Trabajo de fin de curso

Vásquez Velasco, Christian Richard Alberto

18/7/2021

Introducción

Descripción del caso

Se emplearon datos de la campaña de cosecha del cultivo de arándano que se viene realizando en la empresa Agrícola Cerro Prieto, que tiene un área sembrada de cultivo de arándano de 950 hectáreas y 399 hectáreas productoras de arándano hasta la semana 27 de cosecha. Se tienen registros completos del rendimiento obtenido solo en la primera semana de cosecha (semana 27) y se pide crear un modelo predictivo con los datos registrados en la semana 26 y la cosecha que se obtuvo en la semana 27 para ser empleada en las siguientes semanas.

En esta empresa, el cultivo de arándano está distribuido en Sectores, Módulos, Turnos y Válvulas. Cada unidad agrícola descrita anteriormente, engloba la que sigue. En ese sentido, una válvula es la unidad de medida más pequeña y la que se emplea como unidad de observación del presente caso.

Objetivo general: Emplear los conocimientos obtenidos en el curso de Estadística Aplicada a la Agroforestería I sobre la base de datos "Inventario Arandano".

Objetivos específicos:

- Realizar un análisis exploratorio sobre el "Inventario Arándano".
- Evaluar los indicadores estadísticos descriptivos sobre algunas variables del "Inventario Arándano".
- Comparar los Índices de Biodiversidad para las variedades del "Inventario Arándano".
- Generar un muestreo adecuado para el "Inventario Arándano".
- Aplicar un análisis de correlación y regresión sobre el rendimiento por hectárea obtenido en el cultivo de arándano.
- Ejecutar pruebas de hipótesis sobre algunas variables del "Inventario Arándano".
- Realizar un análisis de clasificación no supervisada de las observaciones del "Inventario Arándano".
- Aplciar un análisis de clasificación supervisada de las observaciones del "Inventario Arándano".

Variables del caso

Mediante evaluaciones semanañes se obtienen datos de las siguientes variables:

- Variedad: Variedad de arándano sembrada en una válvula.

- Kg_Total_Ha: Rendimiento de frutos (bayas cosechadas) en una semana obtenido en una válvula entre el Área de la válvula. Esta variable puede estar mal obtenida, debido a que existen válvulas no cosechadas completamente en una semana o que se han cosechado bayas que no debieron cosecharse. Es decir, existirán observaciones con sub registros y otras con sobreregistro.
- Área: Número de hectáreas en una válvula.
- Plantas_Ha: Número de plantas totales por válvula entre el área de la válvula.
- Plantas_Productivas_Ha: Número de plantas productivas (que tienen producción de arándano de forma normal y no tienen daños graves por plagas o enfermedades) por válvula entre el área de la válvula.
- Inicio_Fase_1: Número de frutos promedio por planta en estado fenológico llamado "Inicio crema en Fase 1", que se tratan de aquellos frutos que entran en fase de "envero" o "pintado", lo que significa que estos frutos empiezan a tomar el color morado característico del cultivo de arándano. Predomina el color verde frente al color rosado.
- Inicio_Fase_2: Número de frutos promedio por planta en estado fenológico llamado "Inicio crema en Fase 2", que se caracteriza por tener una mayor coloración del fruto, a comparación del anterior, pero con menores tonalidades verdosas.
- Cremas: Número de frutos promedio por planta en estado fenológico llamado "Crema", que se caracteriza por tener una mayor coloración del rosada a ligeramente morada que se presenta en toda la cobertura del fruto.
- Maduras: Número de frutos promedio por planta en estado fenológico llamado "Madura", que se caracteriza por tener una coloración completamente morada, pero, que no puede ser consumida aún debido a que no tiene el dulzor necesario, que se mide en grados Brix.
- Cosechable: Número de frutos promedio por planta en estado fenológico llamado "Cosechable", que se caracteriza por tener una coloración completamente morada y que ya puede ser consumida por el hombre.
- Prop_Plantas_Productivas: Es la razón del número de plantas productivas entre el número de plantas totales por hectárea.
- Peso_Baya: Es el Peso promedio de bayas en kg.
- Prop_Cosechable: Es la proporción de frutos "Cosechables" que pueden ser cosechados en la siguiente semana.
- Prop_Maduras: Es la proporción de frutos "Maduros" que pueden ser cosechados en la siguiente semana.
- Prop_Cremas: Es la proporción de frutos "Cremas" que pueden ser cosechados en la siguiente semana.
- Prop_Inicio_Fase_2: Es la proporción de frutos en "Inicio Crema fase 2" que pueden ser cosechados en la siguiente semana.
- Prop_Inicio_Fase_1: Es la proporción de frutos en "Inicio Crema fase 1" que pueden ser cosechados en la siguiente semana.
- Kg_Inicio_Fase_1_Ha_Proy: Estimado del Peso de frutos en fase de "Inicio Crema Fase 1" que serían cosechados por hectárea en una válvula en la siguiente semana.
- Kg_Inicio_Fase_2_Ha_Proy: Estimado del Peso de frutos en fase de "Inicio Crema Fase 2" que serían cosechados por hectárea en una válvula en la siguiente semana.
- Kg_Cremas_Ha_Proy: Estimado del Peso de frutos en fase "Crema" que serían cosechados por hectárea en una válvula en la siguiente semana.

- Kg_Maduras_Ha_Proy: Estimado del Peso de frutos en fase “Madura” que serían cosechados por hectárea en una válvula en la siguiente semana.
- Kg_Cosechable_Ha_Proy: Estimado del Peso de frutos en fase “Cosechable” que serían cosechados por hectárea en una válvula en la siguiente semana.
- Cumplimiento: Es la razón del estimado de cosecha total (suma de los kg cosechados en todas las fases en una hectárea) entre los kg cosechados reales por hectárea.
- Dif_abs: Es la diferencia absoluta entre los kg cosechados reales por hectárea y el estimado de cosecha total (suma de los kg cosechados en todas las fases en una hectárea).

Para hacer uso de la información recolectada de campo se tiene que filtrar aquellas válvulas que fueron cosechadas de forma incorrecta, debido a que no alcanzó a cosecharse toda el área por deficit de personal o que se ha cosechado más de lo debido porque le personal ha cosechado aquellas bayas que aún no llegaban a las fases de maduración óptima.

Este procedimiento se realiza calculando cuales son las válvulas con un cumplimiento y diferencia absoluta atípica, eliminándolas de la base de datos, junto a todas las válvulas no cosechadas (con 0 kg cosechados).

Luego de ese procedimiento, se obtiene una base de datos de 63 observaciones.

A esta base de datos se le eliminará todas aquellas variables que no tienen variancia (es decir, tienen valores únicos). Además, se convierte la variable Peso de Baya a unidades en gramos (en vez de kilogramos).

	vars	n	mean	sd	median	trimmed	mad
Variedad*	1	63	4.29	2.55	2.00	4.14	1.48
Kg_Total_Ha	2	63	36.88	37.73	23.32	30.35	22.55
Area	3	63	2.77	0.71	2.94	2.88	0.71
Plantas_Ha	4	63	6185.10	1462.27	5462.00	6051.49	97.85
Plantas_Productivas_Ha	5	63	4928.86	1065.75	4991.00	4932.12	1243.90
Inicio_Fase_1	6	63	13.91	10.68	11.80	12.72	7.31
Inicio_Fase_2	7	63	5.60	5.94	3.53	4.52	3.46
Cremas	8	63	1.70	2.60	0.73	1.10	0.89
Maduras	9	63	0.59	0.74	0.33	0.46	0.40
Cosechable	10	63	0.39	0.86	0.00	0.20	0.00
Prop_Plantas_Productivas	11	63	0.81	0.12	0.75	0.80	0.07
Peso_Baya	12	63	0.00	0.00	0.00	0.00	0.00
Prop_Cosechable	13	63	1.00	0.00	1.00	1.00	0.00
Prop_Maduras	14	63	1.00	0.00	1.00	1.00	0.00
Prop_Cremas	15	63	0.67	0.22	0.65	0.66	0.30
Prop_Inicio_Fase_2	16	63	0.12	0.22	0.03	0.07	0.05
Prop_Inicio_Fase_1	17	63	0.00	0.00	0.00	0.00	0.00
Kg_Inicio_Fase_1_Ha_Proy	18	63	0.00	0.00	0.00	0.00	0.00
Kg_Inicio_Fase_2_Ha_Proy	19	63	9.20	16.00	0.00	5.60	0.00
Kg_Cremas_Ha_Proy	20	63	14.58	20.19	6.05	10.30	8.17
Kg_Maduras_Ha_Proy	21	63	7.31	8.22	5.34	5.96	6.71
Kg_Cosechable_Ha_Proy	22	63	5.17	9.35	0.00	3.22	0.00
Cumplimiento	23	63	204.58	338.82	117.02	140.53	42.57
Dif_abs	24	63	7.34	5.81	7.16	6.87	7.87

	min	max	range	skew	kurtosis	se
Variedad*	1.00	8.00	7.00	0.33	-1.64	0.32
Kg_Total_Ha	1.43	152.76	151.32	1.40	1.04	4.75
Area	0.19	3.44	3.25	-1.50	2.71	0.09
Plantas_Ha	4196.00	9252.00	5056.00	0.95	-0.74	184.23
Plantas_Productivas_Ha	2937.00	8072.00	5135.00	0.14	-0.28	134.27

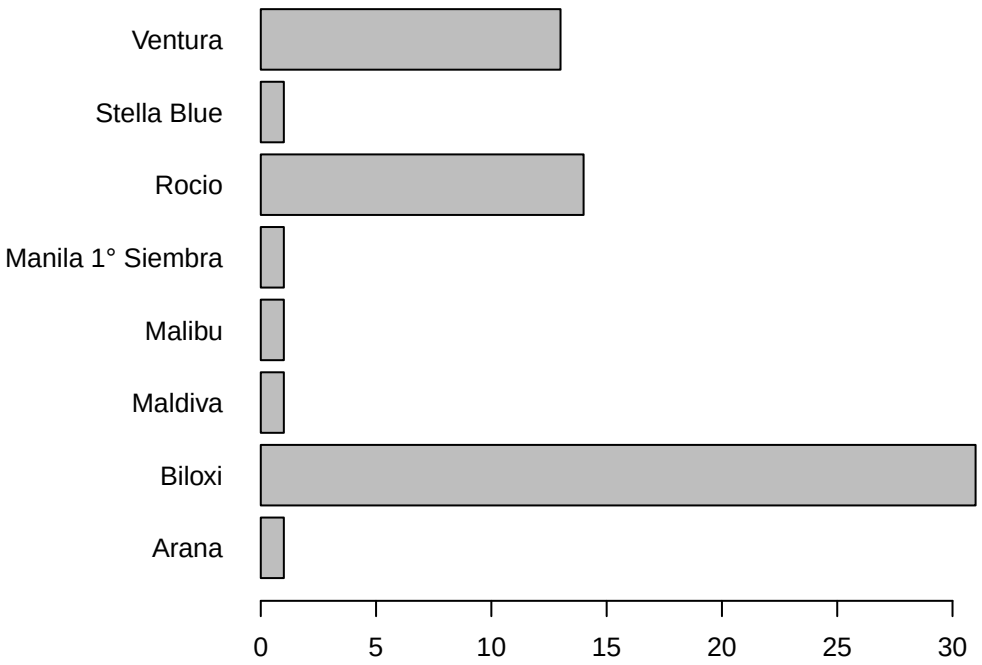
```
Inicio_Fase_1               0.00   52.60   52.60  1.35     2.13   1.35
Inicio_Fase_2               0.00   24.93   24.93  1.65     2.22   0.75
Cremas                      0.00   14.47   14.47  2.84     8.95   0.33
Maduras                     0.00    4.47    4.47  2.59     9.60   0.09
Cosechable                  0.00    5.47    5.47  3.79    17.54   0.11
Prop_Plantas_Productivas    0.62    1.00    0.38  0.40    -1.48   0.01
Peso_Baya                   0.00    0.00    0.00  0.15    -1.77   0.00
Prop_Cosechable             1.00    1.00    0.00   NaN      NaN   0.00
Prop_Maduras                1.00    1.00    0.00   NaN      NaN   0.00
Prop_Cremas                 0.43    1.00    0.57  0.32    -1.51   0.03
Prop_Inicio_Fase_2          0.00    1.00    1.00  2.60     7.04   0.03
Prop_Inicio_Fase_1          0.00    0.00    0.00   NaN      NaN   0.00
Kg_Inicio_Fase_1_Ha_Proy    0.00    0.00    0.00   NaN      NaN   0.00
Kg_Inicio_Fase_2_Ha_Proy    0.00   88.18   88.18  2.53     7.90   2.02
Kg_Cremas_Ha_Proy           0.00   92.58   92.58  1.98     3.58   2.54
Kg_Maduras_Ha_Proy          0.00   43.14   43.14  1.79     4.15   1.04
Kg_Cosechable_Ha_Proy       0.00   53.21   53.21  2.67     9.28   1.18
Cumplimiento               65.12 2574.71 2509.59  5.76    35.93  42.69
Dif_abs                     0.01   20.41   20.39  0.52    -0.85   0.73
```

A esta base de datos resultante le llamaremos “Inventario Arándano”.

```
Informacion del inventario:

'data.frame':   63 obs. of  18 variables:
 $ Variedad                : chr  "Biloxi" "Biloxi" "Biloxi" "Biloxi" ...
 $ Kg_Total_Ha             : num  13.5 12.8 22.2 13.5 12.1 ...
 $ Area                    : num  2.46 2.46 2.46 2.46 2.46 2.46 2.46 2.46 2.48 2.46 ...
 $ Plantas_Ha              : num  5426 5415 5422 5407 5421 ...
 $ Plantas_Productivas_Ha  : num  4298 4963 4481 4580 5024 ...
 $ Inicio_Fase_1           : num  24.1 14.4 12.3 13.8 6.4 ...
 $ Inicio_Fase_2           : num  10.533 0.867 0.667 0.8 2.333 ...
 $ Cremas                  : num  2 0.733 0.733 0.733 1.467 ...
 $ Maduras                 : num  0 0.6 1.2 0.467 0.2 ...
 $ Cosechable              : num  0 0 0 0 0.4 ...
 $ Prop_Plantas_Productivas: num  0.792 0.917 0.827 0.847 0.927 ...
 $ Prop_Cremas             : num  0.45 0.45 0.45 0.45 0.45 0.45 0.45 0.45 0.45 0.45 ...
 $ Prop_Inicio_Fase_2      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Kg_Inicio_Fase_2_Ha_Proy: num  0 0 0 0 0 0 0 0 0 0 ...
 $ Kg_Cremas_Ha_Proy       : num  7.06 2.99 2.7 2.76 6.05 ...
 $ Kg_Maduras_Ha_Proy      : num  0 5.43 9.81 3.9 1.83 ...
 $ Kg_Cosechable_Ha_Proy   : num  0 0 0 0 3.67 ...
 $ Peso_Baya1000           : num  1.82 1.82 1.82 1.82 1.82 ...
  ..- attr(*, "label")= chr "Peso Baya en gramos"
```

```
Primeros registros

  Variedad Kg_Total_Ha Area Plantas_Ha Plantas_Productivas_Ha Inicio_Fase_1
1   Biloxi    13.54191 2.46       5426                  4298      24.06667
2   Biloxi    12.82452 2.46       5415                  4963      14.40000
3   Biloxi    22.15064 2.46       5422                  4481      12.33333
4   Biloxi    13.54191 2.46       5407                  4580      13.80000
5   Biloxi    12.10701 2.46       5421                  5024       6.40000
```

```
6   Biloxi    12.82440 2.46       5412                    5155       8.40000
  Inicio_Fase_2    Cremas   Maduras Cosechable Prop_Plantas_Productivas
1    10.5333333 2.0000000 0.0000000 0.00000000               0.7920443
2     0.8666667 0.7333333 0.6000000 0.00000000               0.9165979
3     0.6666667 0.7333333 1.2000000 0.00000000               0.8265727
4     0.8000000 0.7333333 0.4666667 0.00000000               0.8472180
5     2.3333333 1.4666667 0.2000000 0.40000000               0.9267397
6     5.3333333 0.4000000 0.1333333 0.06666667               0.9525312
  Prop_Cremas Prop_Inicio_Fase_2 Kg_Inicio_Fase_2_Ha_Proy Kg_Cremas_Ha_Proy
1        0.45                  0                        0          7.059412
2        0.45                  0                        0          2.989216
3        0.45                  0                        0          2.698863
4        0.45                  0                        0          2.758599
5        0.45                  0                        0          6.051388
6        0.45                  0                        0          1.693511
  Kg_Maduras_Ha_Proy Kg_Cosechable_Ha_Proy
1           0.000000             0.0000000
2           5.434939             0.0000000
3           9.814049             0.0000000
4           3.901049             0.0000000
5           1.833754             3.6675081
6           1.254453             0.6272263

Total de válvulas: 63

            Arana            Biloxi           Maldiva            Malibu
                1                31                 1                 1
Manila 1° Siembra             Rocio       Stella Blue           Ventura
                1                14                 1                13
```

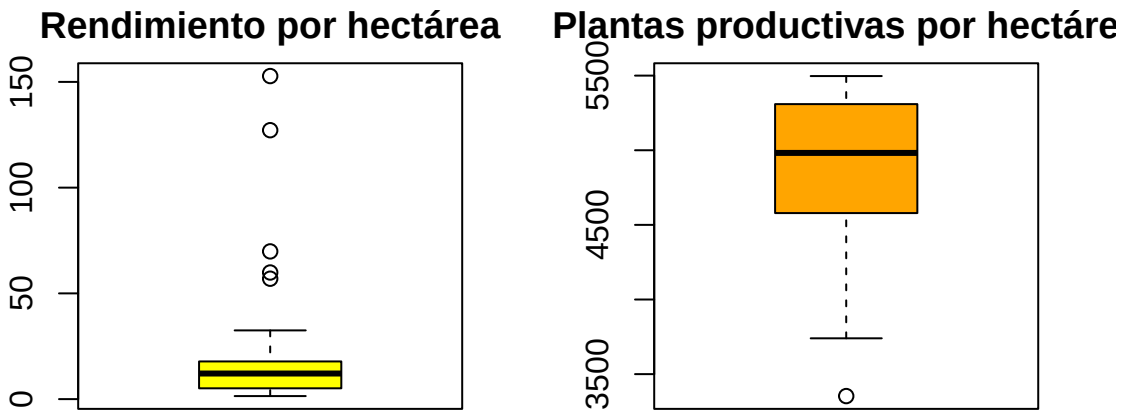


Resultados

Exploracion y estadisticas de la variedad BILOXI

```
  Kg_Total_Ha      Plantas_Productivas_Ha
 Min.   :  1.435   Min.   :3353
 1st Qu.:  5.101   1st Qu.:4579
 Median : 12.107   Median :4982
 Mean   : 23.684   Mean   :4882
 3rd Qu.: 17.813   3rd Qu.:5309
 Max.   :152.757   Max.   :5497
```

Diagrama de Tukey (diagrama de cajas, box plot)




```
$stats
          [,1]
[1,]  1.434797
[2,]  5.100566
[3,] 12.107008
[4,] 17.812748
[5,] 32.500000

$n
[1] 31

$conf
          [,1]
[1,]  8.499591
[2,] 15.714425

$out
[1] 127.16097  69.84625 152.75732  59.87255  56.92724
```

```
$group
[1] 1 1 1 1 1

$names
[1] ""


$stats
     [,1]
[1,] 3740
[2,] 4579
[3,] 4982
[4,] 5309
[5,] 5497

$n
[1] 31

$conf
         [,1]
[1,] 4774.843
[2,] 5189.157

$out
[1] 3353

$group
[1] 1

$names
[1] ""
```

Exploracion y estadisticas de las Variedades de Arándano

Estadisticas

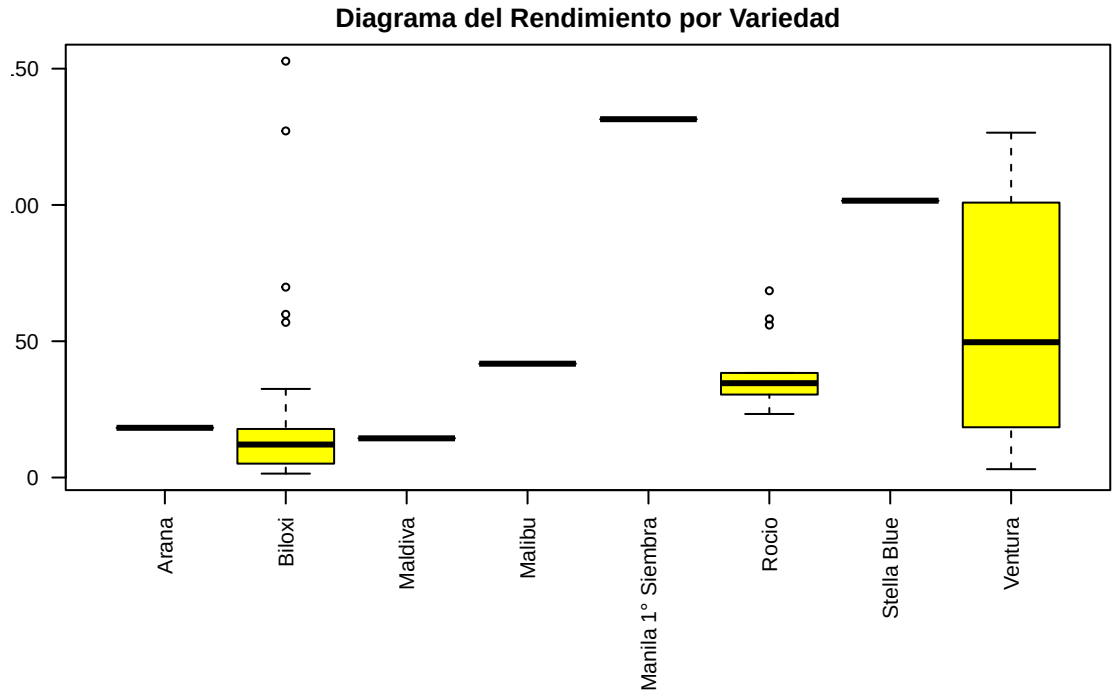
```
           Variedad Kg_Total_Ha Plantas_Productivas_Ha
1             Arana    18.24128               2937.000
2            Biloxi    23.68450               4882.065
3           Maldiva    14.38064               6449.000
4            Malibu    41.76077               6447.000
5 Manila 1° Siembra   131.46334               6276.000
6             Rocio    38.04010               6165.786
7       Stella Blue   101.56593               2938.000
8           Ventura    57.63666               3677.385

           Variedad Kg_Total_Ha Plantas_Productivas_Ha
1             Arana          NA                    NA
2            Biloxi    35.55558               522.0292
3           Maldiva          NA                    NA
4            Malibu          NA                    NA
5 Manila 1° Siembra          NA                    NA
6             Rocio    13.39190               629.6635
```

```
7       Stella Blue          NA                    NA
8           Ventura    43.76915              271.0758
```

Se observa que la media del rendimiento (Kg_Total_Ha) fue mayor en la variedad Manila 1° siembra con 131.46 kg . haˆ-1 y menor en la variedad Maldiva con 14.38 kg . haˆ-1.

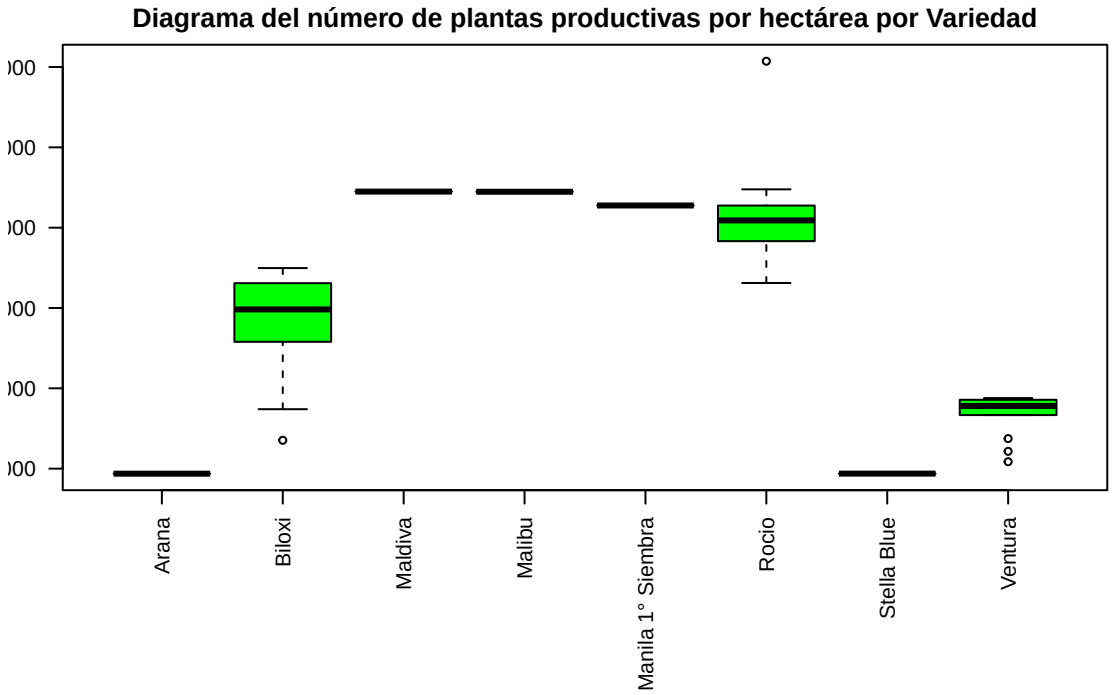
Con respecto a la variancia, debido a que solo las variedades Biloxi, Rocio y Ventura poseen más de una observación, son las únicas que registran variancia.

Diagrama de Tukey para la altura de los arboles de las Variedades



Segun el diagrama de Tukey, se oberva que la variacion del rendimiento no es uniforme y se presentan valores extremos mayormente por encima del conjunto en la variedad Biloxi. La medida central expresada por la mediana es diferente entre las variedades Biloxi, Ventura y Rocio.

Diagrama de Tukey para el número de plantas productivas por hectárea de las Variedades



En el número de plantas productivas la variacion fue heterogenea (tamaño de cajas diferentes), se presentan valores extremos por debajo del conjunto en la Variedad Ventura, mas que en otras variedades. Estas respuestas indican que la variacion es muy alta. La mediana es diferente por Variedad, algunas más distantes de las otras. Esto se debe a que existen variedades con densidades de siembra que permiten mayor número de plantas que otras variedades.

Aplicación de medidas estadísticas

Para establecer la identificacion y diferencias en las medidas, se utilizara el inventario de Arándano.

Se registro informacion de ubicacion y caracteristicas de 8 variedades, de las cuales, solo 3 tienen más de un registro.

```
Variedad
            Arana            Biloxi           Maldiva            Malibu
                1                31                 1                 1
Manila 1° Siembra            Rocio       Stella Blue           Ventura
                1                14                 1                13
```

La descripción de algunas variables permite indicar lo siguiente:

Escala nominal:

- Variedad

Escala de razon

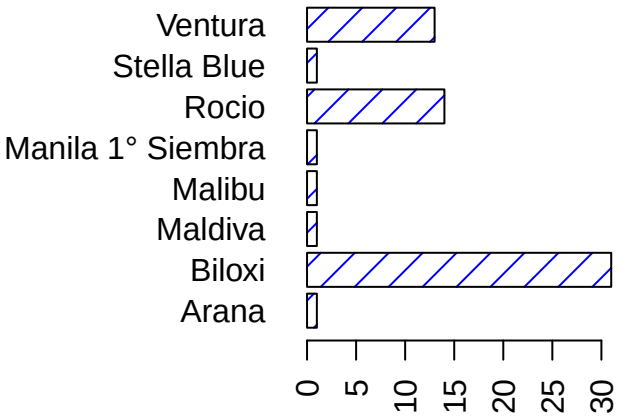
- Kg_Total_Ha
- Cremas
- Maduras

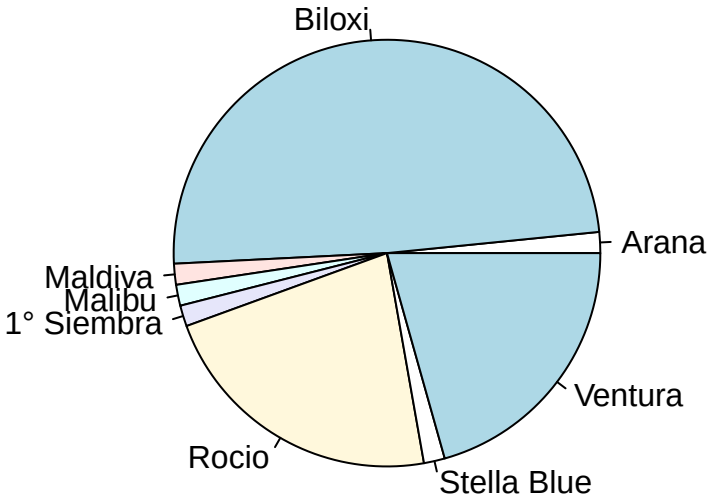
Su identificacion de algunas medidas, permite agrupar, y son consideradas como factores, en el caso de:

- Variedad

Medidas estadisticas

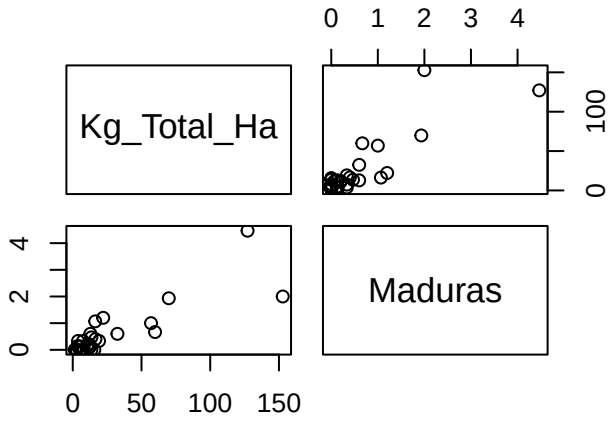
Representacion de las medidas en graficos.





Se puede observar que la variedad con más registros fue Biloxi, seguida de Rocio y Ventura.

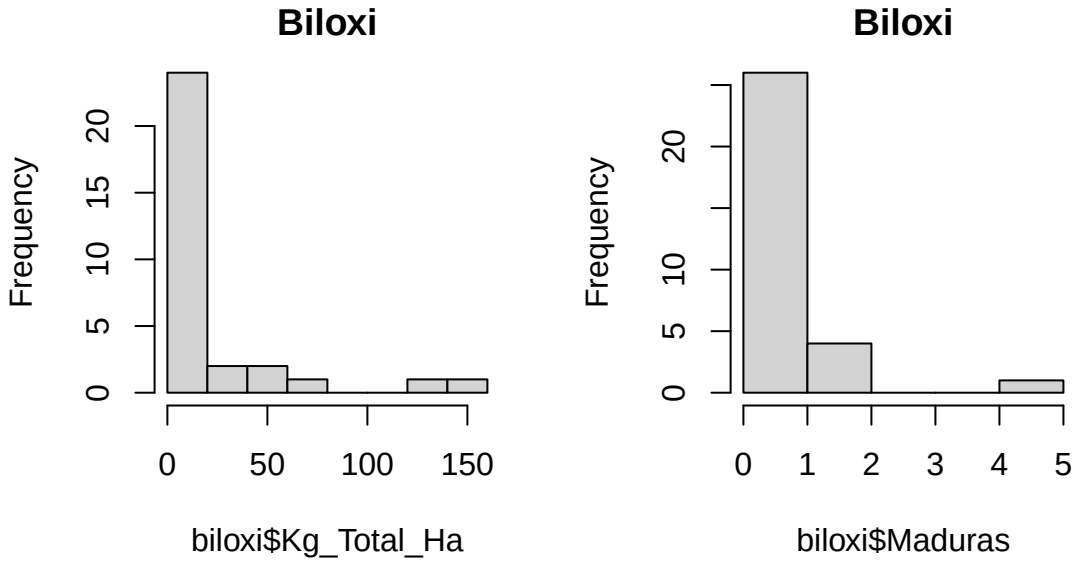
Para mostrar graficos entre variables cuantitativas de escala razón se utilizará el rendimiento y el número de frutos maduros por planta en Arándano.



Se observa que la relación Rendimiento y Número de frutas maduras por planta es directa y positiva. A mayor número de frutos maduros por planta, el rendimiento es mayor.

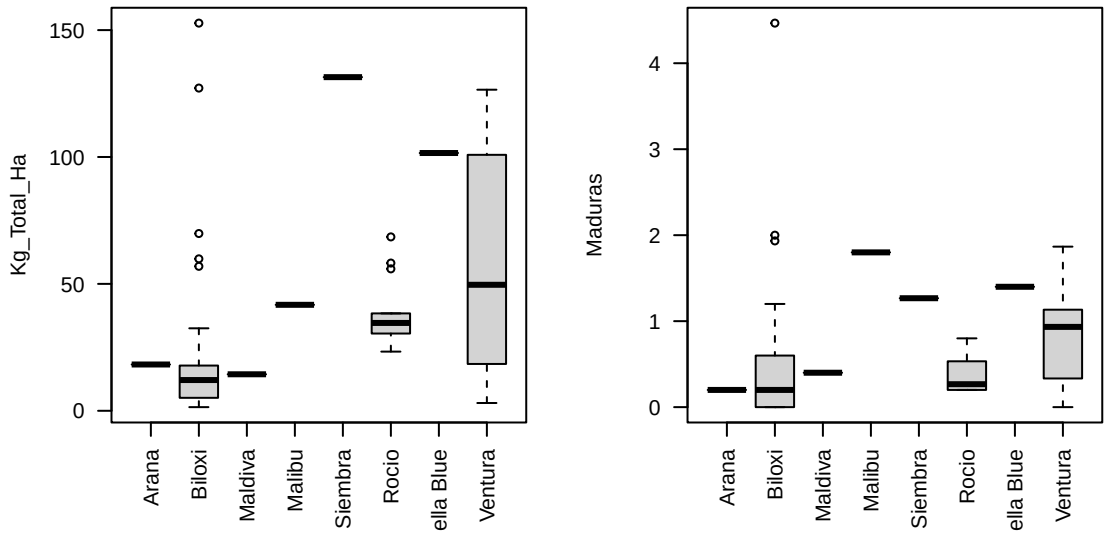
Los histogramas y el diagrama de cajas (Tukey) permiten una mejor descripcion.

BILOXI



Se puede observar que las distribuciones del rendimiento (Kg_Total_Ha) y el número de frutos maduros por planta (Maduras) son asimétricas positivas, con un mayor registro en los valores más bajos de ambas variables.

Diagrama de cajas para el Rendimiento y número de frutos maduros por planta de todas las variedades



Por otro lado, la distribución de las variables rendimiento y número de frutos maduros por planta, son muy similares dentro de cada variedad evaluada, por lo que se puede asumir que ambas variables están altamente correlacionadas.

Indicadores estadísticos.

1. Indicadores de centralidad

a) Promedio. es el punto central de valor continuos y es representativo cuando no se tiene valores extremo.

$$\text{Promedio} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Para ilustración se utilizara el inventario de Arándano en la variedad Biloxi.

```
Promedio de rendimiento de Biloxi: 23.6845 kg . ha^-1.
```

b) Mediana. Una medida robusta de centralidad, útil cuando hay valores extremos (outlairs), su uso frecuente es en variable de tipo ordinal o de escala jerarquía. El valor es obtenido después de ordenar los datos y obtener el punto que divide al conjunto en dos partes iguales.

Si X_i son los valores ordenados del conjunto, y n es el el numero de datos, entonces la mediana es calculada como:

$$\text{Mediana} = \begin{cases} (X_{n/2} + X_{n/2+1})/2 & \text{si n es par} \\ (X_{(n+1)/2}) & \text{si n es impar} \end{cases}$$

Para ilustración se utilizara el inventario de Arándano en la variedad Biloxi.

```
Mediana del rendimiento de Biloxi: 12.10701 kg . ha^-1.
```

c) Moda. Es el valor mas expresivo del conjunto, en datos cualitativos, es el mas abundante, si hay dos,se indica poblacional bimodal. No es frecuente en datos discretos y mucho en continuos. Sin embargo es posible obtener la moda en datos continuos, después de agruparlos.

$$\text{Moda} = \text{Valor mas frecuente del conjunto}$$

```
Moda del rendimiento de Biloxi
```

```
     [-   -]     mode
[1,]  1 26.3 13.90816
```

La moda es 13.91 kg . haˆ-1.

2. Indicadores de dispersión

a) Rango. Es la diferencia de los extremos de los datos. Útil cuando se tiene menos de 10.

$$\text{Rango} = \max(\text{X})\text{- } \min(\text{X})$$

```
Rango del rendimiento de Biloxi: 151.3225
```

b) Desviación estándar (S) Medida de variación obtenida frecuentemente de una muestra de mas de 10 datos.

$$S = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

```
Desviación estándar del rendimiento de Biloxi: 35.55558 kg - haˆ-1.
```

c) Varianza (S^2) Medida de variación obtenida frecuentemente de una muestra de mas de 10 datos.

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

```
Varianza del rendimiento de Biloxi: 1264.199 kg - haˆ-1.
```

d) Covarianza. Mide la variación conjunta de dos variables aleatorias.

$$cov(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

```
Covariancia del rendimiento y número de frutos maduros por planta de Biloxi
```

```
[1] 1264.199
```

La variación conjunta de altura y diámetro de Biloxi es de 1264.199.

e) Rango intercuartil IQR: Medida de variación apropiada cuando hay valores extremos con fines de comparar con otros grupos. Se calcula como la diferencia en los cuartiles 3 y 1

$$IQR = Q3 - Q1$$

```
Rango intercuartil del rendimiento de Biloxi: 12.71218 kg . haˆ-1.
```

El Rango intercuartil es 12.71218, esto también significa, como la mediana es 12.10701, entonces 2l 50% de válvulas de Biloxi tienen un rendimiento que esta disperso entre -0.60517 y 24.81919 kg . haˆ-1. Aproximadamente seria un variación relativa mayor al 100%

f) Coeficiente de variación CV: Medida de variación relativa, útil para comparar variación de medidas con diferente unidad.

$$CV = \frac{S * 100}{\bar{x}}$$

Su valor es normal si esta entre 10, baja variacion es menos de 10 y alta variacion mayor de 20, aceptable hasta 30%

```
Coeficiente de variacion del rendimiento de Biloxi: 150.1217 %.
```

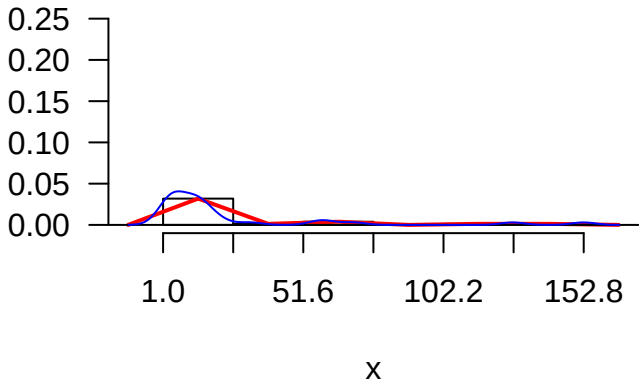
La medida relativa de variación es superior al aceptable, con 150.1217%, por lo tanto se considera como variación heterogénea.

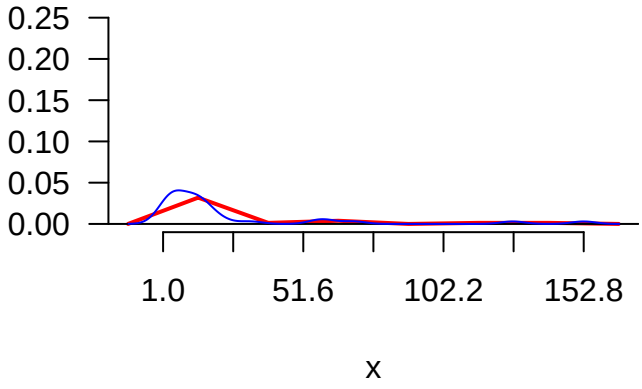
3. Indicadores de la forma de los datos

```
Coeficiente de asimetría del rendimiento de Biloxi: 2.643395
```

El valor es muy lejano a cero y positivo, por lo tanto es asimétrica positiva.

Gráficamente se observa su forma del rendimiento, donde la mayor cantidad de válvulas registradas tiene valores más bajos (entre 1 a 50 kg . haˆ-1).





Tallos y Hojas

En el diagrama de tallos y hojas, se observa igualmente que la mayor cantidad de observaciones posee valores por debajo a un rendimiento de 20 kg . haˆ-1.

```
  The decimal point is 1 digit(s) to the right of the |

   0 | 12234445566781223344667 9
   2 | 23
   4 | 7
   6 | 00
   8 |
  10 |
  12 | 7
  14 | 3
```

Índices de biodiversidad

Índice de diversidad de Margalef (R)

Aplicar el indice en el inventario Arándano.

```
            Arana            Biloxi           Maldiva            Malibu
                1                31                 1                 1
Manila 1° Siembra             Rocio       Stella Blue           Ventura
                1                14                 1                13
```

Se observa del inventario que hay 8 variedades diferentes.

```
Method: Margalef

The index: 1.689542

90 percent confidence interval:
 1.457492 ; 2.299211
```

Según el método de Margalef hay poca diversidad de variedades.

Índice de Shannon- Wienner (H ') (equidad)

```
Method: Shannon

The index: 1.929839

90 percent confidence interval:
 1.570401 ; 2.56764
```

El índice de Shannon indica que existe baja diversidad de variedades.

Índice de Simpson (λ): es un índice de dominancia.

dominancia

```
Method: Simpson.Dom

The index: 0.335349

90 percent confidence interval:
 0.1889062 ; 0.4603175
```

Según el indicador de Simpson para dominancia, existe una dominancia de 33.53 %.

Diversidad

```
Method: Simpson.Div

The index: 0.664651

90 percent confidence interval:
 0.5396825 ; 0.8138
```

Según el indicador de Simpson para diversidad, existe una diversidad de 66.46 %.

Índice de Berger Parker (D):

```
Method: Berger.Parker

The index: 0.4920635

90 percent confidence interval:
 0.2321429 ; 0.6666667
```

Según el método de Berger Parker, existe mayor diversidad que dominancia, pero aparentemente el valor de ambos es casi igual.

Índice de McIntosh (M):

```
Method: McIntosh

The index: 0.4815802

90 percent confidence interval:
 0.3795902 ; 0.6218968
```

Según el índice de McIntosh existe una diversidad de 48.15 %.

Muestreo

Muestra piloto para tener una idea de la variacion

```
[1] 27.47823
```

En el rendimiento de la variedad Biloxi se tiene un rango de 54.10 kg . haˆ-1.

```
Tamaño de muestra optima para Rendimiento: 31
```

Según esta variancia, se necesitaría tomar a toda la población de valvulas de arándano Biloxi para estimar correctamente la media del rendimiento.

Estimar Rendimiento

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.435   5.101  12.107  23.684  17.813 152.757
```

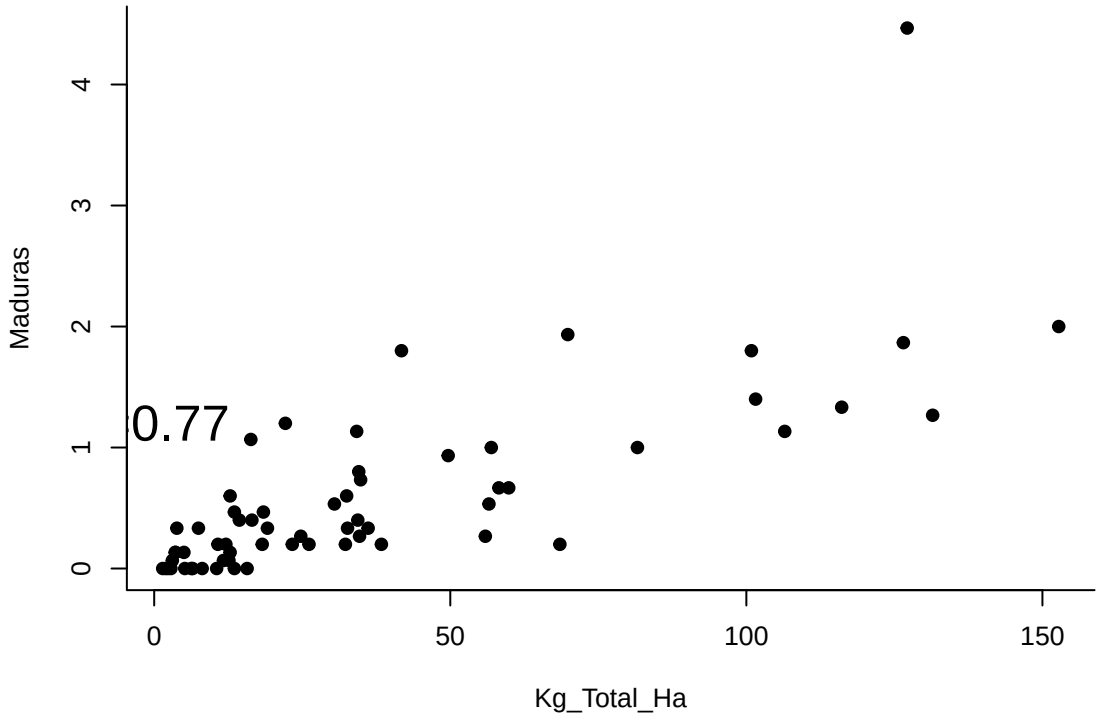
Los verdadero parametros de la poblacion son:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.435   5.101  12.107  23.684  17.813 152.757
```

Medidas de asociación y regresión

Se emplearon todos los datos del Inventario Arandano.

Dispersión y valor de la correlación



La correlación entre el rendimiento (Kg_Total_Ha) y el número de frutos maduros es de 0.77, por lo tanto existe una relación positiva entre estas variables.

Prueba estadistica

```
    Pearson's product-moment correlation

data:  Kg_Total_Ha and Maduras
t = 9.3641, df = 61, p-value = 0.0000000000002069
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6424118 0.8533444
sample estimates:
      cor
0.7679452
```

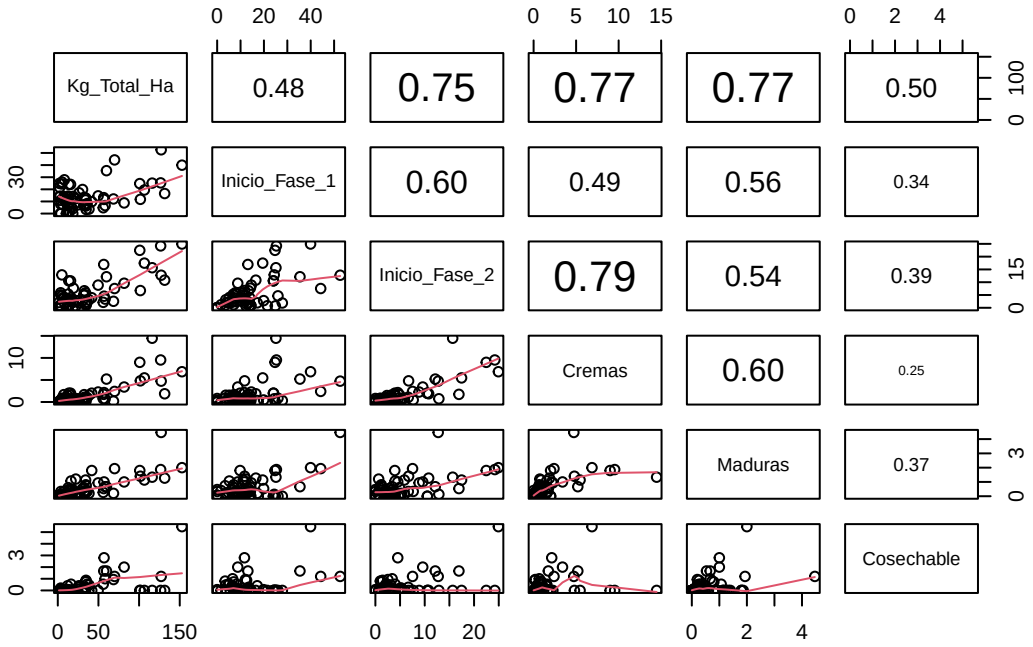
p-value: En términos de riesgo, indica la probabilidad de rechazar la hipótesis nula siendo esta verdadera. nuestro caso probabilidad es 2.069 10^{-13} nos indica que es imposible de equivocarnos en nuestra decisión y el rechazo de la hipótesis nula se da a un nivel de significancia de 0.001.

Por lo tanto existe suficiente evidencia estadística para concluir que el valor de r es estadísticamente diferente de cero y en este caso existe una relación positiva y estadísticamente significativa.

Otros métodos

Método de Spearman

Para una mejor observación del comportamiento de la correlación, se utilizara datos de Rendimiento y número de frutos por planta en diferentes estadíos. Como la prueba de Spearman se puede aplicar a datos que sean por lo menos ordinales, no hay problema en aplicarlos en datos continuos o discretos.



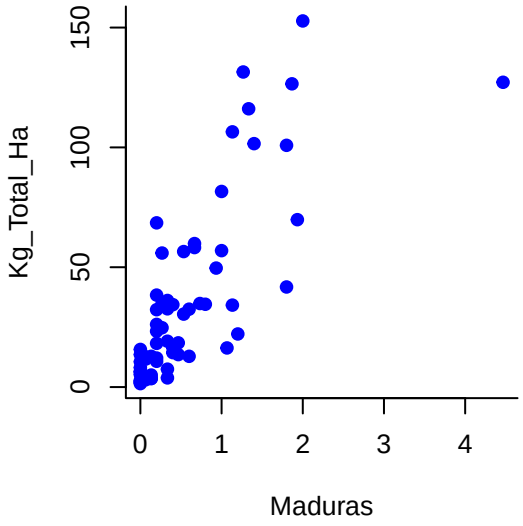
Matriz de correlacion

	Kg_Total_Ha	Inicio_Fase_1	Inicio_Fase_2	Cremas	Maduras	Cosechable
Kg_Total_Ha	1.00	0.48	0.75	0.77	0.77	0.50
Inicio_Fase_1	0.48	1.00	0.60	0.49	0.56	0.34
Inicio_Fase_2	0.75	0.60	1.00	0.79	0.54	0.39
Cremas	0.77	0.49	0.79	1.00	0.60	0.25
Maduras	0.77	0.56	0.54	0.60	1.00	0.37
Cosechable	0.50	0.34	0.39	0.25	0.37	1.00

	Kg_Total_Ha	Inicio_Fase_1	Inicio_Fase_2
Kg_Total_Ha	1.0000000000000000000	0.0000671797174	0.0000000000230349073
Inicio_Fase_1	0.0000671797173950051	1.0000000000000	0.0000024493392536407
Inicio_Fase_2	0.0000000000023034907	0.0000002449339	1.0000000000000000000
Cremas	0.0000000000001769696	0.0000511183686	0.0000000000001998401
Maduras	0.0000000000002069456	0.0000018741493	0.0000402449621539347
Cosechable	0.0000290919248819854	0.0066315147750	0.0013955290677296616

	Cremas	Maduras	Cosechable
Kg_Total_Ha	0.0000000000017696955	0.0000000000002069456	0.00002909192
Inicio_Fase_1	0.0000511183686391384	0.0000018741492688878	0.00663151478
Inicio_Fase_2	0.0000000000001998401	0.0000040244962153935	0.00139552907
Cremas	1.0000000000000000000	0.0000002379909087580	0.04702011712
Maduras	0.0000002379909087580	1.0000000000000000000	0.00312433665
Cosechable	0.0470201171232114134	0.0031243366510902426	1.00000000000

Se observa que la correlación por el método de Spearman entre el rendimiento y el resto de variables fue significativa.



Regresión

Se busca construir un modelo lineal que permita predecir una respuesta mediante un modelo lineal de primer grado. Se entiende que la variable respuesta (dependiente) expuesta a ser predicha, debe ser aleatoria y la variable explicativa (independiente) debe ser fija no sujeta a error.

Para estimar un mejor modelo de regresión, se debe utilizar observaciones lo mas diferentes para contemplar el dominio de la respuesta. Para el siguiente caso se evaluará la influencia del número de frutos maduros por planta sobre el rendimiento del cultivo de arándano.

$$y_i = \alpha + \beta x_i + \epsilon_i$$

x_i: variable independiente (número de frutos maduros).

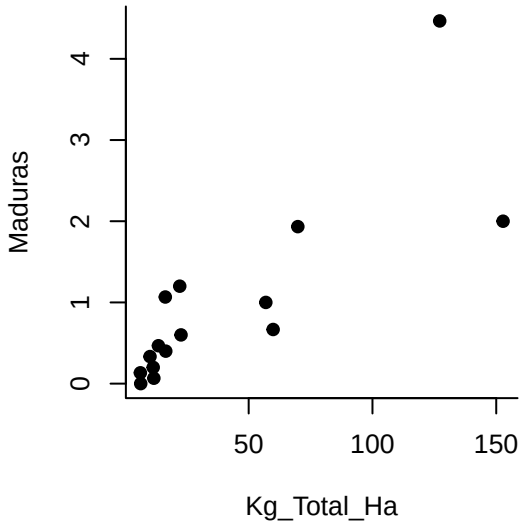
y_i: Variable respuesta (rendimiento), dependiente de x_i.

α: parámetro intercepto tiene unidades del la variable respuesta (kg . haˆ-1).

β: parámetro la pendiente, tasa de cambio de y al cambiar una unidad en x, las unidades es la razón de y e x (kg . haˆ-1 / unidades)

Aplicación

Usar el inventario Arandano y trabajar con la especie Biloxi.



Construcción del modelo

```
Call:
lm(formula = Kg_Total_Ha ~ Maduras, data = especie)

Residuals:
    Min      1Q  Median      3Q     Max
-27.075 -10.043  -5.385  -0.478  79.241

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.977      9.397   0.955 0.356813
Maduras       32.269      6.372   5.064 0.000217 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.44 on 13 degrees of freedom
Multiple R-squared:  0.6636,    Adjusted R-squared:  0.6378
F-statistic: 25.65 on 1 and 13 DF,  p-value: 0.0002169
```

De acuerdo a nuestro análisis.

$\hat{\alpha} = 8.977$

$\hat{\beta} = 32.269$

Ambos son importantes según el pvalue.

$R^2 = 0.6636$ que equivale a 66.36% de explicación del rendimiento esta en función del número de frutos maduros por planta.

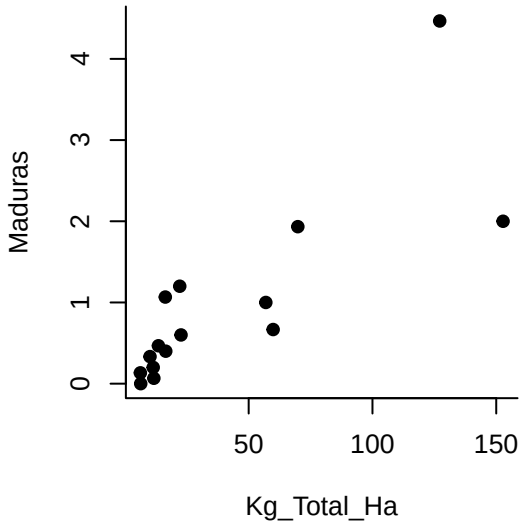
El p-value es de 0.0002169 menor a 0.05, se concluye que el modelo e aceptable y con una explicación de la variancia significativa.

Análisis de la varianza

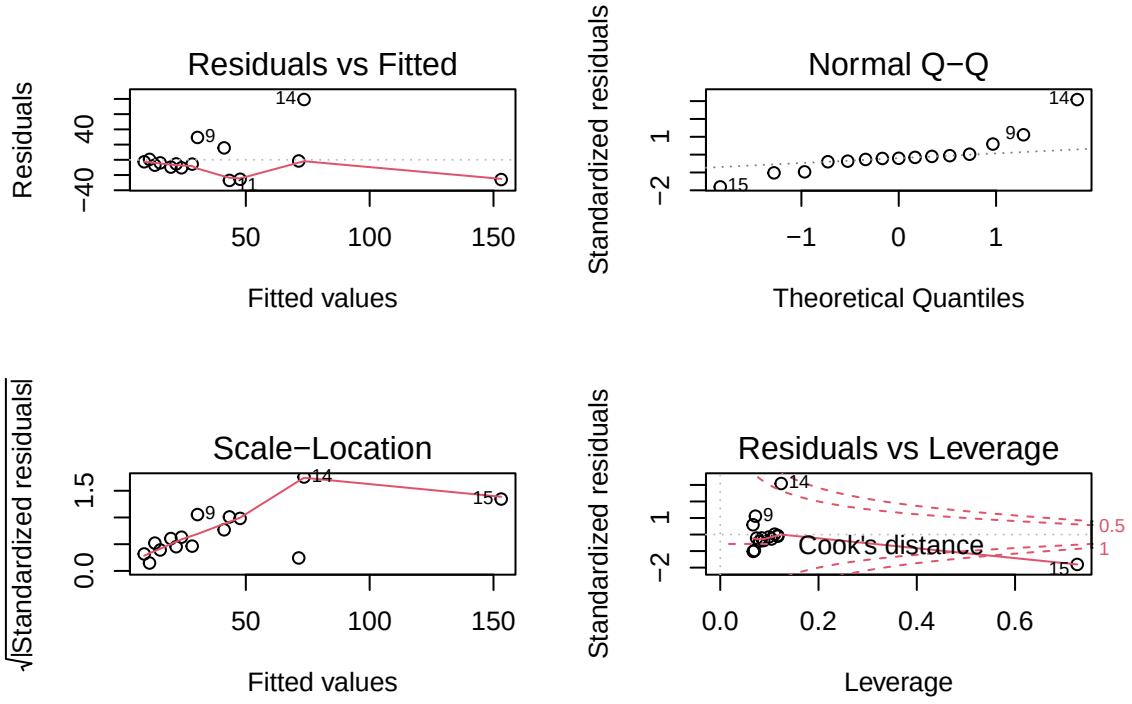
```
Analysis of Variance Table

Response: Kg_Total_Ha
          Df  Sum Sq Mean Sq F value    Pr(>F)
Maduras    1 19306.9 19306.9  25.648 0.0002169 ***
Residuals 13  9785.9   752.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Según el análisis de variancia, a un nivel de confianza de 0.05, la variable Maduras es una buena variable regresora.

Grafico del modelo



analisis del residuo



1. La falta de ajuste

Se observa que no existe una linealidad entre ambas variables, pero, esto se puede deber al bajo tamaño de muestras.

2. La normalidad

Los residuos no son normales y tienen una asimetría positiva.

3. La magnitud del error

El error es heterocedástico, es decir, la variancia no fue constante.

4. Valores extremos

Las observaciones 14 y 15 son consideradas como valores extremos.

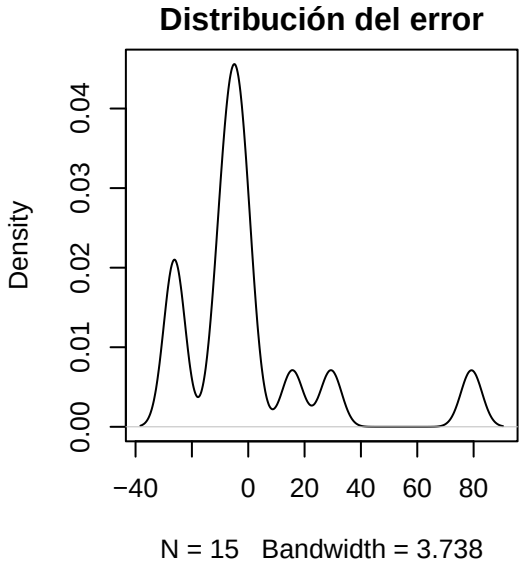
Ajuste del modelo

```
Call:
lm(formula = Kg_Total_Ha ~ Maduras + I(Maduras^2), data = especie)

Residuals:
    Min      1Q  Median      3Q     Max
-36.602  -8.526  -3.044   8.381  63.119

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.183     11.872  -0.268  0.79319
```

```
Maduras        59.417     18.496   3.212  0.00746 **
I(Maduras^2)   -6.503      4.187  -1.553  0.14634
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.06 on 12 degrees of freedom
Multiple R-squared:  0.7199,    Adjusted R-squared:  0.6733
F-statistic: 15.42 on 2 and 12 DF,  p-value: 0.0004826
```

Usar un termino cuadrático no ayuda a mejorar el modelo.

La normalidad de los errores.




```
    Shapiro-Wilk normality test

data:  error
W = 0.77277, p-value = 0.001679
```

el p-value = 0.001679 nos indica que la normalidad no esta presente, el valor es inferior a 0.10.

Predecir el rendimiento con valores propuestos

Predicción del dap utilizado

```
     Maduras Kg_Total_Ha prediccion
1 0.00000000    6.383825    8.97742
2 0.06666667   11.691857   11.12870
3 0.13333333    6.247805   13.27999
4 0.20000000   11.433939   15.43127
5 0.33333333   10.141947   19.73384
6 0.40000000   16.500115   21.88513
```

Se observa un resumen de valores predichos del modelo.

Predicción del dap distinto

Predecir el rendimiento para maduras = 300 y 500.

```
  MadurasNuevo RendimientoNuevo
1          300          9689.76
2          500         16143.62
```

Pruebas de hipótesis

Prueba sobre la media poblacional

La media poblacional es un parametro que mide la centralidad de los datos. Lo que se investiga a travez de una hipotesis, es probar si hay cambios en la medida central.

Prueba de una cola superior

El interes del investigador es probar que la media del DAP es superior al actual.

Prueba la hipótesis que la media del rendimiento de la Variedad Biloxi es superior a 10 kg . haˆ-1.

$H_0 : \mu \leq 10kg.ha^-1$

$H_1 : \mu > 10kg.ha^-1$

$\alpha = 0.05$

Utilice la prueba de t-student

```
    One Sample t-test

data:  Biloxi$Kg_Total_Ha
t = 2.1429, df = 30, p-value = 0.02018
alternative hypothesis: true mean is greater than 10
95 percent confidence interval:
 12.84584      Inf
sample estimates:
mean of x
  23.6845
```

Conclusión, se rechaza la hipótesis nula (pvalue es menor a 0.05), por lo tanto, hay evidencia estadística para afirmar que la media del rendimiento en la variedad Biloxi es superior a 10 kg . haˆ-1.

Prueba de una cola inferior

```
    One Sample t-test

data:  Kg_Total_Ha
t = 3.9241, df = 12, p-value = 0.999
alternative hypothesis: true mean is less than 10
95 percent confidence interval:
     -Inf 79.27252
sample estimates:
```

```
mean of x
 57.63666
```

Conclusión, se acepta la hipótesis nula (pvalue es menor a 0.05), por lo tanto, hay evidencia estadística para afirmar que la media del rendimiento en la variedad Biloxi no es inferior a 10 kg . haˆ-1.

Limites de confianza

```
    One Sample t-test

data:  Kg_Total_Ha
t = 10.628, df = 13, p-value = 0.00000008833
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 30.30785 45.77236
sample estimates:
mean of x
  38.0401
```

Los límites de confianza del rendimiento en la variedad Biloxi son 30.30 y 45.77 kg . haˆ-1.

Prueba de dos colas

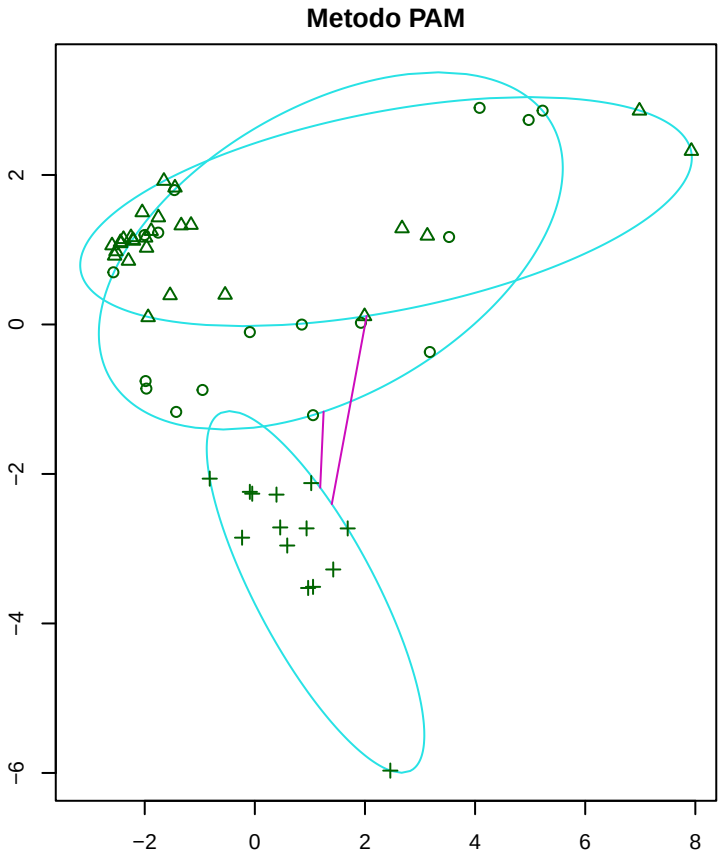
```
    F test to compare two variances

data:  KgVentura and KgRocio
F = 10.682, num df = 12, denom df = 13, p-value = 0.0001489
alternative hypothesis: true ratio of variances is not equal to 1
90 percent confidence interval:
  4.10268 28.41598
sample estimates:
ratio of variances
          10.68199
```

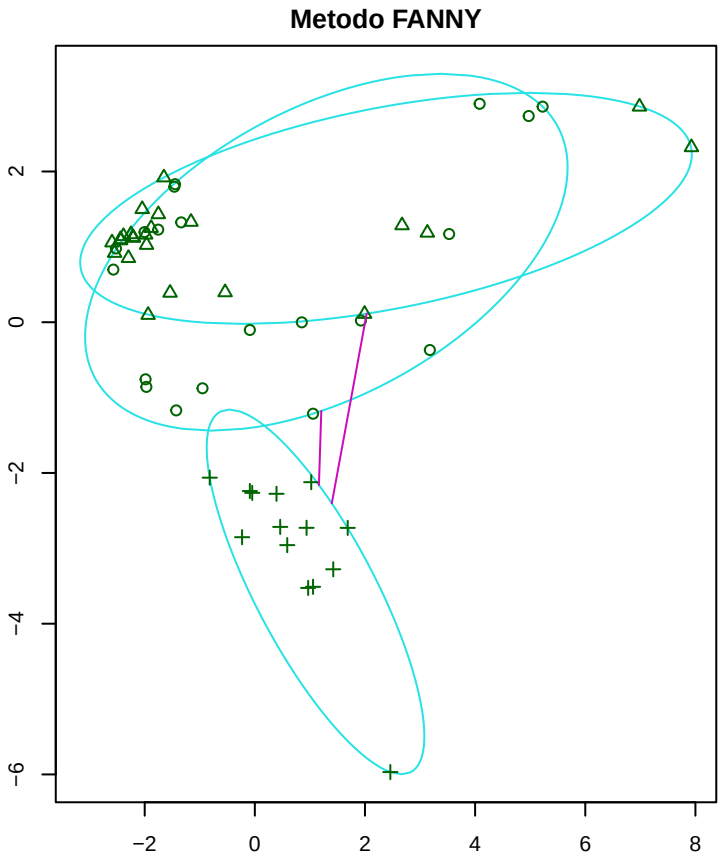
A un nivel de significancia de 0.05, existe suficiente evidencia estadística para concluir que las variancias del rendimiento entre las variedades Ventura y Rocio son diferentes.

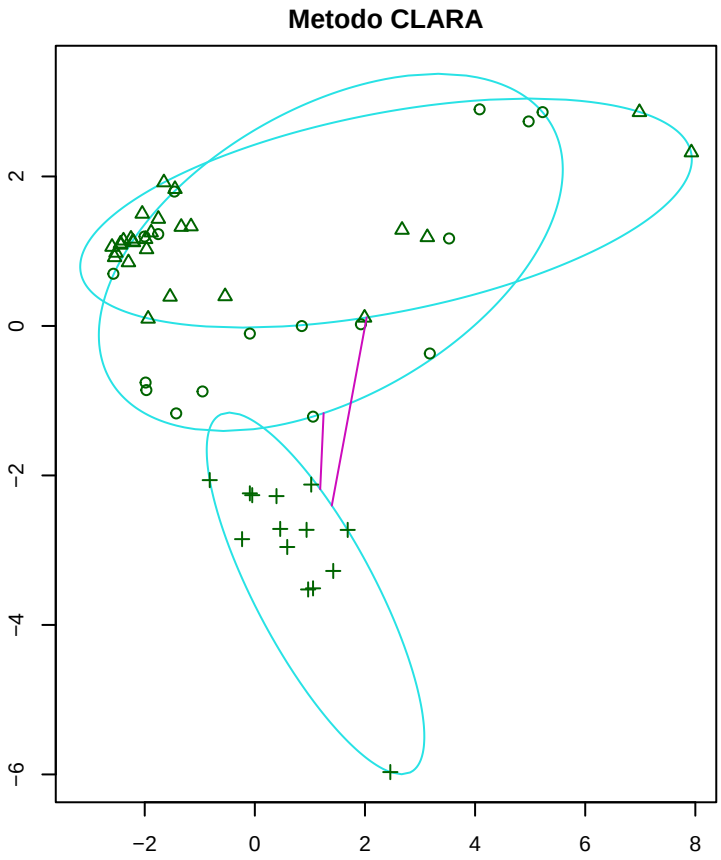
```
    Welch Two Sample t-test

data:  KgVentura and KgRocio
t = 1.5484, df = 14.079, p-value = 0.1437
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.533657 46.726770
sample estimates:
mean of x mean of y
 57.63666  38.04010
```

A un nivel de significancia de 0.05, existe suficiente evidencia estadística para concluir que las medias del rendimiento entre las variedades Ventura y Rocio son diferentes.

Clasificación no supervisada




Metodo FANNY
2
0
−2
−4
−6
−2
0
2
4
6
8



kmeans

La informacion debe ser cuantitativa

```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"

[1] 1 3 3 3 3 3

clase
 1  2  3
17 14 27

 Biloxi   Rocio Ventura
     31      14      13

         clase
variedad   1  2  3
  Biloxi   4  0 27
  Rocio    0 14  0
  Ventura 13  0  0
```

Clasificación supervisada

Analisis discriminante lineal de Fisher

La informacion debe ser cuantitativa.

```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select

 [1] "prior"   "counts"  "means"   "scaling" "lev"     "svd"     "N"
 [8] "call"    "terms"   "xlevels"

Call:
lda(Variedad ~ ., data = base)

Prior probabilities of groups:
   Biloxi     Rocio   Ventura
0.5344828 0.2413793 0.2241379

Group means:
        Kg_Total_Ha     Area Plantas_Ha Plantas_Productivas_Ha Inicio_Fase_1
Biloxi     23.68450 2.629355   5434.581               4882.065     17.815054
Rocio      38.04010 3.341429   8365.857               6165.786      8.914286
Ventura    57.63666 3.160000   5253.154               3677.385     11.984615
        Inicio_Fase_2    Cremas   Maduras Cosechable Prop_Plantas_Productivas
Biloxi       5.339785 1.2688172 0.5290323  0.4408602                0.8979628
Rocio        2.476190 0.6761905 0.3761905  0.5142857                0.7370841
Ventura      9.825641 3.7794872 0.8205128  0.3025641                0.7000365
        Prop_Cremas Prop_Inicio_Fase_2 Kg_Inicio_Fase_2_Ha_Proy
Biloxi    0.5629032         0.03610590                 2.467333
Rocio     0.8397392         0.18532579                16.180892
Ventura   0.6846154         0.09406846                13.537593
        Kg_Cremas_Ha_Proy Kg_Maduras_Ha_Proy Kg_Cosechable_Ha_Proy
Biloxi           8.253905           4.789911              4.170395
Rocio           12.745045           8.028016             10.869557
Ventura         29.631038          10.060112              3.392133
        Peso_Baya1000
Biloxi       1.740645
Rocio        3.594884
Ventura      3.280769

Coefficients of linear discriminants:
                                    LD1            LD2
Kg_Total_Ha               -0.0308583700  0.05075866017
Area                       0.0637687799 -0.44024209936
Plantas_Ha                 0.0018578286 -0.00376681949
Plantas_Productivas_Ha     0.0028852495  0.00003115766
Inicio_Fase_1              0.0005964357 -0.01964836328
Inicio_Fase_2              0.0904544097 -0.02772768334
Cremas                     0.6283895088  0.28847823353
Maduras                   -4.1937665273 -1.81132085248
Cosechable                -0.5068927812  3.22708187037
Prop_Plantas_Productivas -19.8826293242 -2.05188953675
Prop_Cremas                6.9445677123  1.26207683532
Prop_Inicio_Fase_2        -1.6957835765  9.35429992299
Kg_Inicio_Fase_2_Ha_Proy   0.0320913251 -0.10939624222
```

```
Kg_Cremas_Ha_Proy         -0.1165841883 -0.05969386748
Kg_Maduras_Ha_Proy         0.4391923095  0.11762026864
Kg_Cosechable_Ha_Proy      0.0652226751 -0.36183932027
Peso_Baya1000              6.0898873050  5.89348311389

Proportion of trace:
   LD1    LD2
0.8554 0.1446

[1] "class"     "posterior" "x"

[1] Biloxi Biloxi Biloxi Biloxi Biloxi Biloxi
Levels: Biloxi Rocio Ventura

         clase
variedad  Biloxi Rocio Ventura
  Biloxi      31     0       0
  Rocio        0    14       0
  Ventura      0     0      13

% de acierto: 100

% error aparente: 0
```

Se tuvo una tasa de acierto de 100 % al determinar las variedades con todas las variables regresoras de la base de datos del inventario Arandano.