Trabajo Práctico N° 2

Aprendizaje Supervisado

Enunciados:

Parte A – Preprocesamiento de los datos.

```
1) Ingrese a la página web de la Universidad de California.
    * https://archive.ics.uci.edu/ml/datasets/seeds
    * En Download -> Data Folder baje el archivo seeds_dataset.txt.
    * Copie aquí el Abstract (debajo de Download y arriba de la tabla).

2) Busque en la página web e indique aquí:
    * ¿cuáles son las 3 variedades de trigo que se estudiarán?
    * Busque una imagen de granos de trigo.
    * Indique la página web origen de dicha imagen.

3) Abra el archivo seeds_dataset.txt en R como "base".
    * Muestre un head(base).
    * Las variables de la base representan los siguientes atributos de los
      granos de trigo:
        * V1=área de la semilla
        * V2=perímetro de la semilla
        * V3=compactitud
        * V4=largo de la semilla
        * V5=ancho de la semilla
        * V6=coeficiente de asimetría
        * V7=largo de la división frontal de la semilla
        * TPs Data Mining y Machine Learning 3
        * V8=variedad de la semilla (1-kama 2-rosa 3-canadian)
    * Renombre cada variable:
        * names(base)[names(base)=="V1"]="Area"
        * names(base)[names(base)=="V2"]="Perimetro"
        * names(base)[names(base)=="V3"]="Compactitud"
        * names(base)[names(base)=="V4"]="Largo"
        * names(base)[names(base)=="V5"]="Ancho"
        * names(base)[names(base)=="V6"]="Asimetria"
        * names(base)[names(base)=="V7"]="Division"
        * names(base)[names(base)=="V8"]="Variedad"
    * Muestre un head(base) con el cambio de las variables.

4) Transforme a categórica la variable Variedad.
    * Renombre las variedades 1, 2 y 3 como "kama", "rosa" y "canadian".
    * Muestre un head de la base con las variables transformadas.
```

Parte B – Análisis Exploratorio de Datos.

```
1) ¿Cuántas semillas hay por variedad?

2) Realice un gráfico de barras de la variable a predecir Variedad.
    * Indique el código R utilizado.
    * elija un Título.

3) Realice un gráfico de dispersión entre 2 variables:
    * Que no sean Variedad.
    * Coloréelo por la variable Variedad.
    * Agregue una leyenda que indique cuál es cada grupo.
    * Elija las variables en y y x, un título y un pch.
    * Indique el código R utilizado.

4) Con la instrucción base[numFila,] se puede obtener los datos de uno de los
   granos de trigo.
    * Considere los 2 últimos dígitos de su DNI (2numDNI).
    * Muestre aquí el registro correspondiente.
    * ¿De qué variedad es?
```

Parte C - Conjuntos.

```
1) Considere los 3 últimos dígitos de su DNI (3numDNI) para el seteo de
   semilla.
    * particione la base en un conjunto de entrenamiento y uno de testeo.
    * Además:
        * Si su DNI termina en 0, 1, 2 ó 3:
            * Setee p=0.70
        * Si su DNI termina en 4, 5, 6 ó 7:
            * Setee p=0.75
        * Si su DNI termina en 8 ó 9:
            * Setee p=0.80
    * Indique cómo quedó el código R utilizado.

2) Muestre un head y un summary del conjunto de entrenamiento y del conjunto
   de testeo.

3) Realice:
    * summary(base$Variedad)
    * summary(entreno$Variedad)
    * summary(testeo$Variedad)
    * ¿Cuántos registros quedaron por variedad de trigo en el conjunto de
    entrenamiento y en el de testeo?
```

Parte D – Árbol de Decisión.

```
1) Cree un Árbol de Decisión (con libreria rpart) para modelar el problema
   planteado.
    * Escriba arbol<enter>.
    * muestre una captura de pantalla de la información que aparece.

2) Grafique el Árbol de Decisión resultante.

3) ¿Cuántas "hojas" tiene el Árbol de Decisión?

4) Según el Árbol de Decisión creado:
    * ¿cuándo una semilla es de la variedad "rosa"?
    * Indique las reglas siguiendo las ramas desde el nodo raíz hasta las
      hojas "rosa".

5) Testee el Árbol de Decisión.
    * Compare head(pred,10) con head(test$Variedad,10).
    * Vea si la predicción de los 10 primeros elementos coincide con lo
      esperado.

6) Calcule la matriz de confusión utilizando la instrucción confusionMatrix
   de la libreria caret.
    * Muestre una captura de pantalla de la matriz y los resultados obtenidos.

7) Calcule el accuracy según la cantidad de registros bien clasificados.
    * Indique con números la fórmula que usó.
    * Verifique que coincida con el accuracy obtenido por confusionMatrix.

8) ¿Cuál categoría presenta menor sensibilidad?

9) Considere el grano de trigo correspondiente a los últimos 2 dígitos de su
   DNI.
    * Según el Árbol de Decisión, ¿qué variedad es?
    * ¿Coincide la predicción con lo esperado?

10) Indique la predicción de "trigo" con probabilidades, sacando type="class".

11) Indique la regla del Árbol de Decisión para "trigo".
```

Parte E – Red Neuronal.

```
1) Considere los 3 últimos dígitos de su DNI (3numDNI) para el seteo de
   semilla.
    * Cree una Red Neuronal para modelar el problema planteado.
    * con maxit=10000 y size=10.
    * Indique el código R utilizado.

2) Muestre una captura de pantalla de la lista de iteraciones de la Red
   Neuronal.

3) Escriba red<enter> y muestre una captura de pantalla de la información que
   aparece.

4) Indique la cantidad de pesos y la cantidad de iteraciones resultantes.

5) Dibuje la Red Neuronal.
    * Optativo: Cambiar los colores del gráfico de la Red Neuronal.

6) Testee la Red Neuronal.
    * Compare head(pred2,10) con head(testeo$Variedad,10).
    * Vea si la predicción de los 10 primeros elementos coincide con lo
      esperado.

7) Calcule la matriz de confusión utilizando la instrucción confusionMatrix de
   la libreria caret.
    * Muestre una captura de pantalla de la matriz y los resultados obtenidos.

8) ¿Cuál fue el accuracy?

9) Considere el grano de trigo correspondiente a los últimos 2 dígitos de su
   DNI.
    * Según la Red Neuronal, ¿qué variedad es?
    * ¿Coincide la predicción con la variedad esperada?
```

Parte F – Comparación de modelos.

```
1) Cree una tabla con el accuracy de cada modelo, y la sensibilidad y
especificidad de cada modelo por categoría.

2) Compare los resultados obtenidos con el Árbol de Decisión y la Red
   Neuronal.
    * ¿Cuál modelo le parece que resultó mejor?
    * (¿según qué criterio?)
```

Diplomatura en BIG DATA

DATA MINING & MACHINE LEARNING

Escalada Christian, 33.549.575

01-06-2022

Parte A – Preprocesamiento de los datos.

1.1. Ingrese a la página web de la Universidad de California.

Universidad de California

1.2. Descargue el archivo `seeds_dataset.txt`.

seeds_dataset.txt

1.3. Copie aquí el Abstract.

> *Abstract:*
>
> *Measurements of geometrical properties of kernels belonging to three different varieties of wheat. A soft X-ray technique and GRAINS package were used to construct all seven, real-valued attributes.*

2.1. Busque en la página web e indique aquí: ¿cuáles son las 3 variedades de trigo que se estudiarán?

> *Data Set Information:*
>
> *The examined group comprised kernels belonging to three different varieties of wheat: **Kama**, **Rosa** and **Canadian**, 70 elements each, randomly selected for the experiment.*

- Las 3 variedades a estudiar son: **Kama**, **Rosa**, **Canadian**.

2.2. Optativo: busque una imagen de granos de trigo. Indique la página web origen de dicha imagen.

Fuente: depositphotos.com [1]

3.1. Abra el archivo seeds_dataset.txt en R como "base" de la siguiente manera:

```
base=read.table("seeds_dataset.txt",header=F)
```

3.2. Muestre un `head(base)`.

```
head(base)
```

```
##      V1    V2     V3    V4    V5    V6    V7 V8
## 1 15.26 14.84 0.8710 5.763 3.312 2.221 5.220  1
## 2 14.88 14.57 0.8811 5.554 3.333 1.018 4.956  1
## 3 14.29 14.09 0.9050 5.291 3.337 2.699 4.825  1
## 4 13.84 13.94 0.8955 5.324 3.379 2.259 4.805  1
## 5 16.14 14.99 0.9034 5.658 3.562 1.355 5.175  1
## 6 14.38 14.21 0.8951 5.386 3.312 2.462 4.956  1
```

3.3. Renombre cada variable:

```
names(base)[names(base)=="V1"]="Area"
names(base)[names(base)=="V2"]="Perimetro"
names(base)[names(base)=="V3"]="Compactitud"
names(base)[names(base)=="V4"]="Largo"
names(base)[names(base)=="V5"]="Ancho"
names(base)[names(base)=="V6"]="Asimetria"
names(base)[names(base)=="V7"]="Division"
names(base)[names(base)=="V8"]="Variedad"
```

3.4. Muestre un `head(base)` con el cambio de las variables.

```
head(base)
```

```
##    Area Perimetro Compactitud Largo Ancho Asimetria Division Variedad
## 1 15.26     14.84      0.8710 5.763 3.312     2.221    5.220        1
## 2 14.88     14.57      0.8811 5.554 3.333     1.018    4.956        1
## 3 14.29     14.09      0.9050 5.291 3.337     2.699    4.825        1
## 4 13.84     13.94      0.8955 5.324 3.379     2.259    4.805        1
## 5 16.14     14.99      0.9034 5.658 3.562     1.355    5.175        1
## 6 14.38     14.21      0.8951 5.386 3.312     2.462    4.956        1
```

4.1. Transforme a categórica la variable Variedad y renombre las variedades 1, 2 y 3 como "kama", "rosa" y "canadian":

```
base$Variedad=factor(base$Variedad,levels=c(1,2,3),
                     labels=c("kama","rosa","canadian"))
```

4.2. Muestre un head de la base con las variables transformadas.

```
head(base)
```

```
##    Area Perimetro Compactitud Largo Ancho Asimetria Division Variedad
## 1 15.26     14.84      0.8710 5.763 3.312     2.221    5.220     kama
## 2 14.88     14.57      0.8811 5.554 3.333     1.018    4.956     kama
## 3 14.29     14.09      0.9050 5.291 3.337     2.699    4.825     kama
## 4 13.84     13.94      0.8955 5.324 3.379     2.259    4.805     kama
## 5 16.14     14.99      0.9034 5.658 3.562     1.355    5.175     kama
## 6 14.38     14.21      0.8951 5.386 3.312     2.462    4.956     kama
```


1. depositphotos.com↩

Diplomatura en BIG DATA

DATA MINING & MACHINE LEARNING

Escalada Christian, 33.549.575

01-06-2022

Parte B – Análisis Exploratorio de Datos.

1. ¿Cuántas semillas hay por variedad?

```
summary(base$Variedad)
```

```
##    kama    rosa canadian 
##      70      70      70
```

- Existen **70** Semillas por Variedad.

2. Realice un gráfico de barras de la variable a predecir Variedad.

```
pie(table(base$Variedad),main="Variedad  de Semillas por Tipo")
```

3. Realice un gráfico de dispersión:

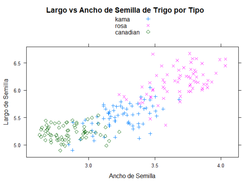
- Entre 2 variables que no sean Variedad.
- coloréelo por la variable Variedad.
- agregue una leyenda que indique cuál es cada grupo.
- Elija las variables en y & x, un título y un pch.

```
# Importamos Librerías:
library(caret)
```

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

```
## Loading required package: lattice
```

```
xyplot(base$Largo~base$Ancho,groups=base$Variedad,base,auto.key=TRUE,
       par.settings=simpleTheme(pch=c(3,4,5)),pch=c(3,4,5),
       main="Largo vs Ancho de Semilla de Trigo por Tipo",
       xlab="Ancho de Semilla", ylab="Largo de Semilla")
```



4.1. Con la instrucción `base[numFila,]` se puede obtener los datos de uno de los granos de trigo.

- Considere los 2 últimos dígitos de su DNI: `(75)` .
- Muestre aquí el registro correspondiente.

```
trigo=base[75,]
trigo
```

```
##     Area Perimetro Compactitud Largo Ancho Asimetria Division Variedad
## 75 16.82    15.51      0.8786 6.017 3.486     4.004    5.041    rosa
```

4.2. ¿De qué variedad es?

- Se trata de la Variedad **ROSA** el Tipo de Semilla que me tocó.

Parte C - Conjuntos.

1.1. Considere los 3 últimos dígitos de su DNI (`3numDNI`) para el seteo de semilla.

- Mis 3 últimos números de mi DNI son: **575.**

1.2. Particione la base en un conjunto de Entrenamiento y uno de Testeo.

- Además:
  - Si su DNI termina en 0, 1, 2 ó 3:
    - Setee **p=0.70**
  - Si su DNI termina en 4, 5, 6 ó 7:
    - Setee **p=0.75**
  - Si su DNI termina en 8 ó 9:
    - Setee **p=0.80**
- Mi último número de DNI es **5**.
- p vá a ser igual a **0.75**.

```
# El código va a quedar:
set.seed(575);particion=createDataPartition(y=base$Variedad,p=0.75,list=FALSE)

entreno= base[particion,]
testeo=  base[-particion,]
```

2. Muestre un head y un summary del conjunto de Entrenamiento y del conjunto de Testeo.

```
# Head Entrenamiento:
head(entreno)
```

```
##    Area Perimetro Compactitud Largo Ancho Asimetria Division Variedad
## 2 14.88    14.57     0.8811 5.554 3.333     1.018   4.956    kama
## 3 14.29    14.09     0.9058 5.291 3.337     2.699   4.825    kama
## 5 16.14    14.99     0.9034 5.658 3.562     1.355   5.175    kama
## 7 14.69    14.49     0.8799 5.563 3.259     3.586   5.219    kama
## 8 14.11    14.10     0.8911 5.420 3.302     2.700   5.000    kama
## 9 16.63    15.46     0.8747 6.053 3.465     2.040   5.877    kama
```

```
# Head Testeo:
head(testeo)
```

```
##     Area Perimetro Compactitud Largo Ancho Asimetria Division Variedad
## 1  15.26    14.84      0.8710 5.763 3.312     2.221    5.220    kama
## 4  13.84    13.94      0.8955 5.324 3.379     2.259    4.805    kama
## 6  14.38    14.21      0.8951 5.386 3.312     2.462    4.956    kama
## 12 14.03    14.16      0.8796 5.438 3.201     1.717    5.001    kama
## 13 13.89    14.02      0.8880 5.439 3.199     3.986    4.738    kama
## 14 13.78    14.06      0.8759 5.479 3.156     3.136    4.872    kama
```

```
# Summary Entrenamiento:
summary(entreno)
```

```
##       Area          Perimetro       Compactitud        Largo      
##  Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899  
##  1st Qu.:12.31   1st Qu.:13.46   1st Qu.:0.8578   1st Qu.:5.263  
##  Median :14.37   Median :14.37   Median :0.8724   Median :5.541  
##  Mean   :14.86   Mean   :14.56   Mean   :0.8711   Mean   :5.629  
##  3rd Qu.:17.22   3rd Qu.:15.71   3rd Qu.:0.8865   3rd Qu.:5.989  
##  Max.   :21.18   Max.   :17.25   Max.   :0.9153   Max.   :6.666  
##      Ancho         Asimetria        Division        Variedad  
##  Min.   :2.648   Min.   :0.7651   Min.   :4.519   kama   :53  
##  1st Qu.:2.947   1st Qu.:2.5985   1st Qu.:5.045   rosa   :53  
##  Median :3.242   Median :3.5980   Median :5.263   canadian:53  
##  Mean   :3.261   Mean   :3.6648   Mean   :5.412                
##  3rd Qu.:3.557   3rd Qu.:4.6940   3rd Qu.:5.877                
##  Max.   :4.033   Max.   :8.4560   Max.   :6.498
```

```
# Summary Testeo:
summary(testeo)
```

```
##       Area          Perimetro       Compactitud        Largo      
##  Min.   :10.74   Min.   :12.73   Min.   :0.8099   Min.   :4.984  
##  1st Qu.:12.26   1st Qu.:13.43   1st Qu.:0.8560   1st Qu.:5.261  
##  Median :14.11   Median :14.21   Median :0.8763   Median :5.479  
##  Mean   :14.82   Mean   :14.55   Mean   :0.8707   Mean   :5.627  
##  3rd Qu.:17.29   3rd Qu.:15.70   3rd Qu.:0.8883   3rd Qu.:5.926  
##  Max.   :20.71   Max.   :17.23   Max.   :0.9183   Max.   :6.675  
##      Ancho         Asimetria        Division        Variedad  
##  Min.   :2.630   Min.   :0.903   Min.   :4.738   kama   :17  
##  1st Qu.:2.933   1st Qu.:2.570   1st Qu.:5.002   rosa   :17  
##  Median :3.232   Median :3.631   Median :5.180   canadian:17  
##  Mean   :3.252   Mean   :3.811   Mean   :5.395                
##  3rd Qu.:3.570   3rd Qu.:4.878   3rd Qu.:5.835                
##  Max.   :3.930   Max.   :8.315   Max.   :6.550
```

3.1. Realice:

- `summary(base$Variedad)`
- `summary(entreno$Variedad)`
- `summary(testeo$Variedad)`

```
summary(base$Variedad)
```

```
##    kama    rosa canadian 
##      70      70      70
```

```
summary(entreno$Variedad)
```

```
##    kama    rosa canadian 
##      53      53      53
```

```
summary(testeo$Variedad)
```

```
##    kama    rosa canadian 
##      17      17      17
```

3.2. ¿Cuántos registros quedaron por variedad de trigo en el conjunto de entrenamiento y en el de testeo?

- En el Conjunto de Entrenamiento quedaron **53** registros por variedad de trigo.
- En el Conjunto de Testeo quedaron **17** registros por variedad de trigo.

Parte D – Árbol de Decisión

1.1. Cree un Árbol de Decisión (con librería rpart) para modelar el problema planteado.

```
# Importamos Librerías:
library(rpart)
```

```
arbol=rpart(Variedad~.,entreno,method="class")
```

1.2. Escriba arbol y muestre una captura de pantalla de la información que aparece.

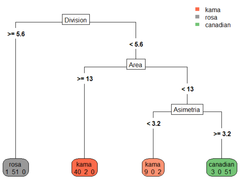
```
arbol
```

```
## n= 159
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 159 106 kama (0.33333333 0.33333333 0.33333333)
##    2) Division>=5.618 52   1 rosa (0.01923077 0.98076923 0.00000000) *
##    3) Division< 5.618 107  54 canadian (0.48598131 0.01869159 0.49532710)
##      6) Area>=13.41 42   2 kama (0.95238095 0.04761905 0.00000000) *
##      7) Area< 13.41 65  12 canadian (0.18461538 0.00000000 0.81538462)
##       14) Asimetria< 3.24 11   2 kama (0.81818182 0.00000000 0.18181818) *
##       15) Asimetria>=3.24 54   3 canadian (0.05555556 0.00000000 0.94444444) *
```

2. Grafique el Árbol de Decisión resultante.

```
# Importamos Librerías:
library(rpart.plot)

#Graficamos el AdD
rpart.plot(arbol,extra=1,type=5)
```



3. ¿Cuántas "hojas" tiene el Árbol de Decisión?

- El Árbol de Decisión tiene 4 hojas.

4.1. Según el Árbol de Decisión creado, ¿cuándo una semilla es de la variedad "rosa"?

- Según mi AdD, la semilla es de variedad **Rosa** cuando la variable predictora **Division** es Mayor o Igual a **5.6** unidades.

4.2. Indique las reglas siguiendo las ramas desde el nodo raíz hasta las hojas "rosa".

- Si **Division** es **Mayor/Igual** a **5.6**, entonces el tipo de semilla es **Rosa**.
- Si **Division** es **Menor** a **5.6**, entonces preguntamos:
  - Si **Área** es **Mayor/Igual** a **13**, entonces el tipo de semilla es **Kama**.
  - Si **Área** es **Menor** a **13**, entonces preguntamos:
    - Si **Asimetría** es **Mayor/Igual** a **3.2**, entonces el tipo de semilla es **Canadian**.
    - Si **Asimetría** es **Menor** a **3.2**, entonces el tipo de semilla es **Kama**.

5.1. Testee el Árbol de Decisión.

```
pred=predict(arbol,testeo,type="class")
```

5.2. Compare: `head(pred,10)` con `head(test$Variedad,10)`.

```
head(pred,10)
```

```
##        1        4        6       12       13       14       17       20
##    kama     kama     kama     kama     kama     kama     kama canadian
##       22       26
##    kama     kama
## Levels: kama rosa canadian
```

```
head(test$Variedad,10)
```

```
##  [1] kama kama kama kama kama kama kama kama kama kama
## Levels: kama rosa canadian
```

5.3. Vea si la predicción de los 10 primeros elementos coincide con lo esperado.

- En líneas Generales podemos afirmar que la predicción del AdD coincide con lo esperado en Test, ya que predijo **Kama** a todos los elementos, excepto a uno que lo clasificó como **Canadian**.

6.1. Calcule la matriz de confusión y muestre los resultados obtenidos.

```
confusionMatrix(pred,test$Variedad)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction kama rosa canadian
##   kama       16    0        4
##   rosa        0   17        0
##   canadian    1    0       13
##
## Overall Statistics
##
##                Accuracy : 0.902
##                  95% CI : (0.7859, 0.9674)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8529
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: kama Class: rosa Class: canadian
## Sensitivity               0.9412     1.0000          0.7647
## Specificity               0.8824     1.0000          0.9706
## Pos Pred Value            0.8000     1.0000          0.9286
## Neg Pred Value            0.9677     1.0000          0.8919
## Prevalence                0.3333     0.3333          0.3333
## Detection Rate            0.3137     0.3333          0.2549
## Detection Prevalence      0.3922     0.3333          0.2745
## Balanced Accuracy         0.9118     1.0000          0.8676
```

7.1. Calcule el Accuracy según la cantidad de registros bien clasificados.

- **Fórmula**:

$$Accuracy = \frac{\text{Diagonal Aciertos}}{Total}$$

- Cantidad de Registros bien clasificados:
  - **Kama**: 16.
  - **Rosa**: 17.
  - **Canadian**: 13.

7.2. Indique con números la fórmula que usó.

$$Accuracy = \frac{(17 + 16 + 13)}{(17 + 17 + 17)}$$

7.3. Verifique que coincida con el Accuracy obtenido por confusionMatrix.

```
## El Accuracy obtenido en este punto es: 0.902 ; Cuyo valor coincide
## con el Acc obtenido por la Matriz de Confusión del punto 6.1.
```

8. ¿Cuál categoría presenta menor sensibilidad?

- La Categoría que presenta menor **Sensibilidad** es la categoría: **Canadian**, que presenta una Sensibilidad de **0.7647**, frente a la Sensibilidad de 1.0 y 0.94 que presentan **kama** y **Rosa** respectivamente.

9.1. Considere el grano de trigo correspondiente a los últimos 2 dígitos de su DNI.

- Mis 2 últimos números de mi DNI son: **75**.

9.2. Según el Árbol de Decisión, ¿qué variedad es?

```
predict(arbol,trigo,type="class")
```

```
##   75
## rosa
## Levels: kama rosa canadian
```

- Según el AdD, la Semilla del Índice **75**, correspondería a la variedad: **Rosa**.

9.3. ¿Coincide la predicción con lo esperado?

```
trigo
```

```
##     Area Perimetro Compactitud Largo Ancho Asimetria Division Variedad
## 75 16.82    15.51      0.8786 6.017 3.486     4.004    5.841    rosa
```

- La Variedad esperada es **Rosa**, por lo tanto podemos afirmar que si **coinciden** ambos resultados.

10. Indique la predicción de "Trigo" con probabilidades.

```
predict(arbol,trigo)
```

```
##          kama       rosa canadian
## 75 0.01923077 0.98076923        0
```

- El AdD Predice que nuestro Trigo (`base[75,]`) va a ser **Rosa** con un **98%** de confianza.

11. Indique la regla del Árbol de Decisión para "Trigo".

```
rpart.predict(arbol, trigo, rules=TRUE)
```

```
##          kama       rosa canadian
## 75 0.01923077 0.98076923        0 because Division >= 5.6
```

- Cuando seteamos `rules=TRUE` , aparece en pantalla `Division >= 5.6` para clasificar nuestra semilla a la variedad **Rosa**.

Parte E – Red Neuronal.

1.1. Considere los 3 últimos dígitos de su DNI (3numDNI) para el seteo de semilla.

- Mis 3 últimos números de mi DNI son: **575**.

1.2. Cree una Red Neuronal para modelar el problema planteado.

- Con:
  - maxIt=10000
  - size=10.

```
# Importamos Librerias:
library(nnet)
library(NeuralNetTools)
```

1.3. Indique el código R utilizado.

```
set.seed(575);red=nnet(Variedad~.,entreno,size=10,maxit=10000)
```

2. Muestre una captura de pantalla de la lista de iteraciones de la Red Neuronal.

```
set.seed(575);red=nnet(Variedad~.,entreno,size=10,maxit=10000)
```

```
## # weights:  113
## initial  value 177.702910
## iter  10 value 108.747857
## iter  20 value 44.811744
## iter  30 value 8.669088
## iter  40 value 0.239508
## iter  50 value 0.002444
## final  value 0.000097
## converged
```

3. Escriba red y muestre una captura de pantalla de la información que aparece.

```
red
```

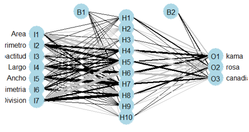
```
## a 7-10-3 network with 113 weights
## inputs: Area Perimetro Compactitud Largo Ancho Asimetria Division
## output(s): Variedad
## options were - softmax modelling
```

4. Indique la cantidad de pesos y la cantidad de iteraciones resultantes.

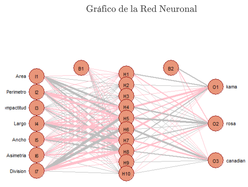
- Cantidad de pesos: **113** weights.
- Cantidad de iteraciones: más de **50** iteraciones realizó la Red Neuronal.

5.1. Dibuje la Red Neuronal.

```
plotnet(red)
```



5.2. Cambiar los colores del gráfico de la Red Neuronal.



6.1. Testee la Red Neuronal.

```
pred2=predict(red,testeo,type="class")
```

6.2. Compare `head(pred2,10)` con `head(testeo$Variedad,10)`.

```
# Observamos la predicción:
head(pred2,10)
```

```
##  [1] "kama"     "kama"     "kama"     "kama"     "kama"     "kama"
## [7] "canadian" "canadian" "kama"     "kama"
```

```
# Observamos Test:
head(testeo$Variedad,10)
```

```
##  [1] kama kama kama kama kama kama kama kama kama kama
## Levels: kama rosa canadian
```

6.3. Vea si la predicción de los 10 primeros elementos coincide con lo esperado.

- La predicción de los primeros 10 elementos en General coincide, salvo 2 elementos que la Red predijo como **Canadian** cuando lo esperado era que predigra **Kama**.

7. Calcule la matriz de confusión y muestre los resultados obtenidos.

```
confusionMatrix(factor(pred2),testeo$Variedad)
```

```
## Warning in confusionMatrix.default(factor(pred2), testeo$Variedad): Levels are
## not in the same order for reference and data. Refactoring data to match.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction kama rosa canadian
##   kama       14    0        2
##   rosa        0   17        0
##   canadian    3    0       15
##
## Overall Statistics
##
##                Accuracy : 0.902
##                  95% CI : (0.7859, 0.9674)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8529
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: kama Class: rosa Class: canadian
## Sensitivity               0.8235     1.0000          0.8824
## Specificity               0.9412     1.0000          0.9118
## Pos Pred Value            0.8750     1.0000          0.8333
## Neg Pred Value            0.9143     1.0000          0.9394
## Prevalence                0.3333     0.3333          0.3333
## Detection Rate            0.2745     0.3333          0.2941
## Detection Prevalence      0.3137     0.3333          0.3529
## Balanced Accuracy         0.8824     1.0000          0.8971
```

8. ¿Cuál fue el Accuracy?

- El Accuracy fue de **0.902**.

9.1. Considere el grano de trigo correspondiente a los últimos 2 dígitos de su DNI.

- Mis 2 últimos números de mi DNI son: 75.

9.2. Según la Red Neuronal, ¿qué variedad es?

```
predict(red,trigo,type="class")
```

```
## [1] "rosa"
```

- Según la Red Neuronal, la Semilla del Índice **75**, correspondería a la variedad: **Rosa**.

9.3. ¿Coincide la predicción con la variedad esperada?

```
trigo
```

```
##     Area Perimetro Compactitud Largo Ancho Asimetria Division Variedad
## 75 16.82     15.51     0.8786 6.017 3.486     4.004     5.841     rosa
```

- La Variedad esperada es **Rosa**, por lo tanto podemos afirmar que si **coinciden** ambos resultados.

Parte F – Comparación de modelos.

1. Cree una tabla con el Accuracy de cada modelo, y la Sensibilidad y Especificidad de cada modelo por categoría.

Variedad	AdD Accuracy	RN Accuracy	AdD Sensibilidad	RN Sensibilidad	AdD Especificidad	RN Especificidad
Kama	0.902	0.902	0.9412	0.8235	0.8824	0.9412
Rosa	0.902	0.902	1.0000	1.0000	1.0000	1.0000
Canadian	0.902	0.902	0.7647	0.8824	0.9706	0.9118

2.1. Compare los resultados obtenidos con el Árbol de Decisión y la Red Neuronal.

2.2. ¿Cuál modelo le parece que resultó mejor?

- Con Respecto al mejor modelo, creo que no hay un claro Ganador.

2.3. ¿Según qué criterio?

- El **criterio** en el cual fundamento mi postura es que:
  - Para ambos modelos, el **Accuracy** es el mismo valor:

$$\Sigma Accuracy_{AdD} = \Sigma Accuracy_{RN} = 0.902$$

  - Con Respecto a la **Sensibilidad**, que es la **tasa de verdaderos positivos**:[1]
    - En General son altas en ambos modelos, eso es bueno ya que: Cuanto mayor es la Sensibilidad, la tasa de **FN** será menor[2], por lo que la cantidad de Semillas que pertenecen a una variedad y fueron clasificadas como otra variedad, se reduce.
    - Para la variedad **Rosa**, ambos modelos presentan el mismo valor.
    - Para la variedad **Kama**, la Sensibilidad del AdD resultó mayor que de la RN.
    - Para la variedad **Canadian**, la Sensibilidad de la RN resultó ser mayor.
    - Pero en el **Global** la Sumatoria de la Sensibilidad en ambos modelos, da el mismo resultado:

$$\Sigma Sensibilidad_{AdD} = \Sigma Sensibilidad_{RN} = 2.7059$$

  - Con Respecto a la **Especificidad**, que es la **tasa de verdaderos negativos**:[3]
  - En General son altas en ambos modelos, eso es bueno ya que: Cuanto mayor es la Especificidad, la tasa de **FP** será menor[4], por lo que la cantidad de Semillas que fueron clasificadas como una determinada variedad, pero en realidad son de otra variedad, se reduce.
    - Para la variedad **Rosa**, ambos modelos presentan el mismo valor.
    - Para la variedad **Kama**, la Especificidad del RN resultó mayor que de la AdD.
    - Para la variedad **Canadian**, la Especificidad de la AdD resultó ser mayor.
    - Pero en el **Global** la Sumatoria de la Especificidad en ambos modelos, da el mismo resultado:

$$\Sigma Especificidad_{AdD} = \Sigma Especificidad_{RN} = 2.853$$

- En base a este breve argumento baso mi criterio de considerar que no hubo un claro Ganador respecto a estos dos modelos de Machine Learning analizados con este Dataset de Semillas de Trigo. Habría que utilizar otros Datasets para volver a comparar sus performances con nuevos datos.

Anexo.

Código:

```
Parte A – Preprocesamiento de los datos.
base=read.table("seeds_dataset.txt",header=F)
head(base)
names(base)[names(base)=="V1"]="Area"
names(base)[names(base)=="V2"]="Perimetro"
names(base)[names(base)=="V3"]="Compactitud"
names(base)[names(base)=="V4"]="Largo"
names(base)[names(base)=="V5"]="Ancho"
names(base)[names(base)=="V6"]="Asimetria"
names(base)[names(base)=="V7"]="Division"
names(base)[names(base)=="V8"]="Variedad"
head(base)
base$Variedad=factor(base$Variedad,levels=c(1,2,3),
                     labels=c("kama","rosa","canadian"))
head(base)

Parte B – Análisis Exploratorio de Datos.
summary(base$Variedad)
pie(table(base$Variedad),main="Variedad  de Semillas por Tipo")
library(caret)
xyplot(base$Largo~base$Ancho,groups=base$Variedad,base,auto.key=TRUE,
       par.settings=simpleTheme(pch=c(3,4,5)),pch=c(3,4,5),
       main="Largo vs Ancho de Semilla de Trigo por Tipo",
       xlab="Ancho de Semilla", ylab="Largo de Semilla")
trigo=base[75,]
trigo
set.seed(575);particion=createDataPartition(y=base$Variedad,p=0.75,list=FALSE)
entreno= base[particion,]
testeo=  base[-particion,]
head(entreno)
head(testeo)
summary(entreno)
summary(testeo)
summary(base$Variedad)
summary(entreno$Variedad)
summary(testeo$Variedad)

Parte D – Árbol de Decisión
library(rpart)
arbol=rpart(Variedad~.,entreno,method="class")
arbol
library(rpart.plot)
rpart.plot(arbol,extra=1,type=5)
pred=predict(arbol,testeo,type="class")
head(pred,10)
head(testeo$Variedad,10)
confusionMatrix(pred,testeo$Variedad)
predict(arbol,trigo,type="class")
predict(arbol,trigo)
rpart.predict(arbol, trigo, rules=TRUE)
library(nnet)
library(NeuralNetTools)
set.seed(575);red=nnet(Variedad~.,entreno,size=10,maxit=10000)
red
plotnet(red)
plotnet(red,
        circle_col = "darksalmon",
        pos_col = "pink",
        neg_col = "grey",
        bord_col = "darkred",
        circle_cex=5,
        alpha_val=1,
        cex_val=0.70,
        max_sp= TRUE)
pred2=predict(red,testeo,type="class")
head(pred2,10)
head(testeo$Variedad,10)
confusionMatrix(factor(pred2),testeo$Variedad)
predict(red,trigo,type="class")
trigo
```

Librerias:

```
library(caret)
library(rpart)
library(rpart.plot)
library(nnet)
library(NeuralNetTools)
```

Bibliografía:

- Centrar Titulos
- Centrar Textos
- How-to-add-whitespace
- Wheet Grains
- Árboles de Decisión I
- Árboles de Decisión II
- Algoritmo CART
- Fórmula Latex
- Plotnet
- Plotnet
- R Colors
- Plots
- Markdown Tables
- Sensibilidad-Especificidad
- Sensibilidad-Especificidad

¡Muchas Gracias!

1. Sensibilidad↵
2. Explicación Sensibilidad↵
3. Especificidad↵
4. Explicación Especificidad↵