FAKULTA MECHATRONIKY,
INFORMATIKY A MEZIOBOROVÝCH
STUDIÍ TUL

Indikace srdeční choroby

Dokumentace k semestrálnímu projektu

Vypracoval: Bc. Ondřej Mach

Předmět: Data Mining

Akademický rok: 2022/23

Vyučovaný semestr: Letní semestr

Jazyk projektu: R

Obsah

1. Úvod

Následující dokument obsahuje dokumentaci popisující celý postup tvorby semestrálního projektu pro předmět Data Mining na téma Indikace srdeční choroby. Obsahem je výběr a nastavení prostředí, použité knihovny, příprava a zpracování dat, tvorba a použití modelů, vizualizace a vyhodnocení jejich úspěšnosti.

První skript (*heart_pt1.R*) je věnovaný analýze a vyhodnocení některých úkolů řešených na hodině. Dále obsahuje Feature selection a přípravu na druhou část.

Druhý skript (*heart_pt2.R*) je zaměřen na modely C&R Tree, Neural net a CHAID (přeněji Ctree – viz. dale).

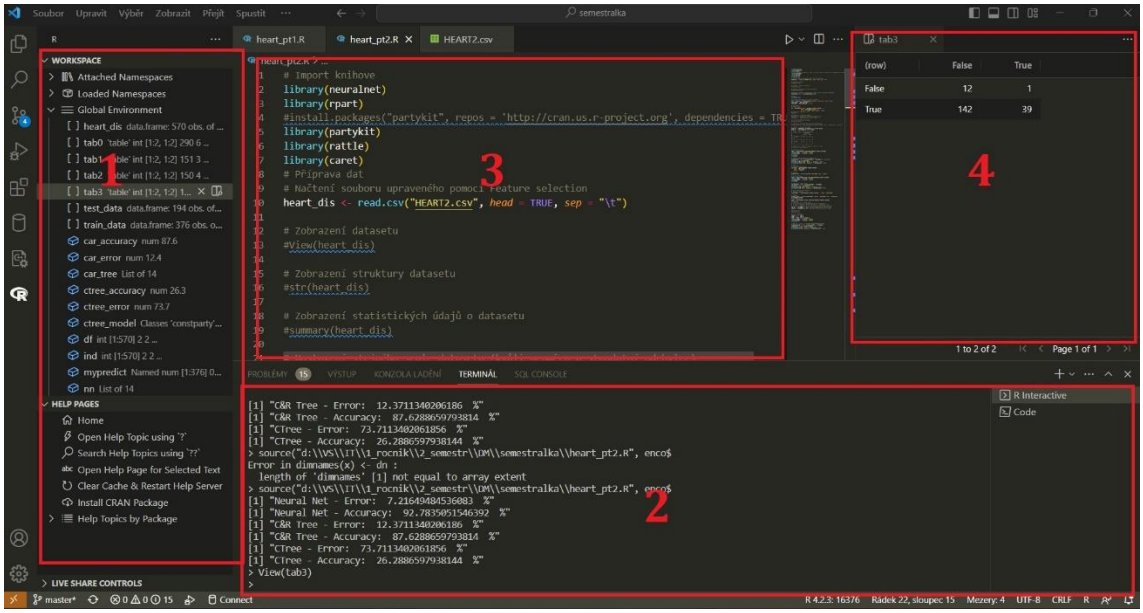
2. Příprava prostředí

Projekt je zpracován v programovacím jazyce R (https://www.r-project.org/) a vyvíjen v editoru Visual Studio Code (https://code.visualstudio.com/). V rámci tohoto editoru je také doporučeno nainstalovat rozšíření pro programovací jazyk R (https://marketplace.visualstudio.com/items?itemName=REditorSupport.r).
Všechny užité nástroje jsou zdarma dostupné online.

Po instalaci uvedených nástrojů je třeba otevřít si složku, kde se bude na projektu pracovat. Pracovní prostředí se skládá ze čtyř hlavních části, kterými jsou:

1. Průzkumník souborů (případně workspace)
2. Příkazový řádek
3. Pole pro psaní kódu
4. Okno pro vizualizaci dat

Všechny části jsou zobrazeny na *Obrázku 1*. Podstatné je zmínit, že v projektu je užíván příkazový řádek klasický (*Code*), a dále také *R Interactive*. Oba se pří zpracování projektu osvědčily v různých situacích.



Obrázek 1: Vývojové prostředí

3. Načtení knihoven

Před využívání knihoven je třeba jednotlivé knihovny nainstalovat. K tomu slouží funkce *install.packages()*, která stáhne balík obsahující knihovnu a nainstaluje ho. V rámci této funkce je třeba zadat zdroj a název. Konkrétní příkazy pak vypadají například takto:

- *install.packages("partykit", repos = 'http://cran.us.r-project.org', dependencies = TRUE)*
- *install.packages("CHAID", repos = "http://R-Forge.R-project.org", type = "source")*

Import knihoven pak pomocí *library()*. Jak již bylo zmíněno, tento projekt se dělí na dvě části:

1. Část – knihovny: *dplyr*, *mlbench*, *caret*, *ggplot*
2. Část – knihovny: *neuralnet, rpart, partykit, rattle, caret*

4. První část

V této části načteme data, získáme určité informace, vykreslíme několik grafů, a nakonec provedeme Feature selection, kdy výstupem budou nová data.

Načtení dat a základní úprava dat

Data načteme z upraveného souboru *HEART.txt*, který je přiložen v projektu. Jedná se o soubor s upravenou hlavičkou pro lepší čitelnost po načtení. Načtení je provedeno pomocí *read.csv()*, jako nový data frame, kde jako separator volíme tabulátor. Dále je možné si je zobrazit pomocí *View()*, puštěním v *R Interactive*.

Po načtení jsou data pomocí *mutate()* z knihovny *dplyr* převedena na datové typy vhodnější pro další práci (typy factor a flag).

Základní úkoly

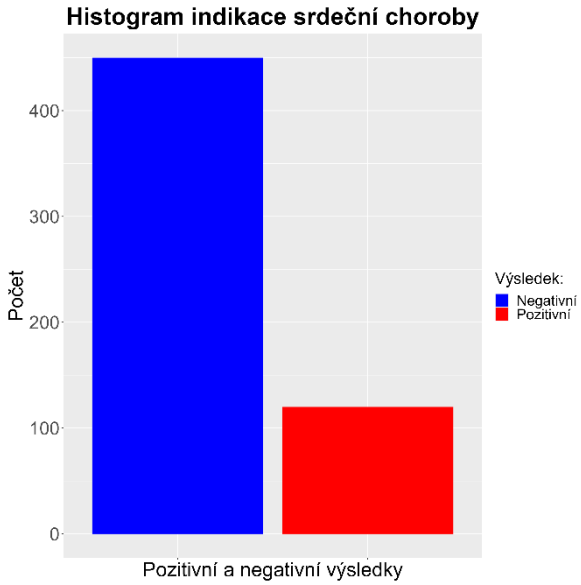
Prvním úkolem bylo určit průměrnou hladinu cholesterolu v krvi u pacientů nad 50 let. Toho dosáhneme vytvořením subsetu s limitující věkem (>= 50) a následným výpočtem průměru tohoto subsetu pomocí *mean()*. Výsledek je 253.2164.

Druhým úkolem bylo zjištění průměrných hodnot, minima a maxima v některých prediktorech dat. Možných řešení je více, nicméně v našem případě byly pomocí *apply()* nasazeny funkce *mean(), min()* a *max()*. Reprezentace je formou dataframu v *Tabulce 1.*

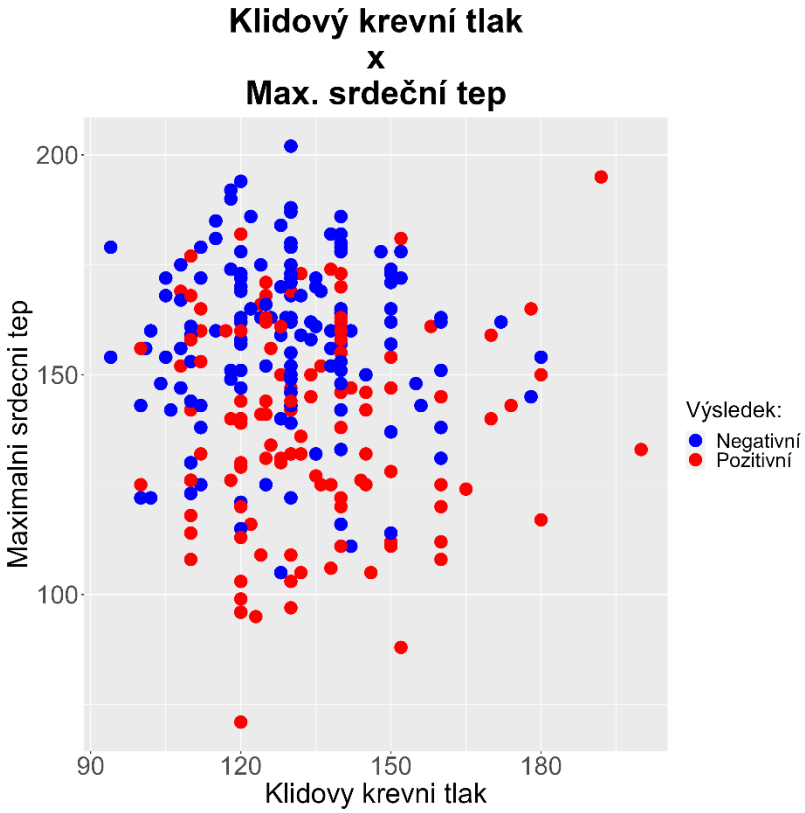
Tabulka 1: Analýza dat

	Průměr	**Min**	**Max**
Věk	53.5245614	29	77.0
Klidový krevní tlak mm	130.0403509	94	200.0
Hladina cholesterolu v krvi mg/dl	246.7929825	126	564.0
Max. srdeční tep	154.2333333	71	202.0
Hodnota stres testu	0.8250877	0	6.2

Třetím a čtvrtým úkolem bylo zobrazení histogramu Indikace srdeční choroby a graf Klidový krevní tlak x Maximální srdeční tep. Oba grafy byly sestaveny pomocí knihovny *ggplot2* a příkazu *ggplot().* Uložení pak proběhlo pomocí *ggsave(),* ve formě obrázků. Výstupy jsou k vidění jako *Obrázek 2* a *Obrázek 3*.



Obrázek 2



Obrázek 3

Feature selection

Na závěr první části byla provedena Feature selection, tj. způsob výběru podstatných prediktorů. Využita byla knihovna *caret* a konkrétní automatická metoda jménem Recursive Feature Elimination. V kódu jako *rfe().* Argumenty jsou kromě prediktorů také control. Ten je v kódu pod názvem *rfeControl()* a využívá funkci Random forest selection.

Po provedení Feature selection byl jako nedůležitý prediktor vyhodnocen „Hladina cukru nad 120 mg/dl“. Proběhlo jeho odstranění a uložení nových dat jako *HEART2.csv,* pomocí funkce *write.table().*

5. Druhá část

Příprava dat

Po načtení knihoven byly data načteny obdobně jako v části první. Nyní byl ovšem použit soubor *HEART2.csv.* Byl založen seed pro reprodukovatelnost modelování a data byla upravena funkcí *scale().* Posledním krokem přípravy bylo rozdělení dat na testovací a trénovací pomocí *sample()*. Zvolený poměr byl 70 % trénovací a 30 % testovací (vedl k optimálnější přesnosti modelů).

Neural Net

Prvním modelem byla neuronová síť. Tento model byl vytvořen užitím funkce *neuralnet()* ze stejnojmenné knihovny. Model se učil na trénovacích datech a následně byl nasazen na testovací. Úspěšnost je hodnocena pomocí tabulky, kterou tvoří matice záměn. Ta je zobrazena v *Tabulce 2*. Podle vzorce byla z matice záměn vypočtena přesnost (**92.78 %**) a chyba (**7.22 %**).

Tabulka 2: Matice záměn Neural Net

	False	**True**
False	151	11
True	3	29

C&R Tree

Zde bylo postupováno obdobně jako u neuronové sítě. Model byl naučen na trénovacích datech – knihovna *rpart*, použitím funkce *rpart()*, ve které bylo nutné zvolit *method = "class".* Následná predikce zde byla provedena funkcí *predict()*, kde obdobně bylo třeba přidat argument *type = "class".* Pro ohodnocení přesnosti byla sestavena matice záměn (*Tabulka 3*) a vypočtena přesnost (**87.63 %**) a chyba (**12.37 %**).

Tabulka 3: Matice záměn C&R Tree

	False	**True**
False	150	20
True	4	20

CHAID (Ctree)

Posledním modelem měl být CHAID. Tento model se bohužel nepodařilo zprovoznit. Chyba se pravděpodobně vázala k potřebě modelu pouze faktorových prediktorů. Bohužel i při pokusech o přetypování a případně kompletní smazání kontinuálních prediktorů nedošlo k vyřešení chyby.

Vyzkoušen byl i Ctree, který by měl být flexibilnější a měl by zvládnout pracovat i typy pro CHAID nevhodnými. Bohužel ani tento model neuspěl a vyústil v pravděpodobně nekorektní matici záměn. Míra přesnosti (**26.29 %**) a chyby (**73.71 %**) vycházela v nepraktických hodnotách.

6. Vyhodnocení a závěr

V rámci úlohy byla zkoumána možnost indikace srdeční choroby. V první části byla data analyzována a následně podrobena Feature section, která vyřadila z datasetu „Hladina cukru nad 120 mg/dl“. Následně byl vytvořen nový dataset a na něm byly testovány modely: Neural Net, C&R Tree a CHAID. První dva modely byly implementovány úspěšně, třetí nikoliv. Místo CHAID byl využit Ctree, ale ani ten bohužel neposkytl uspokojivé výsledky.

Tabulka 4: Shrnutí přesnosti modelů

	Přesnost [%]	**Chyba [%]**
Neural Net	92.78	7.22
C&R Tree	87.63	12.37
Ctree	26.29	73.71