Machine Learning

Trabajo práctico número 1

Integrantes:

- Juan Pablo Dantur, 54623
- Sebastián Ezequiel Guido, 54432
- Juan Cruz Lepore, 55124

Fecha:

26/05/2017

Índice

Resultados

Ejercicio 1

Al analizar los lirios, se encontraron tres especies distintos de ellos, con sus propias características. Comenzando con la especie denominada setosa, se pudo obtener que sus características son:

Característica	Media	Desvío	Mínimo	Máximo	Rango
largo sépalo	5	0	4.3	5.8	1.5
ancho sépalo	3.4	0.3525	2.3	4.4	2.1
largo pétalo	1.5	0.3791	1	1.9	0.9
ancho pétalo	0.2	0.1737	0.1	0.6	0.5

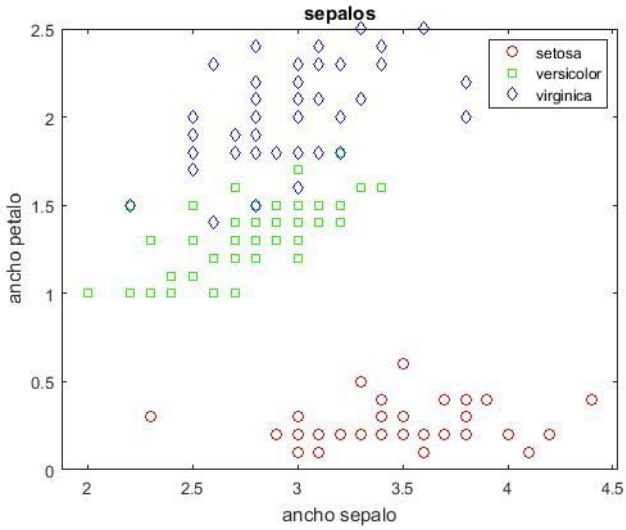
Para la especie versicolor se obtuvo:

Característica	Media	Desvío	Mínimo	Máximo	Rango
largo sépalo	5.9	1.005	4.9	7	2.1
ancho sépalo	2.8	3.005	2	3.4	1.4
largo pétalo	4.35	1.4094	3	5.1	2.1
ancho pétalo	1.3	2.1661	1	1.8	0.8

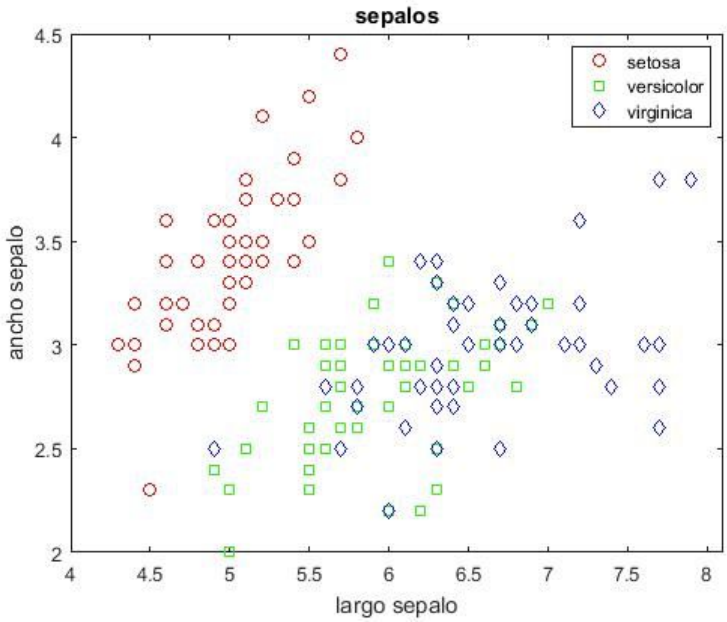
Por último, para la especie de tipo virginica se observó que sus medidas fueron:

Característica	Media	Desvío	Mínimo	Máximo	Rango
largo sépalo	6.5	1.419	4.9	7.9	3
ancho sépalo	3	3.1373	2.2	3.8	1.6
largo pétalo	5.55	1.4188	4.5	6.9	2.4
ancho pétalo	2	2.6450	1.4	2.5	1.1

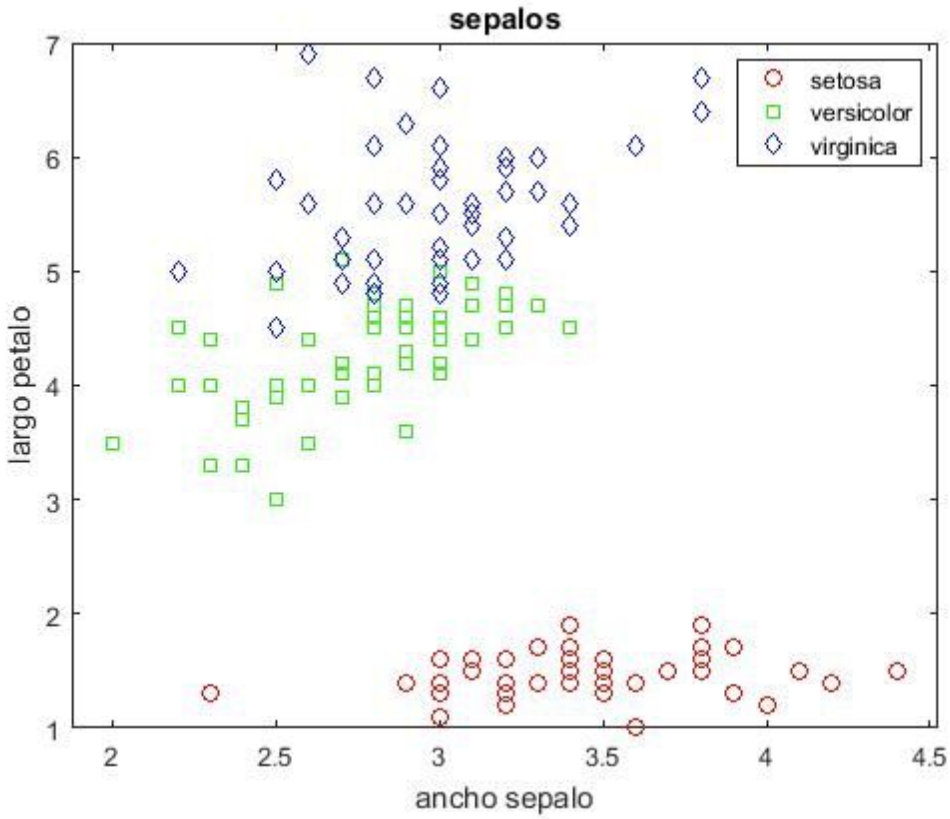
Para continuar con la comparación de las especies y obtener conclusiones más profundas con respecto a cuáles eran los atributos que podrían servir para diferenciar los distintos tipos, se contrastaron las mismas en diversos gráficos que evidenciaban las distribuciones, superposiciones y, de esta manera, diferencias y similitudes entre las tres.



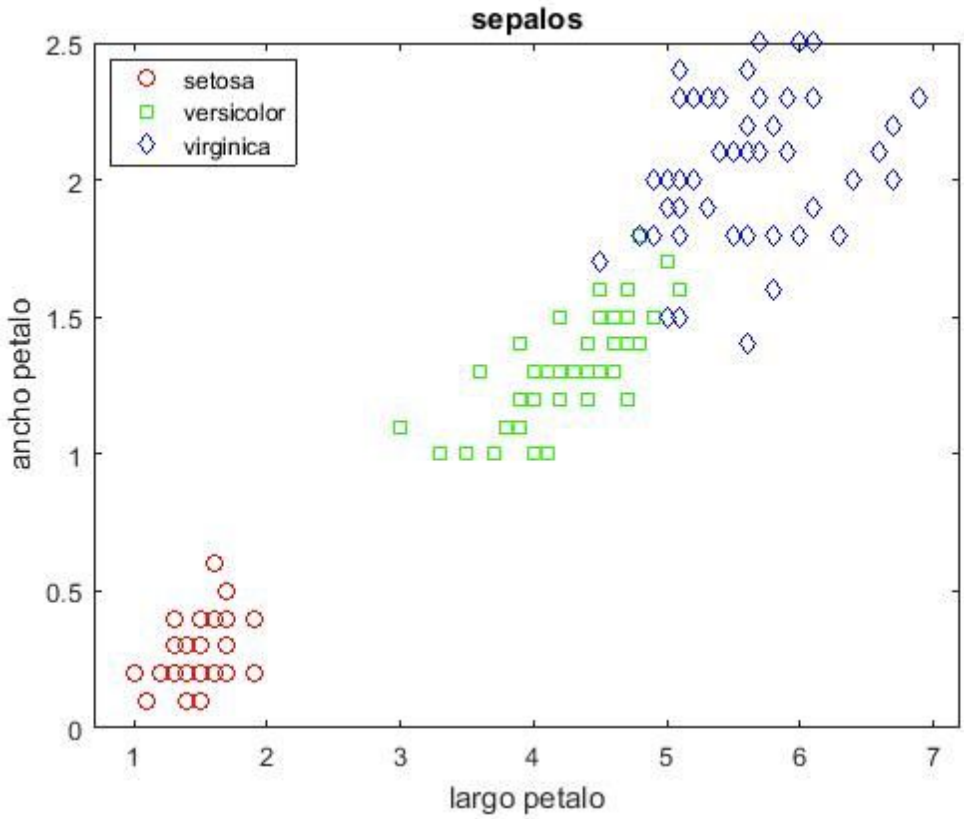
A partir de este gráfico se pudo determinar que tomando los anchos de las dos variables en cuestión, los lirios setosos son fácil y efectivamente clasificados y diferenciados de sus dos especies hermanas, mientras que a pesar de que estas estén diferenciadas en el ancho de los pétalos ligeramente, tienen zonas de superposición que no nos permitiría clasificarlas sin cometer alguna equivocación.



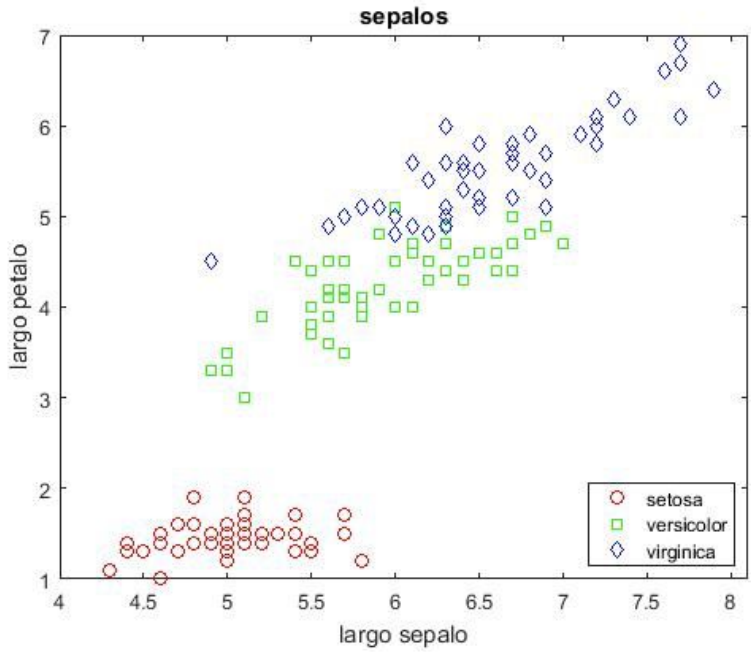
Nuevamente se pudo notar una marcada separación por parte de los lirios setosos de sus hermanos, pero resulta imposible tomar estas medidas para determinar algún tipo de diferencia entre las otras dos especies, puesto que las mismas se encuentran fuertemente superpuestas.



Se puede apreciar cómo los lirios setosos se siguen diferenciando notablemente del resto, estando las dos especies restantes diferenciadas en algunas zonas y mezcladas en otra.

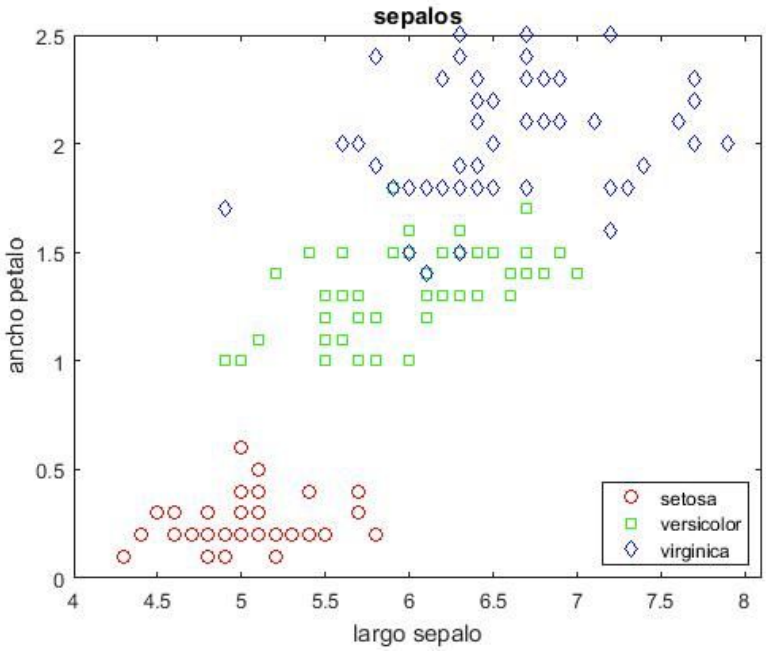


Se repiten los esquemas anteriores, los valores correspondientes a la setosa se encuentran en una zona aislada del resto de los valores, mientras que las dos especies restantes comparten una pequeña zona y se diferencian en el resto.

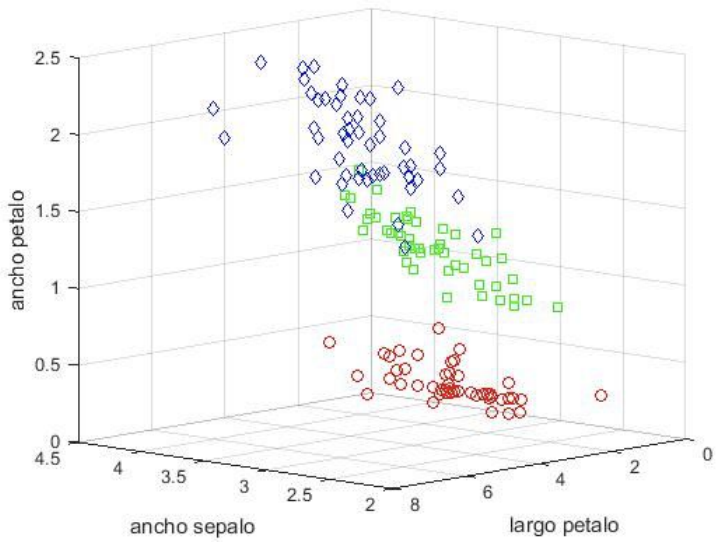


Se puede observar el mismo patrón que en las representaciones anteriores, manteniendo

la zona compartida por las especies versicolor y virginica, justo con sus respectivas zonas exclusivas de cada especie.

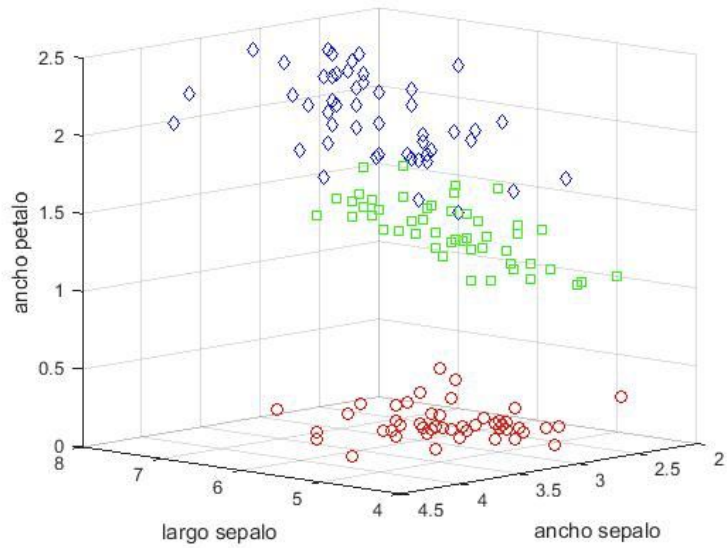


Nuevamente la especie setosa se encontró separada marcadamente de las otras dos, y los anchos de los pétalos podrían haber dado una diferencia entre las otras dos especies, ya que la superposición de los mismos resulta prácticamente nula, a pesar de mantenerse muy poco distantes.

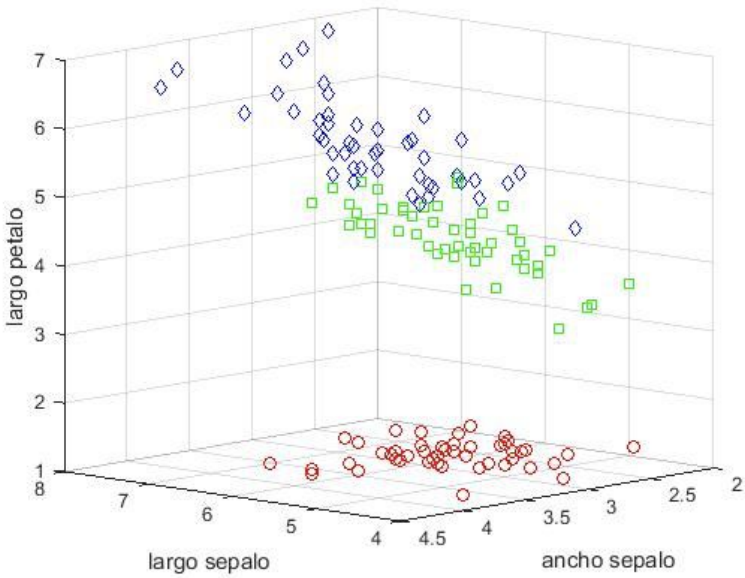


Cuando se tomaron tres variables en conjunto para determinar si la clasificación se aseveraba de alguna manera se reincidió en la marcada distancia de las setosas a las

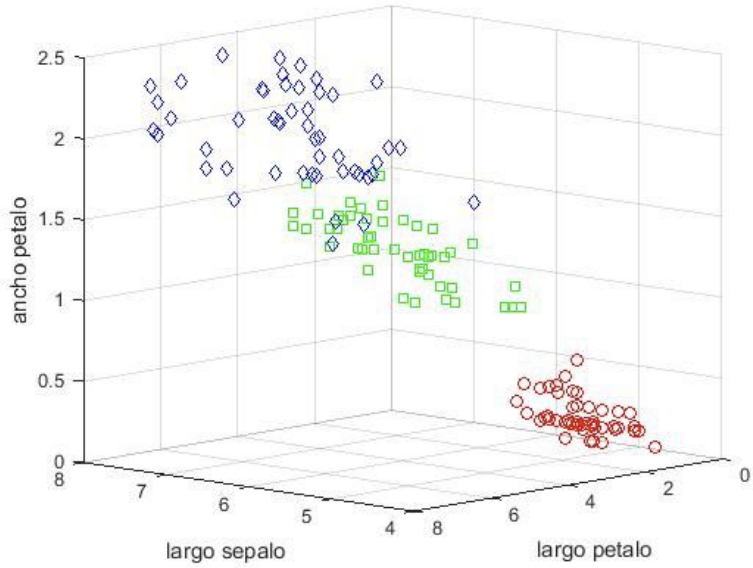
otras dos especies. Sin embargo, tomando estas tres variables de análisis, se pudo distinguir bastante bien entre las dos especies restantes, pero no sin cometer algún error debido a zonas de superposición.



Nuevamente, se repitió el análisis previo, donde la setosa cuenta con sus medidas distinguidas y las dos especies restantes sufren algunas ligeras superposiciones, por lo que no se las logra distinguir con una completa efectividad.



Nuevamente, el análisis resulta análogo, pero en este caso las virginicas de las versicolores resultan aún más difíciles de distinguir, puesto que se encuentran mucho más concentradas que en los casos anteriores.



Por último, en este análisis se vió una conglomeración y agrupamiento más denso por parte de dos de las especies, lo cual permitió distinguirlas bastante efectivamente, a pesar de que las virgínicas se encontraron dispersas y, en algunos casos, chocando con las versicolor.

Para poner a prueba si estas características y cantidad de información resultaba suficiente para categorizar los lirios, se le brindó a un clasificador de Naive Bayes 33 casos de training de cada una de las flores y se lo puso a prueba clasificando las 17 restantes de cada una. La matriz de confusión obtenida fue:

Especie/Especie	setosa	versicolor	virginica
setosa	17	0	0
versicolor	0	17	0
virginica	0	2	15

El porcentaje de datos mal clasificados resultó ser de un 3,92% y nuevamente se encontró consistencia con las predicciones realizadas previamente a partir de las distintas gráficas, las setosas no se confundieron con ninguna de las otras especies. Las versicolor, que a pesar de superponerse con las virginicas, se pudieron clasificar correctamente, dado que las mismas se encontraban más concentradas. Sin embargo, al ser las virginicas las que más dispersas se encontraron, algunas de ellas fueron clasificadas erróneamente.

Ejercicio 2

Los siguientes resultados se obtuvieron suponiendo que los valores obtenidos seguían una distribución normal.

a) **Setosa**

Largo de sépalo:

Mu:	5.005999999999999
Sigma al cuadrado:	0.121764000000000

Ancho de sépalo:

Mu:	3.428000000000001
Sigma al cuadrado:	0.140816000000000

Largo de pétalo:

Mu:	1.462000000000000
Sigma al cuadrado:	0.029556000000000

Ancho de pétalo:

Mu:	0.246000000000000
Sigma al cuadrado:	0.010884000000000

Versicolor

Largo de sépalo:

Mu:	5.936000000000000
Sigma al cuadrado:	0.261104000000000

Ancho de sépalo:

Mu:	2.770000000000000
Sigma al cuadrado:	0.096500000000000

Largo de pétalo:

Mu:	4.260000000000000
Sigma al cuadrado:	0.216400000000000

Ancho de pétalo:

Mu:	1.326000000000000
Sigma al cuadrado:	0.038324000000000

Virginica

Largo de sépalo:

Mu:	6.587999999999998
Sigma al cuadrado:	0.396256000000000

Ancho de sépalo:

Mu:	2.974000000000000
Sigma al cuadrado:	0.101924000000000

Largo de pétalo:

Mu:	5.552000000000000
Sigma al cuadrado:	0.298496000000000

Ancho de pétalo:

Mu:	2.026000000000000
Sigma al cuadrado:	0.073924000000000

b) El error cuadrático medio en cada caso es de:

Setosa

Largo de sépalo:

ECM: 0.121764000000000

Ancho de sépalo:

ECM: 0.140816000000000

Largo de pétalo:

ECM: 0.029556000000000

Ancho de pétalo:

ECM: 0.010884000000000

Versicolor

Largo de sépalo:

ECM: 0.261104000000000

Ancho de sépalo:

ECM: 0.096500000000000

Largo de pétalo:

ECM: 0.216400000000000

Ancho de pétalo:

ECM: 0.038324000000000

Virginica

Largo de sépalo:

ECM: 0.396256000000000

Ancho de sépalo:

ECM: 0.101924000000000

Largo de pétalo:

ECM: 0.298496000000000

Ancho de pétalo:

ECM: 0.073924000000000

c) Los intervalos para las medias, con un 95% de confianza son:

Setosa

Largo de sépalo:

[4.922424659547982 , 5.089575340452017]

Ancho de sépalo:

[3.338123799788596, 3.517876200211406]

Largo de pétalo:

[1.420824248888449, 1.503175751111551]

Ancho de pétalo:

[0.221013066110384, 0.270986933889616]

Versicolor

Largo de sépalo:

[5.813615768879933, 6.058384231120066]

Ancho de sépalo:

[2.695598385822443, 2.844401614177558]

Largo de pétalo:

[4.148584059819066, 4.371415940180934]

Ancho de pétalo:

[1.279112819591029, 1.372887180408970]

Virginica

Largo de sépalo:

[6.437232884188693, 6.738767115811304]

Ancho de sépalo:

[2.897536014505899, 3.050463985494100]

Largo de pétalo:

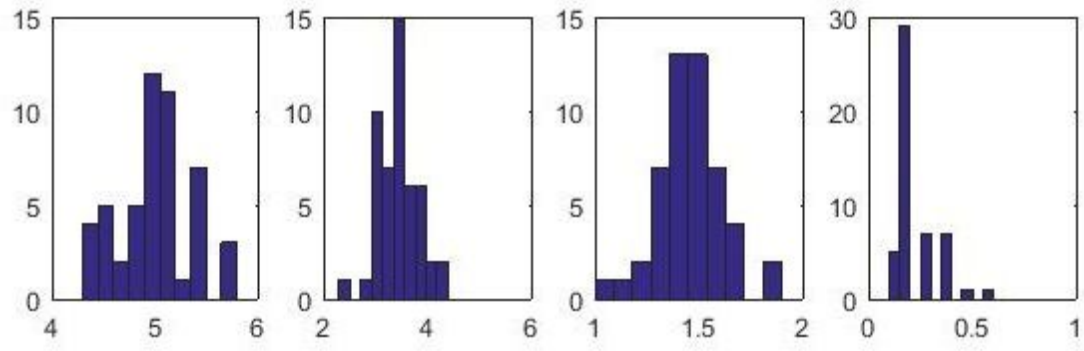
[5.421145711897525, 5.682854288102474]

Ancho de pétalo:

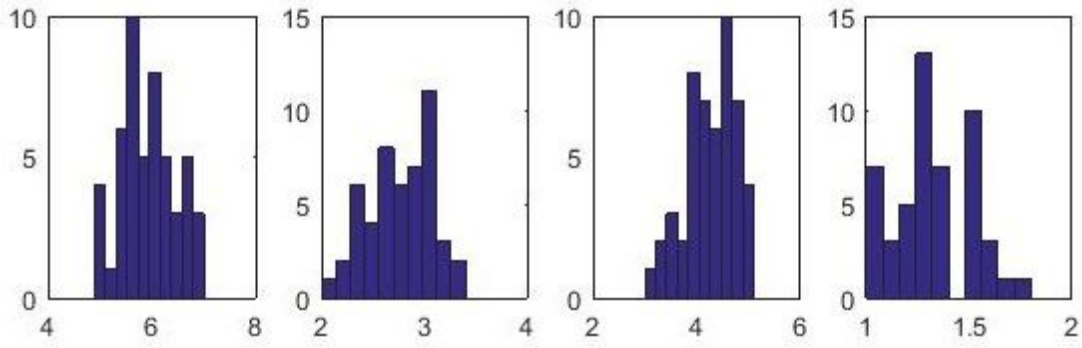
[1.960880444059345, 2.091119555940654]

Ejercicio 3

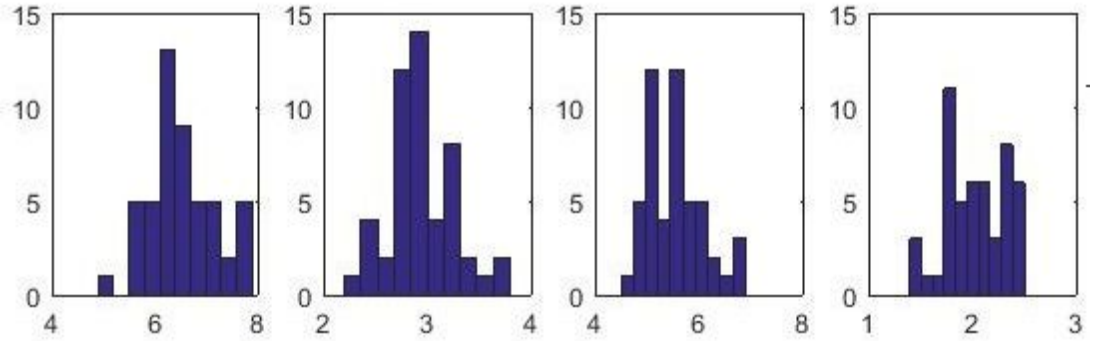
Para cada una de las especies de lirios se realizó un histograma para cada una de sus características, en donde se graficó en el eje vertical la frecuencia de aparición correspondiente a una medida dada en el eje horizontal. Cada uno de los distintos gráficos se corresponden con las variables ‘largo sépalo’, ‘ancho sépalo’, ‘largo pétalo’ y ‘ancho pétalo’ respectivamente.



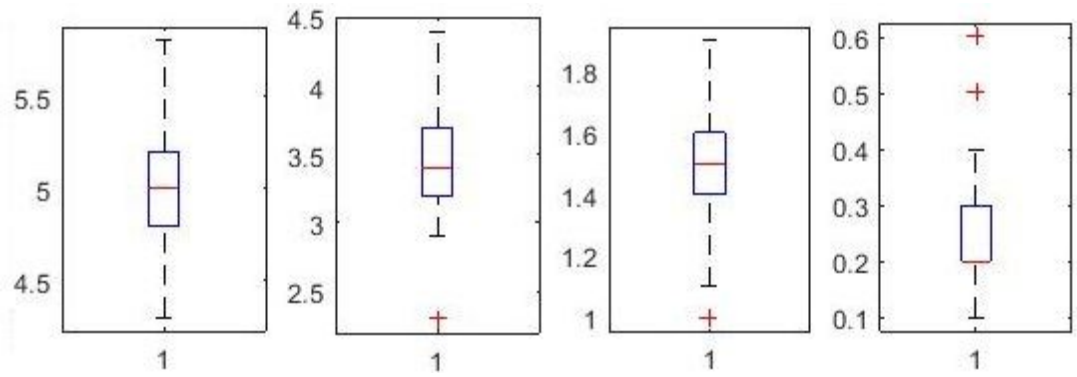
Setosa. Se pudo observar que haber aproximado las variables a una distribución normal no fue demasiado acertado, pues la distribuciones de las medidas no logran representar la curva normal que se hubiera esperado,salvo por el caso de los largos de los pétalos, donde se aproxima más. Sí se pudo notar a partir de estos gráficos también que los distintos valores de la muestra se encontraban agrupados fuertemente dada la cercanía de los picos en cada caso.



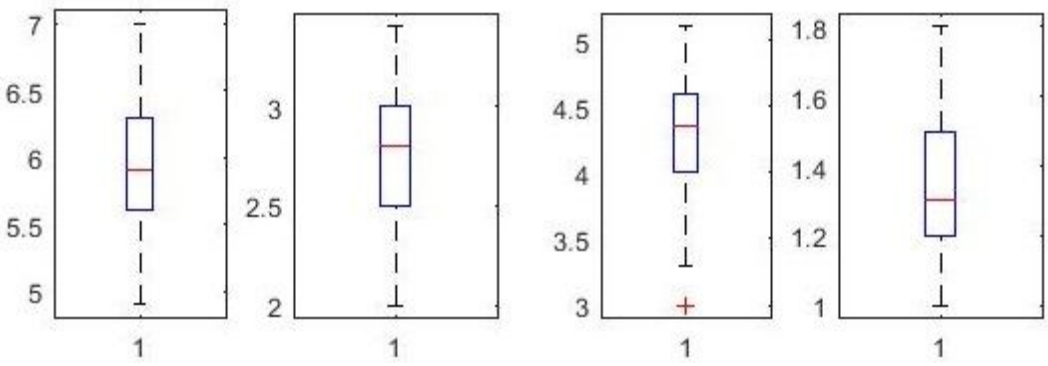
Versicolor. Se obtuvieron curvas que nuevamente no confirmaron la presencia de una distribución normal en sus medidas. Nuevamente los picos de frecuencia solieron estar posicionados cercanos entre sí, por lo que se pudo determinar que, aunque en menor medida que las setosas, se encontraban agrupados.



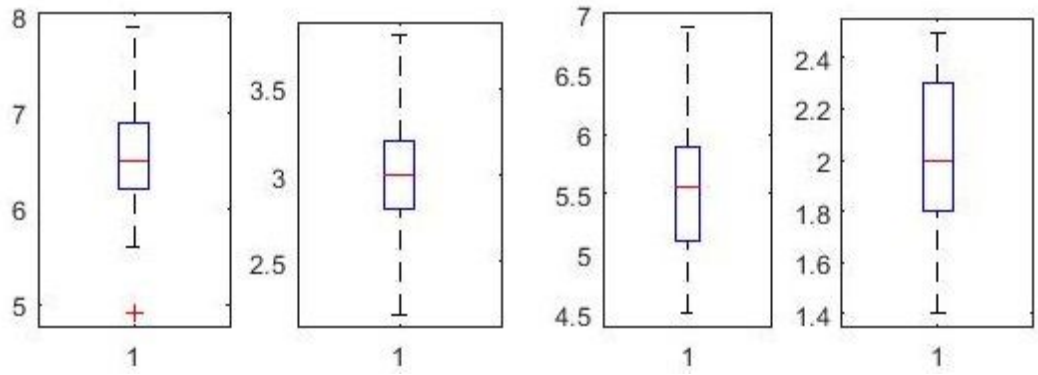
Virginica. Se vió que, nuevamente, y con mayor fuerza que antes, se distanciaban de una distribución normal. En estos casos se vió que la distribución de frecuencias había resultado mucho más dispersa y homogénea que el que en el resto de los casos.



Setosa. En las primeras tres variables se puede ver cómo los valores se encuentran distribuidos de manera casi equitativa por encima y por debajo de la media, pero en el caso del ancho de los pétalos se puede apreciar cómo la presencia de outliers desvía el promedio y lleva a una cantidad desbalanceada de muestras por encima y por debajo de la media.



Versicolor. En este caso todas las variables presentan una distribución más equitativa a lo largo de todos los valores que pueden tomar, aunque presentando un desvío grande en comparación a otros caso para el largo del sépalo.



Virginica. Se puede percibir algo parecido al caso de la especie anterior, pero con un desvío ligeramente menor en el largo de los sépalos y notablemente mayor en el ancho de los pétalos.

Ejercicio 4

a y b)
Se realizaron los siguientes comandos en Matlab:

```
x = [ones(1,26) zeros(1,174)];
[h,p] = ttest(x,0.1,'Tail','right')
```

Como h dio 0, no se puede rechazar la hipótesis nula con un alfa de 0.05, por lo que no hubo un aumento de preferencia hacia la marca.
El valor p es el retornado por el comando de arriba, que dio 0.1049.

c) Lo que se pide en este caso es la probabilidad beta de la prueba de hipótesis. Se ejecutaron los siguientes comandos:

```
pwrout = sampsizepwr('p',0.1,0.15,[],200,'Tail','right');
beta = 1-pwrout
```

Se obtuvo que beta vale 0.3166.

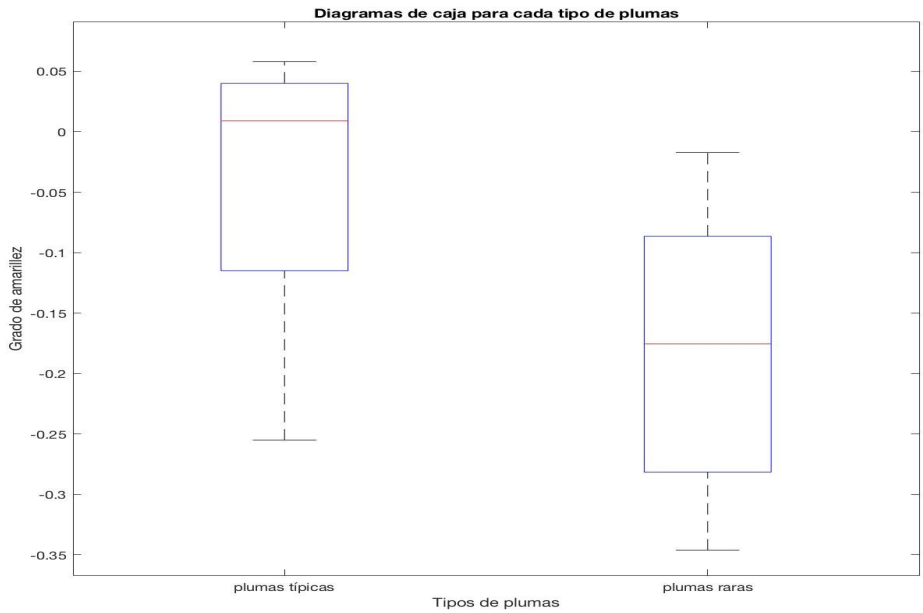
d) Para resolver este punto se utilizó el siguiente comando:

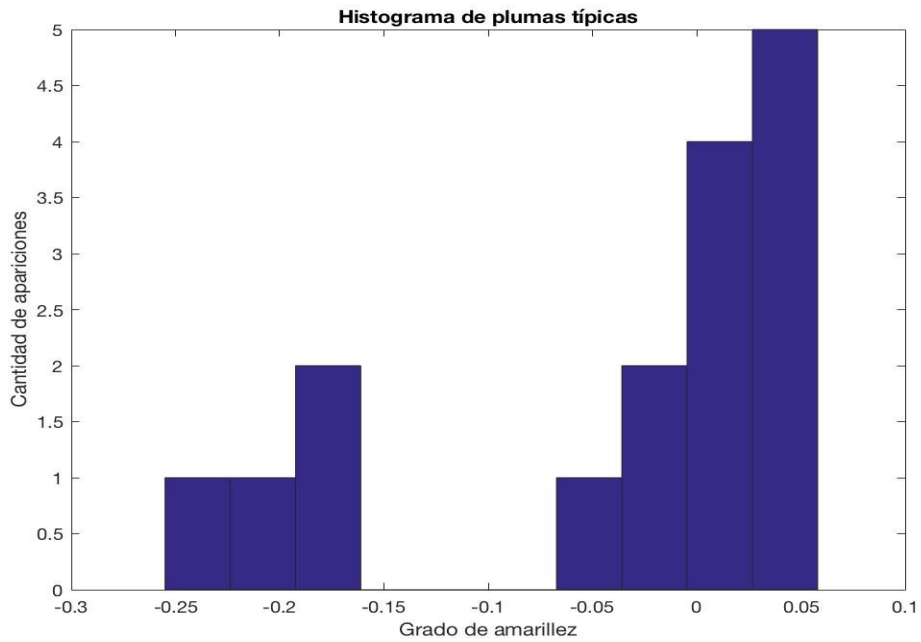
```
nout = sampsizepwr('p',0.1,0.15,0.95,[],'Tail','right')
```

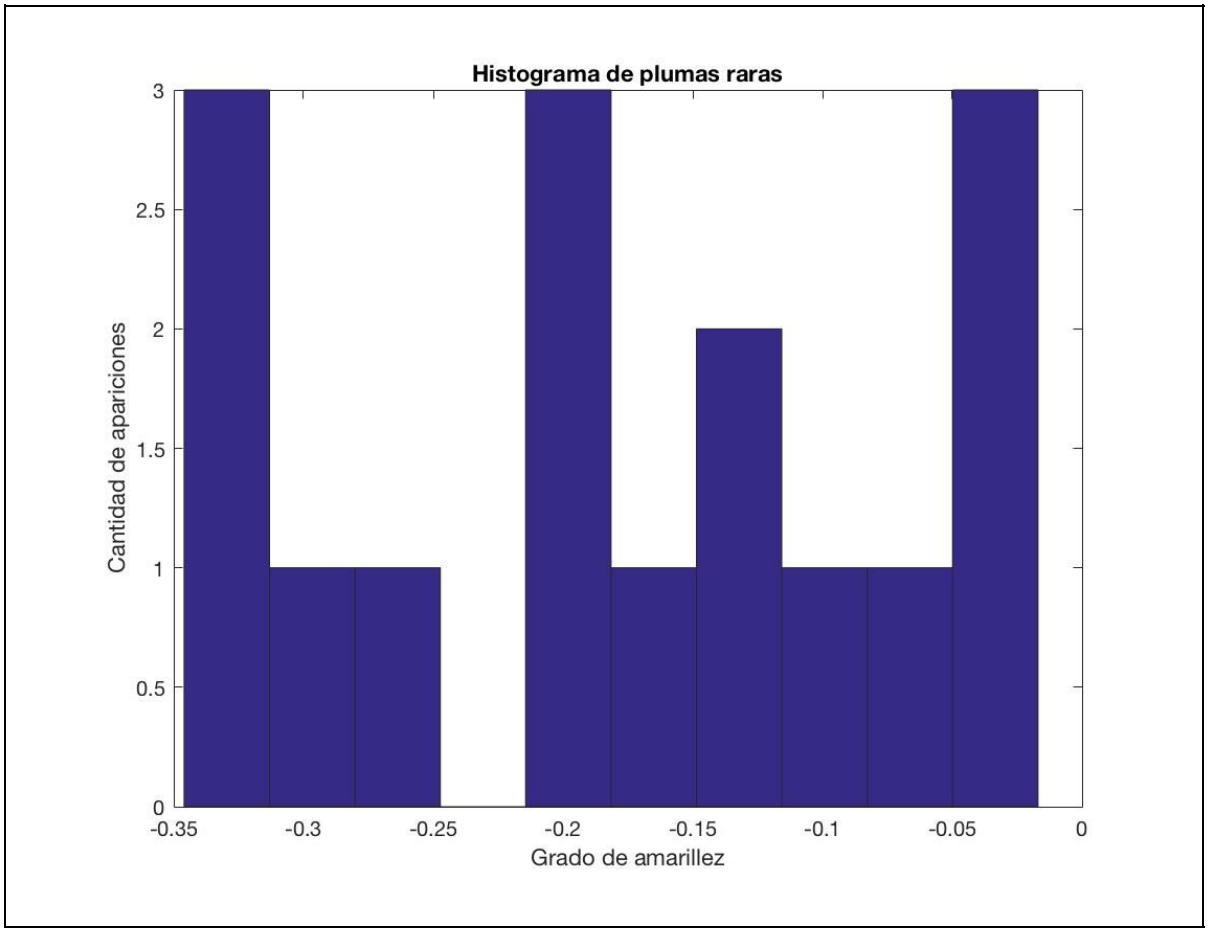
El resultado del mismo es el tamaño de muestra que se debería tomar para que beta valga 0.05, el cual dio 474.

Ejercicio 5

a)
Para resolver este punto se planteó como hipótesis nula que la media de X-Y es 0, siendo X el grado de amarillez de las plumas típicas e Y el grado de amarillez de las plumas raras. Se utilizaron los parámetros default del comando ttest (alfa = 0.05, test de dos colas) y se obtuvo como resultado h=1, lo que indica que se rechaza la hipótesis nula. Por lo tanto, se verifica la hipótesis de los investigadores.
b)


Diagramas de caja para cada tipo de plumas
Grado de amarillez
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
-0.35
plumas típicas
plumas raras
Tipos de plumas

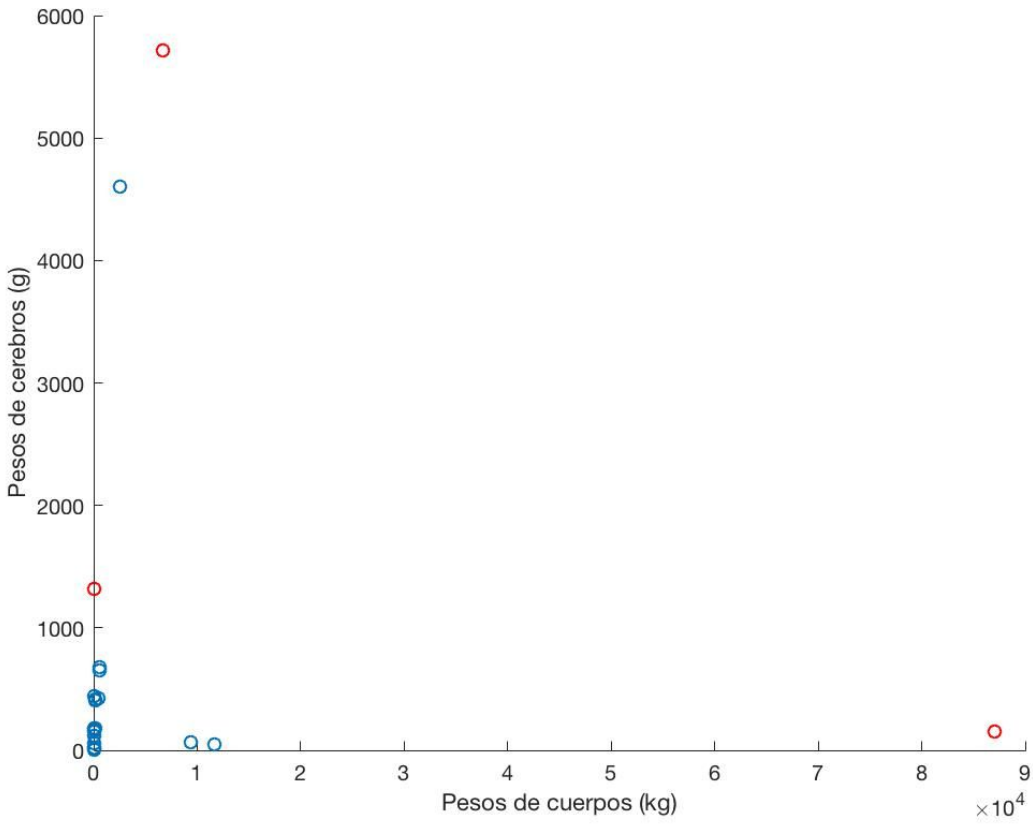

Histograma de plumas típicas
Cantidad de apariciones
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
-0.3
-0.25
-0.2
-0.15
-0.1
-0.05
0
0.05
0.1
Grado de amarillez



Lo que se puede notar en el histograma de las plumas típicas las barras distan mucho de formar una “campana”, y en su respectivo diagrama de caja es notable la diferencia de distancia entre la mediana y los extremos, por lo que no tiene mucho sentido suponer normalidad de los datos. En el caso de las plumas raras, se puede notar una simetría en el diagrama de caja, y el histograma es más parecido a una campana que en el caso anterior, por lo que puede llegar a ser válida la suposición de normalidad.

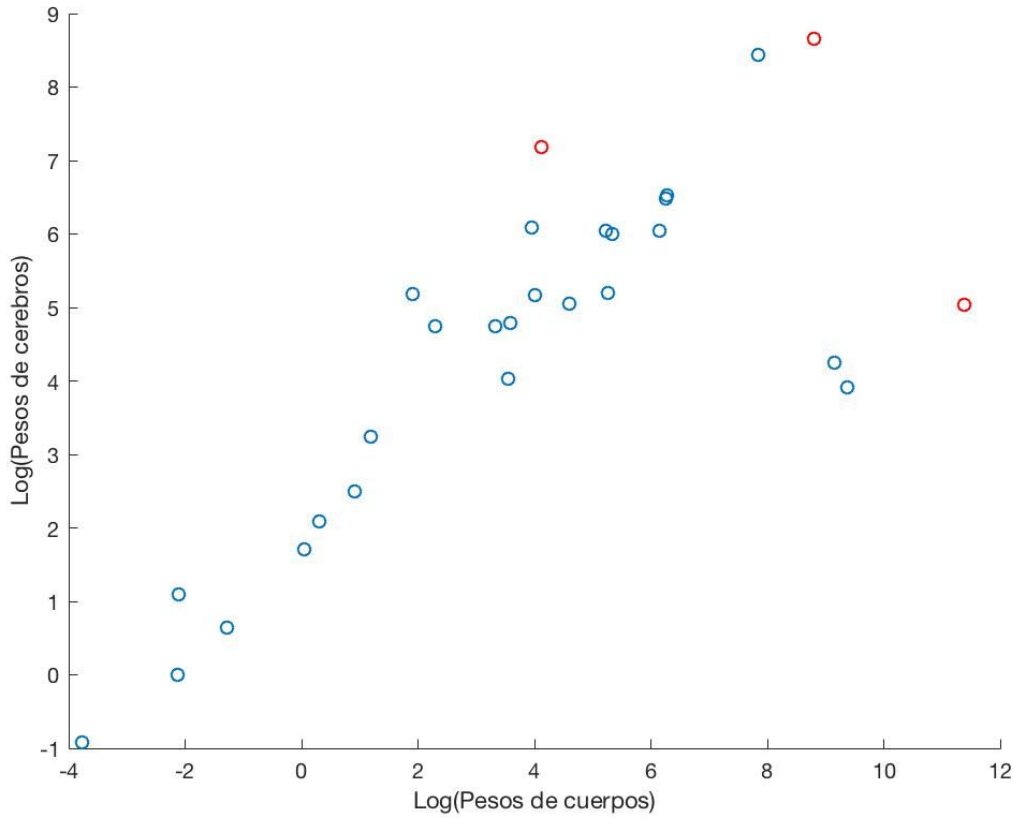
Ejercicio 6

a)



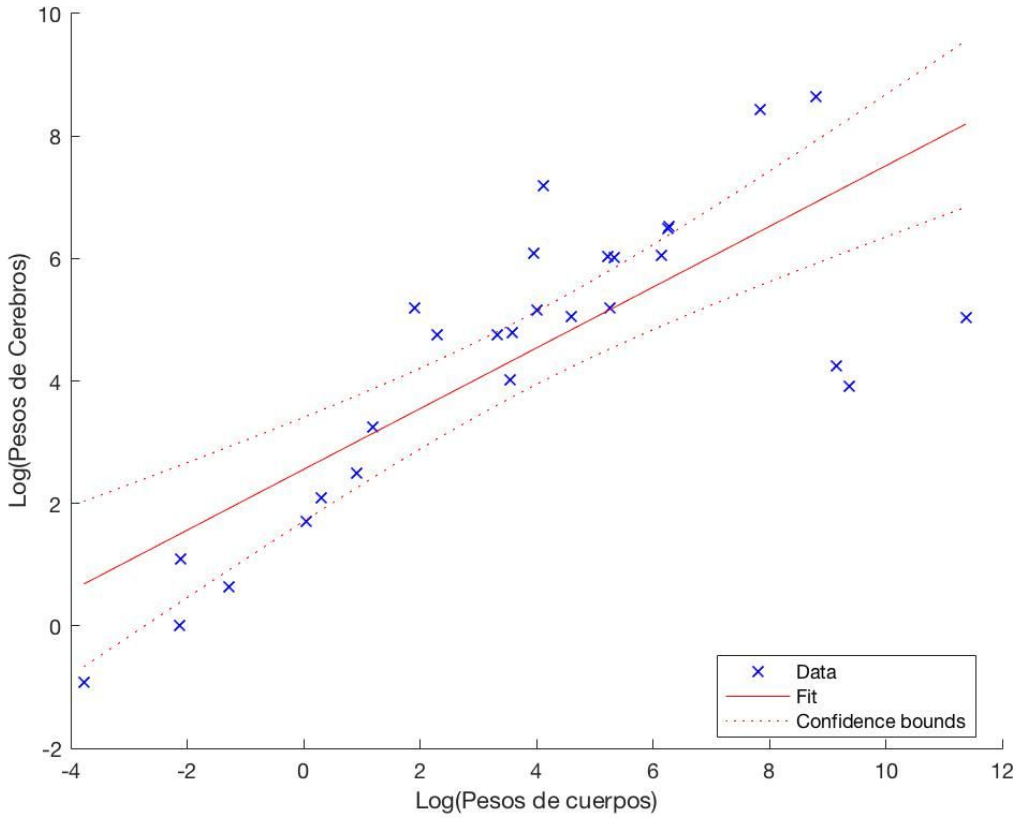
En este gráfico no se puede apreciar una relación lineal en los datos.

b)



En este gráfico se tomó el logaritmo de los valores dados, por lo que se puede apreciar una linealidad en los mismos.

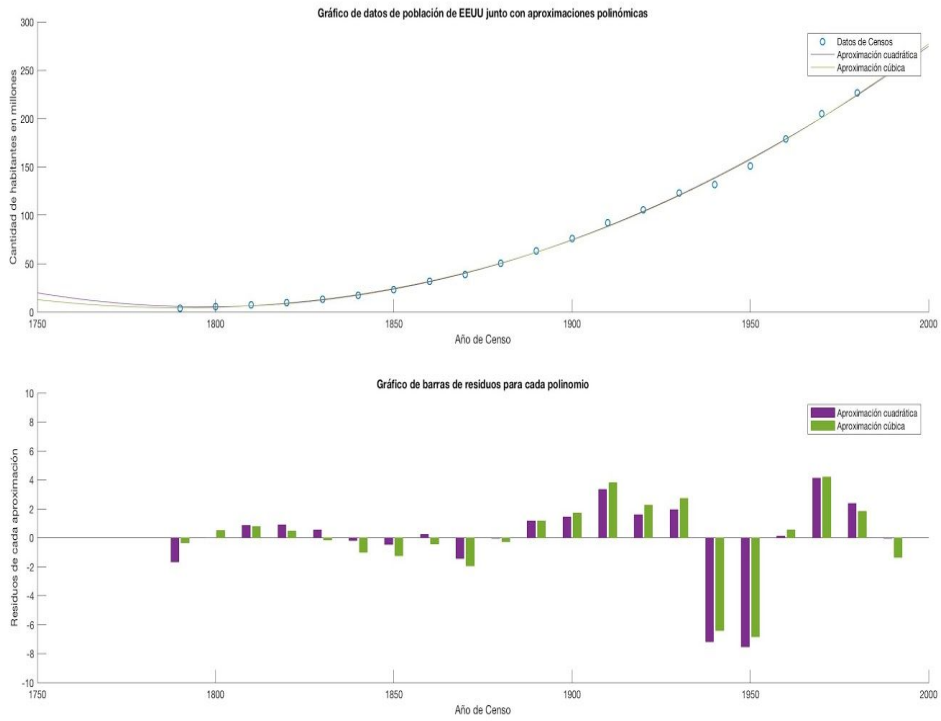
c)



En este gráfico se encuentran los valores del gráfico anterior, junto con la recta de regresión correspondiente.

d) Para este punto, se procedió a comparar el valor de R^2 de la recta de regresión que incluye a los puntos 14,15 y 25 con la recta que no los incluye. En el primer caso R^2 da 0.607610111442041, mientras que en el segundo da 0.680491182163475, lo que da la conclusión de que la recta que no considera los datos 14,15 y 25 ajusta mejor que la recta que sí los considera.

Ejercicio 7



Se calculó el valor de R^2 para cada uno de los polinomios. Se obtuvo un valor de 0.9988 para el polinomio cúbico y un valor de 0.9987 para el polinomio cuadrático, lo que indica que el polinomio cúbico es una mejor aproximación de los datos dados. Para estimar la población en el año 2000, el modelo cúbico fue más preciso, debido a que dio un valor de 277 millones, frente a los 275 millones que dio el polinomio cuadrático.

Ejercicio 8

a) Largo de sépalo:
   - Mu: 5.843333333333335
   - Sigma al cuadrado: 0.681122222222222

   Ancho de sépalo:
   - Mu: 3.057333333333334
   - Sigma al cuadrado: 0.188712888888889

   Largo de pétalo:
   - Mu: 3.758000000000003
   - Sigma al cuadrado: 3.095502666666667

Ancho de pétalo:

Mu: 1.199333333333334

Sigma al cuadrado: 0.577132888888889

b) Con los los valores medidos, se puede obtener la siguiente covarianza muestral:

0.6857	-0.042	1.2743	0.5163
-0.0424	0.19	-0.3297	-0.1216
1.2743	-0.3297	3.1163	1.2956
0.5163	-0.1216	1.2956	0.5810

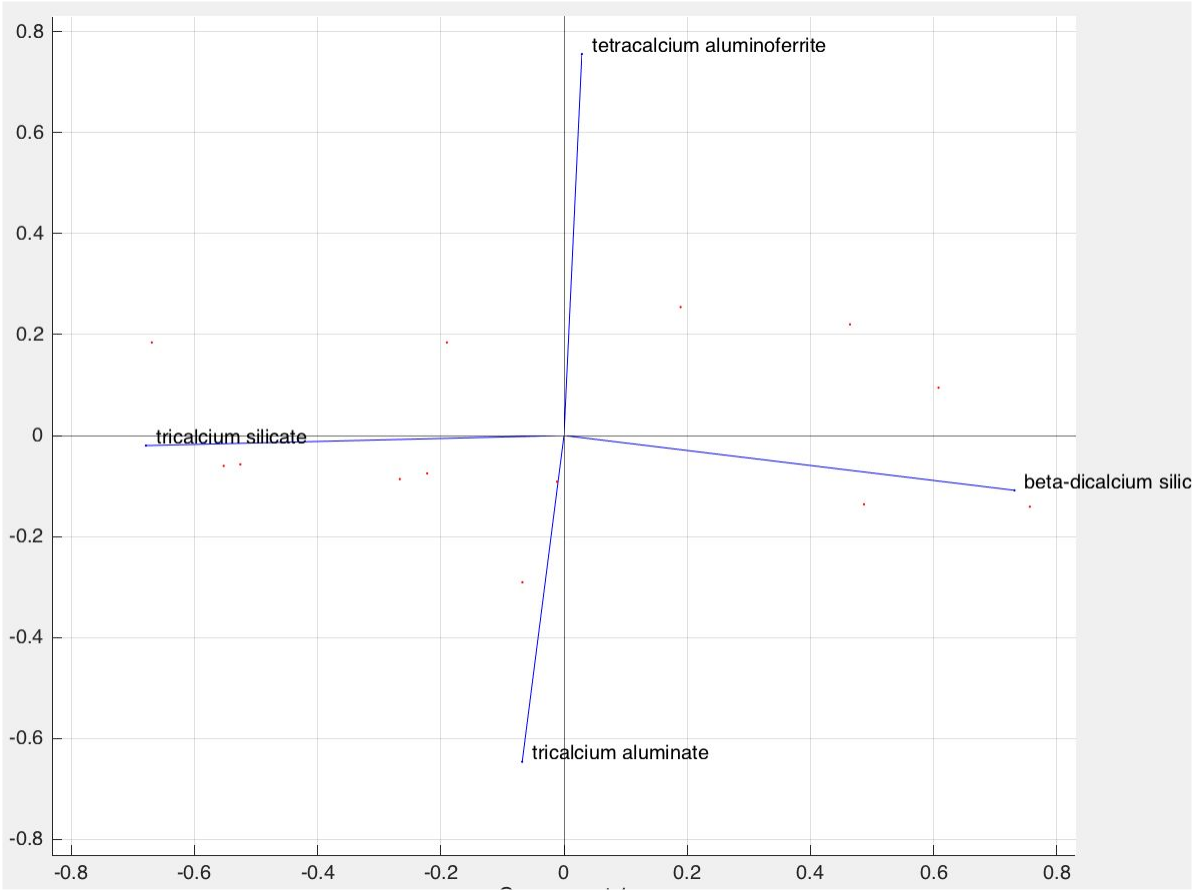
Ejercicio 9

a) La matriz de correlación para los distintos ingredientes resultó:

1	0.2286	-0.8241	-0.2454
0.2286	1	-0.1392	-0.973
-0.8241	-0.1392	1	0.0295
-0.2454	-0.973	0.0295	1

A partir de ella se pudo determinar que únicamente se encontraron variables con algún tipo de correlación cuando se analizaron los segundo y tercer, o tercero y primer ingredientes juntos. En cualquiera de los otros casos, se ve que no hay más que una mínima correlación en módulo, lo cual nos lleva a entender que el conjunto como un todo no se encontraba correlacionado entre variables.

b) Se utilizó el comando *stepwisefit* para aplicar el método de selección hacia adelante, obteniendo como resultado que los parámetros que debían utilizarse eran dos: *tricalcium aluminate* y *beta-dicalcium silicate.* Este resultado se ve respaldado por el siguiente gráfico:



Biplot de los componentes.

Ejercicio 10

Tras realizar un análisis de componentes sobre los ingredientes del dataset ‘hald’, se obtuvo que cada uno de ellos describen la varianza del conjunto total de la siguiente manera:

Componente	Varianza representada	Porcentaje total de la varianza
C1	517.7969	86,6%
C2	67.4964	11.29%
C3	12.4054	2.07%
C4	0.2372	0.004%

A partir de esto se pudo observar que casi la totalidad de la varianza puede ser descripta por tan solo los primeros dos de los componentes principales, alcanzando entre estos dos el 97.89% de la varianza total. Por lo tanto se procedió a realizar un biplot con tan solo estos dos componentes, obteniendo: