

# Métodos multivariados de Análisis de Datos

## Análisis de nutrientes en pizzas

Ricardo Cruz Sánchez  
Rolando Corona Jiménez

CIMAT

June 6, 2019

- 1 Introducción
- 2 Análisis exploratorio.
- 3 Modelos de reducción de dimensiones.
- 4 Análisis por agrupación.
- 5 Modelos de clasificación.
- 6 Regresión Multivariada

# Bromatología

Ciencia encargada del análisis de los nutrientes contenidos en los alimentos. Los estudios relacionados con esta disciplina cobran importancia al considerar que existen cantidades recomendadas en la ingesta diaria de cualquier individuo y el incremento o decremento de las cantidades repercute directamente en la salud del la persona

- Particularmente, la pizza, tiende a ser uno de los alimentos con los cuales se sobrepasan los límites de nutrientes recomendados
- En el presente trabajo, se considera una base de datos relativa a las pruebas nutrimentales realizadas en distintas pizzas y con base a estos datos se pretende realizar un análisis multivariado.

# Repositorio

- <https://github.com/rolandocj/proyecto-pizzas/tree/develop>

- 1 Introducción
- 2 Análisis exploratorio.
- 3 Modelos de reducción de dimensiones.
- 4 Análisis por agrupación.
- 5 Modelos de clasificación.
- 6 Regresión Multivariada

- **Ident:** Variable tipo numérica, la cual corresponde a un identificador para cada pizza.
- **HUMED:** Variable tipo numérica que indica el porcentaje de humedad contenido en la pizza.
- **PROTE:** Variable tipo numérica que indica la cantidad de gramos de proteína contenida en 100g de pizza.
- **GRASA:** Variable tipo numérica que indica la cantidad de gramos de grasa contenida en 100g de pizza.
- **CENIZA:** Variable tipo numérica que indica la cantidad de gramos de ceniza contenida en 100g de pizza.
- **SODIO:** Variable tipo numérica que indica la cantidad de gramos de sodio contenida en 100g de pizza.
- **CARBO:** Variable tipo numérica que indica la cantidad de gramos de carbohidratos contenida en 100g de pizza.
- **CALOR:** Variable tipo numérica que indica la cantidad de calorías contenida en 100g de pizza.

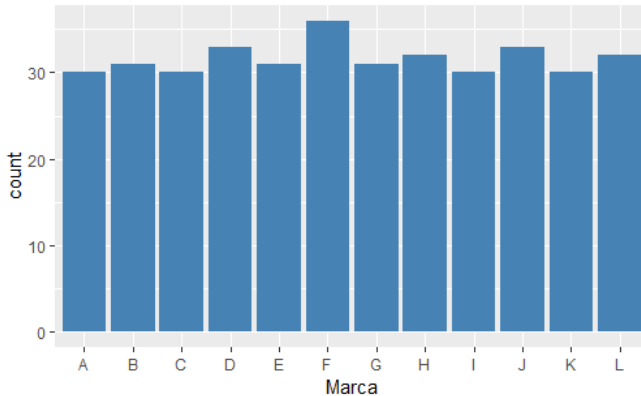


Figure: Conteo de registros por marca



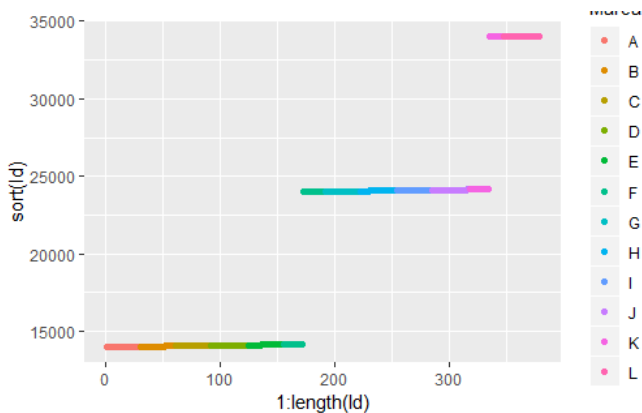


Figure: Comportamiento de la variable Ident de acuerdo a la marca

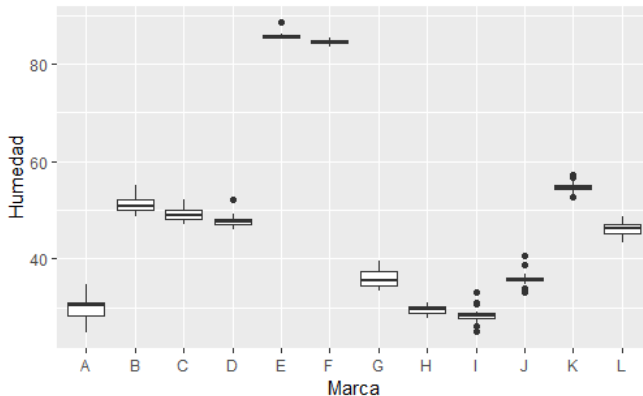


Figure: Comportamiento de la variable Humed de acuerdo a la marca

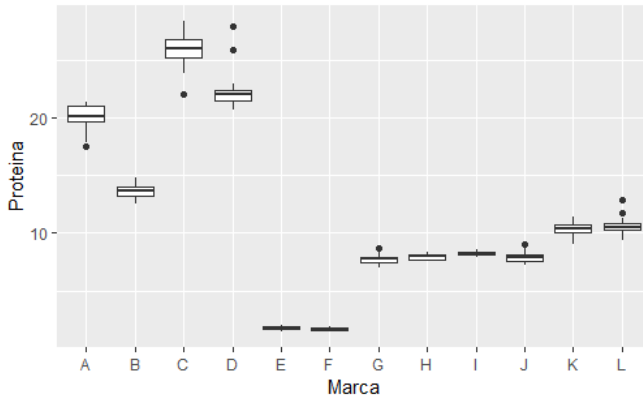


Figure: Comportamiento de la variable Prote de acuerdo a la marca

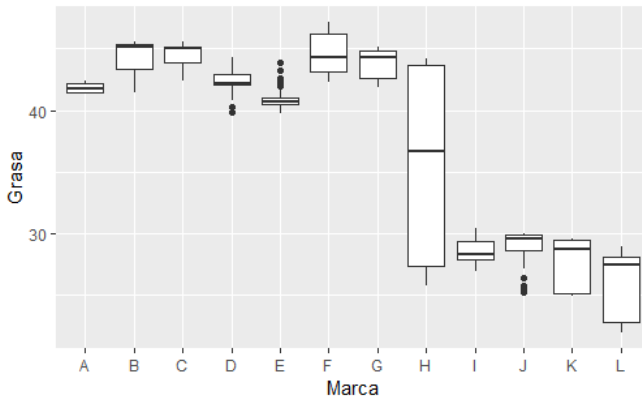


Figure: Comportamiento de la variable Grasa de acuerdo a la marca

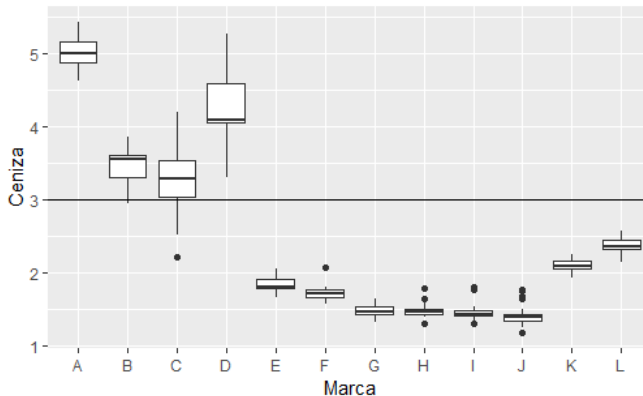


Figure: Comportamiento de la variable CENIZA de acuerdo a la marca

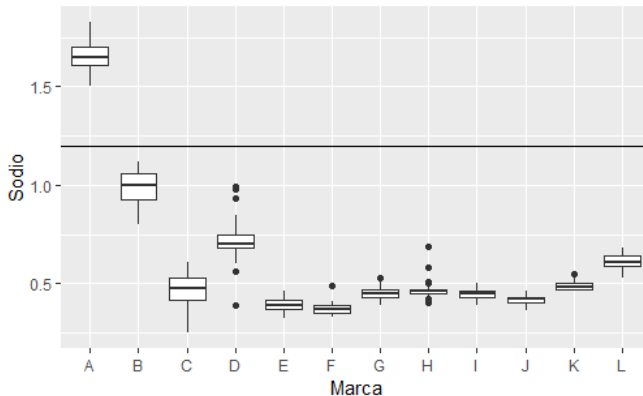


Figure: Comportamiento de la variable sodio de acuerdo a la marca

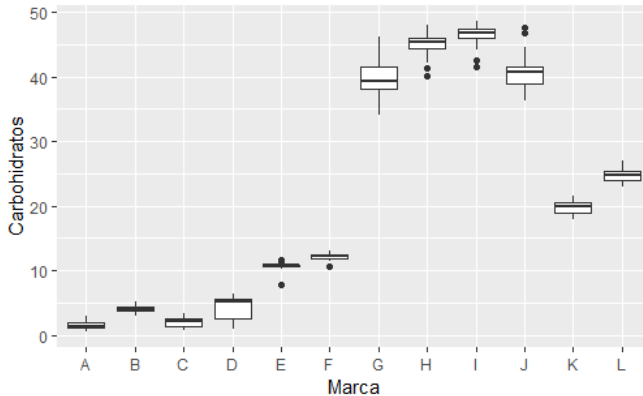


Figure: Comportamiento de la variable Carbo de acuerdo a la marca

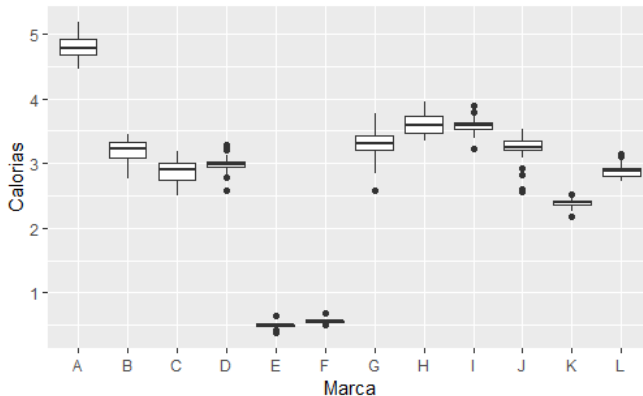


Figure: Comportamiento de la variable Calor de acuerdo a la marca



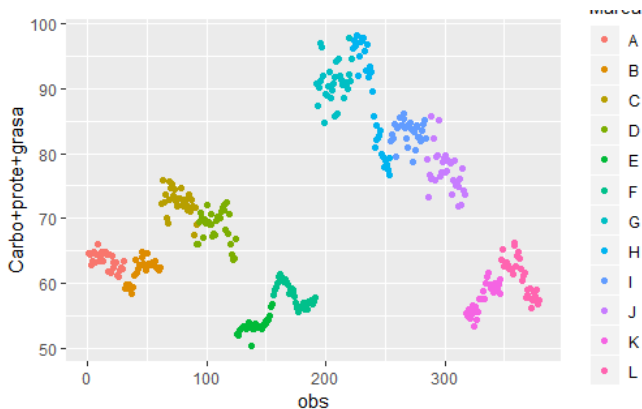


Figure: Grasa+Proteína+Carbohidratos para cada observación

Nutriente	Error promedio
Carbohidratos	1.37
Grasas	0.68
Proteína	0.04

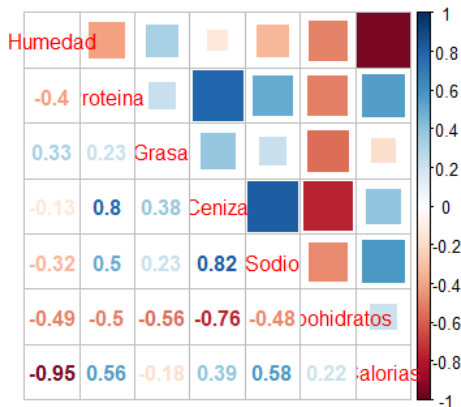


Figure: Correlación de las variables.

- La relación entre humedad y calorías, tal vez corresponda a la presencia del fosfato, pues la presencia de fosfato modifica el valor de estas dos variables y se encuentra en alimentos como el queso y carnes.
- La alta correlación entre Proteínas y cenizas puede ser explicado por la presencia de queso, pues es un alimento el cual es fuente de proteínas y minerales.
- En el caso de la alta correlación entre sodio y cenizas, puede ser explicado por la presencia de tomate, ya que, este alimento eleva el valor de ambas variables.

- 1 Introducción
- 2 Análisis exploratorio.
- 3 Modelos de reducción de dimensiones.
  - Análisis de componentes principales (PCA)
  - Análisis Factorial
- 4 Análisis por agrupación.
- 5 Modelos de clasificación.
- 6 Regresión Multivariada

## Cargas de componentes principales

	PC1	PC2
Humedad	0.21	0.58
Proteína	-0.47	-0.03
Grasa	-0.19	0.41
Ceniza	-0.51	0.15
Sodio	-0.47	-0.02
Carbohidratos	0.32	-0.49
Calorías	-0.34	-0.48
Varianza acumulada	48.64 %	83.59 %

**Table:** Pesos asociados a las primeras dos componentes principales.

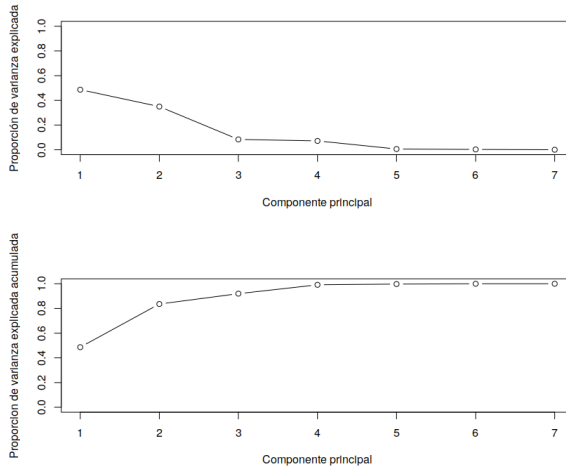


Figure: Varianza explicada por las componentes principales

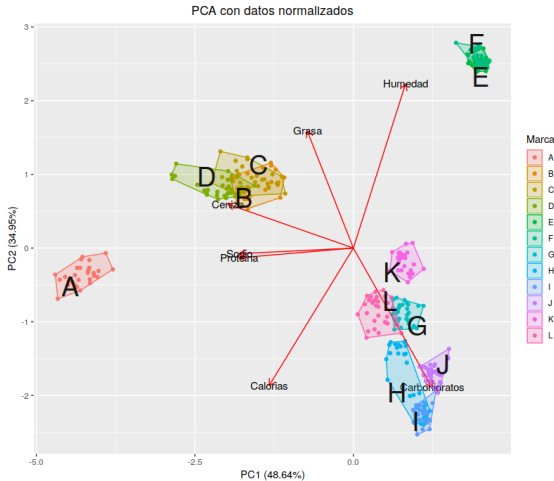


Figure: Biplot PCA



# Resumen de Análisis Factorial

	Factor1	Factor2	Varianza específica	Comunalidades
Humedad	0.06	-1.00	0.01	1.00
Proteína	0.76	0.44	0.23	0.77
Grasa	0.47	-0.30	0.69	0.31
Ceniza	0.94	0.19	0.08	0.92
Sodio	0.73	0.39	0.32	0.68
Carbohidratos	-0.90	0.44	0.01	1.00
Calorías	0.23	0.97	0.01	0.99
Prop. de Varianza	44 %	36.9 %		
Var. acumulada	44 %	80.9 %		

Table: Resultados del análisis factorial.

# Aproximación a R

	Humedad	Proteína	Grasa	Ceniza	Sodio	Carbohidratos	Calorias
Humedad	-0.00	-0.02	0.00	-0.00	0.01	-0.00	-0.00
Proteína	-0.02	0.00	-0.01	-0.01	-0.22	-0.01	-0.04
Grasa	0.00	-0.01	0.00	-0.01	-0.00	-0.00	0.00
Ceniza	-0.00	-0.01	-0.01	-0.00	0.06	0.00	-0.00
Sodio	0.01	-0.22	-0.00	0.06	-0.00	0.01	0.04
Carbohidratos	-0.00	-0.01	-0.00	0.00	0.01	-0.00	-0.00
Calorias	-0.00	-0.04	0.00	-0.00	0.04	-0.00	0.00

Table: Diferencia entre R y  $LL' + \Psi$ , con redondeo a tres dígitos.

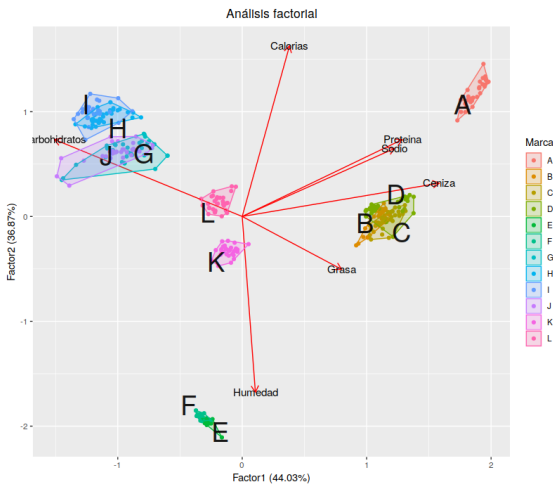


Figure: Biplot usando 2 factores

- 1 Introducción
- 2 Análisis exploratorio.
- 3 Modelos de reducción de dimensiones.
- 4 **Análisis por agrupación.**
  - Elección del número de clusters
  - Asignación de clusters
- 5 Modelos de clasificación.
- 6 Regresión Multivariada

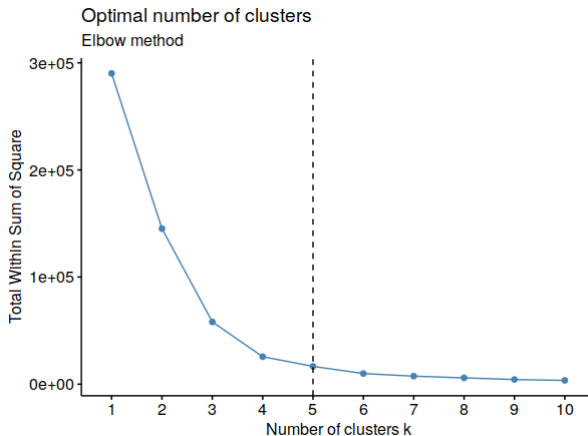


Figure: Suma total de cuadrados entre clústers vs número de clusters

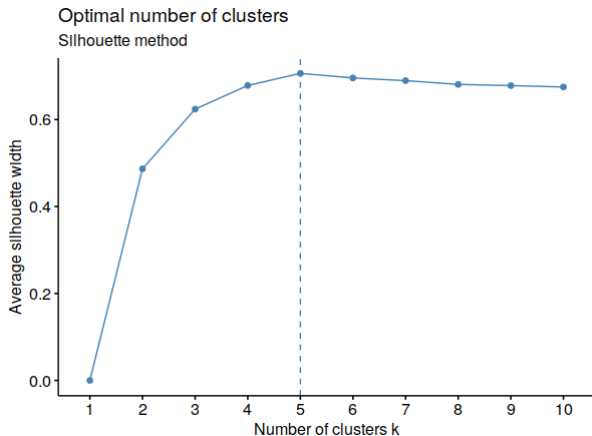


Figure: Silhouette promedio vs número de clusters

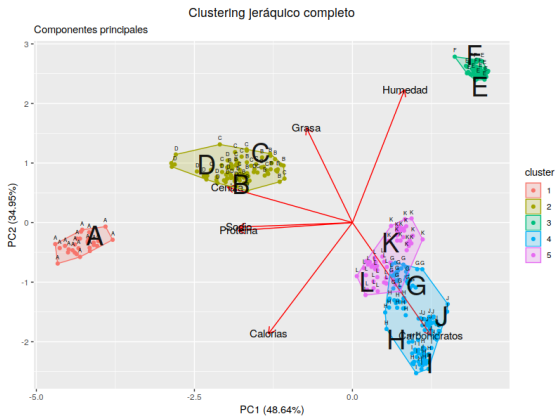


Figure: Clustering jerárquico completo (Representación: PCA)

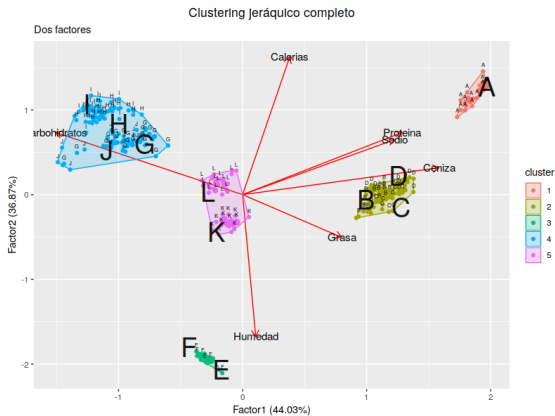


Figure: Clustering jerárquico completo (Representación: Factores)



- 1 Introducción
- 2 Análisis exploratorio.
- 3 Modelos de reducción de dimensiones.
- 4 Análisis por agrupación.
- 5 Modelos de clasificación.
  - LDA
  - Multinomial
  - Validación cruzada con caret
- 6 Regresión Multivariada

	A	B	C	D	E	F	G	H	I	J	K	L
A	9	0	0	0	0	0	0	0	0	0	0	0
B	0	17	0	0	0	0	0	0	0	0	0	0
C	0	0	15	0	0	0	0	0	0	0	0	0
D	0	0	1	19	0	0	0	0	0	0	0	0
E	0	0	0	0	14	2	0	0	0	0	0	0
F	0	0	0	0	1	16	0	0	0	0	0	0
G	0	0	0	0	0	0	12	0	0	0	0	0
H	0	0	0	0	0	0	0	8	9	0	0	0
I	0	0	0	0	0	0	0	2	14	1	0	0
J	0	0	0	0	0	0	0	0	0	13	0	0
K	0	0	0	0	0	0	0	0	0	0	18	0
L	0	0	0	0	0	0	0	0	0	0	0	18

Table: Matriz de confusión LDA train

error de clasificación: 0.08465608

	A	B	C	D	E	F	G	H	I	J	K	L
A	21	0	0	0	0	0	0	0	0	0	0	0
B	0	14	0	0	0	0	0	0	0	0	0	0
C	0	0	15	0	0	0	0	0	0	0	0	0
D	0	0	1	12	0	0	0	0	0	0	0	0
E	0	0	0	0	14	1	0	0	0	0	0	0
F	0	0	0	0	0	19	0	0	0	0	0	0
G	0	0	0	0	0	0	19	0	0	0	0	0
H	0	0	0	0	0	0	0	11	4	0	0	0
I	0	0	0	0	0	0	0	0	13	0	0	0
J	0	0	0	0	0	0	0	0	0	20	0	0
K	0	0	0	0	0	0	0	0	0	0	12	0
L	0	0	0	0	0	0	0	0	0	0	0	14

Table: Matriz de confusión LDA test

error de clasificación: 0.03157895

	A	B	C	D	E	F	G	H	I	J	K	L
A	9	0	0	0	0	0	0	0	0	0	0	0
B	0	17	0	0	0	0	0	0	0	0	0	0
C	0	0	15	0	0	0	0	0	0	0	0	0
D	0	0	0	20	0	0	0	0	0	0	0	0
E	0	0	0	0	16	0	0	0	0	0	0	0
F	0	0	0	0	0	17	0	0	0	0	0	0
G	0	0	0	0	0	0	12	0	0	0	0	0
H	0	0	0	0	0	0	0	17	0	0	0	0
I	0	0	0	0	0	0	0	0	17	0	0	0
J	0	0	0	0	0	0	0	0	0	13	0	0
K	0	0	0	0	0	0	0	0	0	0	18	0
L	0	0	0	0	0	0	0	0	0	0	0	18

**Table:** Matriz de confusión Multinomial train

error de clasificación: 0

	A	B	C	D	E	F	G	H	I	J	K	L
A	21	0	0	0	0	0	0	0	0	0	0	0
B	0	14	0	0	0	0	0	0	0	0	0	0
C	0	0	15	0	0	0	0	0	0	0	0	0
D	0	0	1	12	0	0	0	0	0	0	0	0
E	0	0	0	0	15	0	0	0	0	0	0	0
F	0	0	0	0	1	18	0	0	0	0	0	0
G	0	0	0	0	0	0	19	0	0	0	0	0
H	0	0	0	0	0	0	0	9	6	0	0	0
I	0	0	0	0	0	0	0	0	13	0	0	0
J	0	4	0	0	0	2	0	0	0	14	0	0
K	0	0	0	0	0	0	0	0	0	0	12	0
L	0	0	0	0	0	0	0	0	0	0	0	14

**Table:** Matriz de confusión Multinomial test

error de clasificación: 0.07368421

Clasificador	Accuracy
rn	0.2953684
mr	0.9626203
lda	0.9498792
rf	0.9709806
tree	0.8468115

**Table:** Accuracy usando validación cruzada

- 1 Introducción
- 2 Análisis exploratorio.
- 3 Modelos de reducción de dimensiones.
- 4 Análisis por agrupación.
- 5 Modelos de clasificación.
- 6 **Regresión Multivariada**
  - MANOVA
  - Regresión

$$H_0 : \mu_1 = \dots = \mu_{12}$$



- ANOVA
- MANOVA
- MANOVA  $2^{k-1}$

*J* y *G*. Humedad, Proteína, Calorías y Carbohidratos por separado y aun nivel de significancia del .05.

Se consideraron como variables regresoras a aquellas que son nutrientes, a saber, Proteínas, Grasas, Sodio y Carbohidratos. Mientras que las variables dependientes serán Humedad, Cenizas y Calorías.

- Para las 3 variables dependientes, se encontró que la variables Grasa no es significativa para el modelo
- $R^2$  de 0.96, 0.94 y 0.93