Știința Datelor Laborator 2

ȘD - Scop Laboratorul 2

Scopul laboratarului 2 de Știința Datelor este:

- dezvoltarea capacității de colectare a datelor
- dezvoltarea abilitățlori matematice necesare pentru laboratoarele de ȘD

Colectare

Un process important în știința datelor este acela de colectare a datelor. Vom folosi datele prezentate în https://ec.europa.eu/eurostat/databrowser/view/educ_uoe_enrt01/default/table?lang=en https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL_LINEAR&StrNom=CL_ISCED11&StrLanguageCode=EN&IntCurrentPage=2

Descarcare fisiere

R are o parte de scripting. Putem folosi funcții precum ls, rm, mv. Înainte de a descărca fișiere trebuie să creăm un director unde vom pune aceste fișiere.

```
if (!file.exists("data")) {
    dir.create("data")
}
```

După ce am creat acest director avem nevoie de path-ul fișierul și respectiv url-ul către fișeierul pe care dorim să îl descărcăm.

```
file_path <- file.path("data", "fisier_nou.txt")
file_url <- "https://raw.githubusercontent.com/sdcioc/LaburiSD/main/data/fisier.txt"
```

Acum putem descărca fișierul:

```
download.file(file_url, destfile = file_path, method = "curl")
```

Fisiere locale

O dată ce avem fișieul local putem să îl citim. Vom trece mai departe prin mai multe formate de fișiere. ### TXT format Pentru fișiere txt putem folosi funcția read.table. Separatorul în fișierul descărcat precedent este "," și fișierul are antet.

```
file_path <- file.path("data", "fisier.txt")
# file_url <-
# 'https://raw.githubusercontent.com/sdcioc/LaburiSD/main/data/fisier.txt'
# download.file(file_url, destfile=file_path, method='curl')
data <- read.table(file_path)
head(data)
```

```
##   worktime sector sex isced11 country Year Total
## 1       FT   PRIV   F     ED5      AT 2005    NA
## 2       FT   PRIV   F     ED5      BE 2005    NA
## 3       FT   PRIV   F     ED5      BG 2005    NA
## 4       FT   PRIV   F     ED5      CH 2005    NA
## 5       FT   PRIV   F     ED5      CY 2005    NA
## 6       FT   PRIV   F     ED5      CZ 2005    NA
```

CSV format

Pentru fișiere csv.

```
file_path <- file.path("data", "fisier.csv")
# file_url <-
# 'https://raw.githubusercontent.com/sdcioc/LaburiSD/main/data/fisier.csv'
# download.file(file_url, destfile=file_path, method='curl')
data <- read.csv(file_path)
head(data)
```

```
##   X worktime sector sex isced11 country Year Total
## 1 1       FT   PRIV   F     ED5      AT 2005    NA
## 2 2       FT   PRIV   F     ED5      BE 2005    NA
## 3 3       FT   PRIV   F     ED5      BG 2005    NA
## 4 4       FT   PRIV   F     ED5      CH 2005    NA
## 5 5       FT   PRIV   F     ED5      CY 2005    NA
## 6 6       FT   PRIV   F     ED5      CZ 2005    NA
```

XLSX format

pentru fișiere xlsx

```
library("openxlsx")
file_path <- file.path("data", "fisier.xlsx")
# file_url <-
# 'https://raw.githubusercontent.com/sdcioc/LaburiSD/main/data/fisier.xlsx'
# download.file(file_url, destfile=file_path, method='curl')
data <- read.xlsx(file_path, sheet = 1)
head(data)
```

```
##   X1 worktime sector sex isced11 country Year Total
## 1  1       FT   PRIV   F     ED5      AT 2005    NA
## 2  2       FT   PRIV   F     ED5      BE 2005    NA
## 3  3       FT   PRIV   F     ED5      BG 2005    NA
## 4  4       FT   PRIV   F     ED5      CH 2005    NA
## 5  5       FT   PRIV   F     ED5      CY 2005    NA
## 6  6       FT   PRIV   F     ED5      CZ 2005    NA
```

XML format

```
library("XML")
file_path <- file.path("data", "fisier.xml")
# file_url <-
# 'https://raw.githubusercontent.com/sdcioc/LaburiSD/main/data/fisier.xml'
```

```
# download.file(file_url, destfile=file_path, method='curl')
data <- xmlTreeParse(file_path, useInternal = TRUE)
root_node <- xmlRoot(data)
head(xmlSApply(root_node, xmlValue))
```

```
##                  row                  row                  row
## "FTPRIVFED5AT2005NA" "FTPRIVFED5BE2005NA" "FTPRIVFED5BG2005NA"
##                  row                  row                  row
## "FTPRIVFED5CH2005NA" "FTPRIVFED5CY2005NA" "FTPRIVFED5CZ2005NA"
```

```
head(xpathSApply(root_node, "//country", xmlValue))
```

```
## [1] "AT" "BE" "BG" "CH" "CY" "CZ"
```

JSON format

pentru fișiere json.

```
library("jsonlite")
file_path <- file.path("data", "fisier.json")
# file_url <-
# 'https://raw.githubusercontent.com/sdcioc/LaburiSD/main/data/fisier.json'
# download.file(file_url, destfile=file_path, method='curl')
data <- fromJSON(file_path)
head(data)
```

```
##   worktime sector sex isced11 country Year Total
## 1       FT   PRIV   F     ED5      AT 2005    NA
## 2       FT   PRIV   F     ED5      BE 2005    NA
## 3       FT   PRIV   F     ED5      BG 2005    NA
## 4       FT   PRIV   F     ED5      CH 2005    NA
## 5       FT   PRIV   F     ED5      CY 2005    NA
## 6       FT   PRIV   F     ED5      CZ 2005    NA
```

Baze de date

O altă sursă pentru informații o reprezintă bazele de date. Voi prelua una dintre cele mai folosite în știința datelor MySQL, dar sunt variante și pentru MongoDB, Cassandra, Oracle etc. o conecxiune se face cu dbConnect oferinf parola si calea la care să se lege. Pot fi trimise în două feluri cereriile către baza de date prin getqueery care returnează răspunsul. Sau prin send querry care trimite doar cerrea către baza de date iar răspunsul poae fi luat mai târziu pe părți. Efificient din punct de vedere al memoriei. La final se eliberea întrebarea din memoria bazei de date. Și Ne deconectăm de la baza de date.

RMySql

```
library("RMySQL")
```

```
## Loading required package: DBI
```

```
uDBC <- dbConnect(MySQL(), user = "genome", host = "genome-mysql.cse.ucsc.edu")
response <- dbGetQuery(uDBC, "show databases;")
head(response, n = 5)
```

```
##   Database
## 1  acaChl1
## 2  ailMel1
## 3  allMis1
## 4  allSin1
## 5  amaVit1
```

```
question <- dbSendQuery(uDBC, "show databases;")
answer <- fetch(question, n = 10)
dbClearResult(question)
```

```
## [1] TRUE
```

```
print(answer)
```

```
##    Database
## 1   acaChl1
## 2   ailMel1
## 3   allMis1
## 4   allSin1
## 5   amaVit1
## 6   anaPla1
## 7   ancCey1
## 8   angJap1
## 9   anoCar1
## 10  anoCar2
```

```
dbDisconnect(uDBC)
```

```
## [1] TRUE
```

Pagini Web

Următoarea sursă sunt paginiile web pe care putem să le descărcăm complet în scripturi. Sunt 3 variante, prin a face o conexiune url la pagina și a citi toate liniile într-o variabilă., de a folosi XML pentru a citi pagina formatată html sau de a folosi biblioteca httr care permite autentificare în plus și respectiv păstrarea de cookies.

url

```
con = url("https://ocw.cs.pub.ro/courses/pr/labs/04")
codhtml <- readLines(con)
codhtml[1:10]
```

```
##  [1] ""
##  [2] "    "
##  [3] ""
##  [4] "    <!DOCTYPE html PUBLIC \"-//W3C//DTD XHTML 1.0 Transitional//EN\""
##  [5] " \"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd\">"
##  [6] "<html xmlns=\"http://www.w3.org/1999/xhtml\" xml:lang=\"en\""
##  [7] " lang=\"en\" dir=\"ltr\">"
##  [8] "<head>"
##  [9] "  <meta http-equiv=\"Content-Type\" content=\"text/html; charset=utf-8\" />"
## [10] "  <title>"
```

```
close(con)
```

xml

```
library("XML")
url <- "https://ocw.cs.pub.ro/courses/pr/labs/04"
codhtml <- htmlTreeParse(url, useInternalNodes = TRUE)
root_node <- xmlRoot(codhtml)
head(xmlSApply(root_node, xmlValue))
```

```
##                                       body
## "https://ocw.cs.pub.ro/courses/pr/labs/04"
```

httr

```
library("httr")
# html <- GET(url, authenticate('', ''))
url <- "https://ocw.cs.pub.ro/courses/pr/labs/04"
html <- GET(url)
html_content <- content(html, as = "text")
codhtml <- htmlParse(html_content, asText = TRUE)
root_node <- xmlRoot(codhtml)
head(xmlSApply(root_node, xmlValue), n = 1)
```

```
##
## "\n    Laboratorul 04. OSPFv3 advanced    [CS Open CourseWare]\n  /*<![CDATA[*/var NS='pr:labs';var .
```

```
# pagina <- handle(url) #pentru păstrare cookies
# GET(handle=pagina, path='/')
```

Interfață de programare a aplicațiilor

O altă metodă de colectare a datelor de pe platforme mai complexe este prin intermediul interfețelor de programare a aplicațiilor. Mare parte din ele se fac prin Oauth, dar am pus link-uri pentru fiecare implementare pentru diferite platforme. Se poate căuta „platformă R programming API package" , este un mod de analiza chiar datele voastre personale sau a prieteniilor.

Facebook[https://cran.r-project.org/web/packages/Rfacebook/], Twitter [https://www.rdocumentation.org/packages/twitteR], Google [https://cran.r-project.org/web/packages/googleAuthR/vignettes/setup.html], Github[https://cran.r-project.org/web/packages/httr/vignettes/api-packages.html]

Test de clasificare binară

Se consider un test de clasificare binară un test care depistează un atribut ce împarte o pupulație în două. Cele mai populare test sunt cele din medicină care spun dacă ești pozitiv pentru un anumit virus.

Definiții matematice

Fie $X = \{$toți oameni care au făcut un test covrig$\}$. Fie $B = \{x \in X \mid x$ are Covrig$\}$ și $S = \{x \in X \mid x$ nu are Covrig$\}$. Fie T un test atunci $P = \{x \in X \mid x$ a ieșit pozitiv la testul $T\}$ și $N = \{x \in X \mid x$ a ieșit negativ la testul $T\}$. Avem următoarele proprietăți: $X = B \cup S, B \cap S = \emptyset, X = P \cup N, P \cap N = \emptyset$. În general $\mid B \mid = \mid S \mid$.

Denumire	Formulă	Descriere
AP	$B \cap P$	adevărat pozitivi
AN	$S \cap N$	adevărat negativi
FN	$B \cap N$	falși negativi
FP	$S \cap P$	falși pozitivi
pB	$\mid B \mid / \mid X \mid$	proporția din populație care au boala
pS	$\mid S \mid / \mid X \mid$	proporția din populație care sunt sănătoși
p(AP)	$\mid B \cap P \mid / \mid B \mid$	rata de adevărat pozitivi
p(FP)	$\mid S \cap P \mid / \mid S \mid$	rata de falși pozitivi
p(AN)	$\mid S \cap N \mid / \mid S \mid$	rata de adevărat negativi
p(FN)	$\mid B \cap N \mid / \mid B \mid$	rata de falși negativi

Matrice de confuzie

Matricea de confuzie este o matrice 2x2 unde coloane reprezintă partea din populație care este bolnavă sau sănătoasă iar rânduri rezultatul la test. fiecare celulă conține numărul de oameni care fac parte dintr-o anumită parte a populație cu un anumit rezultat la test.

Test	Bolnavi	Sănătoși
Pozitiv	96	2
Negativ	4	98

În R pentru a calcula matricea de confuzie se folosește funcția table. Un exemplu putem da pe setul de date mtcars care conține date despre mai mașini. Propun un test teoretic în care în funcție de cai putere spunem dacă mașina poate faci mai mult de 20 de mile cu un galon de combustibil. Îns etul de date există și adevarată valoare, dar vom verifica dacă tesul nostru este bun.

```
data <- table(mtcars$hp < 110, mtcars$mpg > 20)
dimnames(data) <- list(c("<110", ">=110"), c(">20", "<=20"))
print(data)
```

```
##       >20 <=20
## <110   17    4
## >=110   1   10
```

```
pozitive_rate = data[1, 1]/sum(data[, 1])
print(pozitive_rate)
```

```
## [1] 0.9444444
```

Termeni Statistici

Media

Media: se calculeaza suma tuturor valorilor si se imparte la numarul total de intrari de setul de date

```
x <- c(1, 2, 3, 4, 5, 1, 2, 3, 1, 2, 4, 5, 2, 3, 1, 1, 2, 3,
    5, 6)  # our data set
mean.result = mean(x)  # calculate mean
print(mean.result)
```

```
## [1] 2.8
```

Mediana

Mediana se obtine ordonand crescator numerele din setul de date si extragand valoarea de la pozitia din mijloc. In cazul sirurile cu numar par de elemente, se face media aritmetica a elementelor din pozitiile de mijloc.

```
x <- c(1, 2, 3, 4, 5, 1, 2, 3, 1, 2, 4, 5, 2, 3, 1, 1, 2, 3,
    5, 6)  # our data set
median.result = median(x)  # calculate median
print(median.result)
```

```
## [1] 2.5
```

Variația

Variatia ne spune cat de departate sunt numerele din serie de medie.

```
x <- c(1, 2, 3, 4, 5, 1, 2, 3, 1, 2, 4, 5, 2, 3, 1, 1, 2, 3,
    5, 6)  # our data set
variance.result = var(x)  # calculate variance
print(variance.result)
```

```
## [1] 2.484211
```

Deviația Standard

Deviatia standard este radicalul variatiei si e o masura prin care se cuantifica dispersia setului de date. A aparut din cauza dezavantajelor masurarii variatiei: - Se exprimă cu unităţile de măsurăale variabilei, ridicate la pătrat - Are în general valori foarte mari comparativ cu media.

```
x <- c(1, 2, 3, 4, 5, 1, 2, 3, 1, 2, 4, 5, 2, 3, 1, 1, 2, 3,
    5, 6)  # our data set
sd.result = sqrt(var(x))  # calculate standard deviation
print(sd.result)
```

```
## [1] 1.576138
```

Modulul

Clasa sau valoarea corespunzătoare frecvenţei maxime dintr-o serie statistică se numeşte mod.

```r
mode <- function(x) {
    ux <- unique(x)
    ux[which.max(tabulate(match(x, ux)))]
}

x <- c(1, 2, 3, 4, 5, 1, 2, 3, 1, 2, 4, 5, 2, 3, 1, 1, 2, 3,
    5, 6)  # our data set

mode.result = mode(x)  # calculate mode (with our custom function named 'mode')
print(mode.result)
```

```
## [1] 1
```

Cuantile

În statistică și probabilitate, cuantilele sunt puncte tăiate împărțind intervalul unei distribuții de probabilitate în intervale continue cu probabilități egale sau împărțind observațiile într-un eșantion în același mod. Exemplu: - Decile. Pe eşantioane mai mari de multe sute de indivizi. Sunt 9 decile, fiecare corespunzând unui procent de 10%, 20%, ..., 90% din eşantion. Decila a 5-a este mediana. - Centilele. Folosite, în studii pe mii de cazuri, de obicei de un interes mai larg, naţional, internaţional şi sunt corespunzătoare procentelor de 1%, 2%, ..., 99% din lot. Centila a 50-a este mediana.

```r
quantile(x <- rnorm(1001))  # Extremes & Quartiles by default
```

```
##          0%         25%         50%         75%        100%
## -3.20065667 -0.71696694 -0.04992333  0.63206102  3.77151233
```

```r
quantile(x, probs = c(0.1, 0.5, 1, 2, 5, 10, 50, NA)/100)
```

```
##        0.1%        0.5%          1%          2%          5%         10%
## -2.93663782 -2.52876208 -2.38956431 -2.15060203 -1.69309883 -1.25132522
##         50%
## -0.04992333          NA
```

Rată de creștere

Există doup tipuri de cresteri: - creștere discretă, atunci când schimbările apar la intervale specifice - creștere continua: atunci când schimbările sunt permanente

Rată de creștere discretă : $a * (1 + r)^t$

Rată de creștere continuă : $a * e^{(r*t)}$

Utilitate : Exemplu dobândă bancă

Dobândă discretă este simplă. Este pe intervale fixe de timp în general ani. Câți bani voi avea dacă depun 1000 de lei la o dobândă de 5% pe an după 2 ani.

```r
initial = 1000
rate = 0.05
time = 2
initial * ((1 + rate)^time)
```

```
## [1] 1102.5
```

Dobânda continuă a apărut datorită lui Euler care și întrebat prietenul bancher: Dar nu ai vrea tu să îmi dai jumăte din dobândă pe jumăte din perioadă de mai multe ori până perioada a ajuns la o secundă. Considerând dobânda inițială de r% pe an și n numărul decâte ori a fost înjumătățită dobând și timpul rezultă suma finală ar fi $a * (1 + \frac{r}{2^n})^{2^n * t} = a * (1 + \frac{1}{2^n/r})^{\frac{2^n}{r} * (r*t)}$. Înlocuim $m = \frac{2^n}{r}$ și rezultă că suma finală o să aibă forma $a * (1 + \frac{1}{m})^{m*(r*t)}$ trecând în limită de m la infinit rezultă că suma finală are forma $a * e^{(r*t)}$. Deci pentru o dobândă continuă de 5% pe termen de 2 ani cu suma initială de 1000 de lei vom avea o sumă finală mai mare decât la dobânda discretă, deci Euler și-a păcălit prietenul.

```
initial = 1000
rate = 0.05
time = 2
initial * exp(rate * time)
```

```
## [1] 1105.171
```

Exemplu: Cresterea populatiei

In functie de tipul de populatie, cresterea poate fi continua, precum in cazul unei colonii de bacterii, unde putem prezice cresterea populatiei dupa o anumita perioada de timp, dar poate fi si discreta, cum e in cazul tigrilor, care au o perioada de imperechere

Exemplu: Bursa

Valoarea indecsilor de la bursa se schimba in fiecare zi, pare o modificare continua, dar nu exista o rata predictibila. Vedem o multime de salturi si pentru a putea obtine informatii concludente, se fac studii anuale. De obicei, bursa se descrie printr-o crestere discreta anuala.

Probabilități

Pentru a explica mai bine termenii de statistici ne vom folosi de 2 albume de muzică VOL. 3 - L'ESPERANCE HITS 2008 [https://www.lautarul.shop/en/manele/1827-vol-lesperance-hits.html] vol 3 și FORTZA MANELE [https://www.lautarul.shop/en/manele/780-fortza-manele.html].

Nr. melodie	L'ESPERANCE HITS 2008	FORTZA MANELE
1	Jean de la Craiova - Tu esti femeia care-mi place	Alberto voce de diamant - Nu mai pot
2	George de la Stefanesti Dorel de la Popesti - Iti dau partea mea de fericire	Florin Salam - Eu sunt bomba nucleara
3	Mihaita Piticul - Ce inima de gheata	Copilul de Aur - Daca ramai in viata mea
4	Nicolae Guta - Ce le-as face la dusmani	Danezu - Arunc miliarde
5	Adrian Minune - Si cand vad patul gol	Denisa - Inima mea
6	Stefan de la Barbulesti - Ma rog la tine	Sorina - Nu mai pot sa ma indragostesc
7	Neluta Neagu - De-ar putea dusmani	Liviu Pustiu - Am scoala de mafiot
8	Sorinel Pustiul - Danseaza cu mine macar cinci minute	Octavian Francezul - Cunosc omul dupa fata
9	Mihaita Piticul - Acum imi ceri sa fim amici	Vali Vijelie si R. Printisorul - Taicutul meu
10	Stefan de la Barbulesti - Inima mea plange	Adrian Copilul Minune si Mihaita Piticu - Nevasta mea
11	Neluta Neagu - Sunt un barbat luxos	Sorinel Pustiu - Ce dimineata trista
12	Adrian Minune - Cat te-ai schimbat	Nicolae Guta si Play A.J. - Cum te misti, asa vorbesti

Introducem datele în R pentru a le putea procesa mai târziu.

```
album_1 <- data.frame(autor = c("Jean de la Craiova", "George de la Stefanesti",
    "Mihaita Piticul", "Nicolae Guta", "Adrian Minune", "Stefan de la Barbulesti",
    "Neluta Neagu", "Sorinel Pustiul", "Mihaita Piticul", "Stefan de la Barbulesti",
    "Neluta Neagu", "Adrian Minune"), melodie = c("Tu esti femeia care-mi place",
    "Iti dau partea mea de fericire", "Ce inima de gheata", "Ce le-as face la dusmani",
    "Si cand vad patul gol", "Ma rog la tine", "De-ar putea dusmani",
    "Danseaza cu mine macar cinci minute", "Acum imi ceri sa fim amici",
    "Inima mea plange", "Sunt un barbat luxos", "Cat te-ai schimbat"))
album_2 <- data.frame(autor = c("Alberto voce de diamant", "Florin Salam",
    "Copilul de Aur", "Danezu", "Denisa", "Sorina", "Liviu Pustiu",
    "Octavian Francezul", "Vali Vijelie", "Adrian Minune", "Sorinel Pustiu",
    "Nicolae Guta"), melodie = c("Nu mai pot", "Eu sunt bomba nucleara",
    "Daca ramai in viata mea", "Arunc miliarde", "Inima mea",
    "Nu mai pot sa ma indragostesc", "Am scoala de mafiot", "Cunosc omul dupa fata",
    "Taicutul meu", "Nevasta mea", "Ce dimineata trista", "cum te misti, asa vorbesti"))
```

Definiția unei probabilități

Def. Gradul de încredere în valaorea de adevăr a unui eveniment.

Fie x un eveniment

Notație: $P(x)$ = prbabilitatea ca x să fie adevărat . $P(x) \in [0, 1]$.

Folosind albumul L'ESPERANCE HITS 2008 probabilitatea ca să ascultând o melodie de pe album să asculți o manea este 1.

Fie $!x$ evenimentul în care x nu este adevărat atunci: $P(x) + P(!x) = 1$.

Folosind albumul L'ESPERANCE HITS 2008 probabilitatea ca să asculți o melodie de Adrian Minune sau sa ascultți o melodie de alt cântereș este 1.

Def. $P(x) = \frac{\text{numărul de experimente în care x a avut loc}}{\text{număr total de experimente}}$

Folosind albumul L'ESPERANCE HITS 2008 probabilitatea ca să asculți o melodie de Adrian Minune este numărul de melodii cântate de adrian minune pe numărul total de melodii de pe album, adică $\frac{2}{12} = \frac{1}{6} = 0.1666$.

```
sum(album_1$autor == "Adrian Minune")/length(album_1$autor)
```

```
## [1] 0.1666667
```

```
sum(album_1$autor == "Adrian Minune")/length(album_1$autor) +
    sum(album_1$autor != "Adrian Minune")/length(album_1$autor)
```

```
## [1] 1
```

Distribuție de probabilități/Principiul Indiferenței

Def. Distribuția de probabilități este o colecție de evenimente care sunt exhaustive (cel puțin un eveniment este adevărat) și exclusive (cel mult un eveniment este adevărat).

Folosind albumul L'ESPERANCE HITS 2008 noi la un moment dat sigur ascutăm o singură melodie de pe album (cel puțin un eveniment este adevărat și cel mult un eveniment este adevărat).

Principiul Indiferenței:

Fie $X = \{x_1, x_2, ..., x_n\}$ cu $n =\mid X \mid$ atunci dacă nu știm nimic specific despre evenimentele x_i putem considera că $\frac{1}{n}$

Folosind albumele noastre, probabilitatea ca noi să ascultăm o melodie de pe unul dintre albume este $\frac{1}{2} = 0.5$ sau cum se mai spune jumătate tu jumătate eu.

Operații

Fie X, Y două evenimente sau grupuri de evenimente

Și

$P(X$ și $Y) =$ probabilitatea ca evenimentele X și Y să fie adevărate în același timp

$P(X$ și $Y) = P(X) * P(Y) \Leftrightarrow$ X și Y sunt indepedente adică $X \cap Y = \emptyset$

Folosind albumul L'ESPERANCE HITS 2008 hai să calculăm probabilitatea ca noi ascultând două melodii aleatoriu de pe album una după alta să fie cantate de Adrian Minune.

```
(sum(album_1$autor == "Adrian Minune")/length(album_1$autor)) *
    (sum(album_1$autor == "Adrian Minune")/length(album_1$autor))
```

```
## [1] 0.02777778
```

Notație: $P(X$ și $Y) = P(X, Y) = P(Y$ și $X) = P(Y, X) = P(X \cap Y) = P(Y \cap X)$

Sau

$P(X \quad \text{sau} \quad Y)$ = probabilitatea ca evenimentul X sau evenimentul Y să fie adevărat.

$P(X \quad \text{sau} \quad Y) = P(X, !Y) + P(!X, Y) + P(X, Y) = P(X) * P(!Y) + P(!X) * P(Y) + P(X, Y) = P(X) * (1 - P(Y)) + (1 - P(X)) * P(Y) + P(X) * P(Y) = P(X) - P(X) * P(Y) + P(Y) - P(X) * P(Y) + P(X) * P(Y)$.

$P(X \quad \text{sau} \quad Y) = P(X) + P(Y) - P(X \cap Y)$.

Folosind albumul L'ESPERANCE HITS 2008 hai să calculăm probabilitatea ca noi ascultând o melodie aleatoriu de pe album fie cantată de Adrian Minune sau de Nicolae Guță. De reținut că nu exitstă nici o melodie pe album să fie cântat și de Adrian Minune și de Nicolae Guță.

```
(sum(album_1$autor == "Adrian Minune")/length(album_1$autor)) +
    (sum(album_1$autor == "Nicolae Guta")/length(album_1$autor))
```

```
## [1] 0.25
```

Dependența

$P(X \mid Y)$ = probabilitatea ca X să fie adevărat dacă știm că Y este adevărat

$P(X \mid Y) = \frac{\text{experimentele în care X și Y sunt adevărate}}{\text{experimentele în care Y este adevărat}} = \frac{P(X,Y)}{P(Y)}$

Folosind cele două albume, putem calcula probabilități depedente precum: probabilitatea ca ascultând o melodie de pe albumul L'ESPERANCE HITS 2008 să fie cântat de Adrian Minune sau probabilitatea ca ascultând o melodie de pe albulmul FORTZA MANELE să fie cântată de Adrian Minune.

```
(sum(album_1$autor == "Adrian Minune")/length(album_1$autor))
```

```
## [1] 0.1666667
```

```
(sum(album_2$autor == "Adrian Minune")/length(album_2$autor))
```

```
## [1] 0.08333333
```

Regula Produsului $P(X, Y) = P(X \mid Y) * P(Y)$

Folosind cele două albume, putem calcula probabilitatea să ascultăm o melodie cântată de Adrian Minune și să fie o melodie de pe albumul L'ESPERANCE HITS 2008 în același timp.

```
(sum(album_1$autor == "Adrian Minune")/length(album_1$autor)) *
    0.5
```

```
## [1] 0.08333333
```

Regula Sumei Fie Y o distribuție de probabilități și X un eveniment atunci $P(X) = \sum_{Y_i \in Y} P(X \mid Y_i) * P(Y_i)$

Folosind cele două albume, putem calcula probabilitatea să ascultăm o melodie cântată de Adrian Minune.

```
prob_adrian <- (sum(album_1$autor == "Adrian Minune")/length(album_1$autor)) *
    0.5 + (sum(album_2$autor == "Adrian Minune")/length(album_2$autor)) *
    0.5
print(prob_adrian)
```

```
## [1] 0.125
```

Teorema 1 $P(X \mid Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(Y,X)}{P(Y)} = \frac{P(Y|X)*P(X)}{P(Y)}$

Folosind cele două albume, putem calcula probabilitatea să ascutăm melodia după albumul L'ESPERANCE HITS 2008 știind că ascutăm o melodie de Adrian Minune. X = L'ESPERANCE HITS 2008 și Y = Adrian Minune.

```
((sum(album_1$autor == "Adrian Minune")/length(album_1$autor)) *
    0.5)/prob_adrian
```

```
## [1] 0.6666667
```

Corolar Teorema 1 + Regula Sumei $P(Y_i \mid X) = \frac{P(X|Y_i)*P(Y_i)}{P(X)} = \frac{P(X|Y_i)*P(Y_i)}{\sum_{Y_j \in Y} P(X|Y_j)*P(Y_j)}$

Folosind cele două albume, noi ascultăm aleatoriu melodii de pe unul dintre cele două albume. Știind că am ascultat 2 melodii consecutive cântate de Adrian minune care este probabilitatea să ascultăm melodii de pe albumul L'ESPERANCE HITS 2008.

```
prob_adrian_esperance <- sum(album_1$autor == "Adrian Minune")/length(album_1$autor)
prob_2_adrian_esperance <- prob_adrian_esperance * prob_adrian_esperance
prob_adrian_fortza <- sum(album_2$autor == "Adrian Minune")/length(album_2$autor)
prob_2_adrian_fortza <- prob_adrian_fortza * prob_adrian_fortza
prob_esperance_2_adrian <- (prob_2_adrian_esperance * 0.5)/(prob_2_adrian_esperance *
    0.5 + prob_2_adrian_fortza * 0.5)
print(prob_esperance_2_adrian)
```

```
## [1] 0.8
```

Teorema 2 Fie $X^{'} = X+x$ astfel încât $P(X^{'} \mid !X) = 0$ (Dacă X nu are loc atunci $X^{'}$ nu poate avea loc) și $P(Y_i \mid X^{'}) = \frac{P(X^{'}|Y_i)*P(Y_i)}{P(X^{'})}$ atunci $P(Y_i \mid X^{'}) = \frac{P(x|Y_i)*P(Y_i|X)}{\sum_{Y_j \in Y} P(x|Y_j)*P(Y_j|X)}$

Pare greu dar folosind cele două albume sperăm să putem explica. Considerăm X ca fiind ascultarea a doua 2 melodii consecutive cântate de Adrian Minune. Considerăm că $X^{'}$ este acultarea a trei melodii consecutive cântate de Adrian Minune. Se observă că nu putem asculta trei melodii consecutive cântate de Adrian Minune dacă nu am ascultat 2 deja consecutiv, deci se verifică $P(X^{'} \mid !X) = 0$. Să spunem că avem de calculat probabilitatea că ascultăm melodii aleatorii cu repetiție de pe albumul L'ESPERANCE HITS 2008 știind că au fost 3 melodii consecutive cântate de Adrian Minune. Noi știm deja rezultate pentru 2 melodii cosnecutive, să îl numim un experiment precedent. Probabilitatea ascultăm melodii de pe albumul L'ESPERANCE HITS 2008 este probabilitatea de a asculta o melodie cântată de Adrian Minune de pe albumul L'ESPERANCE HITS 2008 înmulțit cu probabilitatea de că albumul de pe care am ascultat primele două melodii consecutive cu adrian minune să fie L'ESPERANCE HITS 2008. Totul supra sumei dintre probabilitatea prezentată precedent și cea ca acest evenimentul de mai sus să se întâmple pe celălalt album.

```
prob_fortza_2_adrian <- 1 - prob_esperance_2_adrian
prob_esperance_3_adrian <- (prob_adrian_esperance * prob_esperance_2_adrian)/(prob_adrian_esperance *
    prob_esperance_2_adrian + prob_adrian_fortza * prob_fortza_2_adrian)
print(prob_esperance_3_adrian)
```

```
## [1] 0.8888889
```

Teorema 3 Probabilitatea ca s evenimente să se înâmple din n încercări.

Fie X un eveniment și $P(X)$ probabilitatea ca X să se înâmple. Făcând un experiment de n ori putem calcula probabilitate $\binom{n}{s} P(X)^s * (1 - P(X))^{n-s}$

Folsind albumul L'ESPERANCE HITS 2008, ascultând 50 de melodii aleatoriu după album care este probabilitatea ca fix 14 să fie cântate de Adrian Minune.

```
binom_probability <- function(p, s, n) {
    choose(n, s) * p^s * (1 - prob_adrian_esperance)^(n - s)
}
binom_probability(prob_adrian_esperance, 14, 50)
```

```
## [1] 0.01688453
```

Utilitate: Test Covid

Unele din cel mai bune teste de covid are următoarele valori: p(AP)=98%, p(FN)=2%, p(FP)=4%, p(TN)=96%, Populație_România=16 (mil) Bolnavi_România=276802 (05.11.2020). Vom calcula probabiliatea de a fi bolnav dacă ai făcut 3 teste de covid dintre care unul singur a fost pozitiv (adevărat ar fi în acest caz, confuzie de nume). Putem crea matricea de confuzie și vom calcula valorii de mai sus cum am fi avut doar matricea de confuzie.

```
data <- table(1:2, 1:2)
dimnames(data) <- list(c("T", "F"), c("P", "N"))
data[1, 1] <- 98
data[2, 1] <- 2
data[1, 2] <- 4
data[2, 2] <- 96
print(data)
```

```
##    P  N
## T 98  4
## F  2 96
```

```
t_pozitive_rate <- data[1, 1]/sum(data[, 1])
f_pozitive_rate <- data[1, 2]/sum(data[, 2])
f_negative_rate <- data[2, 1]/sum(data[, 1])
t_negative_rate <- data[2, 2]/sum(data[, 2])
```

Calculăm probabilități să avem un test "true" și două "false" dacă suntem pozitive (bolnavi), respectiv negativi (sănătoși).

```
probability_tff_p <- t_pozitive_rate * f_negative_rate * f_negative_rate
probability_tff_n <- f_pozitive_rate * t_negative_rate * t_negative_rate
```

Suntem două variante de a calcula probabilitatea la a fi bolanv având rezultatele testelor. Prima este când luăm un om aleatoriu din populația României și face testul, al doilea când persoana este suspectă fiind în contact intră în categoria oameniilor căror li se face teste. Prima dată calculăm probabilitatea să fii bolnav dacă ești cetățean român. A doua este luat numărul de bolnavi (chiar dacă sunt cei ce au ieșit "true" (pozitvi la test)) relativ la nunmărul de teste făcute la ziua respectivă. Calculăm probabilitatea să fi bolnav dacă ai rezultatele respective la teste ca probabilitatea de avea rezultatele respective daca ești bolnav înmulțit cu probabilitatea să fi bolnav supra probabilitatea de avea rezultatele respective daca ești bolnav înmulțit cu probabilitatea să fi bolnav plus probabilitatea de avea rezultatele respective daca ești sanatos înmulțit cu probabilitatea să fi sanatos.

```
romania_population <- 1.6e+07
romania_positive <- 276802
romania_proability_p <- romania_positive/romania_population
probability_p_tff_ro <- (probability_tff_p * romania_proability_p)/(probability_tff_p *
    romania_proability_p + probability_tff_n * (1 - romania_proability_p))
test_number <- 30000
```

```
test_p <- 9000
test_probability_p <- 9000/30000
probability_p_tff_test <- (probability_tff_p * test_probability_p)/(probability_tff_p *
    test_probability_p + probability_tff_n * (1 - test_probability_p))
print(probability_p_tff_ro)
```

```
## [1] 0.0001871676
```

```
print(probability_p_tff_test)
```

```
## [1] 0.004536617
```

Utilitate: Multiple Intrusion detection systems

O utilitate în Rețelistică este acea de folosire de mai multe IDS-uri la intrarea în rețea de la diferești vendori. Întrebare cu cât este mai bună prevenția dacă folosim mai multe IDS-uri decât unul singur. Trebuie să calculăm probabilitatea ca un virus să treacă de două IDS-uri. Fie două IDS-uri cu următoarele propirietăți: IDS1 : p(AP)=97%, p(FN)=3%, p(FP)=4%, p(TN)=96% IDS1 : p(AP)=96%, p(FN)=4%, p(FP)=2%, p(TN)=98% Considerăm grosolan că 1% din pachetele trimise către echipamente sunt malițioase.

```
t_pozitive_rate_1 <- 0.97
f_negative_rate_1 <- 0.03
f_pozitive_rate_1 <- 0.04
t_negative_rate_1 <- 0.96
t_pozitive_rate_2 <- 0.96
f_negative_rate_2 <- 0.04
f_pozitive_rate_2 <- 0.02
t_negative_rate_2 <- 0.98
malicious_porbability <- 0.01
```

Trebuie să calculăm probabilitatea ca un pachet mailițios să treacă de ambele ids-uri, adică să primească "false" de la ambele (false negative) și probabilitatea ca un pachet mailițios să treacă de primul IDS.

```
test_ff_malitios <- f_negative_rate_1 * f_negative_rate_2
test_ff_ok <- t_negative_rate_1 * t_negative_rate_2
malitios_ff <- (test_ff_malitios * malicious_porbability)/(test_ff_malitios *
    malicious_porbability + test_ff_ok * (1 - malicious_porbability))
print(malitios_ff)
```

```
## [1] 1.288378e-05
```

```
malitios_f <- (f_negative_rate_1 * malicious_porbability)/(f_negative_rate_1 *
    malicious_porbability + t_negative_rate_1 * (1 - malicious_porbability))
print(malitios_f)
```

```
## [1] 0.000315557
```

Deci este de 100 de ori mai puțin probabil să treacă de ambele decât de unul singur.

Eurostat

Eurostat este o bază de date europene cu statistici despre țăriile din Europa (nu doar UE). Un beneficiu al limajului R este că această bază de date paote fi accesată direct din limbajul R cu ajutorul mai multor bibloteci. Puteți găsi datele pe care puteți accesa pe [https://ec.europa.eu/eurostat/data/database].

```
# install.packages(c('rgdal', 'RColorBrewer', 'sp',
# 'GISTools', 'classInt', 'maptools'))
# install.packages('SmarterPoland')
library("rgdal")
```

```
## Loading required package: sp
```

```
## rgdal: version: 1.5-18, (SVN revision 1082)
## Geospatial Data Abstraction Library extensions to R successfully loaded
## Loaded GDAL runtime: GDAL 3.1.1, released 2020/06/22
## Path to GDAL shared files: /Library/Frameworks/R.framework/Versions/4.0/Resources/library/rgdal/gdal
## GDAL binary built with GEOS: TRUE
## Loaded PROJ runtime: Rel. 6.3.1, February 10th, 2020, [PJ_VERSION: 631]
## Path to PROJ shared files: /Library/Frameworks/R.framework/Versions/4.0/Resources/library/rgdal/proj
## Linking to sp version:1.4-4
## To mute warnings of possible GDAL/OSR exportToProj4() degradation,
## use options("rgdal_show_exportToProj4_warnings"="none") before loading rgdal.
```

```
library("RColorBrewer")
library("sp")
library("GISTools")
```

```
## Loading required package: maptools
```

```
## Checking rgeos availability: TRUE
```

```
## Loading required package: MASS
```

```
## Loading required package: rgeos
```

```
## rgeos version: 0.5-5, (SVN revision 640)
##  GEOS runtime version: 3.8.1-CAPI-1.13.3
##  Linking to sp version: 1.4-2
##  Polygon checking: TRUE
```

```
library("classInt")
library("maptools")
library("SmarterPoland")
```

```
## Loading required package: ggplot2
```

```
## Loading required package: htmltools
```

```
# temp <- tempfile(fileext = '.zip')
# download.file('http://epp.eurostat.ec.europa.eu/cache/GISCO/geodatafiles/NUTS_2010_60M_SH.zip',
# temp) unzip(temp)
EU_NUTS <- readOGR(dsn = "./NUTS_2010_60M_SH/data", layer = "NUTS_RG_60M_2010")
```

```
## Warning in OGRSpatialRef(dsn, layer, morphFromESRI = morphFromESRI, dumpSRS =
## dumpSRS, : Discarded datum European_Terrestrial_Reference_System_1989 in CRS
## definition: +proj=longlat +ellps=GRS80 +towgs84=0,0,0,0,0,0,0 +no_defs
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "/Users/ciocirlan/Downloads/Data Science Curs/LaburiSD/NUTS_2010_60M_SH/Data", layer: "NUTS_
## with 1920 features
## It has 4 fields
```

```
EU_NUTS <- spTransform(EU_NUTS, CRS("+proj=merc +a=6378137 +b=6378137 +lat_ts=0.0 +lon_0=0.0 +x_0=0.0 +
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO", prefer_proj
```

```
## = prefer_proj): Discarded ellps WGS 84 in CRS definition: +proj=merc +a=6378137
## +b=6378137 +lat_ts=0 +lon_0=0 +x_0=0 +y_0=0 +k=1 +units=m +nadgrids=@null
## +wktext +no_defs +type=crs
```

```
## Warning in showSRID(uprojargs, format = "PROJ", multiline = "NO", prefer_proj =
## prefer_proj): Discarded datum World Geodetic System 1984 in CRS definition
```

```
plot(EU_NUTS)
```

În primă fază am descărcat o hartă cu țări din euroap și regiuniile lor administrative. Acum urmează să descărcăm date textuale. Vom folosi un set de date care arată rata somajului per regiune in functie de sex,vârstă, și an. Pentru fiecare regiune vom face o cloana cu una din comvinațiile de sex,vârstă și an. În tabela EurostatTOC putem accesa taote bazele de date prin codul de pe coloana numită "code".

```
EurostatTOC <- getEurostatTOC()
head(EurostatTOC)
```

```
##                                                                 title      code
## 1                                                Database by themes      data
## 2                                     General and regional statistics   general
## 3             European and national indicators for short-term analysis   euroind
## 4                    Business and consumer surveys (source: DG ECFIN)     ei_bcs
## 5                                 Consumer surveys (source: DG ECFIN) ei_bcs_cs
## 6                                       Consumers - monthly data  ei_bsco_m
##      type last.update.of.data last.table.structure.change data.start data.end
## 1  folder
## 2  folder
## 3  folder
## 4  folder
## 5  folder
## 6 dataset          29.10.2020                 29.10.2020    1980M01  2020M10
##   values
## 1     NA
## 2     NA
## 3     NA
## 4     NA
## 5     NA
## 6     NA
```

```
data <- getEurostatRCV(kod = "lfst_r_lfu3rt")
head(data)
```

```
##                       unit    age sex  geo time value
## PC_Y15-24_F_AT_2019     PC Y15-24   F   AT 2019   7.8
## PC_Y15-24_F_AT1_2019    PC Y15-24   F  AT1 2019  11.8
## PC_Y15-24_F_AT11_2019   PC Y15-24   F AT11 2019    NA
## PC_Y15-24_F_AT12_2019   PC Y15-24   F AT12 2019   9.2
## PC_Y15-24_F_AT13_2019   PC Y15-24   F AT13 2019  14.5
## PC_Y15-24_F_AT2_2019    PC Y15-24   F  AT2 2019    NA
```

```
library("data.table")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:rgeos':
##
##     intersect, setdiff, union
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
mapdata <- dcast(setDT(data), geo ~ time + age + sex)
EU_NUTS@data <- inner_join(EU_NUTS@data, mapdata, by = c(NUTS_ID = "geo"))
my_colours <- brewer.pal(5, "RdPu")
breaks <- classIntervals(EU_NUTS@data[["2012_Y20-64_T"]], n = 5,
    style = "fisher", unique = TRUE)$brks
```

```
## Warning in classIntervals(EU_NUTS@data[["2012_Y20-64_T"]], n = 5, style =
## "fisher", : var has missing values, omitted in finding classes
```

```
plot <- plot(EU_NUTS, col = my_colours[findInterval(EU_NUTS@data[["2012_Y20-64_T"]],
    breaks, all.inside = TRUE)], axes = FALSE, border = NA)
```

```
plot
```

```
## NULL
```

Exerciții

1. Pornind de la baza de date din capitolul Eurostat, calculati urmatoarele:

- media varstei barbatilor someri din zona AT
- mediana varstei femeilor somere din zona AT1

2.

Biroul australian de statistică a estimat că în 2012, în Australia erau 22.683.600 de locuitori, cu o creștere de 1.6% față de anul trecut. Considerând că această rată de creștere e constantă de-a lungul anilor, calculați care era populația in 2007.

3. Folosind ca exemplu albumele de la capitolul 'Probabilitați', calculați:

- Care este probabillitatea ca, ascultând 50 de melodii aleatorii de pe cele 2 albume, fix 5 sa fie cântate de Adrian Minune sau de Vali Vijelie?
- Care este probabillitatea ca o melodie cântata de Adrian Minune sau de Mihaiță Piticul sa fie o melodie de pe albumul L'ESPERANCE HITS 2008 în același timp.

4. Considerând informațiile din capitolul 'Multiple Intrusion detection systems'.

- Care este probabilitatea ca primul IDS să dea false positive și al doilea false negative în cazul unui pachet de date?
- Care este probabilitatea ca un pachet legitim să fie considerat virus?