APRENDIZAJE ESTADÍSTICO

2º CUATRIMESTRE 2020

Trabajo Práctico - Clasificación

AUTOR

Giampieri Mutti, Leonardo - #102 358
<lgiampieri@fi.uba.ar>

DOCENTE

García, Jemina María

Índice

1. Introducción

En el siguiente trabajo, se analiza el set de datos **Iris**, con el fin de clasificar a las muestras en las diferentes especies de plantas, utilizando métodos de análisis discriminantes, **LDA** y **QDA**. A continuación, se muestra un pequeño extract del set.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
133	6.40	2.80	5.60	2.20	virginica
15	5.80	4.00	1.20	0.20	setosa
32	5.40	3.40	1.50	0.40	setosa
27	5.00	3.40	1.60	0.40	setosa
20	5.10	3.80	1.50	0.30	setosa
68	5.80	2.70	4.10	1.00	versicolor
87	6.70	3.10	4.70	1.50	versicolor
28	5.20	3.50	1.50	0.20	setosa
50	5.00	3.30	1.40	0.20	setosa
16	5.70	4.40	1.50	0.40	setosa
134	6.30	2.80	5.10	1.50	virginica
132	7.90	3.80	6.40	2.00	virginica
57	6.30	3.30	4.70	1.60	versicolor
72	6.10	2.80	4.00	1.30	versicolor
45	5.10	3.80	1.90	0.40	setosa

La primera columna simplemente muestra los ids de las muestras tomadas al azar del set. El resto de las variables explicativas indican, en centímetros, el largo y el ancho de los sépalos y los pétalos para cada especie de planta. Con esta información, se busca entrenar modelos para poder crear fronteras entre los datos. Estas fronteras son formas cuadráticas o hiperplanos, dependiendo del método utilizado, que segmentan el espacio en el cual los datos están embebidos según la clase a la que pertenezcan. Para poder construirlas, se toma el enfoque Bayesiano, en el cual se asume una probabilidad a priori para cada dato. Esto es, se le asigna una probabilidad de pertenecer a una clase. Además, se asume la distribución conjunta de las variables explicativas dentro de cada población, lo que nos permite construir las funciones discriminantes basadas en las distribuciones a priori. Las fronteras son las reglas de clasificación.

2. Primera visualización de datos

En esta sección se busca tener una idea de como se ven los datos.

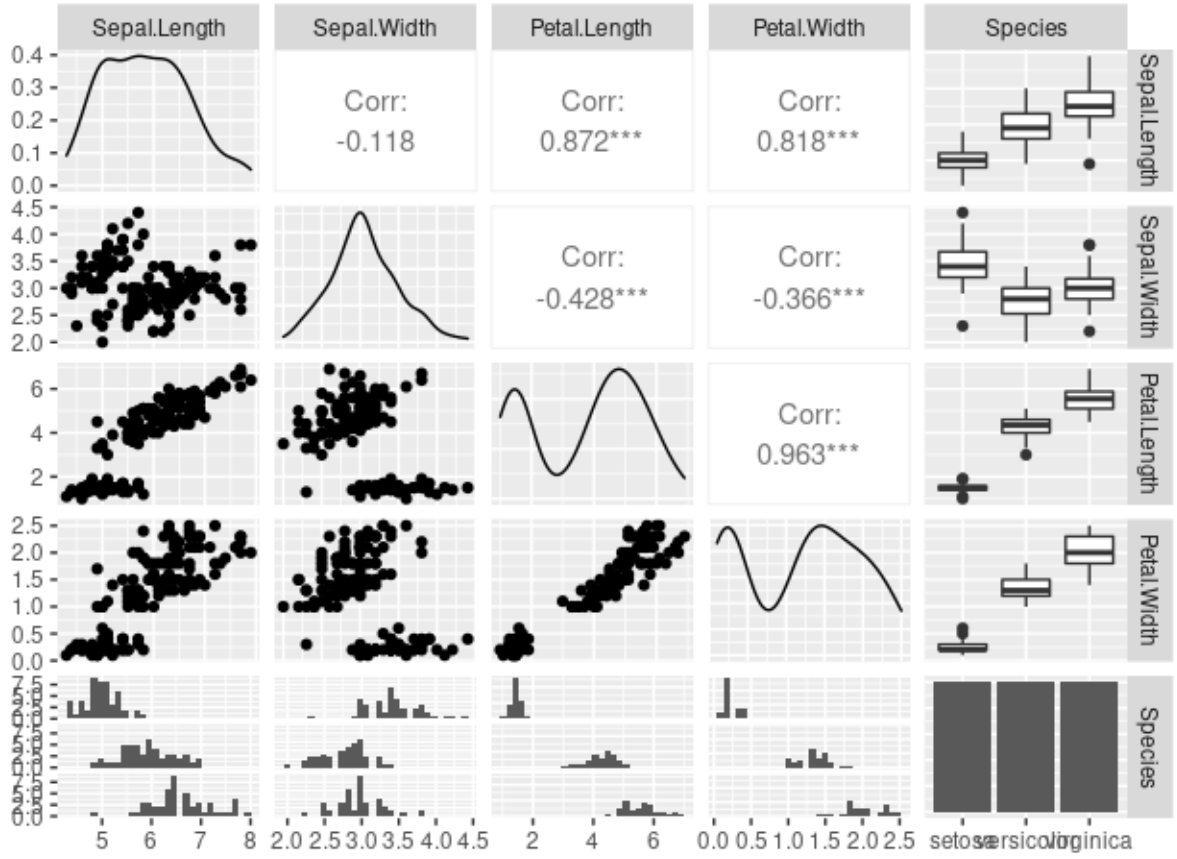

Figura 2.1

En la diagonal principal del pairplot, se pueden observar los gráficos de densidad de las variables explicativas para las poblaciones. Si bien no explican el comportamiento propio de cada población particular, muestran en general la concentración de los datos para cada variable. Esto se ve mucho más claro en los boxplots de la derecha. Zoom.

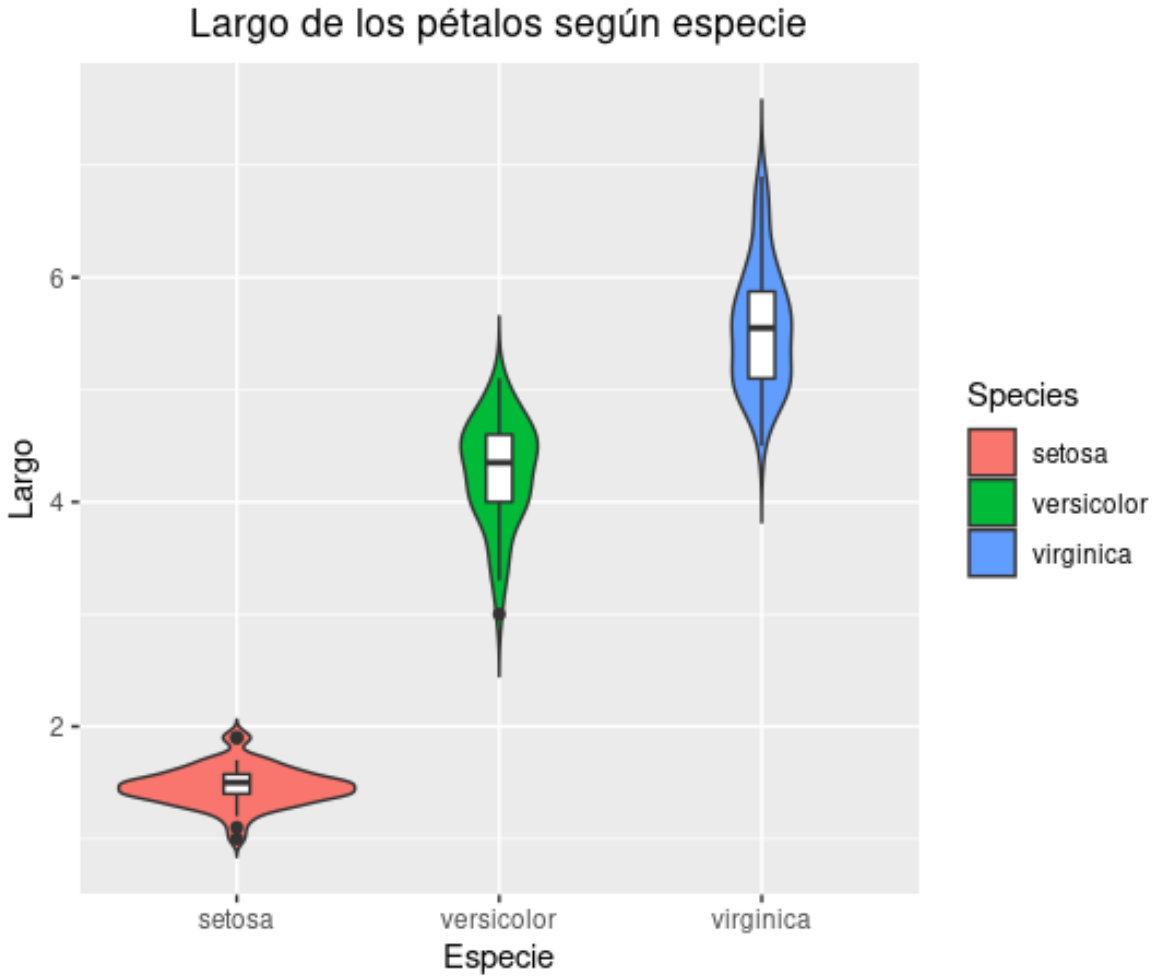

Figura 2.2

Por ejemplo, *Petal.Length* parace ser una variable muy importante para separar a las especies, ya que para *Setosa* es mucho menor que para *Versicolor* y *Virginica*.

El gap del medio se da debido a la diferencia que hay entre el largo de los pétalos de Setosa y Versicolor, que es exáctamente lo que se visualiza como diferencia vertical en el violin plot para esas dos especies.

Hay un pico en los valores más chicos, correspondientes a las plantas de especie Setosa, y un pico más ancho en los valores más grandes, correspondiente a las especies Versicolor y Virginica. El ancho se da a que hay una mayor concentración de datos para ese rango, ya que la diferencia del largo de los pétalos es menos marcada entre Versicolor y Virginica. De todas maneras, vemos que esta variable es indispensable para poder diferenciar las especies, ya que Setosa queda definida por el largo de sus pétalos. Una separación podría darse por esta variable: si el largo de los pétalos es menor a 2cm, entonces la planta corresponde a la especie Setosa. De esta manera, y sin mucho esfuerzo, se consigue un clasificador que le pega por lo menos a un tercio de los datos del set.

3. Clasificación lineal

En esta primera clasificación, se utilizan todas las columnas del set de datos, asumiendo que la varianza entre las clases es la misma. Por ende, el método que usamos es **LDA** para 3 clases. No se pueden visualizar las regiones ya que los datos estan embebidos en $\mathbb{R}^4$, pero lo que si se puede hacer es ver la proyección sobre los ejes principales. Para ver la performance del modelo se calcula la matriz de confusión.

3.1. LDA - Error aparente

Para entrenar el modelo, utilizamos todo el set de entrenamiento, y se predice sobre el mismo. Las probabilidades a priori son de $1/3$ para cada clase, es decir, equiprobables, y se asume que la distribución a priori de los datos en cada clase es normal. Las proyecciones obtenidas son las siguientes.

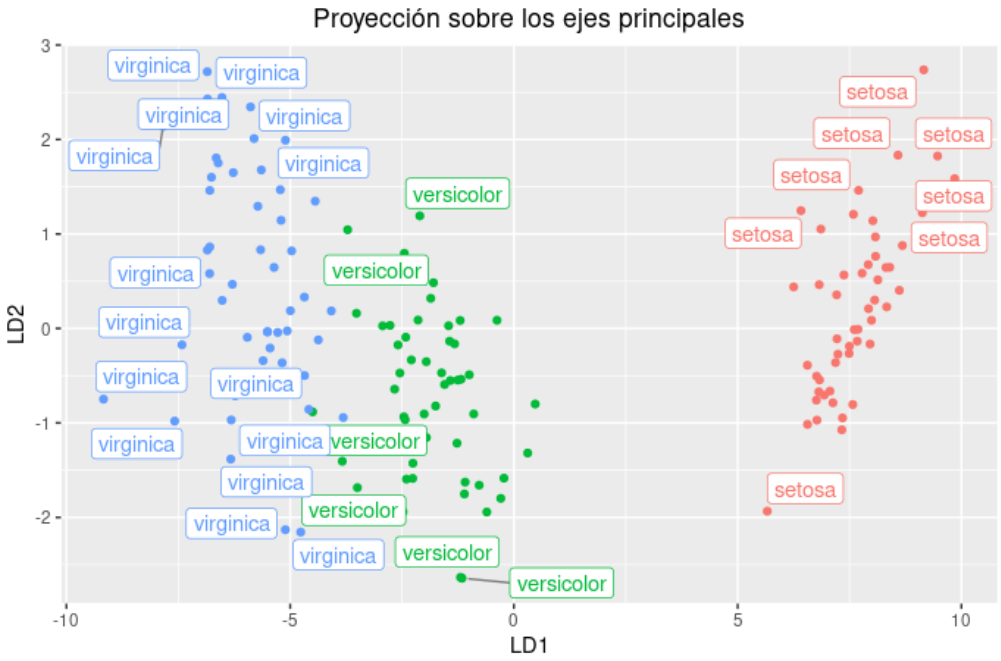

Figura 3.1

Se nota claramente cómo LDA logra clusterizar al conjunto de datos en 3 regiones, siendo Setosa la más distinguible. Como se vió antes, un buen criterio para discernir esta especie del resto es el largo de sus pétalos. De hecho, el mejor en cuanto a los datos disponibles del set. *LD1* y *LD2* son los ejes principales sobre los cuales LDA proyecta a los datos. La cantidad siempre es $k-1$, donde k es el cardinal del conjunto de clases. Estos ejes se buscan con la idea de maximizar la varianza entre clases y minimizar la varianza en los clusters. Además, se busca que la distancia de estos clusters a un punto central sea máxima, para que esten lo más separados posibles. De proyectar sobre LD1, se vería muy poco overlapping, seria de algunos puntos de Virginica y Versicolor. Sin embargo, de proyectar sobre LD2, habría muy

poca separación. Por esta razón, el eje principal es LD1. Sin embargo LD2 es importante, ya que reduce el overlapping al que se hizo referencia antes.

La matriz de confusión del modelo es la siguiente.

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	1
Virginica	0	2	49

El modelo clasifica mal 3 datos, 2 de los cuales asigna a la clase Virginica siendo estos Versicolor, y 1 de la clase Virginica que asigna a la clase Versicolor. El error aparente entonces es de 2 %.

El inconveniente de este modelo es que se testea con los mismos datos que entrena, con lo cual no se puede generalizar la exactitud del modelo fuera del set de entrenamiento. Esto puede introducir sesgo de selección y overfitting. Es decir, podemos entrenar el modelo y obtener altísimos resultados en cuanto a la precisión pero al momento de generalizar, el modelo no logra reproducir los mismos resultados. Para remediar esto, se usa cross-validation.

3.2. LDA - Cross-Validation

Se valida ahora el modelo mediante validación cruzada. La idea es separar el set de entrenamiento en dos subconjuntos disjuntos, el subset de entrenamiento y el subset de validación. El modelo predice sobre el subset de validación, con datos que nunca vió, y de tal manera se generaliza el poder predictivo del mismo.

El split sobre el set es de 70/30, realizado al pseudo azar.

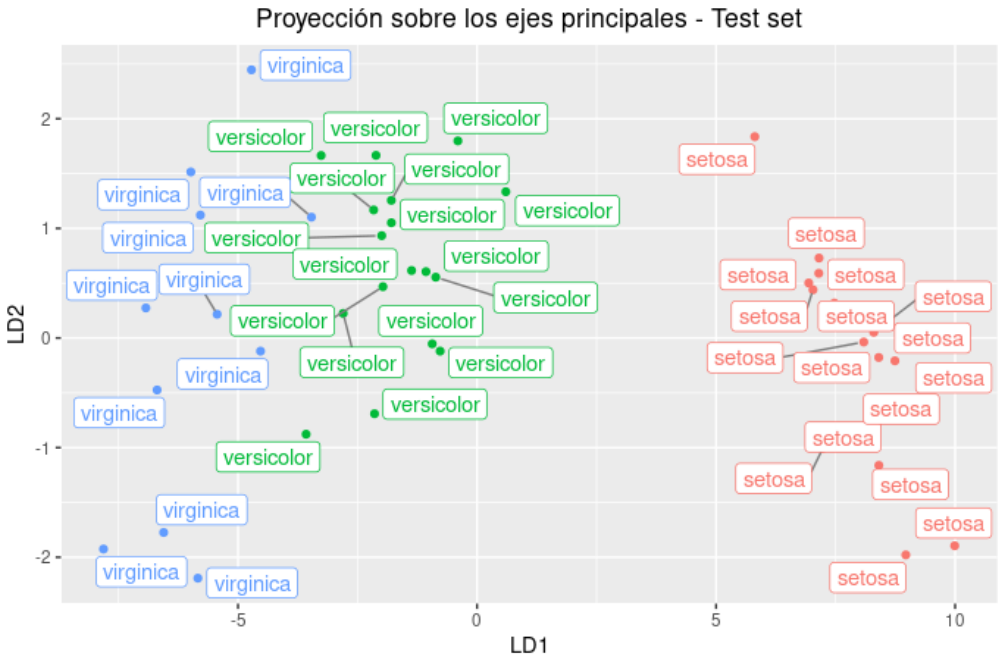

Figura 3.2

La forma en la cual separa los datos es muy similar a la del gráfico anterior. El error cometido sobre el set de test se muestra en la matriz de confusión.

	Setosa	Versicolor	Virginica
Setosa	17	0	0
Versicolor	0	16	1
Virginica	0	1	10

En este caso, el modelo comete sobre el set de test 2 errores, con lo cual el error aparante de testeo es de 4,4 %.

En el set de train, el error cometido es menor.

	Setosa	Versicolor	Virginica
Setosa	33	0	0
Versicolor	0	32	0
Virginica	0	1	39

El modelo comete solo un error, por lo que el error aparante es de 0,95 %, lo cual es menor que el error cometido por el modelo sin hacer el split de validación. Lo que hay que tener en cuenta es que la precisión en el set de validación es menor que en el set de entrenamiento. Esto significa que el modelo overfittea los datos un poco. Sin embargo, el error que se calcula sobre este modelo es independiente del set de entrenamiento, lo cual es deseable, ya que permite saber si el modelo generaliza bien. El intervalo de confianza de nivel 95 % para la precisión sobre el set de test es $(0{,}8485, 0{,}9946)$. De tener más datos para testear, se achicaría el largo del mismo.

4. Clasificación cuadrática

En esta sección, se analiza el entrenamiento de un modelo descartando la homocedasticidad entre clases. La separación ahora se dará por los cortes de nivel de una forma cuadrática.

4.1. QDA - Error aparente

Primero, se entrena y se testea al modelo sobre el set de entrenamiento. La matriz de confusión para este modelo es la siguiente.

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	48	1
Virginica	0	2	49

Se cometen los mismos errores que con LDA sin CV, por ende el error aparante de este modelo es 2 %. Hasta el momento parace no agregar más información.

4.2. QDA - Cross-Validation

Se prueba QDA pero con el set de validación para ver cómo generaliza el modelo. De nuevo el split que se hace sobre el set de entrenamiento es de 70/30.

	setosa	versicolor	virginica
setosa	17	0	0
versicolor	0	9	0
virginica	0	1	18

El modelo solo comete un error, por lo que el error aparante es 0,29 %, menor que el que comete LDA. Tomando este dato, se concluye que este es el mejor modelo, pero no por mucho. Para elegir entre los modelos se debe hacer un balance entre qué tan bueno es su desarrollo y el esfuerzo de máquina requerido por cada algoritmo. En este caso no se nota porque los datos son pocos, de modo que se elige éste como el mejor modelo.

4.3. K-Fold Cross-Validation

Ahora se prueban varios modelos, segmentando el set de entrenamiento en k particiones disjuntas. El modelo entrena sobre $k-1$ de éstas y valida sobre la partición restante. El error aparante se calcula promediando el error cuadrático medio que se comete en cada partición.

Se descarta $k = 1$, ya que equivale a testear el modelo directamente sobre el set de test sin previamente entrenar, por lo cual se esperaría que devuelva las probabilidades a priori para cada especie.

4.3.1. LDA

Haciendo K-Fold con LDA, se obtuvo el mejor modelo con $k = 145$. Esto muestra una tendencia hacia *leave-one-out-CV*, por ende los resultados del modelo serán más variados y menos sesgados. La precisión obtenida es de 0,9885621 sobre los folds de validación.

4.3.2. QDA

Con QDA, se obtuvo el mejor fold con $k = 135$, nuevamente cercano a *loocv*. La precision que se obtuvo es de 0,9833333 sobre los folds de validación, un poco menor que con LDA.

5. Sepal Width v. Petal Width

En esta sección, se busca entrenar modelos utilizando solo dos variables del set, el largo de los pétalos y de los sépalos. A diferencia de los modelos anteriormente testeados, se hace primero una verificación sobre los supuestos del análisis discriminante. LDA y QDA se construyen sobre el supuesto de que los datos de entrada se distribuyen de manera normal. Las funciones discriminantes se derivan asumiendo que las funciones de densidad provienen de distribuciones normales. Para validar los modelos, se debe checkear que se cumplen los supuestos. Asi se ven los datos en el plano.

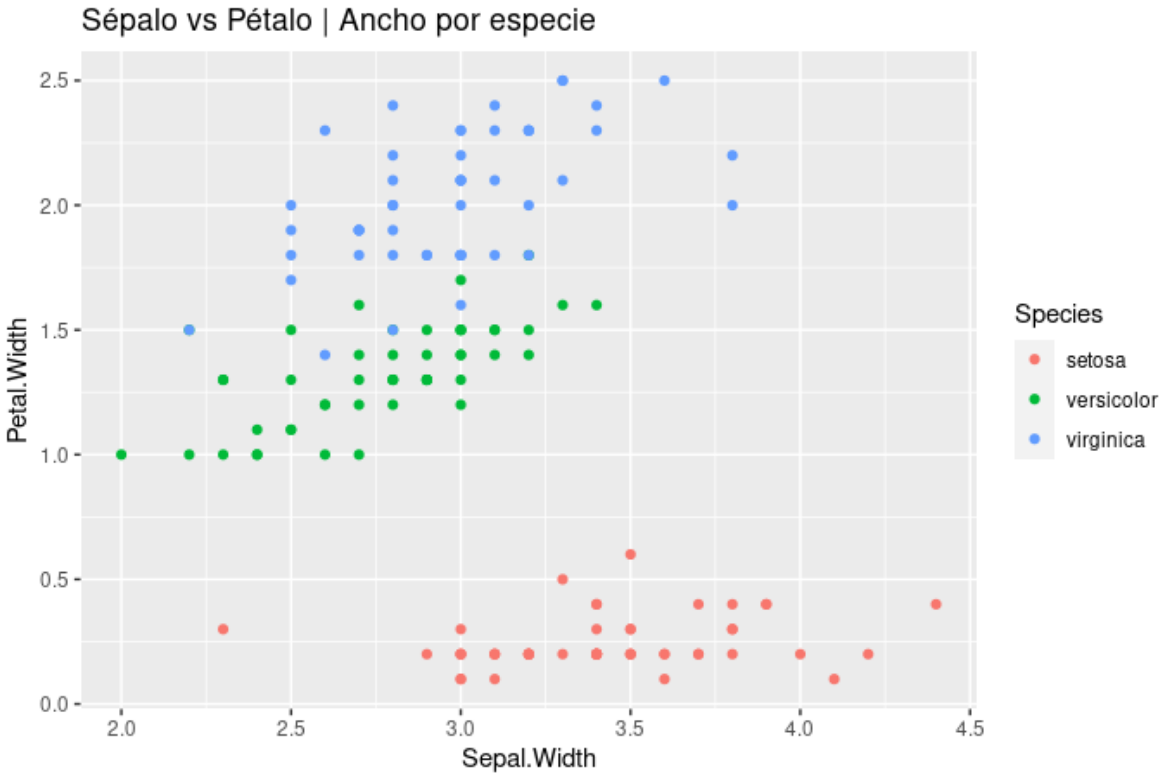

Figura 5.1

Verificamos si la variable aleatoria $(Sepal.Width, Petal.Width)$ es normal bivariada para cada especie.

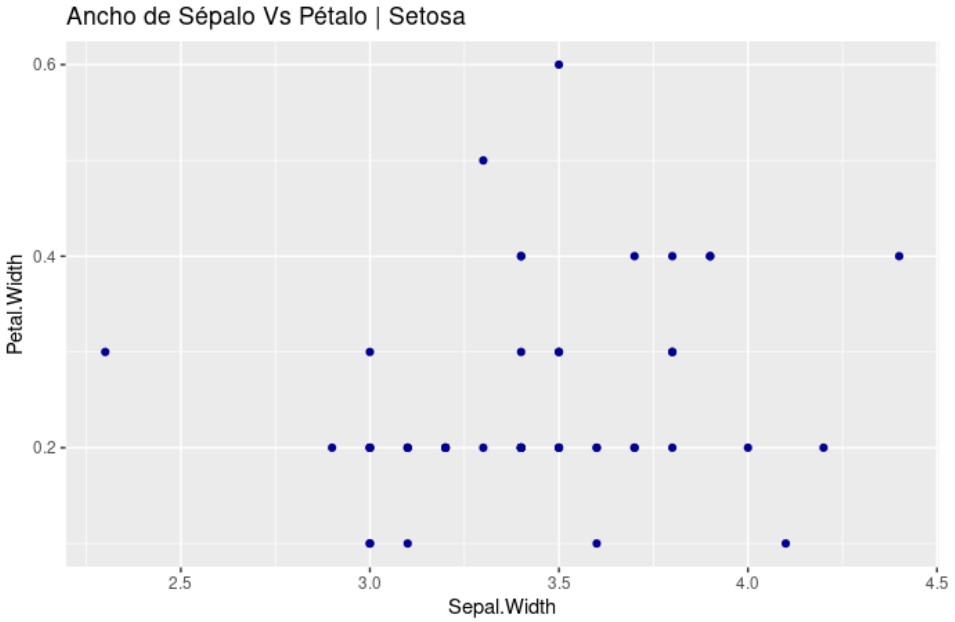

Figura 5.2

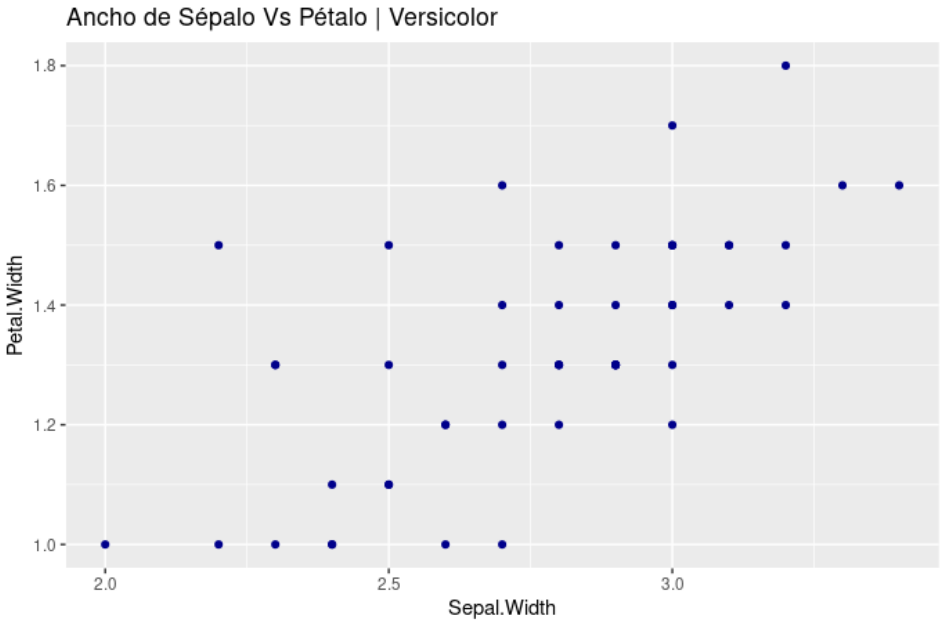

Figura 5.3

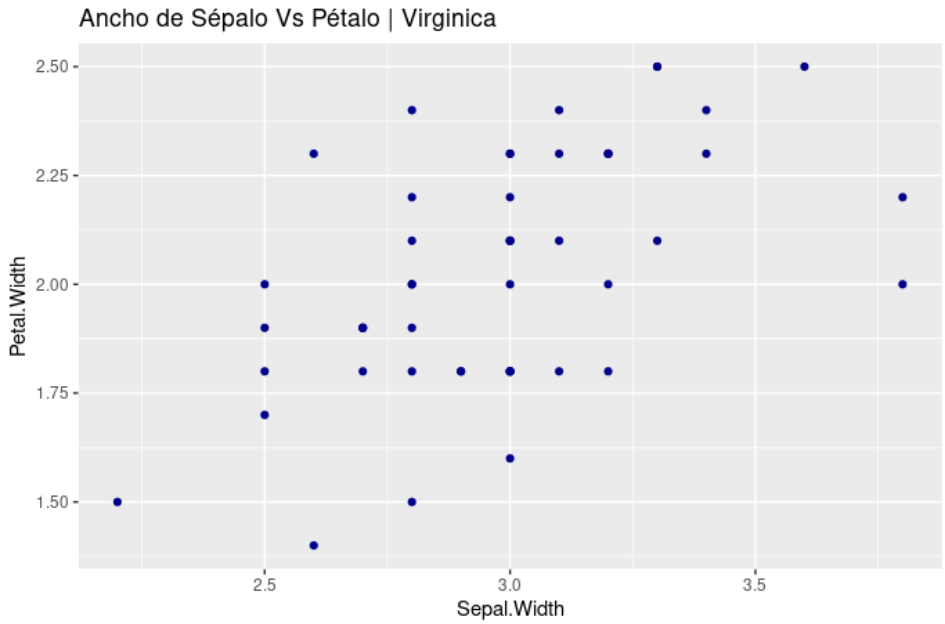

Figura 5.4

Una manera interesante de pensar si los datos pueden provenir de una normal bivariada es pensar en los cortes de nivel de la gráfica. Estos corresponden a elipses. Los datos deberían ajustarse a la forma de una elipse, orientada de cualquier manera.

En el caso de la especie setosa, es dificil ver esta elipse. Muchos de los datos se mantienen constantes para el valor de $Petal.Width = 0{,}2$. Esperariamos ver mas concentración de datos sobre el centroide del cluster, de ser normal bivariada. La primera impresión es que la distribución no es normal bivariada para setosa.

Para versicolor, los datos parecen ajustarse más a una elipse, con el eje de simetría mayor en diagonal, una recta con pendiente positiva. La concentración de los datos es mayor sobre uno de los focos de la misma, por lo que se pueden pensar que la distribución tendría una asimetría negativa sobre el plano del eje mayor. En principio, se podría pensar que la distribucion es normal, pero la asimetría es visible.

En el caso de virginica, se puede pensar nuevamente en la elipse con el eje mayor en diagonal con pendiente positiva. Ahora los datos estarían concentrados sobre su centro, tirando hacía el foco izquierdo. Se puede pensar que es normal, pero es visible la asimetría.

Se muestran ahora los mapas de contorno y de calor de las densidades para cada especie, en busca de las elipses.

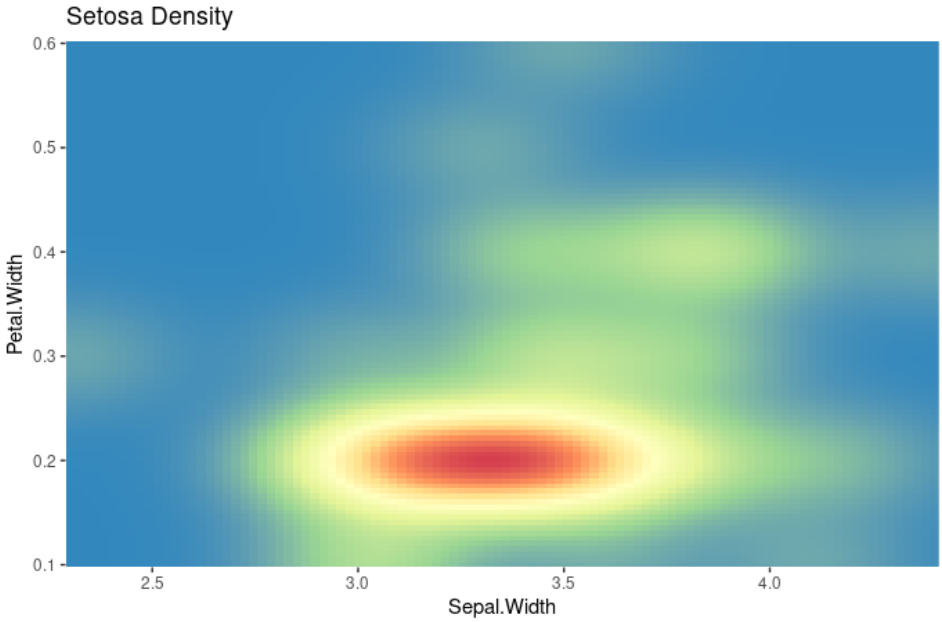

Figura 5.5

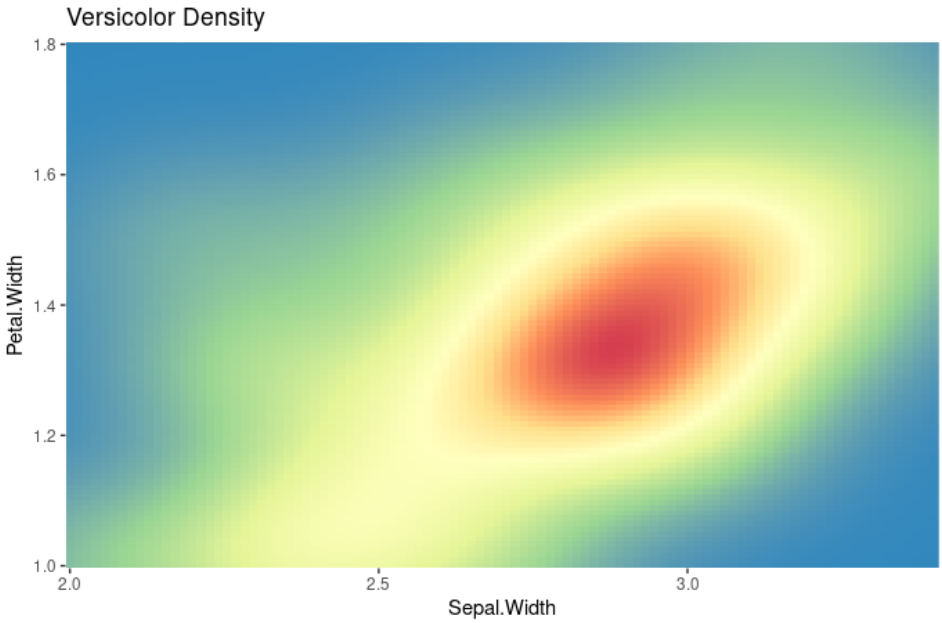

Figura 5.6

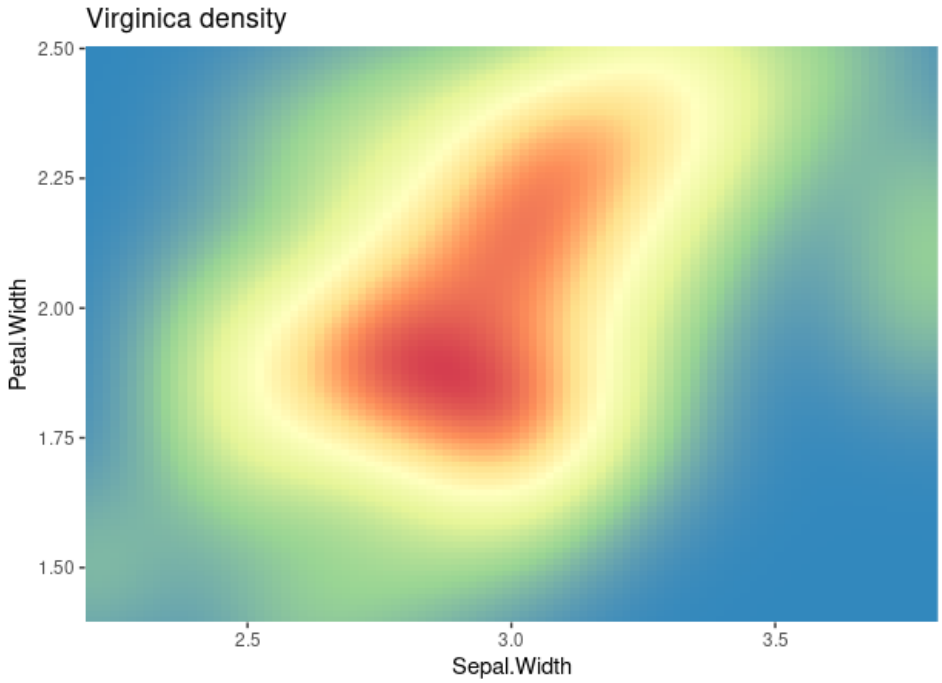

Figura 5.7

Los colores cálidos muestran valores mayores de la densidad, los frios lo contrario.

En el caso de setosa, el contorno parece una elipse achatada. Esto tiene sentido, ya que sobre planos paralelos al eje x hay mayor varianza que sobre aquellos paralelos al eje y.

Para versicolor, se puede ver un poco más clara la elipse, con su eje mayor en pendiente positiva, pero deformada hacia el foco izquierdo, ya que como se vio en el scatter, los datos se encuentran concentrados hacia el foco derecho.

Para virginica, se puede ver un zapallo anco, con su eje mayor de pendiente positiva. Se puede notar que el la tendencia de los datos es hacia el foco izquierdo, contrario a versicolor, y más marcado incluso, de ahí la forma de zapallo.

Finalmente se muestran los conjuntos de nivel de las densidades.

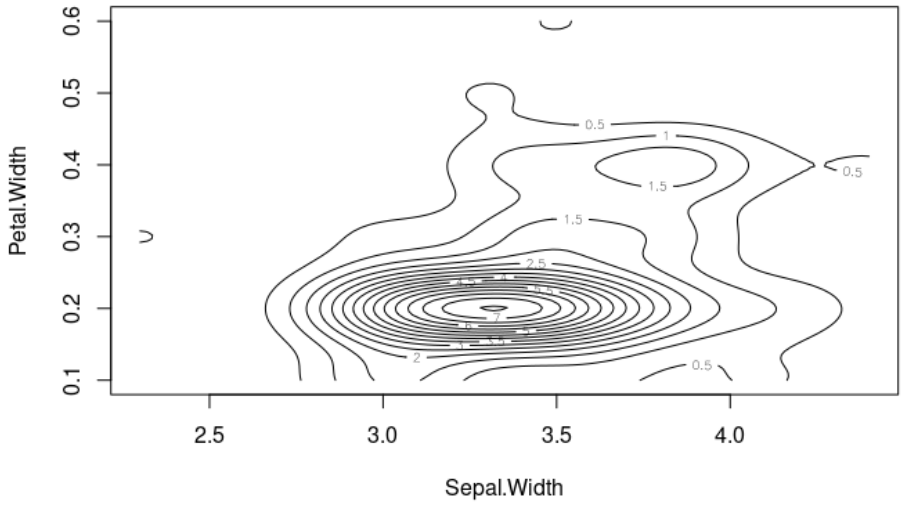

Figura 5.8

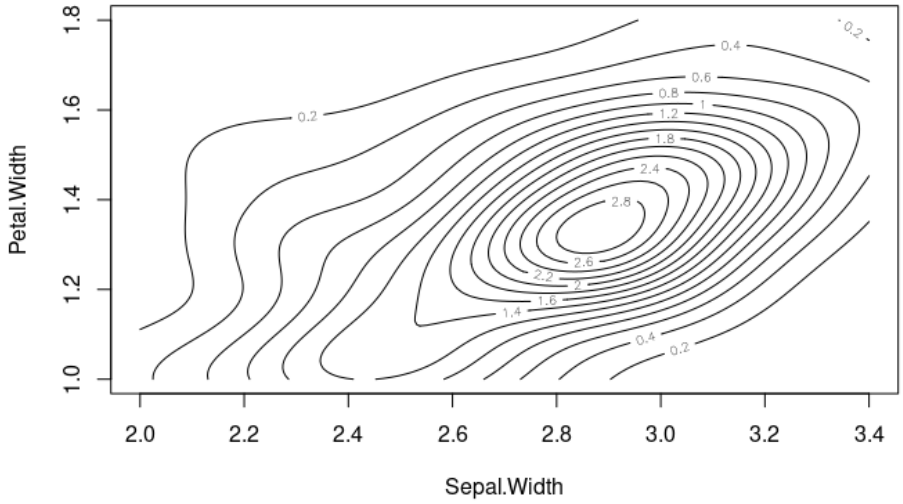

Figura 5.9

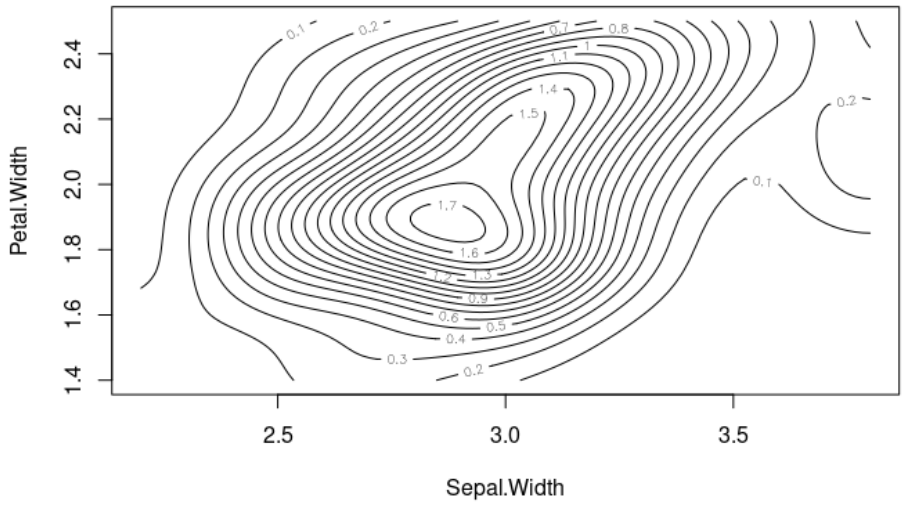

Figura 5.10

En las 3 figuras hay formas elípticas deformadas. Hasta el momento no parece que ninguna de las variables aleatorias cumplan ser normales bivariadas, de modo que se busca otra manera de verificar la normalidad (al menos, de no descartar la normalidad).

Para esto, se recurre a teoría de probabilidades. Si una variable conjunta $X = (X1, X2)$ se distribuye como una normal bivariada, entonces sus marginales son normales, de modo que si alguna de las marginales no es normal, seguramente X no será normal. Se muestran las densidades para cada variable.

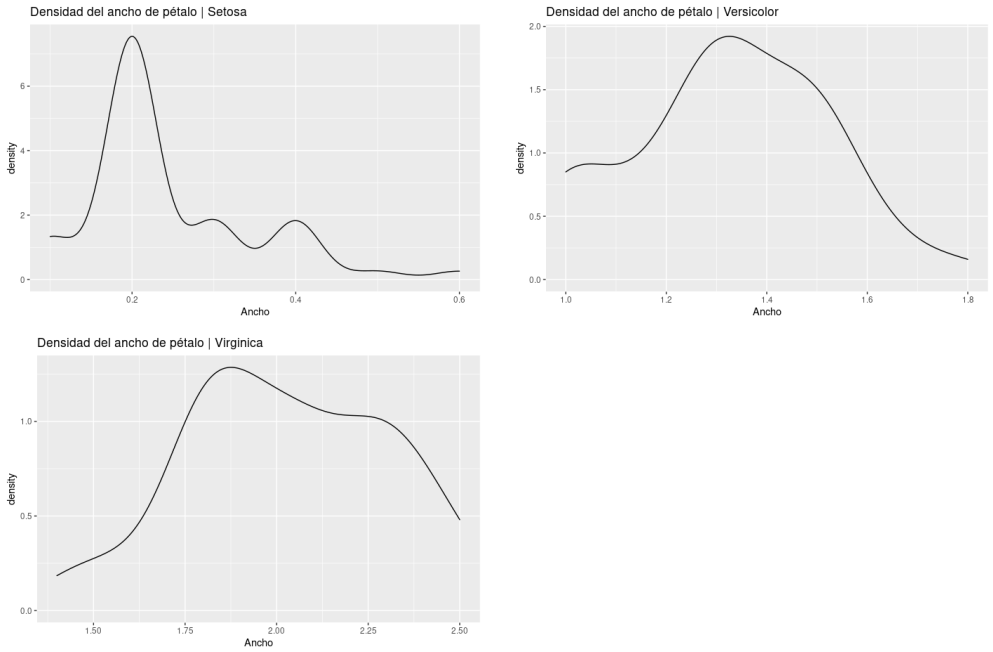

Figura 5.11

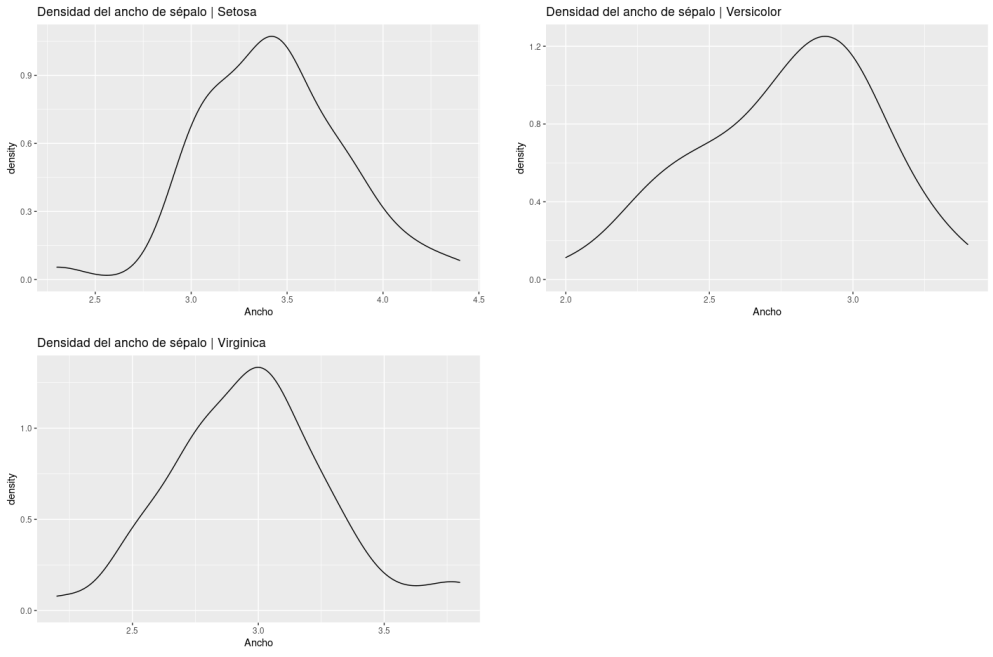

Figura 5.12

Ahora se nota, por ejemplo, que el ancho de los pétalos para setosa no puede ser normal, ya que esa densidad no se asemeja a una campana. Ídem, aunque menos claro, el ancho de

los sépalos para la especie versicolor. Para formalizar, se testea normalidad usando el test de Shapiro-Wilk, con una probabilidad de error de tipo I, $\alpha = 0{,}05$. La hipótesis nula es que la distribución de los datos de entrada es normal univariada.

	p-value
Setosa	8.659e-07
Versicolor	0.02728
Virginica	0.08695

Tabla 1: Test de normalidad para pétalos según especie

	p-value
Setosa	0.2715
Versicolor	0.338
Virginica	0.1809

Tabla 2: Test de normalidad para sépalos según especie

Con un nivel de significación de 0,05, rechazamos la hipótesis nula, esto es, que la distribución sea normal para el ancho de los pétalos de las especies setosa y versicolor. Sobre el resto de las variables no se puede decir nada. Por ende, se concluye que ni setosa ni versicolor pueden seguir una distribución normal bivariada. Para virginica, no hay evidencia suficiente para que los tests decidan.

5.1. Modelos

En esta sección, se asume que para las 3 especies, la distribución de $(Sepal.Width, Petal.Width)$ es normal bivariada. Primero, se asume que los datos no son homocedásticos, de modo que el modelo propuesto es QDA.

5.1.1. QDA

Luego de construir la regla de decisión, se busca clasificar al dato $\mathbf{x} = (3{,}5, 1{,}75)$.
Se muestran las probabilidades a posteriori para $\mathbf{x}$.

	Setosa	Versicolor	Virginica
1	0.00	0.78	0.22

Bajo el principio de parsimonia, se clasifica al dato en la clase Versicolor.
Vemos como QDA separa el espacio para las 3 regiones.

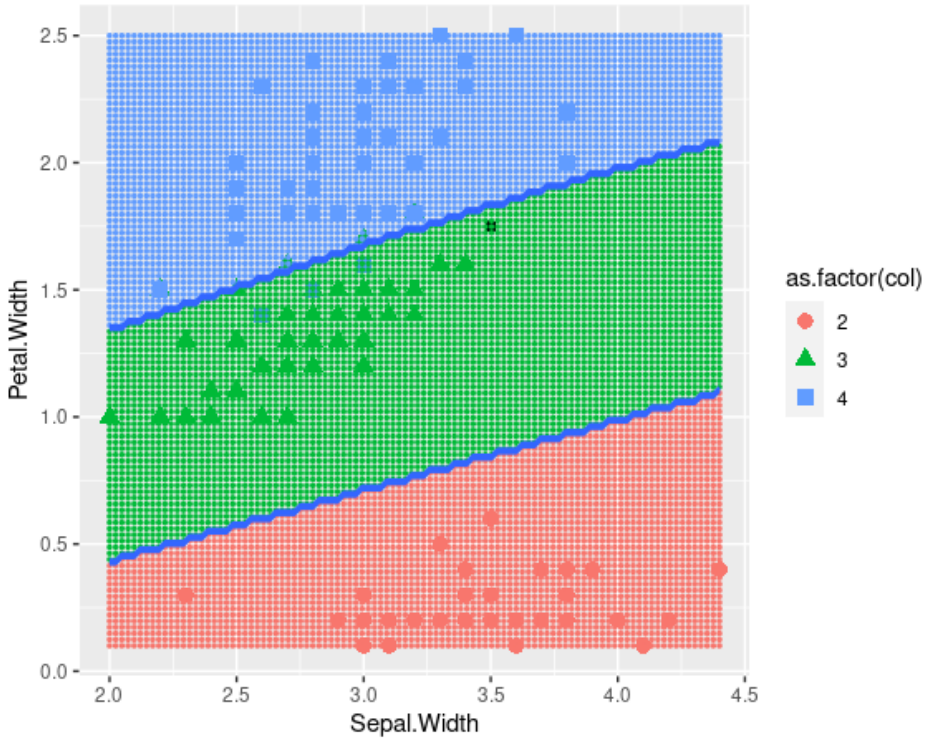

Figura 5.13

La región azul corresponde a virginica, la verde a versicolor, y la roja a setosa. El punto negro corresponde a **x**, y cae sobre la región de versicolor. Se puede apreciar, por más que sea sutil, cómo se curva la region verde en sus bordes.

La matriz de confusión para este modelo es la siguiente.

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	46	3
Virginica	0	4	47

5.1.2. LDA

El próximo modelo asume que las matrices de covarianzas son iguales para las 3 poblaciones. Las estimaciones de las matrices de covarianza son las siguientes.

	Sepal.Width	Petal.Width
Sepal.Width	0.14	0.01
Petal.Width	0.01	0.01

Tabla 3: Matriz de covarianza estimada - Setosa

	Sepal.Width	Petal.Width
Sepal.Width	0.10	0.04
Petal.Width	0.04	0.04

Tabla 4: Matriz de covarianza estimada - Versicolor

	Sepal.Width	Petal.Width
Sepal.Width	0.10	0.05
Petal.Width	0.05	0.08

Tabla 5: Matriz de covarianza estimada - Virginica

Se puede ver que no es precisamente el caso que sean iguales, pero de todas maneras se puede ajustar el modelo. La consecuencia directa de la no homocedasticidad y la no normalidad es que no se logra el máximo de la varianza entre grupos.

Para **x**, estas son las probabilidades a posteriori del modelo LDA.

	Setosa	Versicolor	Virginica
1	0.00	0.72	0.28

Al igual que QDA, LDA clasifica a **x** como Versicolor. Las probabilidades son apenas distintas comparando con el modelo anterior, pero por el formateo a dos cifras significativas se ven iguales.

LDA separa al espacio de la siguiente manera.

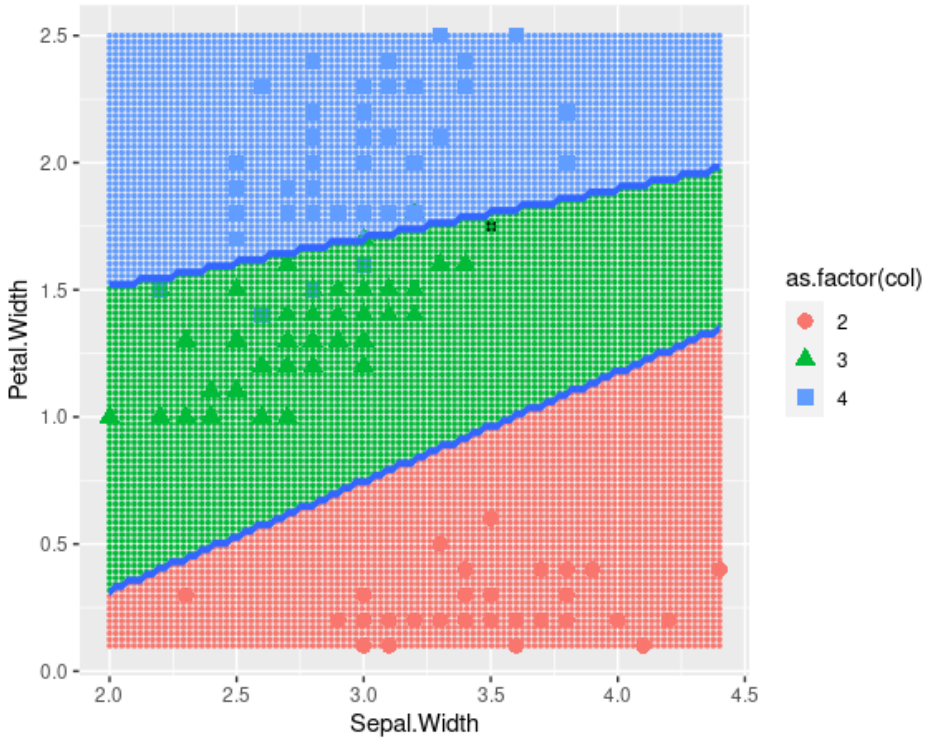

Figura 5.14

En este caso, el punto esta más cerca de la frontera. Entre ambos modelos hay diferencias sutiles sobre las clases Virginica y Versicolor. Setosa se clasifica exáctamente igual en ambos casos. La matriz de confusión para este modelo es la siguiente.

	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	49	4
Virginica	0	1	46

En este caso, LDA comete 3 errores menos que QDA. 3,3 % de los datos fueron mal clasificados.

Se estima el error por validación cruzada. El split sobre el set de entrenamiento es de 70/30, al azar. El error cometido por validación cruzada es 0,022.

6. Conclusión

Se vió que en muchos de los modelos, LDA igualó e incluso superó a QDA como clasificador, pero que ambos tienen buen score, incluso droppeando 2 de las variables explicativas y sin cumplir estrictamente la normalidad de los datos o la heterocedasticidad entre clases en caso de QDA.

7. Anexo

Repositorio en donde se encuentran el código y los plots utilizados.
https://github.com/leogm99/Aprendizaje_Estadistico_tp2