Diplomatura en Ciencia de Datos

Módulo 2: Machine Learning - TP N°2

Profesor: Marcela Leticia Riccillo

Nombre: Ivan Kim

DNI: 37.687.040

Parte A – Preprocesamiento de los datos

1) Ingrese a la página web de la Universidad de California UCI
https://archive.ics.uci.edu/ml/datasets/seeds

Copie aquí el resumen que dice "Measurements of…"

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.

Baje el archivo zip con el botón Download. Copie fuera del zip al archivo seeds_dataset.txt.

2) Busque en la página web, en el apartado "Additional Information:" e indique aquí: ¿cuáles son las 3 variedades de trigo que se estudiarán?

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

Las 3 variedades de trigo que se estudiarán en este trabajo son: Kama, Rosa y Canadian.

Fuente: https://www.diet-health.info/es/recetas/ingredientes/in/wl9405-grano-de-trigo

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

base=read.table("seeds_dataset.txt",header=FALSE)
Muestre un 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

Muestra las variables de trigos sin clasificar

4) 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
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"]="VariedadDeSemilla"

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

	Area	Perimetro	Compactitud	Largo	Ancho	Asimetria	Division	VariedadDeSemilla
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

Se clasificaron los diferentes tipos de trigo según el enunciado

5) Transforme a categórica la variable VariedadDeSemilla y renombre las variedades 1, 2 y 3 como “kama”, “rosa” y “canadian”:
base$VariedadDeSemilla=factor(base$VariedadDeSemilla,levels=c(1,2,3),labels=c("kama","rosa","canadian"))

Muestre un head(base) para ver cómo quedaron las variables transformadas.

```
   Area Perimetro Compactitud Largo Ancho Asimetria Division VariedadDeSemilla
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
```

Las variedades de semillas fueron clasificados según el número indicado:
1 = Kama, 2 = Rosa y 3 = Canadian.

Parte B – Análisis Exploratorio de Datos

1) ¿Cuántas semillas hay en total y por variedad?
dim(base)
summary(base$VariedadDeSemilla)

```
> dim(base)
[1] 210   8
> summary(base$VariedadDeSemilla)
    kama     rosa canadian
      70       70       70
```

Se pueden ver las dimensión de la base que son 210 filas y 8 columnas
En el resumen se puede ver que hay 70 semillas en cada variedad

2) Realice un gráfico de sectores de la variable a predecir VariedadDeSemilla. Elija un Título. pie(table(base$VariedadDeSemilla),main="Titulo")

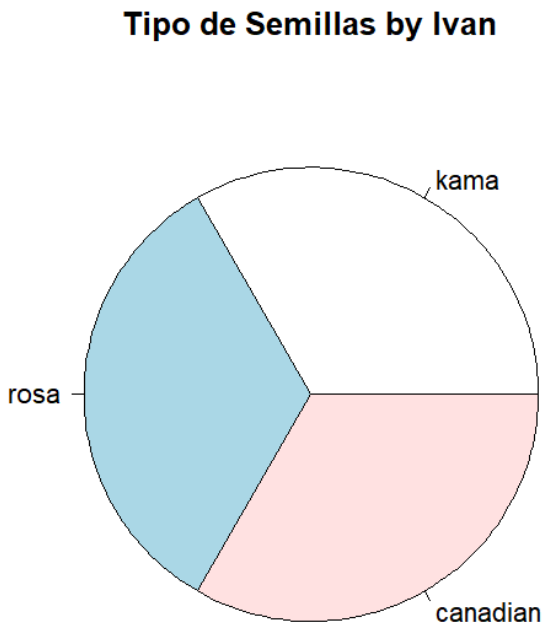
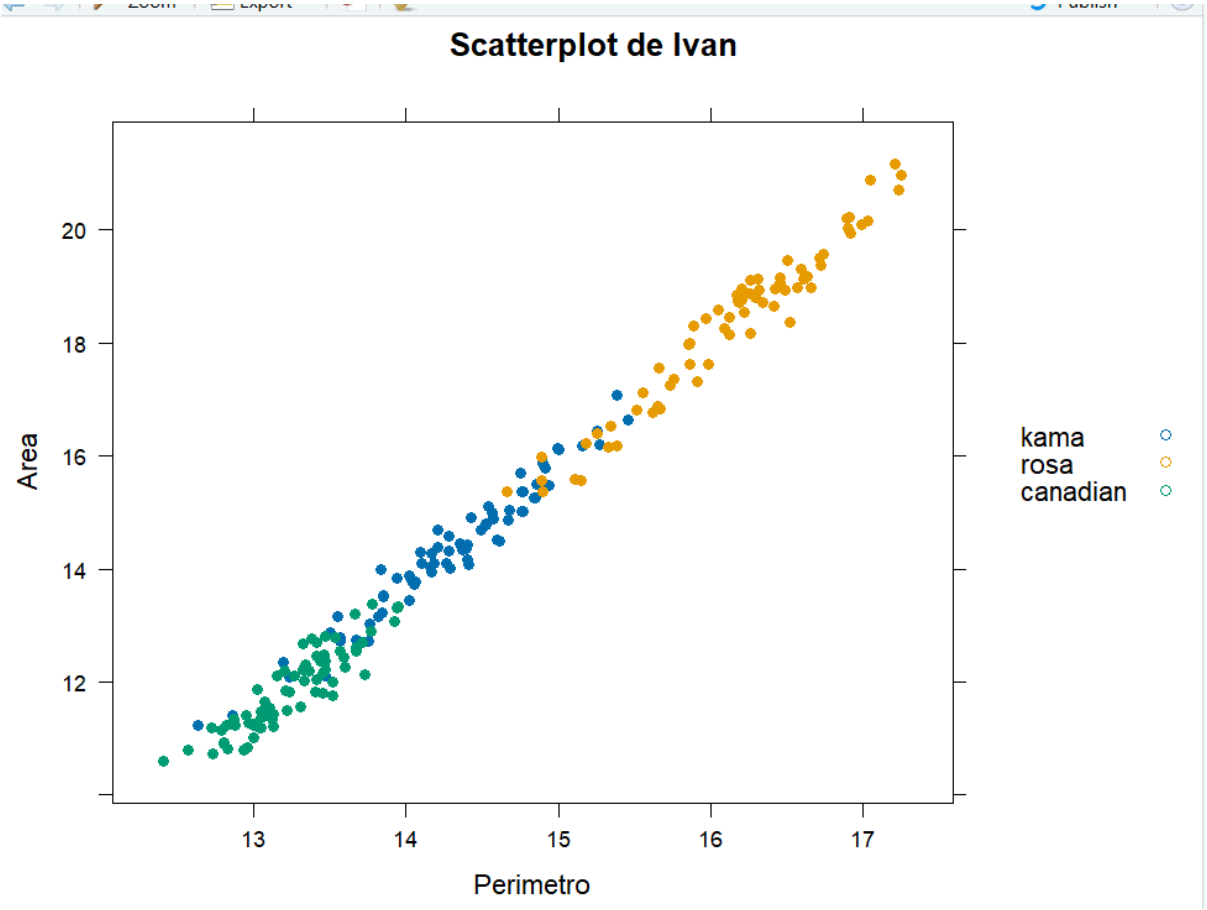


Gráfico de torta que representa en partes iguales los 3 tipos de semillas dentro de la base.

3) Realice un gráfico de dispersión entre 2 variables (que no sean VariedadDeSemilla) y coloréelo por la variable VariedadDeSemilla (agregue una leyenda que indique cuál es cada grupo).

```
library(caret)
xyplot(Vy~Vx,groups=VariedadDeSemilla,base,auto.key=TRUE,main="Título",pch=numero)
```



Aquí se puede visualizar la relación entre el área y el perímetro dentro dentro del scatterplot marcando un tendencia lineal ascendente.

a) Para el título ingrese su nombre, como “Gráfico de Marcela” (o sea "Gráfico de SuNombre").

El título elegido fue “Scatterplot de Iván”

b) Indique el código R utilizado.

```
dim(base)
summary(base$VariedadDeSemilla)

pie(table(base$VariedadDeSemilla),main="Tipo de Semillas by Ivan")

library(caret)
xyplot(Area ~ Perimetro,groups = VariedadDeSemilla,
       data = base, auto.key = TRUE,main = "Scatterplot de Ivan",
       pch = 16)
```

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) y muestre aquí el registro correspondiente. trigo=base[2numDNI,] trigo ¿De qué variedad es?

```
trigo=base[40,]
base[40,]
```

```
> base[40,]
    Area Perimetro Compactitud Largo Ancho Asimetria Division VariedadDeSemilla
40 14.28     14.17      0.8944 5.397 3.298     6.685    5.001              kama
>
```

El trigo que se muestra en el número de fila 40 (último 2 dígitos de mi DNI) es de variedad kama.

Parte C - Conjuntos

1) Considere su DNI (completo) para el seteo de semilla y particione la base en un conjunto de entrenamiento y uno de testeo, utilizando la instrucción createDataPartition de la librería caret.

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

set.seed(DNI);particion=createDataPartition(y=base$VariedadDeSemilla,p=asignado,list=FALSE)
entreno=base[particion,]
testeo=base[-particion,]

Indique cómo quedó el código R utilizado.

```
library(caret)
set.seed(37687040);particion=createDataPartition(y=base$VariedadDeSemilla,p=0.70,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(entreno)

```
> head(entreno)
    Area Perimetro Compactitud Largo Ancho Asimetria Division VariedadDeSemilla
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
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
7 14.69     14.49      0.8799 5.563 3.259     3.586    5.219             kama
```

summary(entreno)

```
> summary(entreno)
      Area         Perimetro      Compactitud        Largo           Ancho         Asimetria        Division     VariedadDeSemilla
 Min.   :10.59   Min.   :12.41   Min.   :0.8081   Min.   :4.899   Min.   :2.641   Min.   :0.8551   Min.   :4.519   kama    :49
 1st Qu.:12.21   1st Qu.:13.46   1st Qu.:0.8558   1st Qu.:5.253   1st Qu.:2.926   1st Qu.:2.4830   1st Qu.:5.045   rosa    :49
 Median :14.46   Median :14.40   Median :0.8724   Median :5.541   Median :3.245   Median :3.5980   Median :5.263   canadian:49
 Mean   :14.87   Mean   :14.57   Mean   :0.8708   Mean   :5.631   Mean   :3.260   Mean   :3.6366   Mean   :5.415
 3rd Qu.:17.17   3rd Qu.:15.70   3rd Qu.:0.8874   3rd Qu.:5.979   3rd Qu.:3.557   3rd Qu.:4.5410   3rd Qu.:5.879
 Max.   :21.18   Max.   :17.23   Max.   :0.9183   Max.   :6.675   Max.   :4.033   Max.   :8.3150   Max.   :6.550
>
```

head(testeo)

```
> head(testeo)
    Area Perimetro Compactitud Largo Ancho Asimetria Division VariedadDeSemilla
4  13.84     13.94      0.8955 5.324 3.379     2.259    4.805             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
15 13.74     14.05      0.8744 5.482 3.114     2.932    4.825             kama
16 14.59     14.28      0.8993 5.351 3.333     4.185    4.781             kama
18 15.69     14.75      0.9058 5.527 3.514     1.599    5.046             kama
>
```

summary(testeo)

```
> summary(testeo)
      Area         Perimetro      Compactitud        Largo           Ancho         Asimetria        Division     VariedadDeSemilla
 Min.   :10.74   Min.   :12.63   Min.   :0.8107   Min.   :4.902   Min.   :2.630   Min.   :0.7651   Min.   :4.703   kama    :21
 1st Qu.:12.64   1st Qu.:13.41   1st Qu.:0.8598   1st Qu.:5.298   1st Qu.:3.022   1st Qu.:2.6560   1st Qu.:5.042   rosa    :21
 Median :14.11   Median :14.18   Median :0.8779   Median :5.516   Median :3.201   Median :3.6910   Median :5.178   canadian:21
 Mean   :14.79   Mean   :14.53   Mean   :0.8716   Mean   :5.622   Mean   :3.255   Mean   :3.8486   Mean   :5.392
 3rd Qu.:17.34   3rd Qu.:15.71   3rd Qu.:0.8877   3rd Qu.:5.955   3rd Qu.:3.570   3rd Qu.:4.9350   3rd Qu.:5.795
 Max.   :20.97   Max.   :17.25   Max.   :0.9077   Max.   :6.563   Max.   :3.991   Max.   :8.4560   Max.   :6.498
>
```

3) Realice un
summary(base$VariedadDeSemilla)
summary(entreno$VariedadDeSemilla)
summary(testeo$VariedadDeSemilla)

```
> summary(base$VariedadDeSemilla)
    kama     rosa canadian
      70       70       70
> summary(entreno$VariedadDeSemilla)
    kama     rosa canadian
      49       49       49
> summary(testeo$VariedadDeSemilla)
    kama     rosa canadian
      21       21       21
```

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

En el de entreno quedó 49 de cada uno y el de testeo quedó 21 de cada uno

Parte D - Arbol de Decision

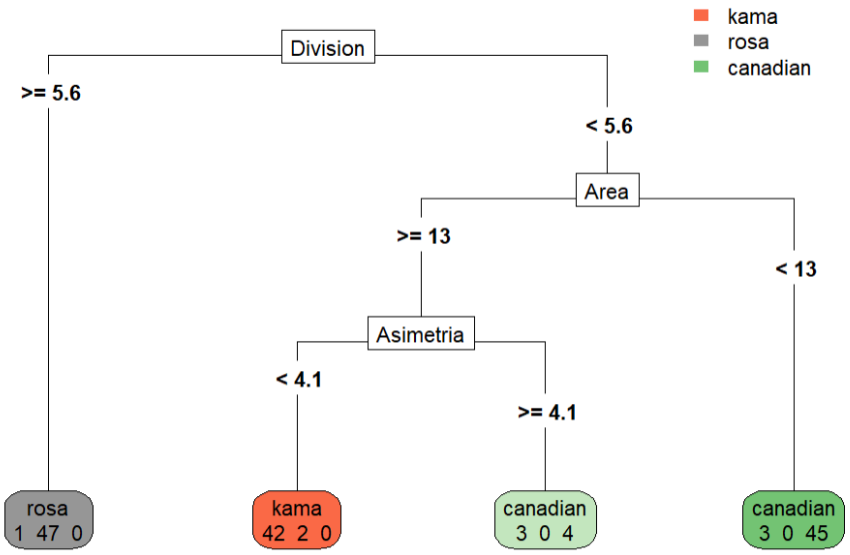
1) Cree un Árbol de Decisión (con librería rpart) para modelar el problema planteado.

arbol=rpart(VariedadDeSemilla~.,entreno)

Escriba print(arbol) y muestre una captura de pantalla de la información que aparece.

```
1) root 147 98 kama (0.33333333 0.33333333 0.33333333)
  2) Division>=5.5755 48  1 rosa (0.02083333 0.97916667 0.00000000) *
  3) Division< 5.5755 99 50 canadian (0.48484848 0.02020202 0.49494949)
    6) Area>=12.71 51  6 kama (0.88235294 0.03921569 0.07843137)
     12) Asimetria< 4.1365 44  2 kama (0.95454545 0.04545455 0.00000000) *
     13) Asimetria>=4.1365 7  3 canadian (0.42857143 0.00000000 0.57142857) *
    7) Area< 12.71 48  3 canadian (0.06250000 0.00000000 0.93750000) *
>
```

2) Grafique el Árbol de Decisión resultante con la instrucción rpart.plot de la librería rpart.plot
library(rpart.plot)
rpart.plot(arbol,extra=1,type=5)



3) ¿Cuántas “hojas” tiene el Árbol de Decisión?

Se visualiza 4 hojas dentro del AdD.

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”).

El AdD interpreta que las semillas con una división frontal grande mayor o igual a 5.6 pertenecen a la variedad rosa, sin importar los valores de las demás variables.

5) Calcule la matriz de confusión utilizando la instrucción confusionMatrix de la librería caret. Muestre una captura de pantalla de los resultados completos (la matriz de confusión, accuracy y tablas).

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

```
Confusion Matrix and Statistics

          Reference
Prediction kama rosa canadian
  kama       15    0        1
  rosa        0   21        0
  canadian    6    0       20

Overall Statistics

               Accuracy : 0.8889
                 95% CI : (0.7844, 0.9541)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8333

 Mcnemar's Test P-Value : NA
```

```
Statistics by Class:

                     Class: kama Class: rosa Class: canadian
Sensitivity               0.7143      1.0000           0.9524
Specificity               0.9762      1.0000           0.8571
Pos Pred Value            0.9375      1.0000           0.7692
Neg Pred Value            0.8723      1.0000           0.9730
Prevalence                0.3333      0.3333           0.3333
Detection Rate            0.2381      0.3333           0.3175
Detection Prevalence      0.2540      0.3333           0.4127
Balanced Accuracy         0.8452      1.0000           0.9048
>
```

6) La cantidad de elementos de la matriz de confusión es igual a la cantidad de elementos de testeo (o sea dim(testeo)).

```
> dim(testeo)
[1] 63  8
>
```

Sumando todos los elementos de la matriz de confusión 15 + 21 + 20 + 6 + 1 = 63.

Sume la cantidad de elementos de la diagonal de la matriz de confusión y divida el resultado por dim(testeo).
Muestre la cuenta con números y muestre que es igual al accuracy.

Sumando la diagonal de la matriz 15 + 21 + 20 = 56 / la cantidad de registros total (63)

Accuracy = 56 / 63 = 0.8889

Como se puede ver, da el mismo valor manualmente como lo calcula R.

7) Vea la tabla Statistics by Class debajo de la matriz de confusión e indique cuál clase presenta menor sensibilidad.

La que presenta menor sensibilidad es la semilla “Kama” con 0,7143.

Esto quiere decir que el modelo tuvo más dificultad para identificar correctamente las semillas de esa variedad en comparación a las otras.

8) Considere el grano de trigo correspondiente a los últimos 2 dígitos de su DNI.
Según el Árbol de Decisión, ¿qué variedad es?
predict(arbol,trigo,type="class")

```
> trigo <- base[40, ]
> trigo$VariedadDeSemilla
[1] kama
Levels: kama rosa canadian
> predict(arbol,trigo,type="class")
      40
canadian
Levels: kama rosa canadian
```

Según el árbol es “canadian”

¿Coincide la variedad predicha por el Árbol de Decisión con la variedad original de la semilla?

Según el Número de fila 40 (último número de mi DNI) dice que es “Kama” y la predicción dice “Canadian” por ende no coinciden.

Parte E – Red Neuronal

1) Considere su DNI (completo) para el seteo de semilla y cree una Red Neuronal (con librería nnet) para modelar el problema planteado con maxit=10000 y cantidad de neuronas en la capa oculta size=25.

library(nnet)
set.seed(DNI);red=nnet(VariedadDeSemilla~.,entreno,size=25,maxit=10000)

Indique el código R utilizado.

library(nnet)
set.seed(37687040);red=nnet(VariedadDeSemilla~.,entreno,size=25,maxit=10000)

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

```
> set.seed(37687040);red=nnet(VariedadDeSemilla~.,entreno,size=25,maxit=10000)
# weights:  278
initial  value 226.998287
iter  10 value 102.332748
iter  20 value 65.575811
iter  30 value 17.210818
iter  40 value 10.219389
iter  50 value 1.401163
iter  60 value 0.017918
iter  70 value 0.001276
iter  80 value 0.000235
final  value 0.000008
converged
```

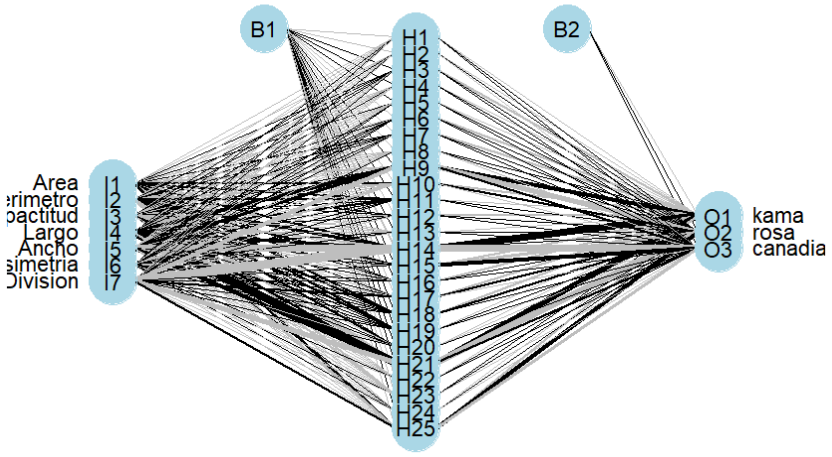
3) Escriba print(red) y muestre una captura de pantalla de la información que aparece.

```
> print(red)
a 7-25-3 network with 278 weights
inputs: Area Perimetro Compactitud Largo Ancho Asimetria Division
output(s): VariedadDeSemilla
options were - softmax modelling
>
```

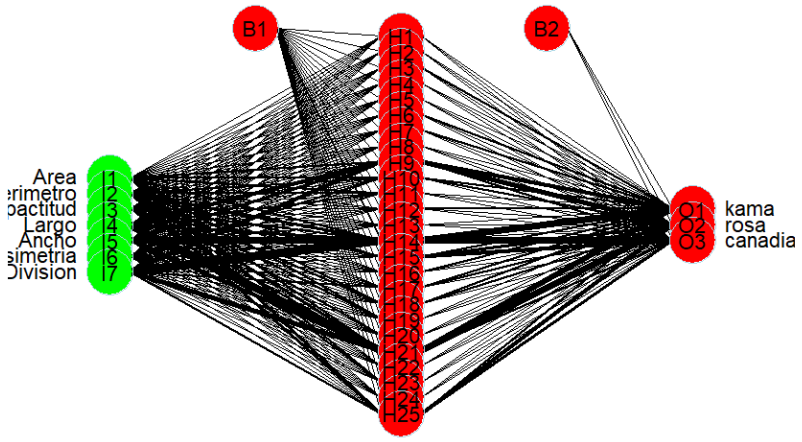
4) Indique la cantidad de pesos y la cantidad de iteraciones resultantes.

La cantidad de pesos son de 278 y con 80 iteraciones resultantes.

5) Dibuje la Red Neuronal
library(NeuralNetTools)
plotnet(red)



Optativo: Cambiar los colores del gráfico de la Red Neuronal.




```
plotnet(red,
        circle_col = list("green", "red"),
        pos_col = "black",
        neg_col = "black")
```

6) Calcule la matriz de confusión utilizando la instrucción confusionMatrix de la librería caret. Muestre una captura de pantalla de los resultados

completos (la matriz de confusión, accuracy y tablas).

pred2=predict(red,testeo,type="class")
confusionMatrix(factor(pred2),testeo$VariedadDeSemilla)

```
> confusionMatrix(factor(pred2),testeo$VariedadDeSemilla)
Confusion Matrix and Statistics

          Reference
Prediction kama rosa canadian
  kama       19    0        3
  rosa        0   21        0
  canadian    2    0       18

Overall Statistics

               Accuracy : 0.9206
                 95% CI : (0.8244, 0.9737)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.881

 Mcnemar's Test P-Value : NA
```

```
Statistics by Class:

                     Class: kama Class: rosa Class: canadian
Sensitivity               0.9048      1.0000          0.8571
Specificity               0.9286      1.0000          0.9524
Pos Pred Value            0.8636      1.0000          0.9000
Neg Pred Value            0.9512      1.0000          0.9302
Prevalence                0.3333      0.3333          0.3333
Detection Rate            0.3016      0.3333          0.2857
Detection Prevalence      0.3492      0.3333          0.3175
Balanced Accuracy         0.9167      1.0000          0.9048
```

7) ¿Cuál fue el accuracy?

El accuracy es de 0,9206 o 92,06%.

8) Considere el grano de trigo correspondiente a los últimos 2 dígitos de su DNI. Según la Red Neuronal, ¿qué variedad es?
predict(red,trigo,type="class")

```
> trigo <- base[40, ]
> predict(red, trigo, type = "class")
[1] "kama"
> trigo$VariedadDeSemilla
[1] kama
Levels: kama rosa canadian
```

La variedad es “kama”

¿Coincide la predicción con la variedad esperada?

Si, a diferencia con el AdD, esta red neuronal coincide con la predicción de variedad esperada. (Semilla tipo “kama”)

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. La tabla esperada no es hacerla con R, sino una tabla tipo Word.

Modelo	Accuracy	Sensibilidad (*kama*)	Especificidad (*kama*)	Sensibilidad (*rosa*)	Especificidad (*rosa*)	Sensibilidad (*canadian*)	Especificidad (*canadian*)
Árbol de Decisión	0.8889	0.7143	0.9762	1.0000	1.0000	0.9524	0.8571
Red Neuronal	0.9206	0.9048	0.9286	1.0000	1.0000	0.8571	0.9524

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

Como se puede ver la red neuronal tuvo un mejor resultado global pero el árbol de decisión fue más fuerte en una clase que en otras y más simple de interpretar.

A mi parecer el modelo de red neuronal fue el más eficiente, ya que en el accuracy fue mayor.

ANEXO CÓDIGO R

El código se usó en este orden.

```
getwd()
base <- read.table("...", header = FALSE)
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"]="VariedadDeSemilla"
head(base)

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

dim(base)
summary(base$VariedadDeSemilla)

pie(table(base$VariedadDeSemilla),main="Tipo de Semillas by Ivan")

library(caret)
xyplot(Area ~ Perimetro,groups = VariedadDeSemilla,
      data = base, auto.key = TRUE,main = "Scatterplot de Ivan",
      pch = 16)

trigo=base[40,]
base[40,]

library(caret)
set.seed(37687040);particion=createDataPartition(y=base$VariedadDeSemilla,p=0.70,list=F
ALSE)
entreno=base[particion,]
testeo=base[-particion,]

head(entreno)
summary(entreno)
head(testeo)
summary(testeo)

summary(base$VariedadDeSemilla)
```

```
summary(entreno$VariedadDeSemilla)
summary(testeo$VariedadDeSemilla)

library(rpart)
arbol=rpart(VariedadDeSemilla~.,entreno)
print(arbol)

library(rpart.plot)
rpart.plot(arbol,extra=1,type=5)

pred=predict(arbol,testeo,type="class")
confusionMatrix(pred,testeo$VariedadDeSemilla)

dim(testeo)

trigo <- base[40, ]
trigo$VariedadDeSemilla
predict(arbol,trigo,type="class")

library(nnet)
set.seed(37687040);red=nnet(VariedadDeSemilla~.,entreno,size=25,maxit=10000)
print(red)

library(NeuralNetTools)
plotnet(red)

plotnet(red,
      circle_col = list("green", "red"),
      pos_col = "black",
      neg_col = "black")

library(caret)
pred2=predict(red,testeo,type="class")
confusionMatrix(factor(pred2),testeo$VariedadDeSemilla)


trigo <- base[40, ]
predict(red, trigo, type = "class")

trigo$VariedadDeSemilla
```