Podniková analytika (Pochopenie a príprava dát - riešenie)

Úlohy na pochopenie a prípravu dát

- Stiahnite dáta **El Nino** dostupné na https://archive.ics.uci.edu/ml/machine-learning-databases/el_nino-mld/elnino.gz a uložte ich do pracovného adresára.

```
fileUrl<-"https://archive.ics.uci.edu/ml/machine-learning-
databases/el_nino-mld/elnino.gz"
download.file(fileUrl, destfile = "elnino.gz")
```

- Načítajte tieto dáta do premennej **elnino** použitím príkazu *read.table*.

```
elnino = read.table("elnino.gz", header=FALSE)
```

- Na webovej stránke https://archive.ics.uci.edu/ml/datasets/El+Nino sú dostupné informácie o datasete **El Nino** a taktiež popis jeho atribútov. Prečítajte si túto časť (hlavne **Attribute Information**) a podľa toho doplňte názvy stĺpcov.

```
colnames(elnino)    <-    c("buoy","day","latitude","longitude",
"zon.winds","mer.winds","humidity","airtemp","s.s.temp")
```

- Vypíšte informácie o dátach a ich atribútoch a upravte ich typ podľa potreby.

```
str(elnino)
elnino$zon.winds = as.numeric(as.character(elnino$zon.winds))
elnino$mer.winds = as.numeric(as.character(elnino$mer.winds))
elnino$humidity = as.numeric(as.character(elnino$humidity))
elnino$airtemp = as.numeric(as.character(elnino$airtemp))
elnino$s.s.temp = as.numeric(as.character(elnino$s.s.temp))
```

- Skontroluje dátový typ a počet chýbajúcich hodnôt každého atribútu.

```
str(elnino)
summary(elnino)
```

- Vypíšte početnosť záznamov pre jednotlivé dni (atribút **Day**) a zobrazte tieto početnosti aj pomocou grafu (*barplot*).

```
table(elnino$day)
barplot(table(elnino$day))
```

- Zobrazte v grafe (*boxplot*) rozdelenie hodnôt atribútu **Teplota vzduchu** (Air Temperature) pomocou systému vykrasľovania *Base* a *Lattice*.

```
library(lattice)
boxplot(elnino$airtemp)
bwplot(elnino$airtemp)
```

- Zobrazte pomocou *histogramu* hodnoty atribútu, ktorý informuje o **Teplote hladiny mora** (Sea Surface Temperature).

```
hist(elnino$s.s.temp, breaks = 10)
```

- Vytvorte nový data.frame s názvom **elnino_2**, ktorý bude obsahovať skopírované dáta **elnino.**

```
elnino_2 = elnino
```

- Doplňte v **elnino_2** chýbajúce hodnoty atribútov podávajúcich informácie o **Vetroch** (Zonal Winds, Meridional Winds) tak, že ich nahradíte priemerom daného stĺpca.

```
elnino_2$zon.winds     <-     ifelse(is.na(elnino_2$zon.winds),
mean(elnino_2$zon.winds,na.rm = TRUE),elnino_2$zon.winds)
elnino_2$mer.winds     <-     ifelse(is.na(elnino_2$mer.winds),
mean(elnino_2$mer.winds,na.rm = TRUE),elnino_2$mer.winds)
```

- Doplňte v **elnino_2** chýbajúce hodnoty atribútov **Relatívna vlhkosť** (Relative Humidity) a **Teplota vzduchu** (Air Temperature) tak, že ich nahradíte hodnotou mediánu daného stĺpca.

```
elnino_2$humidity      <-      ifelse(is.na(elnino_2$humidity),
median(elnino_2$humidity,na.rm = TRUE),elnino_2$humidity)
elnino_2$airtemp       <-       ifelse(is.na(elnino_2$airtemp),
median(elnino_2$airtemp,na.rm = TRUE),elnino_2$airtemp))
```

- V dataframe **elnino_2** vymažte atribút **Teplota hladiny mora (**Sea Surface Temperature) a nakoniec vypíšte informácie (sumár) o týchto dátach.

```
elnino_2$s.s.temp <- NULL
summary(elnino_2)
```

- Vytvorte dataframe s názvom **elnino_1**, ktorý bude obsahovať dáta **elnino** očistené od všetkých chýbajúcich hodnôt.

```
elnino_1 = na.omit(elnino)
```

- Vytvorte globálny pohľad na závislosti medzi numerickými atribútmi v dátach **elnino_1** (použite funkciu *pairs*).

```
pairs(~buoy+day+latitude+longitude+zon.winds+mer.winds+humidity
+airtemp+s.s.temp, data=elnino_1)
```

- Vypíšte hodnoty vyjadrujúce silu korelácie medzi atribútmi v dátach **elnino_1** pomocou korelačnej matice.

```
cor(elnino_1)
```

- Zobrazte silu korelácie medzi atribútmi v dátach **elnino_1** pomocou špeciálneho typu grafu (*korelogram*).

```
corrgram(elnino_1, order=TRUE, lower.panel=panel.shade,
         upper.panel=panel.pie, text.panel=panel.txt)

corrgram(elnino_1, order=TRUE, lower.panel=panel.ellipse,
         upper.panel=panel.pts,           text.panel=panel.txt,
         diag.panel=panel.minmax)
```

- Vytvorte v dátach **elnino_1** nové atribúty s názvami **zon.winds_1** a **mer.winds_1**, ktoré budú odvodené z atribútov **Zon.winds** a **Mer.winds** podľa podmienok popísaných v informáciách o atribútoch (Attribute Information) na https://archive.ics.uci.edu/ml/datasets/El+Nino .

```
elnino_1$zon.winds_1 = ifelse(elnino_1$zon.winds < 0,"West", "East")
elnino_1$mer.winds_1 = ifelse(elnino_1$mer.winds < 0,"South", "North")
```

- Usporiadajte dataframe **elnino_1** takým spôsobom, že atribút **zon.winds** sa bude nachádzať pri **zon.winds_1** a rovnako aj **mer.winds** pri atribúte **mer.winds_1.**

```
elnino_1 <- elnino_1[,c(1,2,3,4,5,10,6,11,7,8,9)]
```

- Opäť vypíšte informácie o dátach **elnino_1** a upravte dátový typ novo vytvorených atribútov na faktor.

```
summary(elnino_1)
elnino_1$zon.winds_1 = as.factor(elnino_1$zon.winds_1)
elnino_1$mer.winds_1 = as.factor(elnino_1$mer.winds_1)
```

- Vytvorte kontingenčnú tabuľku s názvom **tbl** z dvoch nominálnych atribútov v dátach **elnino_1** a sledujte závislosť medzi nimi pomocou Chi-kvadrát testu.

```
tbl = table(elnino_1$zon.winds_1, elnino_1$mer.winds_1)
chisq.test(tbl)
```

- Pre nasledujúce úlohy (normalizáciu a diskretizáciu) nainštalujte a načítajte balíčky s názvami *clusterSim, arules*.

```
install.packages("clusterSim")
library(clusterSim)
install.packages("arules")
library(arules)
```

- Normalizujte na nulový stred atribút **Relatívna vlhkosť (**Humidity) a atribút **Teplota vzduchu** (Air Temperature) na rozsah <-1,1> (*Poznámka*: použite funkciu **data.Normalization**, typ normalizácie zvoľte podľa strany č. 12 v dokumente https://cran.r-project.org/web/packages/clusterSim/clusterSim.pdf)

```
elnino_1$humidity <- data.Normalization(elnino_1$humidity, type = "n4", normalization = "columns")
elnino_1$airtemp <- data.Normalization(elnino_1$airtemp, type = "n5", normalization = "columns")
```

- Vytvorte nový atribút v dátach **elnino_1** s názvom **s.s.temp_1**, kde budú hodnoty (s.s.temp) rozdelené do 4 rovnako veľkých intervalov (Ekvidištančná diskretizácia). Vypíšte taktiež hranice intervalov spolu s početnosťou záznamov, ktoré do nich spadajú.

```
elnino_1$s.s.temp_1 <- cut(elnino_1$s.s.temp, breaks = 4, dig.lab = 10)
table(cut(elnino_1$s.s.temp, breaks = 4, dig.lab = 10))
```

alebo

```
table(elnino_1$s.s.temp_1)
```

- Vytvorte nový atribút v dátach **elnino_1** s názvom **s.s.temp_2**, kde budú hodnoty (s.s.temp) rozdelené do 3 intervalov vyjadrených triedou od 1,2,3. Hranice intervalov zvoľte 20,24,28 a 32. Vypíšte taktiež početnosti záznamov, ktoré spadajú do tried 1,2,3.

```
elnino_1$s.s.temp_2 <- findInterval(elnino_1$s.s.temp, c(20,24,28,32))
table(findInterval(elnino_1$s.s.temp, c(20,24,28,32)))
```

- Vytvorte nový atribút v dátach **elnino_1** s názvom **s.s.temp_3**, kde budú hodnoty (s.s.temp) pridelené do 5 intervalov s rôznou šírkou, ale rovnakou početnosťou záznamov (Ekvifrekvenčná diskretizácia). Vypíšte taktiež početnosti záznamov, ktoré spadajú do vytvorených intervalov.

```
elnino_1$s.s.temp_3 <- discretize(elnino_1$s.s.temp, "frequency",categories = 5)
table(discretize(elnino_1$s.s.temp, "frequency", categories = 5))
```

alebo

```
table(elnino_1$s.s.temp_3)
```

- Vyberte z dát **elnino_1** len vzorku 100 náhodných záznamov a uložte ich do dataframeu s názvom **sample_data**.

```
sample_data <- elnino_1[sample(100),]
```