informací, pokud je mezi prediktory zařazena proměnná, která je silně asociována s vysvětlovanou proměnnou. Problémem stromů

je i relativně velký *overfitting* nad daty trénovací množiny. Řešením některých problémů je vybudování náhodného lesu nad trénovacími daty, eventuálně pruning stromů.

2.7 Závěr

Byly natrénovány a otestovány klasifikační stromy pro některé zvolené závislé proměnné. Predikční přesnost naučených stromů je relativně vysoká, topologie stromů je dobře slučitelná s klinickými znalostmi.

2.8 Implementace řešení v R

Zde je uveden kód v jazyce R, ve kterém byly zpracovávány veškeré výpočty a rovněž generovány diagramy.

```
################################################################################
################################################################################
################################################################################

## instaluji a loaduji balíčky ------------------------------------------------

invisible(
    lapply(c(
            "xtable",
            "openxlsx",
            "rpart",
            "rpart.plot",
            "RColorBrewer",
            "rattle"
        ),
        function(package){

            if(!(package %in% rownames(installed.packages()))){

                install.packages(
                    package,
                    dependencies = TRUE,
                    repos = "http://cran.us.r-project.org"
                )

            }

            library(package, character.only = TRUE)

        }
```

```
    )
)


## ------------------------------------------------------------------------

################################################################################

## nastavuji handling se zipováním v R ----------------------------------------

Sys.setenv(R_ZIPCMD = "C:/Rtools/bin/zip")


## ------------------------------------------------------------------------

################################################################################

## nastavuji pracovní složku --------------------------------------------------

while(!"script.R" %in% dir()){
    setwd(choose.dir())
}

mother_working_directory <- getwd()


## ------------------------------------------------------------------------

################################################################################

## vytvářím posložky pracovní složky ------------------------------------------

setwd(mother_working_directory)

for(my_subdirectory in c("vstupy", "vystupy")){

    if(!file.exists(my_subdirectory)){

        dir.create(file.path(

            mother_working_directory, my_subdirectory

        ))

    }

}
```

```
## ------------------------------------------------------------------------

################################################################################

## loaduji data ----------------------------------------------------------

setwd(paste(mother_working_directory, "vstupy", sep = "/"))

my_data <- read.csv(

    file = "Adamek06.csv",
    header = TRUE,
    sep = ";",
    check.names = FALSE,
    colClasses = "character"

)

setwd(mother_working_directory)


## ------------------------------------------------------------------------

################################################################################
################################################################################
################################################################################

## preprocessing ---------------------------------------------------------

#### vytvářím vektor indexů těch proměnných, které nejsou příliš řídké
#### (nemají tolik chybějících hodnot) ----------------------------------------

less_sparse_variables <- c(

    1:6,                            ## identifikátory, osobní informace
    8, 10, 17, 18, 19, 20, 25, 27,  ## skupina proměnných o otci
    32, 34, 41, 42, 43, 44, 49, 51, ## skupina proměnných o matce
    55,                             ## lékař
    56, 57, 58,                     ## osobní informace
    59, 60, 61,                     ## zátěž
    62,                             ## kouření
    68, 69, 70,                     ## alkohol
    71,                             ## alergie
    74,                             ## ICHS
    90,                             ## ICHPT
```

```
    97,                                 ## CMP
    103,                                ## diabete mellitus
    109,                                ## hypertenze
    114,                                ## hyperlipoproteinémie
    119,                                ## aneurysma
    123,                                ## komorbidity
    128:135,                            ## symptomatologie
    136,                                ## léčen dietou?
    141,                                ## léčen farmakologicky?
    146:158,                            ## fyzikální vyšetření, status praesens
    165:170,                            ## laboratorní nález
    171, 174:182, 191, 194,             ## EKG
    197,                                ## ambulance, ve které vyšetřen
    198,                                ## datum vyšetření
    199, 200                            ## věk úmrtí matky a otce

)


meaningful_variables <- c(

    4, 5,                               ## identifikátory, osobní informace
    10,                                 ## skupina proměnných o otci
    34,                                 ## skupina proměnných o matce
    56, 58,                             ## osobní informace
    59, 60,                             ## zátěž
    62,                                 ## kouření
    68, 69, 70,                         ## alkohol
    74,                                 ## ICHS
    90,                                 ## ICHPT
    97,                                 ## CMP
    103,                                ## diabete mellitus
    109,                                ## hypertenze
    114,                                ## hyperlipoproteinémie
    119,                                ## aneurysma
    123,                                ## komorbidity
    128:135,                            ## symptomatologie
    136,                                ## léčen dietou?
    141,                                ## léčen farmakologicky?
    146:158,                            ## fyzikální vyšetření, status praesens
    165:170,                            ## laboratorní nález
    171, 174:182, 191, 194#,             ## EKG
    #199, 200                           ## věk úmrtí matky a otce

)
```

```
response_variables <- c(

    "ICHS",
    "ICHPT",
    "CMP",
    "DM",
    "HT",
    "HLP",
    "AneurysmataAorty",
    "JineChoroby",
    "Dusnost",
    "BolestHrud",
    "Palpitace",
    "Otoky",
    "Synkopa",
    "Kasel",
    "Hemoptyza",
    "Klaudikace",
    "LecDietou",
    "LecLeky"

)


## ------------------------------------------------------------------------

#### převádím proměnné na numerické, či na kategorické ------------------------

for(i in 1:dim(my_data)[2]){

    if(
        suppressWarnings(
            all(
                !is.na(
                    as.numeric(
                        gsub(
                            ",",
                            ".",
                            as.character(
                                my_data[!is.na(my_data[, i]), i]
                            )
                        )
                    )
                )
            )
        )
    ){
```

```
        my_data[, i] <- suppressWarnings(
            as.numeric(
                gsub(
                    ",",
                    ".",
                    as.character(my_data[, i])
                )
            )
        )

    }else{

        my_data[which(my_data[, i] == "yes"), i] <- "ano"
        my_data[which(my_data[, i] == "no"), i] <- "ne"

        my_data[, i] <- as.factor(my_data[, i])

    }

}


## ------------------------------------------------------------------------

#### první dvě proměnné jsou kategorické (identifikátory) ----------------------

for(i in 1:2){

    my_data[, i] <- as.factor(as.character(my_data[, i]))

}


#### třetí a 198. proměnná jsou datumy -----------------------------------------

for(i in c(3, 198)){

    my_data[, i] <- as.Date(as.character(my_data[, i]), "%d.%m.%Y")

}


#### některé proměnné jsou numerické, i přesto že obsahují volný text;
#### převádím je i za cenu ztráty některých hodnot -----------------------------

for(i in c(
```

```
    68, 69, 70,
    146:149,
    151:157,
    165:170,
    173:181,
    199:200

)){

    my_data[, i] <- suppressWarnings(as.numeric(as.character(my_data[, i])))

}


## ------------------------------------------------------------------------

#### handling s chybějícími hodnotami ----------------------------------------
#### v případě kategorické proměnné nahrazuji chybějící hodnotu modusovou
#### hodnotou, v případě numerické proměnné průměrnou hodnotou ----------------

my_data_wo_NA <- my_data

for(i in 1:dim(my_data)[2]){

    if(class(my_data[, i]) == "numeric"){

        if(any(is.na(my_data[, i]))){

            my_data_wo_NA[which(is.na(my_data_wo_NA[, i])), i] <- mean(
                my_data[, i],
                na.rm = TRUE
            )

        }

    }

    if(class(my_data[, i]) == "factor"){

        if(any(my_data[, i] == "")){

            my_data_wo_NA[, i] <- as.character(my_data_wo_NA[, i])

            if(length(my_data_wo_NA[, i][my_data_wo_NA[, i] != ""]) > 0){

                my_data_wo_NA[which(my_data_wo_NA[, i] == ""), i] <- max(
```

```
                    my_data_wo_NA[, i][my_data_wo_NA[, i] != ""]
                )

            }

            my_data_wo_NA[, i] <- as.factor(my_data_wo_NA[, i])

        }

    }


}


## ------------------------------------------------------------------------

################################################################################
################################################################################
################################################################################

## začínám s analýzami ---------------------------------------------------

################################################################################

#### rozhodovací stromy -------------------------------------------------

################################################################################

## helper funkce ---------------------------------------------------------

getMyAccuracy <- function(my_table){

    # '''
    # vrací přesnost pro konfuzní matici "my_table"
    # '''

    return(sum(diag(my_table)) / sum(my_table))

}


## ------------------------------------------------------------------------

################################################################################

## zkouším rozhodovací stromy --------------------------------------------
```

```
#### nejdříve rozděluji data na trénovací a testovací množinu -----------------

#### do trénovací množiny zahrnu 70 % dat -------------------------------------

train_set_portion <- 0.7

set.seed(2017)


#### vytvářím množinu indexů pozorování, která budou zahrnuta do trénovací
#### množiny -----------------------------------------------------------------

train_set_indices <- sample(
    c(1:dim(my_data)[1]),
    floor(dim(my_data)[1] * train_set_portion),
    replace = FALSE
)


#### vytvářím trénovací a testovací množinu -----------------------------------

train_set <- my_data_wo_NA[train_set_indices, ]
test_set <- my_data_wo_NA[setdiff(c(1:dim(my_data)[1]), train_set_indices), ]


#### učím stromy --------------------------------------------------------------

size_table <- setNames(
    c(0.8, 1, 1, 1, 0.6, 0.4, 1, 0.4, 0.8, 1, 0.6, 1, 1, 1, 1, 1, 0.8, 0.8),
    response_variables
)


for(my_variable in response_variables){

    ## formuluji závislost proměnné na ostatních ----------------------------

    my_formula <- paste(

        my_variable,
        " ~ ",
        paste(setdiff(
            colnames(train_set[, response_variables]),
            my_variable
        ), collapse = " + "),
        sep = ""
```

```
)


## nechám vyrůst daný strom -------------------------------------------

eval(
    parse(
        text = paste(
            "my_tree",
            " <- ",
            "rpart(",
            "formula = ",
            my_formula,
            ", ",
            "data = train_set[, meaningful_variables]",
            ")",
            sep = ""
        )
    )
)


## ukládám strom do souboru -------------------------------------------

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

cairo_ps(
    file = paste(my_variable, "_tree.eps", sep = ""),
    width = 8 * size_table[my_variable],
    height = 5 * size_table[my_variable],
    pointsize = 18
)

par(mar = c(0, 0, 0, 0))

prp(
    my_tree,
    varlen = 20,
    faclen = 13,
    yes.text = "ano",
    no.text = "ne"
)

dev.off()

setwd(mother_working_directory)
```

```
    ## konfuzní matice a přesnost ----------------------------------------

    predict(object = my_tree, newdata = test_set, type = "class")

    my_table <- table(
        test_set[, my_variable],
        predict(object = my_tree, newdata = test_set, type = "class"),
        dnn = list("skutečné hodnoty", "predikované hodnoty")
    )

    print(
        "######################################################################"
    )

    print(
        paste("Proměnná ", my_variable, sep = "")
    )

    print(
        xtable(
            my_table,
            align = rep("", ncol(my_table) + 1),
            digits = 3
        ),
        floating = FALSE,
        tabular.environment = "tabular",
        hline.after = NULL,
        include.rownames = TRUE,
        include.colnames = TRUE
    )

    print(getMyAccuracy(my_table))


}


## ------------------------------------------------------------------------

##############################################################################
##############################################################################
##############################################################################
```

3 Reference

[1] BUDERA, P., Z. STRAKA, P. OSMANCIK, T. VANEK, S. JELINEK, J. HLAVICKA, R. FOJT, P. CERVINKA, M. HULMAN, M. SMID, M. MALY a P. WIDIMSKY. Comparison of cardiac surgery with left atrial surgical ablation vs. cardiac surgery without atrial ablation in patients with coronary and/or valvular heart disease plus atrial fibrillation: final results of the PRAGUE-12 randomized multicentre study. *European Heart Journal* [online]. 2012, **33**(21), 2644–2652. Dostupné z: doi:10.1093/eurheartj/ehs290

[2] WILSON, P. W. F., R. B. D'AGOSTINO, D. LEVY, A. M. BELANGER, H. SILBERSHATZ a W. B. KANNEL. Prediction of Coronary Heart Disease Using Risk Factor Categories. *Circulation* [online]. 1998, **97**(18), 1837–1847. Dostupné z: doi:10.1161/01.cir.97.18.1837

[3] KANNEL, William B. Some lessons in cardiovascular epidemiology from Framingham. *The American Journal of Cardiology* [online]. 1976, **37**(2), 269–282. Dostupné z: doi:10.1016/0002-9149(76)90323-4

[4] WIRTH, Rüdiger. CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*. 2000, s. 29–39.

[5] R CORE TEAM. *R: A Language and Environment for Statistical Computing* [online]. Vienna, Austria: R Foundation for Statistical Computing, 2016. Dostupné z: https://www.R-project.org/

[6] BERKA, Petr. *Dobývání znalostí z databází*. Praha: Academia, 2003. ISBN 80-200-1062-9.