TEC | Tecnológico de Costa Rica

Aprendizaje Automático

Tecnológico de Costa Rica

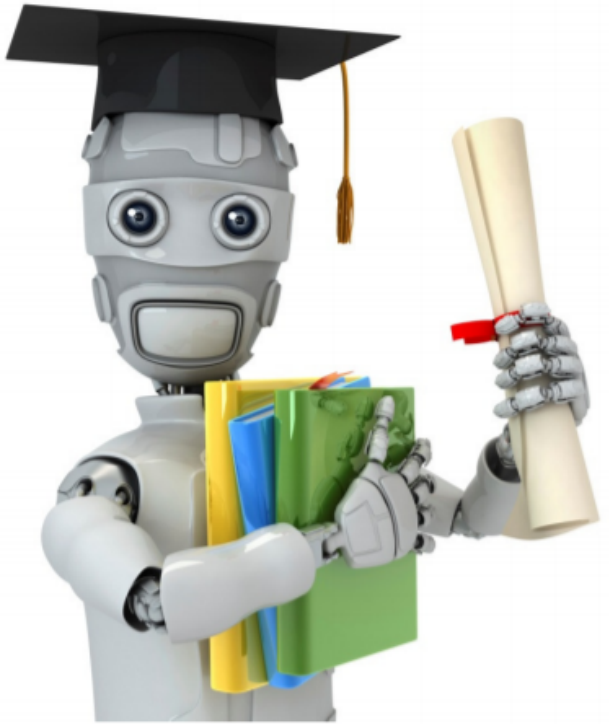
Programa de Ciencia de Datos

Frans van Dunné

Agenda

TEC | Tecnológico de Costa Rica

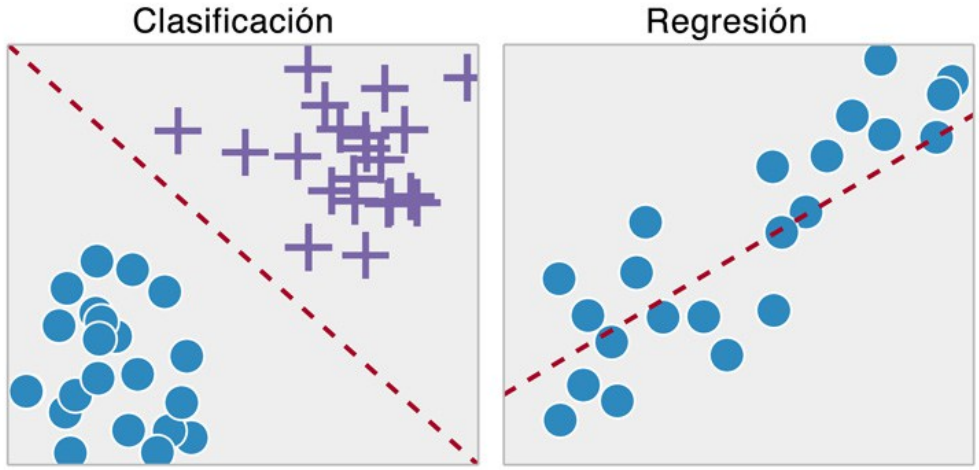
- **Aprendizaje Automático**
  - Métodos supervisados.
  - Clasificación
    - Arboles de decisión
    - Random forest
    - Regresión Logística



Aprendizaje supervisados

TEC | Tecnológico de Costa Rica

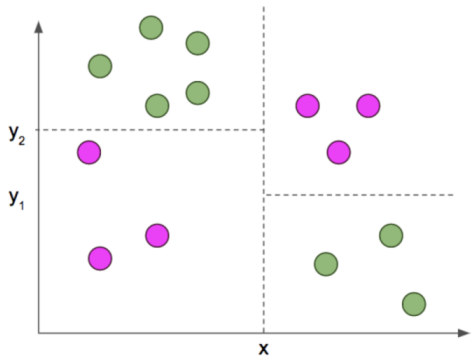
- **Algoritmos de clasificación**
  - Estos algoritmos tienen como objetivo determinar cuál es la **clase**, de las que ya se tiene conocimiento, a la que debe pertenecer una nueva muestra, teniendo en cuenta la información que se puede extraer del conjunto de entrenamiento.

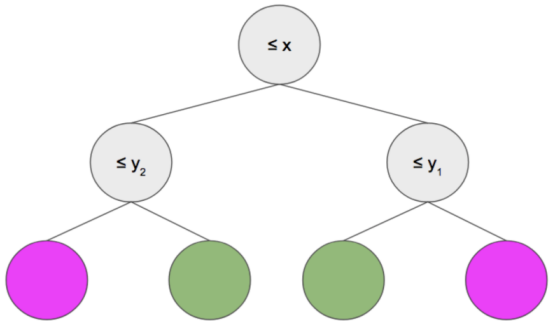


Arboles de decisión

TEC | Tecnológico de Costa Rica

- El enfoque *classification and regression tree (CART)* fue desarrollado por Breiman (1984).
- Son un tipo de algoritmos de aprendizaje supervisado (i.e., existe una variable objetivo predefinida).
- Principalmente usados en problemas de clasificación.
- Las variables de entrada y salida pueden ser categóricas o continuas.
- Divide el espacio de predictores (variables independientes) en regiones distintas y no sobrepuestas.

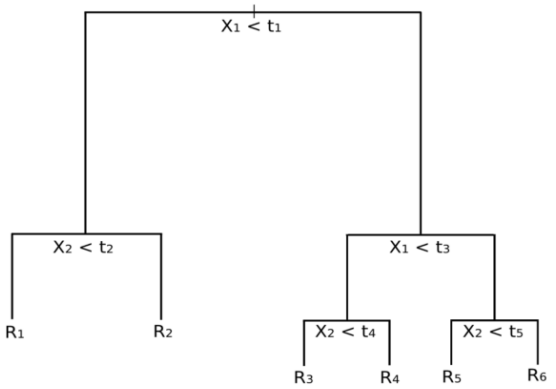




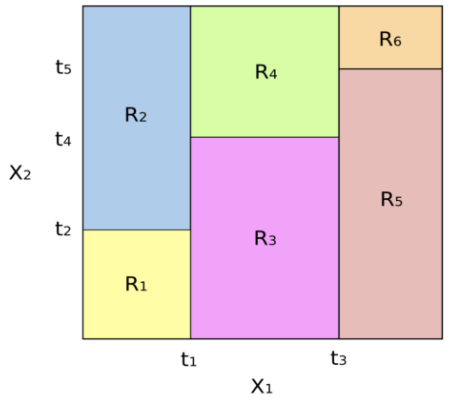
Arboles de decisión

TEC | Tecnológico de Costa Rica

- Se divide la población o muestra en conjuntos homegéneos basados en la variable de entrada más significativa.
- La construcción del árbol sigue un enfoque de división binaria recursiva (top-down greddy approach). Greedy -> analiza la mejor variable para ramificación sólo en el proceso de división actual.



A Decision Tree with six separate regions



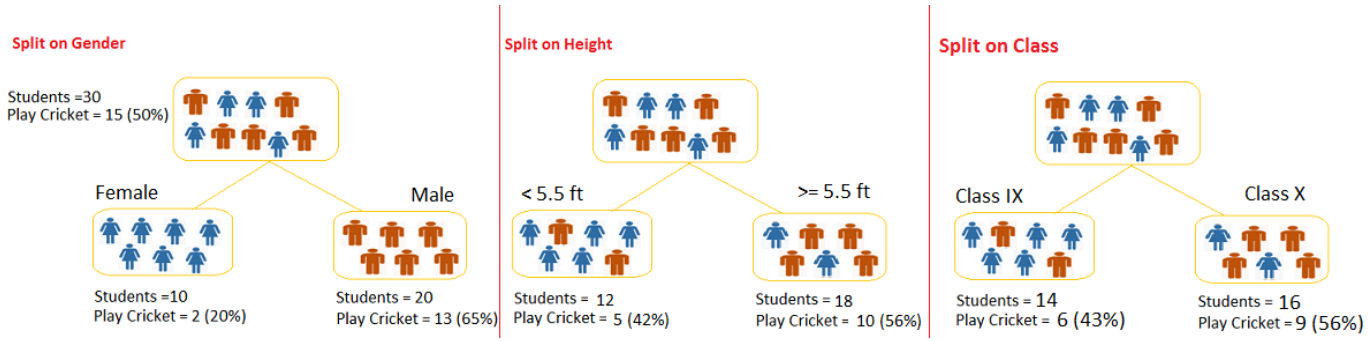
The resulting partition of the subset of $\mathbb{R}^2$ into six regional "blocks"

$$\text{RSS} = \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{y_{R_m}})^2$$

Arboles de decisión

TEC | Tecnológico de Costa Rica

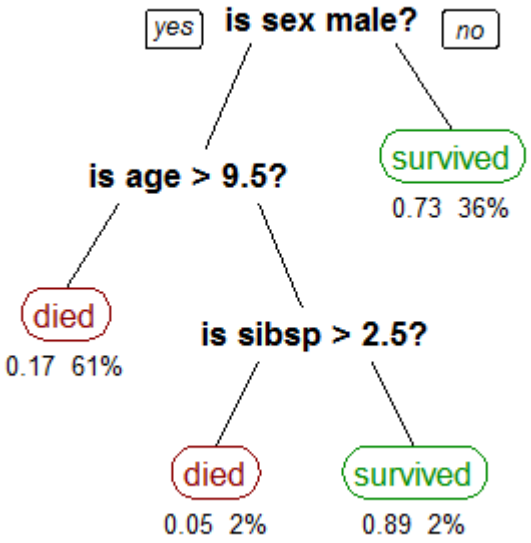
- **Ejemplo**
  - 30 estudiantes
  - 3 variables: Género (hombre/mujer), Clase (IX/X) y Altura (5 a 6 pies).
  - 15 estudiantes juegan cricket en su tiempo libre
  - Crear un modelo para predecir quien jugará cricket
  - Segregar estudiantes basados en todos los valores de las 3 variables e identificar aquella variable que crea los conjuntos más homogéneos de estudiantes y que a su vez son heterogéneos entre ellos.



Ventajas

TEC | Tecnológico de Costa Rica

- Fácil de entender
- Útil en exploración de datos: identificar importancia de variables a partir de cientos de variables.
- Menos limpieza de datos: outliers y valores faltantes no influencian el modelo (A un cierto grado)
- El tipo de datos no es una restricción
- Es un método no paramétrico (i.e., no hay suposición acerca del espacio de distribución y la estructura del clasificador)

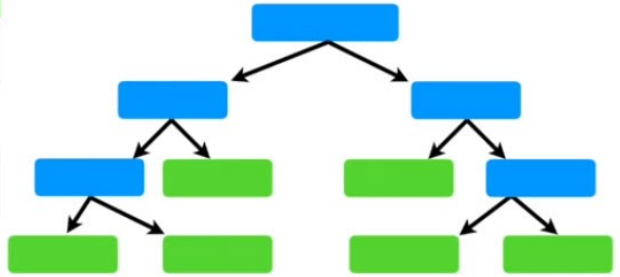


Desventajas

TEC | Tecnológico de Costa Rica

- Sobreajuste
- Pérdida de información al categorizar variables continuas
- Precisión: métodos como SVM y clasificadores tipo ensamblador a menudo tienen tasas de error 30% más bajas que CART (Classification and Regression Trees)
- Inestabilidad: un pequeño cambio en los datos puede modificar ampliamente la estructura del árbol. Por lo tanto la interpretación no es tan directa como parece.

Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...



Random Forest

TEC | Tecnológico de Costa Rica

- Los métodos tipo ensamblador están formados de un **grupo de modelos predictivos** que permiten alcanzar una mejor precisión y estabilidad del modelo. Estos proveen una mejora significativa a los modelos de árboles de decisión.

- **Por qué surgen los ensambladores de árboles?**
  - Así como todos los modelos, un árbol de decisión también sufre de los problemas de sesgo y varianza. Es decir, cuánto en promedio son los valores predecidos diferentes de los valores reales (sesgo) y cuan diferentes serán las predicciones de un modelo en un mismo punto si muestras diferentes se tomaran de la misma población' (varianza).

  - Al construir un árbol pequeño se obtendrá un modelo con baja varianza y alto sesgo. Normalmente, al incrementar la complejidad del modelo, se verá una reducción en el error de predicción debido a un sesgo más bajo en el modelo. En un punto el modelo será muy complejo y se producirá un sobre-ajuste del modelo el cual empezará a sufrir de varianza alta.

  - El modelo óptimo debería mantener un balance entre estos dos tipos de errores. A esto se le conoce como "trade-off" (equilibrio) entre errores de sesgo y varianza. El uso de ensambladores es una forma de aplicar este "trade-off".

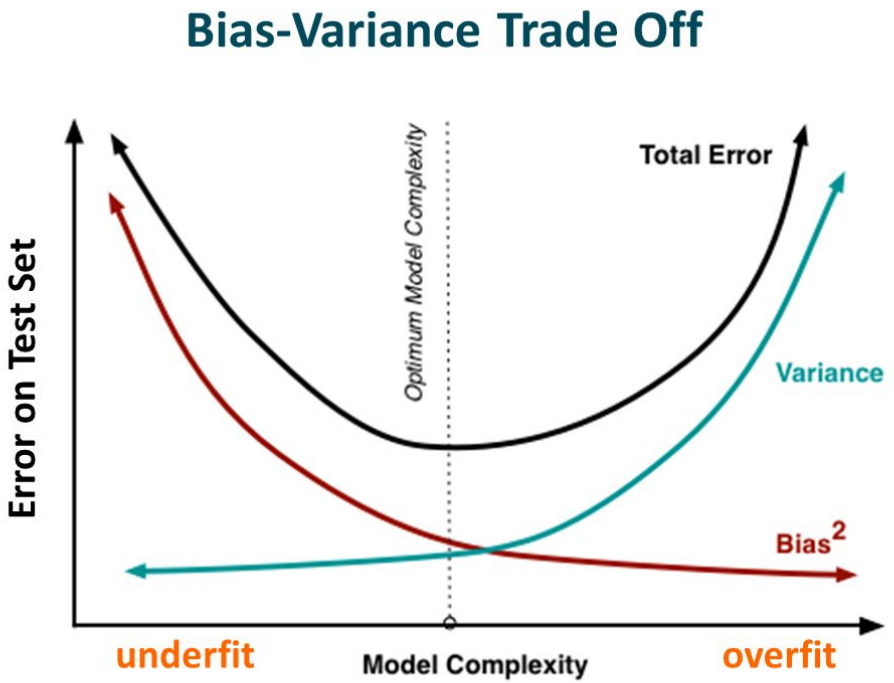
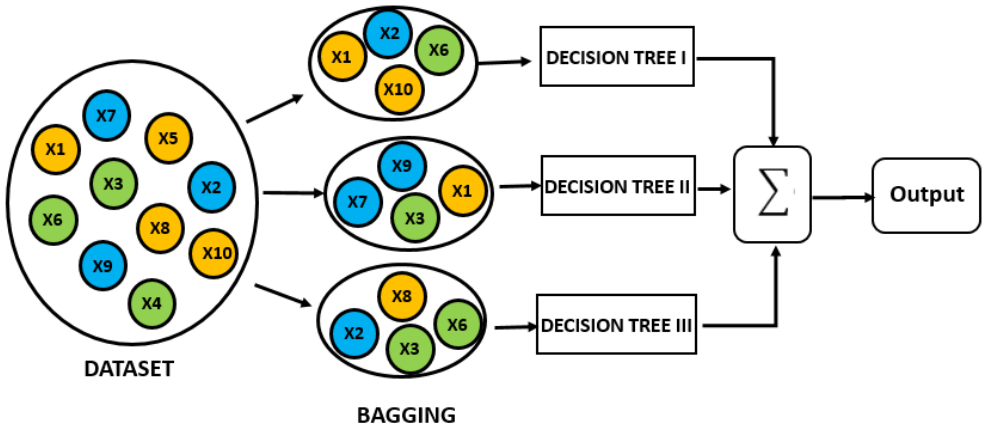


image credit: scott.fortmann-roe.com

¿Qué es el proceso de *bagging* y cómo funciona?

TEC | Tecnológico de Costa Rica

- Bagging es una técnica usada para reducir la varianza de las predicciones a través de la combinación de los resultados de varios clasificadores, cada uno de ellos modelados con diferentes subconjuntos tomados de la misma población.
- En resumen:
  - Crear múltiples subconjuntos de datos
  - Construir múltiples modelos
  - Combinar los modelos.



Random Forest

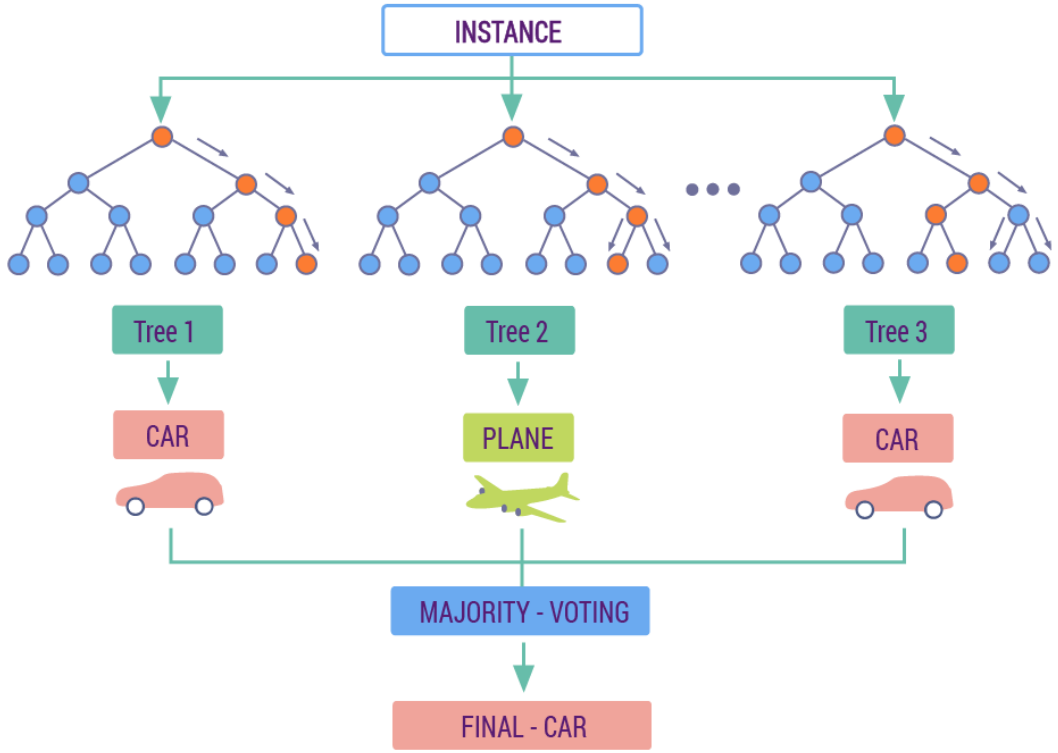
TEC | Tecnológico de Costa Rica

- Random Forest se considera como la “panacea” en todos los problemas de ciencia de datos.
- Util para regresión y clasificación.
- Un grupo de modelos “débiles”, se combinan en un modelo robusto.
- Sirve como una técnica para reducción de la dimensionalidad.
- Se generan múltiples árboles (a diferencia de CART).
- Cada árbol da una clasificación (vota por una clase). Y el resultado es la clase con mayor número de votos en todo el bosque (forest).
- Para regresión, se toma el promedio de las salidas (predicciones) de todos los árboles.

Ventajas

TEC | Tecnológico de Costa Rica

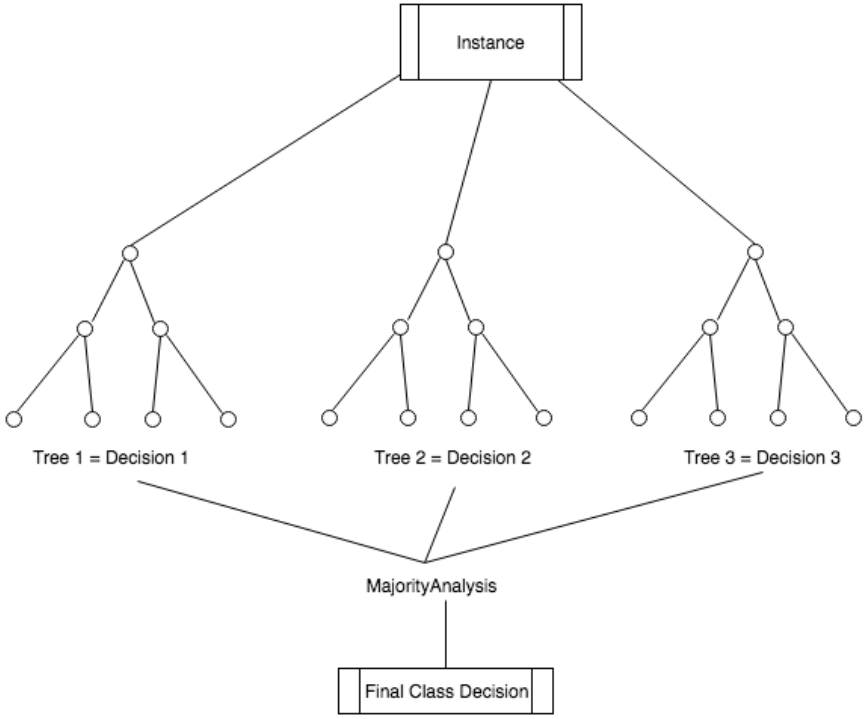
- Existen muy pocas suposiciones y por lo tanto la preparación de los datos es mínima.
- Puede manejar hasta miles de variables de entrada e identificar las más significativas. Método de reducción de dimensionalidad.
- Una de las salidas del modelo es la **importancia de variables**.
- Incorpora métodos efectivos para estimar valores faltantes.
- Es posible usarlo como método no supervisado (clustering) y detección de outliers.



Desventajas de Random Forest

TEC | Tecnológico de Costa Rica

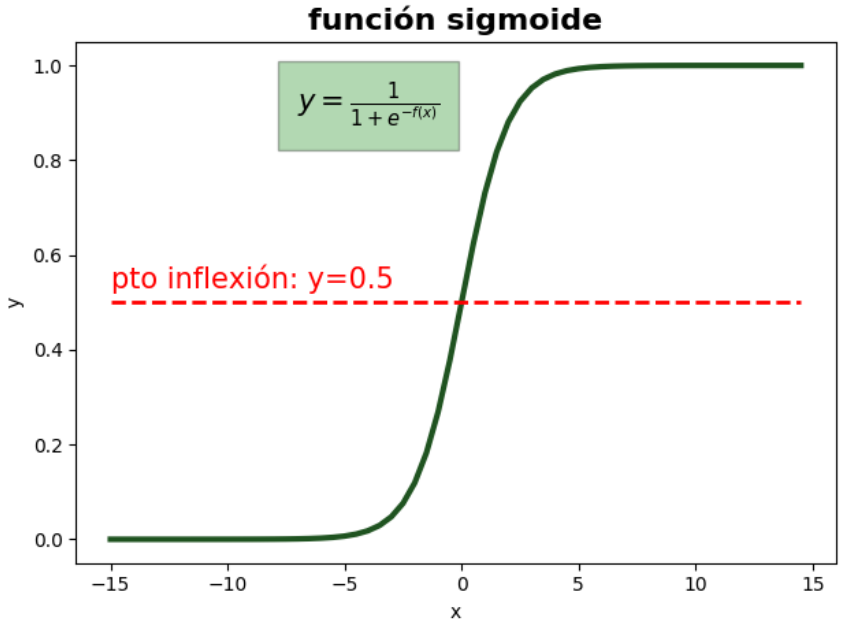
- Pérdida de interpretación
- Bueno para clasificación, no tanto para regresión. Las predicciones no son de naturaleza continua.
- En regresión, no puede predecir más allá del rango de valores del conjunto de entrenamiento.
- Poco control en lo que hace el modelo (modelo caja negra para modeladores estadísticos)



Regresión Logística

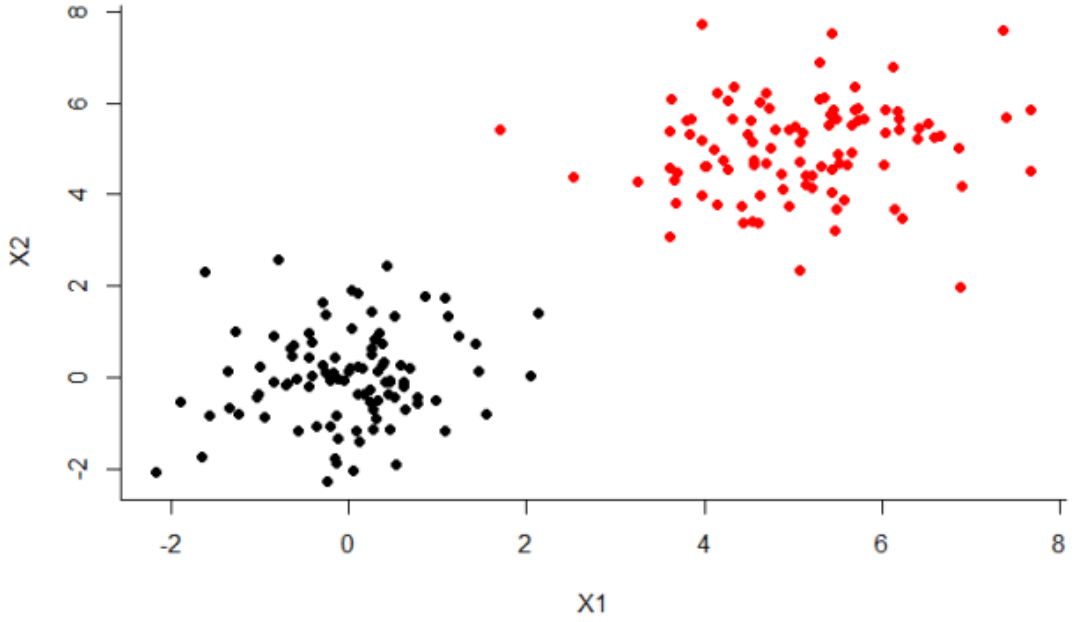
TEC | Tecnológico de Costa Rica

- La regresión logística es un procedimiento cuantitativo de gran utilidad para problemas donde la variable dependiente toma valores en un conjunto finito
- Ejemplo:
  - Un paciente muere o no antes del alta.
  - Una persona deja o no de fumar después de un tratamiento.
  - En un estudio retrospectivo un individuo es caso o control.
  - Un paciente positivo al VIH está o no en el estado IV.



Regresión Logística

- La ***Regresión Logística*** no se utiliza para variables numéricas, sino que se emplea para **predecir** el resultado de una **variable categórica** en función de variables independientes. Debido a su sencillez y rápida aplicación, este algoritmo suele utilizarse con frecuencia para problemas de ***Clasificación Binaria*** y ***Clasificación Multiclase*** **con fronteras lineales**.



Bibliografía

TEC | Tecnológico de Costa Rica

- https://blog.gfi.es/algoritmos-entrenamiento-machine-learning/
- https://unmonoqueteclea.github.io/machine-learning/regresionlogistica/
- http://ligdigonzalez.com/aprendizaje-supervisado-logistic-regression/
- https://bigdatadummy.com/2017/09/19/regresion-logistica/
- https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb
- https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
- https://www.geeksforgeeks.org/decision-tree/