Inteligencia Artificial y Machine Learning
Certificación Avanzada en Ciencia de Datos

Trabajo Práctico 2

Aprendizaje Supervisado

Profesora: Dra Marcela Riccillo

Alumno: Federico José Zarategui

Dni: 27.691.042

En este ejercicio se pide realizar modelos para predecir variedades de granos de trigo. Primero se realiza un Análisis Exploratorio de los Datos para entender la base, luego se particiona la base en un conjunto de entrenamiento y uno de testeo, después con estos conjuntos se realiza el modelado de un Árbol de Decisión y el modelado de una Red Neuronal. Finalmente se compara la eficiencia de ambos modelos.

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

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

Las 3 variedades de trigo que se estudiaran son:

- Kama
- Rosa
- Canadian

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

https://www.loscanastos.me/post/2018/08/07/grano-de-trigo-y-sus-beneficios

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

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
```

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 de la base con 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

Parte B – Análisis Exploratorio de Datos

1) ¿Cuántas semillas hay en total y por variedad?

```
dim(base)
```

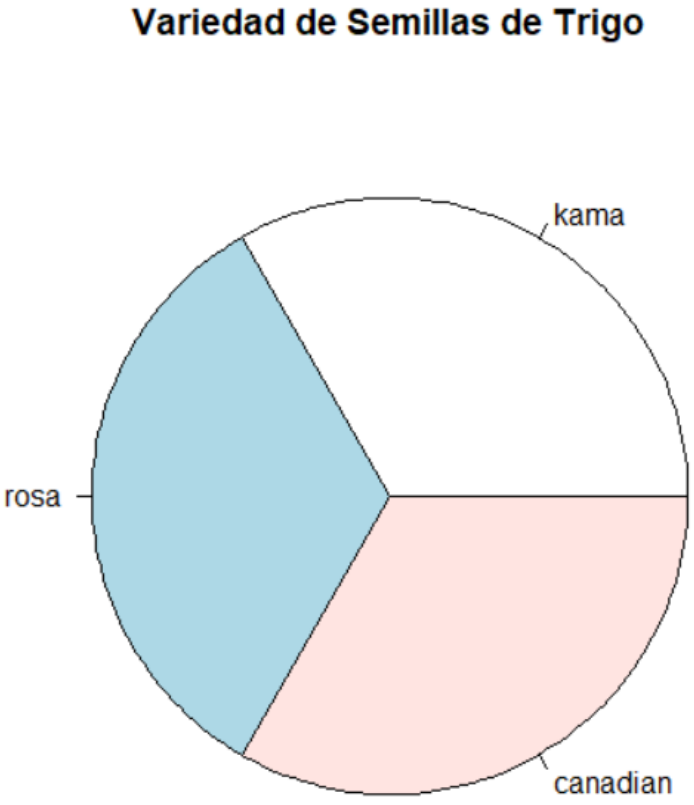
```
Hay 210 semillas.
```

summary(base$VariedadDeSemilla)

```
Hay 70 semillas de cada variedad.
70 Kama
70 Rosa
70 Canadian
```

2) Realice un gráfico de sectores de la variable a predecir VariedadDeSemilla. Elija un Título.

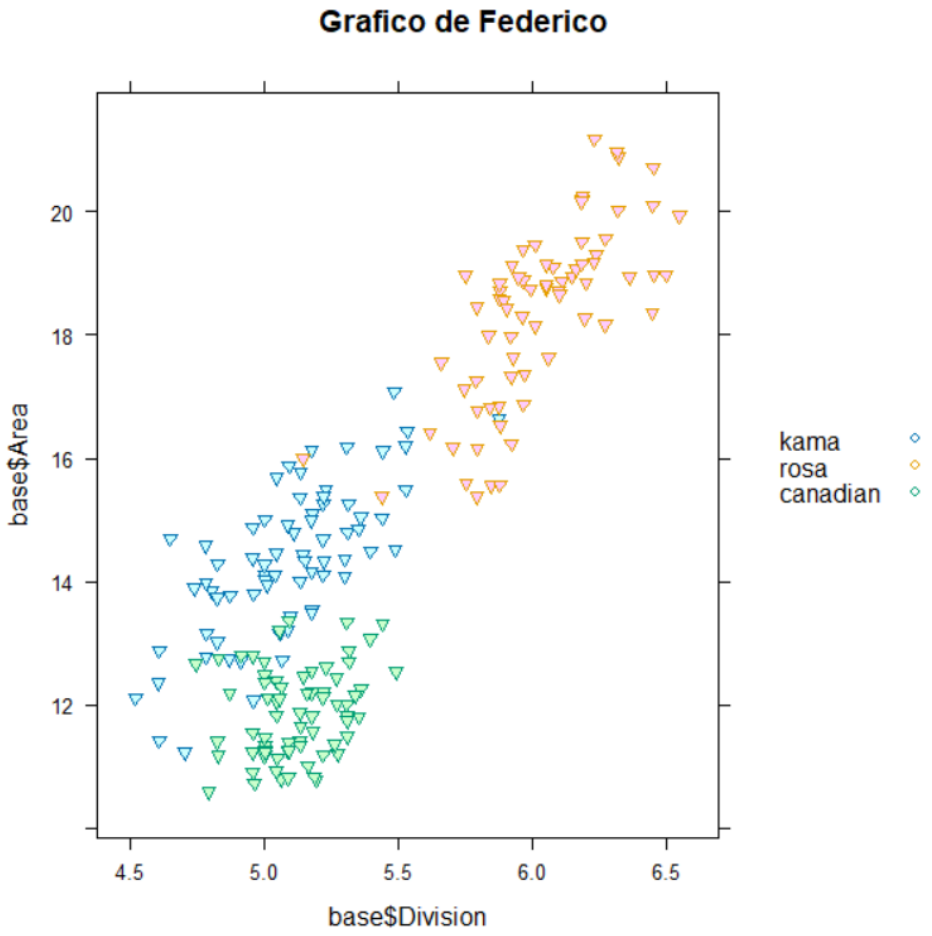
pie(table(base$VariedadDeSemilla),main="Titulo")



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)
```

a) Para el título ingrese su nombre, como “Gráfico de Marcela”.



b) Indique el código R utilizado.

```
xyplot(base$Area~base$Division,groups=VariedadDe
Semilla,base,auto.key=TRUE,main="Grafico de
Federico",pch=25)
```

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?

```
El numero 42 es de tipo Kama.
```

	Area	Perimetro	Compactitud	Largo	Ancho	Asimetria	Division	VariedadDeSemilla
42	13.5	13.85	0.8852	5.351	3.158	2.249	5.176	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.

```
set.seed(27691042);particion=createDataPartition(y=base$VariedadDeSemilla,p=0.7,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)

	Area	Perimetro	Compactitud	Largo	Ancho	Asimetria	Division	VariedadDeSemilla
1	15.26	14.84	0.8710	5.763	3.312	2.221	5.220	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
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

summary(entreno)

Area	Perimetro	Compactitud	Largo	Ancho	Asimetria	Division
Min. :10.59	Min. :12.41	Min. :0.8082	Min. :4.899	Min. :2.630	Min. :0.7651	Min. :4.605
1st Qu.:12.46	1st Qu.:13.46	1st Qu.:0.8579	1st Qu.:5.279	1st Qu.:2.967	1st Qu.:2.6640	1st Qu.:5.045
Median :14.34	Median :14.28	Median :0.8726	Median :5.516	Median :3.231	Median :3.6380	Median :5.222
Mean :14.84	Mean :14.56	Mean :0.8712	Mean :5.627	Mean :3.258	Mean :3.6838	Mean :5.398
3rd Qu.:17.10	3rd Qu.:15.65	3rd Qu.:0.8877	3rd Qu.:5.989	3rd Qu.:3.557	3rd Qu.:4.6065	3rd Qu.:5.818
Max. :21.18	Max. :17.25	Max. :0.9153	Max. :6.581	Max. :4.033	Max. :8.4560	Max. :6.498

VariedadDeSemilla
kama :49
rosa :49
canadian:49

head(testeo)

	Area	Perimetro	Compactitud	Largo	Ancho	Asimetria	Division	VariedadDeSemilla
2	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	kama
4	13.84	13.94	0.8955	5.324	3.379	2.259	4.805	kama
7	14.69	14.49	0.8799	5.563	3.259	3.586	5.219	kama
12	14.03	14.16	0.8796	5.438	3.201	1.717	5.001	kama
14	13.78	14.06	0.8759	5.479	3.156	3.136	4.872	kama
17	13.99	13.83	0.9183	5.119	3.383	5.234	4.781	kama

summary(testeo)

Area	Perimetro	Compactitud	Largo	Ancho	Asimetria	Division
Min. :10.80	Min. :12.57	Min. :0.8081	Min. :4.902	Min. :2.668	Min. :1.018	Min. :4.519
1st Qu.:11.94	1st Qu.:13.37	1st Qu.:0.8539	1st Qu.:5.208	1st Qu.:2.845	1st Qu.:2.311	1st Qu.:5.063
Median :14.52	Median :14.49	Median :0.8763	Median :5.563	Median :3.259	Median :3.533	Median :5.263
Mean :14.87	Mean :14.57	Mean :0.8706	Mean :5.633	Mean :3.261	Mean :3.739	Mean :5.432
3rd Qu.:17.98	3rd Qu.:15.86	3rd Qu.:0.8874	3rd Qu.:5.979	3rd Qu.:3.612	3rd Qu.:4.965	3rd Qu.:5.920
Max. :20.88	Max. :17.23	Max. :0.9183	Max. :6.675	Max. :4.032	Max. :7.524	Max. :6.550

VariedadDeSemilla
kama :21
rosa :21
canadian:21

3) Realice un :

summary(base$VariedadDeSemilla)
summary(entreno$VariedadDeSemilla)
summary(testeo$VariedadDeSemilla)

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

```
Quedaron 49 registros de cada tipo de
semilla en el conjunto de entrenamiento y 21
registros de cada tipo de semilla en el
conjunto de testeo.
```

Parte D – Árbol de Decisión

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

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

Escriba arbol<enter> y muestre una captura de pantalla de la
información que aparece.

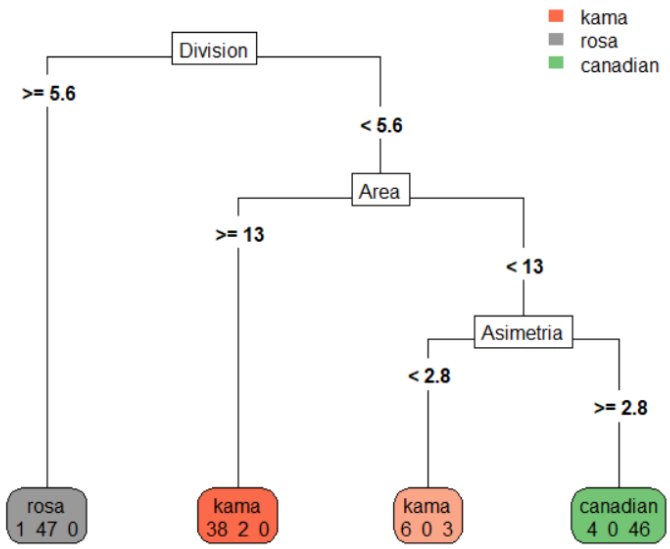
```
n= 147

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 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>=13.41 40  2 kama (0.95000000 0.05000000 0.00000000) *
     7) Area< 13.41 59 10 canadian (0.16949153 0.00000000 0.83050847)
      14) Asimetria< 2.764 9  3 kama (0.66666667 0.00000000 0.33333333) *
      15) Asimetria>=2.764 50  4 canadian (0.08000000 0.00000000 0.92000000) *
```

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



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

```
El árbol tiene 4 hojas.
```

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”).

```
Una semilla será considerada variedad
“rosa” cuando el valor de Division sea
mayor o igual a 5,6
```

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       20    0        1
  rosa        0   21        0
  canadian    1    0       20

Overall Statistics

               Accuracy : 0.9683
                 95% CI : (0.89, 0.9961)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9524

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: kama Class: rosa Class: canadian
Sensitivity               0.9524      1.0000           0.9524
Specificity               0.9762      1.0000           0.9762
Pos Pred Value            0.9524      1.0000           0.9524
Neg Pred Value            0.9762      1.0000           0.9762
Prevalence                0.3333      0.3333           0.3333
```

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

```
20+21+20=61
61/63= 0.968254
Accuracy : 0.9683
```

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

```
La menor sensibilidad la presentan las
clases Kama y Canadian, ya que ambas
tienen un valor de 0,9524
```

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")

¿Coincide la predicción con lo esperado?

```
Si, coinciden la predicción y lo esperado.
```

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.

```
set.seed(27691042);red=nnet(VariedadDeSemilla~.,entreno,size=25,maxit=10000)
```

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

```
# weights:  278
initial  value 191.771716
iter  10 value 103.449290
iter  20 value 50.910280
iter  30 value 20.980311
iter  40 value 8.383222
iter  50 value 4.678425
iter  60 value 0.059865
iter  70 value 0.023864
iter  80 value 0.005441
iter  90 value 0.002335
iter 100 value 0.000475
iter 110 value 0.000345
iter 120 value 0.000242
iter 130 value 0.000216
final  value 0.000081
converged
```

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

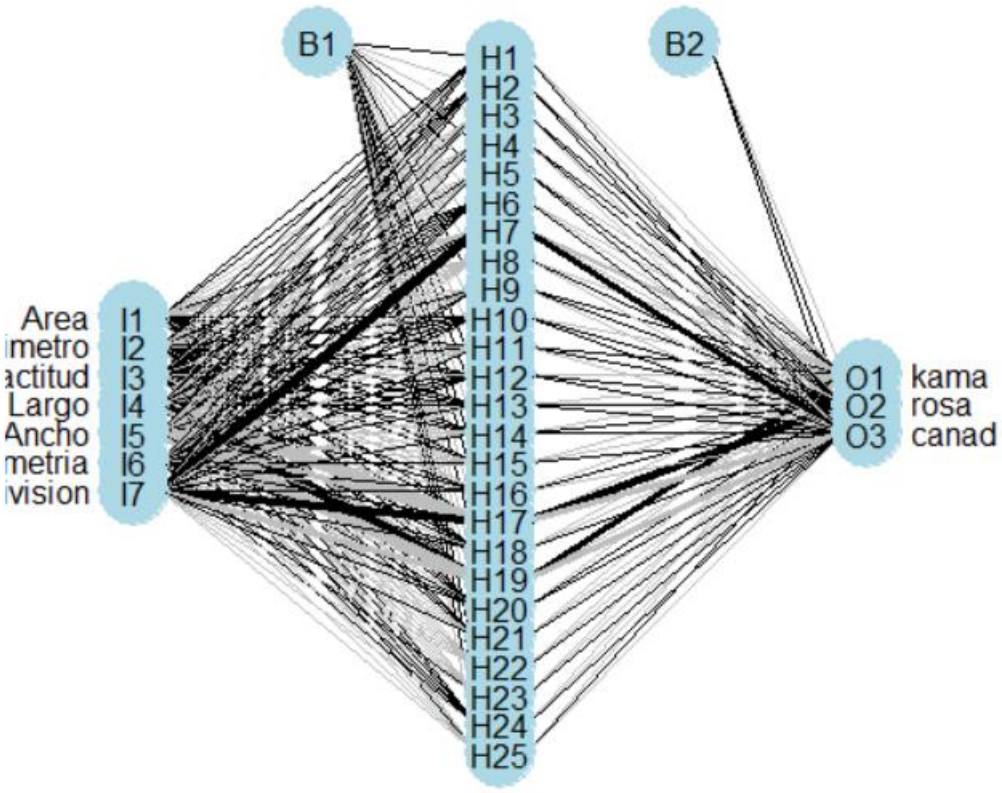
```
a 7-25-3 network with 278 weights
inputs: Area Perimetro Compactitud Largo Ancho Asimetria Division
output(s): VariedadDeSemilla
options were - softmax modelling
```

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

```
pesos:  278
iteracciones: 130
```

5) Dibuje la Red Neuronal

```
library(NeuralNetTools)
plotnet(red)
```



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

```
          Reference
Prediction kama rosa canadian
  kama       17    0        0
  rosa        0   21        0
  canadian    4    0       21

Overall Statistics

               Accuracy : 0.9365
                 95% CI : (0.8453, 0.9824)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9048

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: kama Class: rosa Class: canadian
Sensitivity               0.8095      1.0000          1.0000
Specificity               1.0000      1.0000          0.9048
Pos Pred Value            1.0000      1.0000          0.8400
Neg Pred Value            0.9130      1.0000          1.0000
Prevalence                0.3333      0.3333          0.3333
Detection Rate            0.2698      0.3333          0.3333
Detection Prevalence      0.2698      0.3333          0.3968
Balanced Accuracy         0.9048      1.0000          0.9524
```

7) ¿Cuál fue el accuracy?

```
El accuracy es de 0,9365
```

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")

¿Coincide la predicción con la variedad esperada?

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

	AdD	RN
	Sensibilidad	Sensibilidad
Kama	0,9524	0,8095
Rosa	1,0000	1,0000
Canadian	0,9524	1,0000
Accuracy	0,9683	0,9365

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

En este caso resulto mejor el Árbol de Decisión ya que es mayor el accuracy, es decir que tuvo más aciertos sobre el total.

Anexo: Código R utilizado

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

head(base)

names(base)[names(base)=="V1"]="Area"

names(base)[names(base)=="V2"]="Perimetro"
```

```
names(base)[names(base)=="V3"]="Compacti
tud"

names(base)[names(base)=="V4"]="Largo"

names(base)[names(base)=="V5"]="Ancho"

names(base)[names(base)=="V6"]="Asimetri
a"

names(base)[names(base)=="V7"]="Division
"

names(base)[names(base)=="V8"]="Variedad
DeSemilla"

head(base)

base$VariedadDeSemilla=factor(base$Varie
dadDeSemilla,levels=c(1,2,3),labels=c("k
ama","rosa","canadian"))

head(base)

dim(base)

summary(base$VariedadDeSemilla)

pie(table(base$VariedadDeSemilla),main="
Variedad de Semillas de Trigo")

library(caret)
```

```
xyplot(base$Area~base$Division,groups=Va
riedadDeSemilla,base,auto.key=TRUE,main=
"Grafico de Federico",pch=25)

trigo=base[42,]

trigo

set.seed(27691042);particion=createDataP
artition(y=base$VariedadDeSemilla,p=0.7,
list=FALSE)

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,
method="class")

árbol

library(rpart.plot)

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

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

confusionMatrix(pred,testeo$VariedadDeSe
milla)

library(caret)

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

confusionMatrix(pred,testeo$VariedadDeSe
milla)

dim(testeo)

20+21+20

61/63

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


library(nnet)
```

```
set.seed(27691042);red=nnet(VariedadDeSe
milla~.,entreno,size=25,maxit=10000)

red

library(NeuralNetTools)

plotnet(red)

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

confusionMatrix(factor(pred2),testeo$Var
iedadDeSemilla)

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