Introducción

1EST17 - Aprendizaje Estadístico I

Enver G. Tarazona

2022-09-03

Objetivos de Aprendizaje del curso

1. **Conocimiento.** El estudiante tiene conocimiento sobre los modelos y métodos de aprendizaje estadístico más populares que se utilizan para *predicción* e *inferencia* en ciencia y tecnología. El énfasis está en los modelos estadísticos de tipo regresión, clasificación y aprendizaje no supervisado.
2. **Habilidades.** El estudiante puede, en base a un conjunto de datos existente, elegir un modelo estadístico adecuado, aplicar métodos estadísticos sólidos y realizar los análisis usando software estadístico. El estudiante puede presentar, interpretar y comunicar los resultados de los análisis estadísticos y sabe qué conclusiones se pueden sacar de los análisis y cuáles son las limitaciones.

Sobre el curso

Enfoque: Teoría estadística **y** hacer análisis

- El curso se enfoca en la **teoría estadística**, pero aplicamos todos los modelos y la teoría utilizando (principalmente) funciones disponibles en R y conjuntos de datos reales.

- Es importante que el estudiante al final del curso **pueda analizar todo tipo de datos** (cubiertos en el curso), no solo entender la teoría.

- Y viceversa: el estudiante también debe **comprender** el modelo, los métodos y los algoritmos utilizados.

Estrategia pedagógica

- Dividir los temas del curso en unidades modulares con enfoque específico.
- Esto (esperamos) facilita el aprendizaje?
- Algunas semanas sin clases, dedicadas a evaluaciones (revisar cronograma).

Métodos de aprendizaje, actividades y evaluación

- Clases, ejercicios y actividades grupales (evaluación continua).
- Examen parcial y final (individual)
  - Aprox. el 50% de la nota está enfocada en teoría y el entendimiento de los modelos.
  - El puntaje restante evalúa la aplicación e interpretación de los modelos usando R.
- Fórmula de Evaluación : $\text{NF} = 0.4\text{EC} + 0.30\text{EP} + 0.30\ \text{EF}$ (ver silabo)

Módulo 1

Objetivos del primer módulo

- Una introducción al aprendizaje estadístico. ¿Qué es?
- Tipos de problemas que veremos
- Introducir notación y conceptos básicos preliminares

¿Qué es el aprendizaje estadístico?

- Se refiere a *un amplio conjunto de herramientas para comprender los datos* (libro de texto, p. 1).

- Distinción principal: *Supervisado* versus *aprendizaje no supervisado*.

- La "cadena'' del aprendizaje estadístico:

modelo $\rightarrow$ método $\rightarrow$ algoritmo $\rightarrow$ análisis $\rightarrow$ interpretación

- Tanto **predicción** como **inferencia** (comprender $\rightarrow$ sacar conclusiones).

- El aprendizaje estadístico es **una disciplina estadística**, pero los límites son cada vez más borrosos.

Aprendizaje estadístico versus "Aprendizaje automático'' (Machine Learning)

- El aprendizaje automático se centra más en la parte algorítmica del aprendizaje y es una *disciplina de las ciencias de la computación*.

- Pero muchos métodos/algoritmos son comunes a ambos campos.

Aprendizaje Estadístico vs. "Ciencia de Datos''

Ciencia de los datos

- Objetivo: extraer conocimiento y comprensión de los datos.
- Requiere una combinación de estadística, matemáticas, cálculo, ciencias de la computación e informática.

Esto abarca todo el proceso de

1. adquisición/extracción (scraping) de datos
2. pasar de datos no estructurados a datos estructurados
3. configurar un modelo de datos
4. implementar y realizar análisis de datos
5. interpretación y comunicación de resultados

En el aprendizaje estadístico no trabajaremos los dos primeros anteriores (adquisición y no estructurado a estructurado).

R for Data Science: Referencia importante sobre R.

Problemas que aprenderás a resolver

Hay **tres tipos principales de problemas** discutidos en este curso:

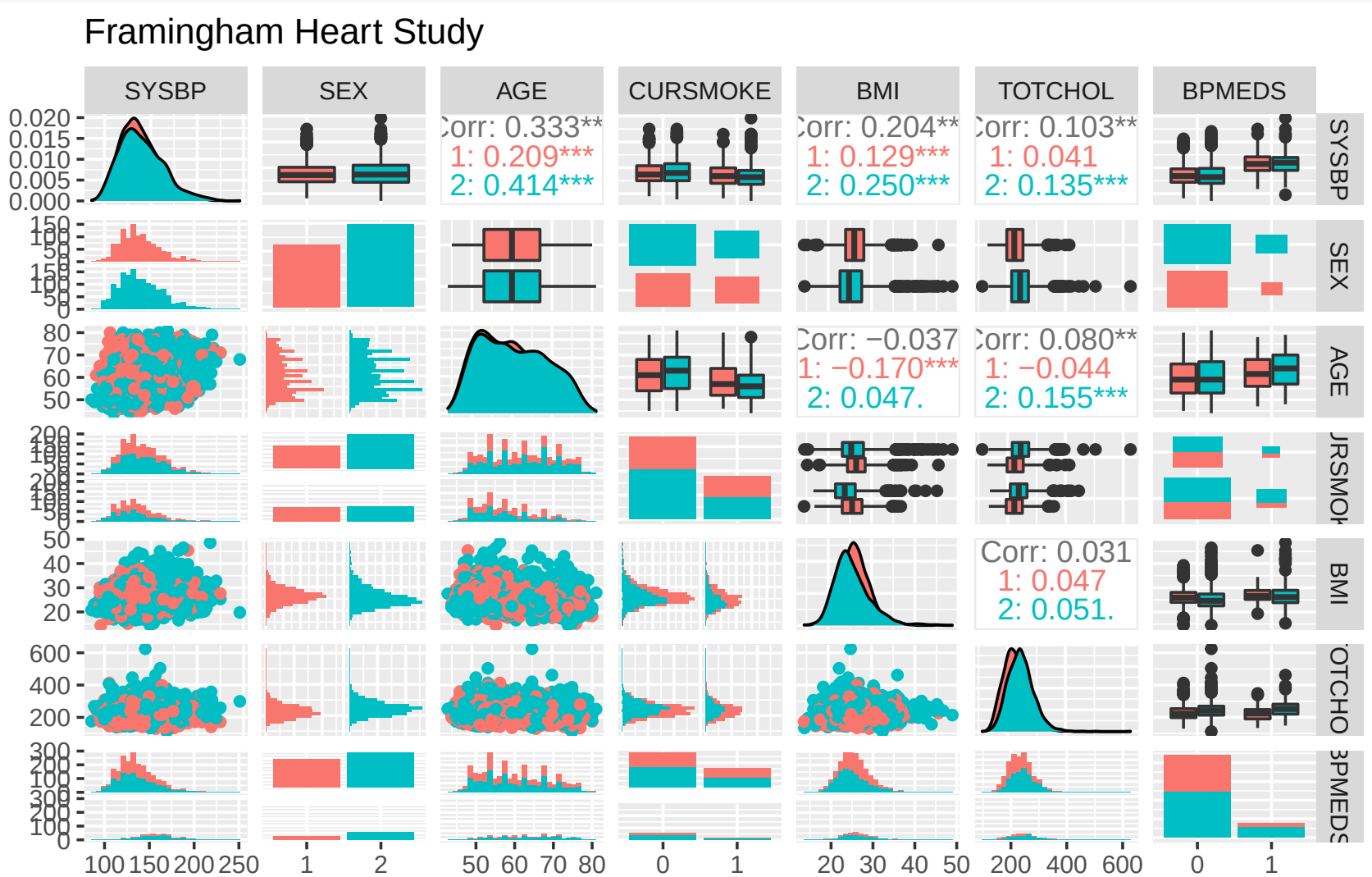
- Regresión (supervisado)
- Clasificación (supervisado)
- Métodos no supervisados

utilizando datos de ciencia, tecnología, industria, economía/finanzas, ...

Ejemplo 1: Regresión (Etiología de la ECV)

- El Framingham Heart Study investiga las causas subyacentes de las enfermedades cardiovasculares (ECV) (consulte https://www.framinghamheartstudy.org/).
- Objetivo: modelar la presión arterial sistólica (`SYSBP`) usando datos de $n = 2600$ personas.
- Para cada persona en el conjunto de datos tenemos mediciones de las siguientes siete variables.

- `SYSBP` presión arterial sistólica (mmHg),
- `SEX` 1=masculino, 2=femenino,
- `AGE` edad (años),
- `CURSMOKE` Tabaquismo actual en el examen: 0 = no fumador actual, 1 = fumador actual,
- `BMI` índice de masa corporal
- `TOTCHOL` colesterol total sérico(mg/dl),
- `BPMEDS` uso de medicamentos antihipertensivos en el examen: 0 = no los usa actualmente, 1 = los usa actualmente.



¿Que muestra la gráfica?

Rojo: masculino; turquesa: femenino

- Diagonal: gráfico de densidad (generalización del histograma), o gráfico de barras.
- Diagonales inferiores: diagrama de dispersión, histogramas
- Diagonales superiores: correlaciones, boxplots o barplots

Usamos `sexo` para colorear el gráfico.

Etiología de la ECV

La pregunta: **¿Cuáles son los factores que causan la PAS alta?**

$\rightarrow$ estamos interesados en la *inferencia* (explicación), ¡no en la predicción!

- Se ajustó un *modelo de regresión lineal normal múltiple* al conjunto de datos con

$$-\frac{1}{\sqrt{\texttt{SYSBP}}}$$

como respuesta (salida) y todas las demás variables como covariables (entradas).

- Los resultados se utilizan para formular hipótesis sobre la etiología de las ECV, que se estudiarán en nuevos ensayos.

```r
modelB = lm(-1/sqrt(SYSBP) ~ SEX + AGE + CURSMOKE + BMI + TOTCHOL + BPMEDS,
    data = thisds)

summary(modelB)
```

```
##
## Call:
## lm(formula = -1/sqrt(SYSBP) ~ SEX + AGE + CURSMOKE + BMI + TOTCHOL +
##     BPMEDS, data = thisds)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0207366 -0.0039157 -0.0000304  0.0038293  0.0189747
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.106e-01  1.342e-03 -82.413  < 2e-16 ***
## SEX2        -2.989e-04  2.390e-04  -1.251 0.211176
## AGE          2.378e-04  1.434e-05  16.586  < 2e-16 ***
## CURSMOKE1   -2.504e-04  2.527e-04  -0.991 0.321723
## BMI          3.087e-04  2.955e-05  10.447  < 2e-16 ***
## TOTCHOL      9.288e-06  2.602e-06   3.569 0.000365 ***
## BPMEDS1      5.469e-03  3.265e-04  16.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005819 on 2593 degrees of freedom
## Multiple R-squared:  0.2494, Adjusted R-squared:  0.2476
## F-statistic: 143.6 on 6 and 2593 DF,  p-value: < 2.2e-16
```

Ejemplo 2: Clasificación (plantas de iris)

El conjunto de datos de la flor del `iris` es un conjunto de datos multivariante muy famoso introducido por el estadístico y biólogo británico Ronald Fisher en 1936.

El conjunto de datos contiene

- **tres especies de plantas** {setosa, virginica, versicolor}
- **cuatro características medidas** para cada muestra correspondiente:
  - `Sepal.Length`
  - `Sepal.Width`
  - `Petal.Length`
  - `Petal.Width`.

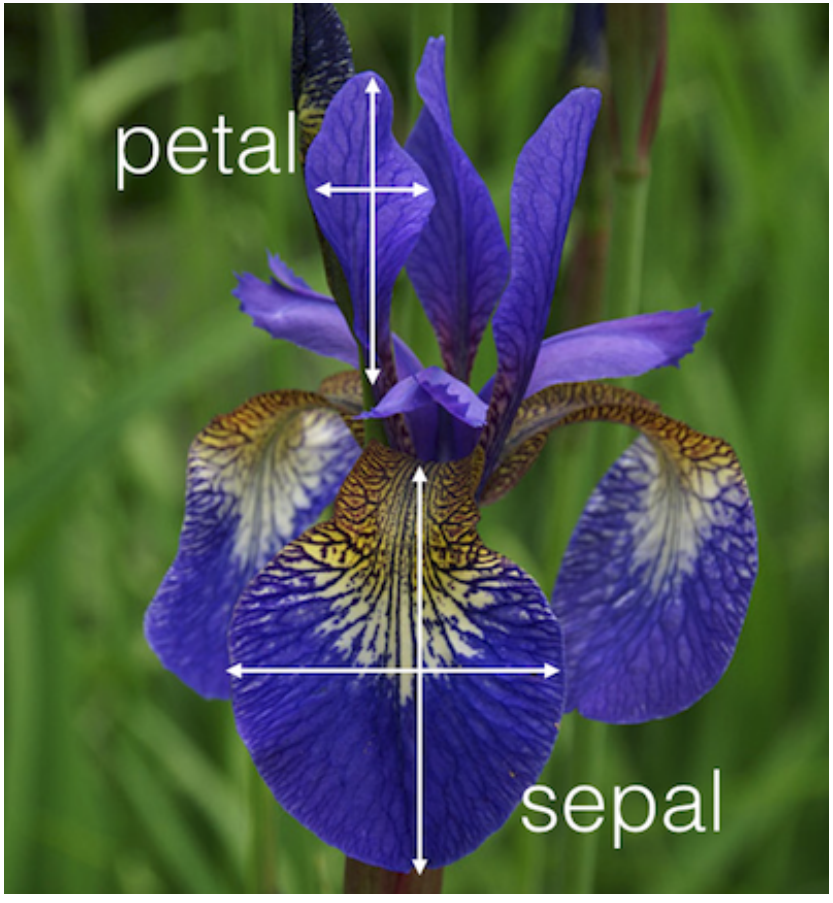


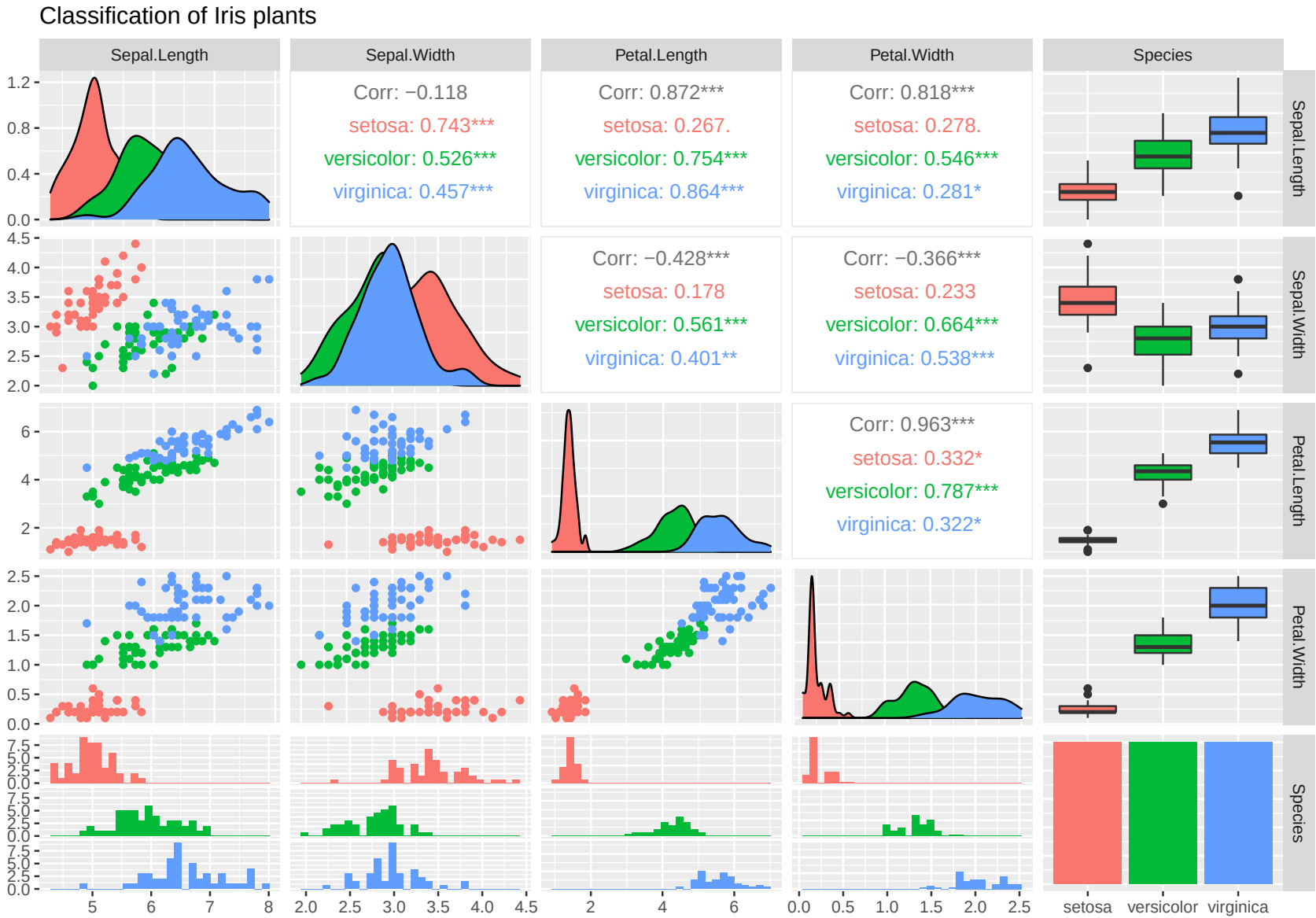
Figure 1: Planta de iris con hojas de sépalo y pétalo

http://blog.kaggle.com/2015/04/22/scikit-learn-video-3-machine-learning-first-steps-with-the-iris-dataset/

```
head(iris)
```

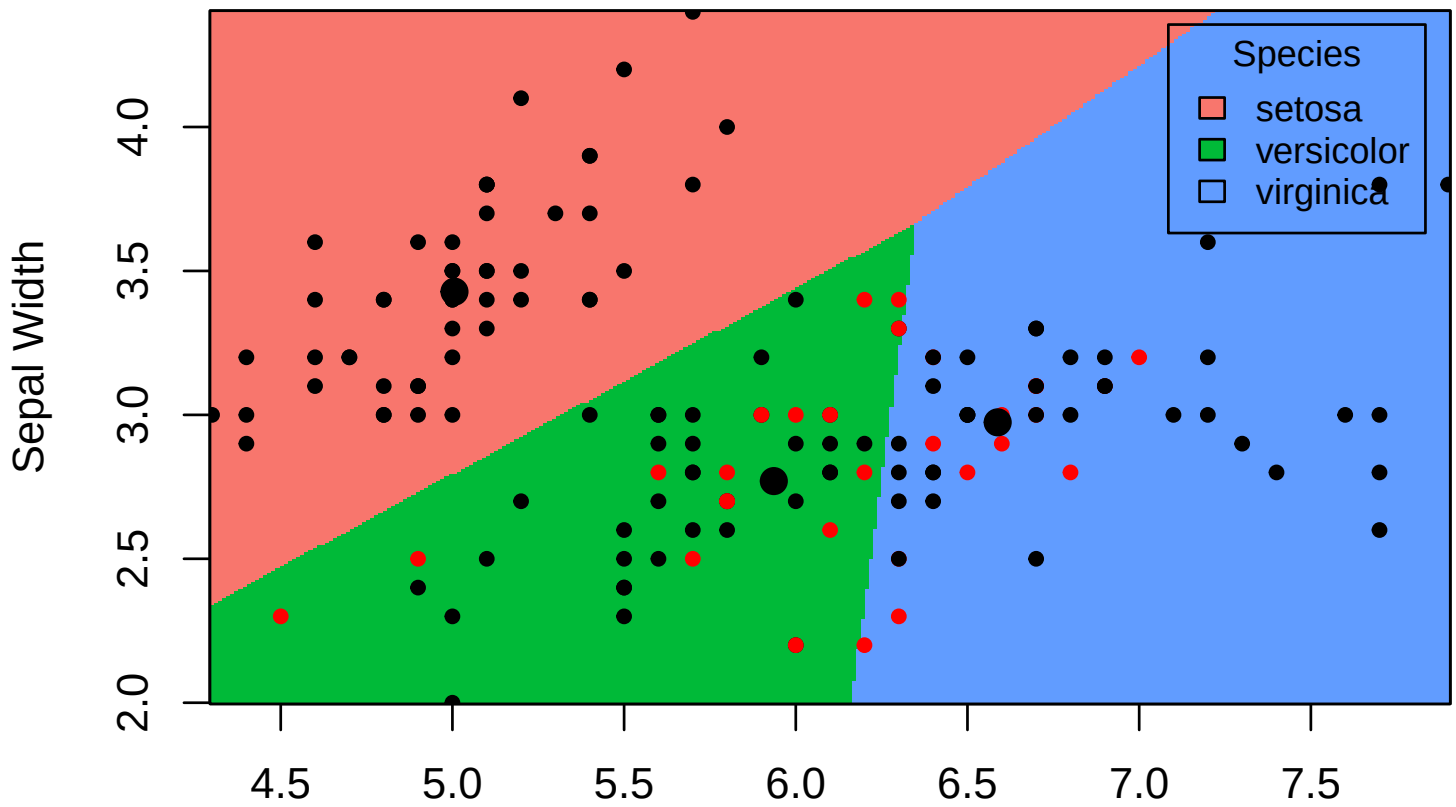
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Objetivo: clasificar correctamente las especies de una planta de iris a partir de la longitud y el ancho del sépalo.



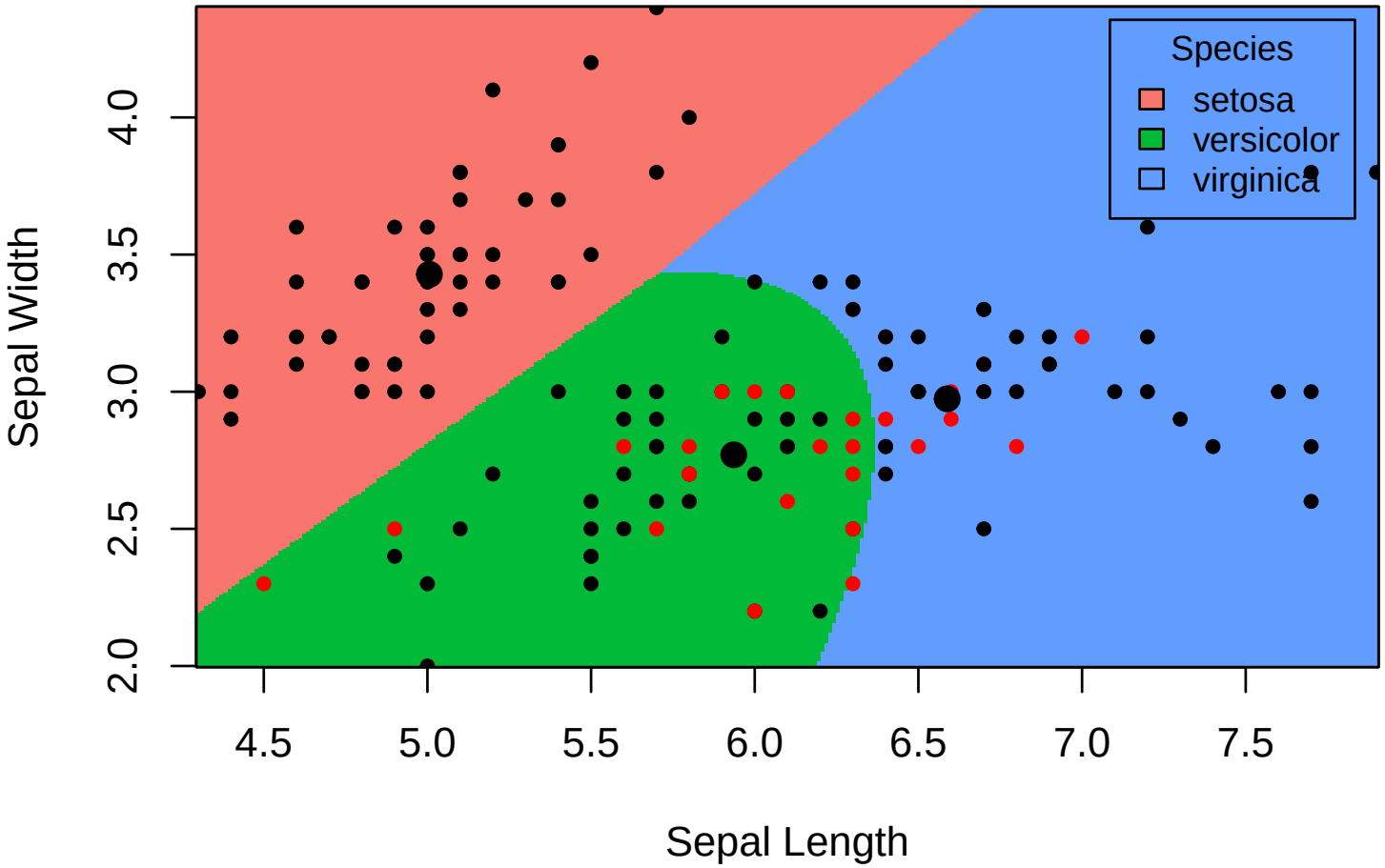
Fronteras lineales

En esta gráfica, los pequeños puntos negros representan plantas de lirio clasificadas correctamente, mientras que los puntos rojos representan clasificaciones erróneas. Los grandes puntos negros representan las medias de clase.



Fronteras no lineales

A veces, una frontera no lineal es más adecuada.



Ejemplo 3: Métodos no supervisados (expresion genica)

- Se estudió la relación entre el consumo máximo de oxígeno innato y la expresión génica del músculo esquelético.
- Las ratas fueron seleccionadas artificialmente por su alta y baja capacidad de carrera (HCR y LCR, respectivamente).
- Las ratas se mantuvieron sedentarias o entrenadas.
- Se identificaron transcripciones significativamente relacionadas con la capacidad de correr y el entrenamiento.
- Los mapas de calor que muestran el nivel de expresión de las transcripciones más significativas se presentaron gráficamente
- Este es un *análisis de conglomerados jerárquico* con medida de distancia de correlación de Pearson.

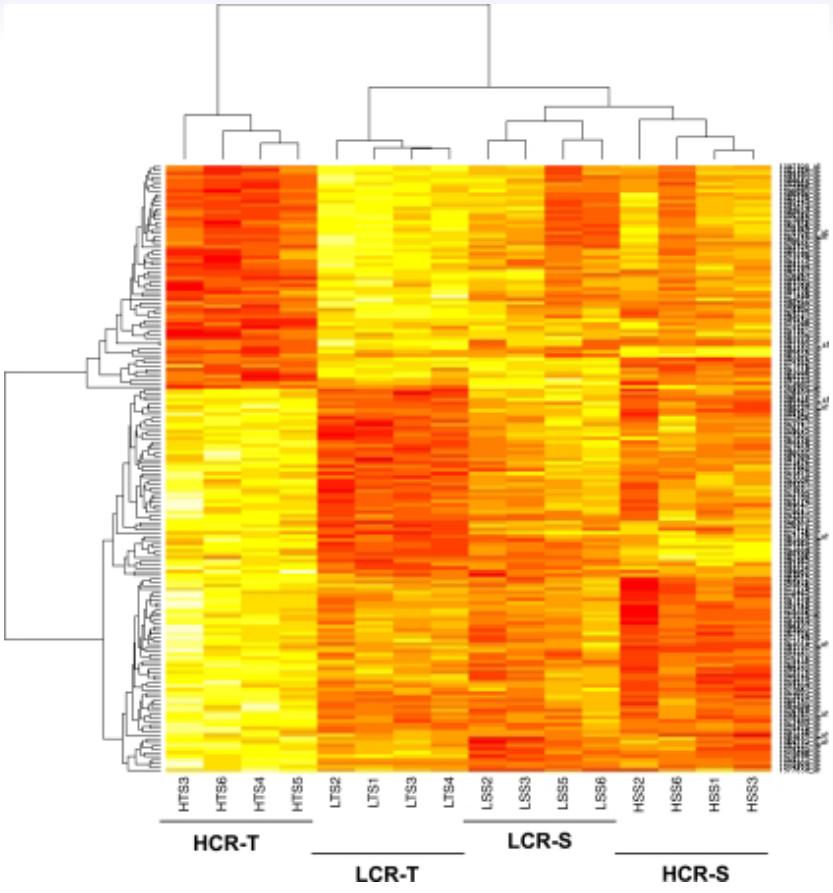


Figure 2: Mapa de calor de las transcripciones más significativas. Las transcripciones con una expresión alta se muestran en rojo y las transcripciones con una expresión baja se muestran en amarillo.

More: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2585023/