1. Pochopenie cieľa
   - Popíšete, prečo ste si Vaše dáta zvolili, čo v nich chcete sledovať, resp. čo Vás zaujíma atď.
   - Definujte cieľov dolovania v dátach (konkrétnych, na ktoré dokážete v závere odpovedať)
2. Pochopenie dát
   - Spomeniete odkiaľ sú Vaše dáta, kde sa dajú stiahnuť, aké obsahujú atribúty, popíšete každý atribút, dátové typy, môžete to potom zhrnúť napríklad do jednej tabuľky ako ste zmenili dátové typy
   - Do tejto fázy pridáte pri pochopení dát aj nejaké grafy, napr. početnosti záznamov pre jednotlivé kategórie cieľového atribútu, zobrazenie distribúcie hodnôt numerického atribútu napr. pomocou boxplotu, taktiež porovnanie boxplotov vedľa seba napr. pre rôzne kategórie nominálneho atribútu, pri nominálnych atribútoch zobraziť nejaké barploty, atď.
   - Každý jeden graf musíte popísať, čiže čo ste ním chceli sledovať a taktiež nejaký záver čo ste sa z neho dozvedeli
   - Sledovanie korelácií medzi numerickými atribútmi, taktiež medzi nominálnymi atribútmi sledovanie závislosti pomocou Chi-kvadrát testu. Všetko objasniť, čo daný výsledok znamenal. Môžete pridať aj nejaký corrgram atď.
   - Pri pochopení dát už môžete spomenúť aj početnosť NA hodnôt, ale ešte ich nemusíte mazať, resp. dopĺňať
   - Bolo na to zamerané cvičenie č.7, kde máte vzorové ukážky na Pochopenie dát
3. Príprava dát
   - Hlavná časť je úprava NA hodnôt, čiže objasniť ako s nimi budete pracovať a prečo ste si zvolili daný spôsob spracovania dát
   - Ak máte veľa NA hodnôt tak po ich úprave môžete tiež niektoré časti z Pochopenia dát vykonať opäť a pozrieť si, či sa nejaké súhrnné štatistiky zmenili
   - Experimentovať so zmenou dátových typov, napr. numerický atribút skúsiť zmeniť na intervaly (kategórie) a potom v modelovaní porovnávať výsledky
   - Tiež máte nejaké ukážky na prípravu dát v cvičení č.7
4. Modelovanie
   - Na úvod spomeňte aké techniky, algoritmy budete používať, ďalej či sa zameriate na predikciu numerického atribútu alebo klasifikáciu záznamov do tried atď.
   - Vykonajte rôzne rozdelenia do trénovacej a testovacej množiny
   - Vykonajte rôzne kombinácie výberu atribútov na vytváranie modelu
   - Vykreslite a tiež popíšte vytvorené modely
5. Vyhodnotenie
   - Posúdenie modelov, analýza výsledkov a ich dôležitosti
   - Záverečné vyhodnotenie vytvorených modelov ich porovnanie podľa ukazovateľov (napr. presnosť, chybovosť atď.)
6. Nasadenie
   - RShiny, ktoré bude obsahovať viacero okien (kariet)
   - Ukážka peknej RShiny aplikácie jedného diplomanta, ktorý končil minulý rok - https://dpmatfiak.shinyapps.io/PodporaDiagnostikyPch/ (Samozrejme nemusí byť až na takej úrovni, ale skúste sa s tým dostatočne pohrať. Niečo na spôsob ako je karta Rozhodovacie stromy a Zhlukovanie, ale tiež tam pridajte aj tie pre vytváranie grafov jednotlivých atribútov a výpis dát).

Na záver malá ukážka RMarkdown dokumentu zo zadania z minulého roku

Pochopenie dat

Prvotny zber dat

Dáta použité v tejto práci sú voľne dostupné na internete, na webovej stránke UCI Machine Learning Repository. Na ich stránke sa aktuálne nachádza 351 datasetov, ktoré slúžia či už študentom, učiteľom alebo analytikom ako podklad pre strojové učenie alebo dolovanie v dátach. Dáta boli zozbierané na klinike Cleveland Clinic Foundation v roku 1988, autorom je Robert Detrano, M.D., Ph.D. Nachádzajú sa v nich medicínske informácie o ľuďoch, ktorí či už trpeli alebo netrpeli na chorobu srdca.

Po stiahnutí dát a ich načítaní bolo potrebné priradiť k stĺpcom ich názvy, ktoré boli uvedené na tej istej stránke.

Nazvy stlpcov po stiahnuti dat.

```
##  [1] "V1"  "V2"  "V3"  "V4"  "V5"  "V6"  "V7"  "V8"  "V9"  "V10" "V11"
## [12] "V12" "V13" "V14"
```

Nazvy stlpcov po zmene nazvov.

```
##  [1] "age"      "sex"      "cp"       "trestbps" "chol"     "fbs"
##  [7] "restecg"  "thalach"  "exang"    "oldpeak"  "slope"    "ca"
## [13] "thal"     "num"
```

Po krátkom prezretí dát bolo zistené, že množstvo atribútov má priradený zlý dátový typ. Tieto chybne definované dátové typy atribútov bolo nutné čo najskôr zmeniť, nemalo by zmysel vykonávať ďalšie analýzy a skúmanie.

Povodne datove typy atributov.

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex     : num  1 1 1 1 0 1 0 0 1 1 ...
##  $ cp      : num  1 4 4 3 2 2 4 4 4 4 ...
##  $ trestbps: num  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol    : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs     : num  1 0 0 0 0 0 0 0 0 1 ...
```