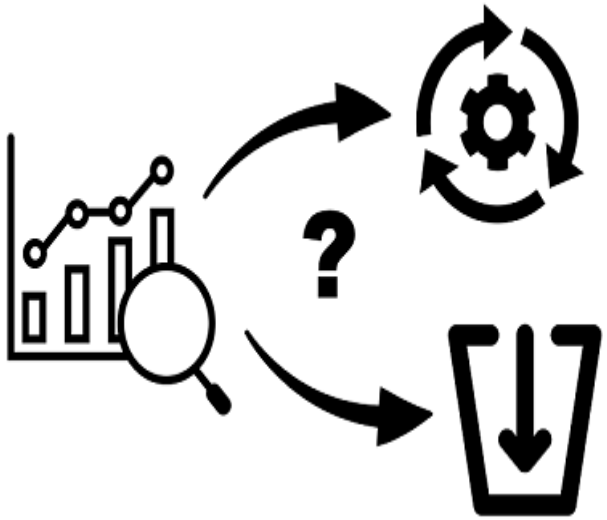
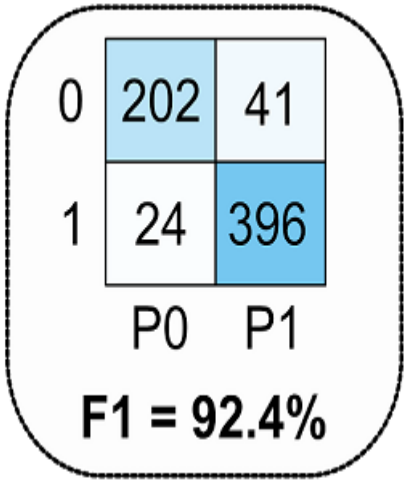




Fundamentos de Data Science

Unidad 4: Extracción de Conocimiento

Semana 13 – La Modelización: Evaluación de Modelos Analíticos

1. Regresión Logística

La regresión logística es una técnica estadística multivariante que nos permite estimar la relación existente entre una variable dependiente (dicotómica) y un conjunto de variables independientes cuantitativas o cualitativas.

Objetivo primordial es modelar cómo influyen las variables regresoras en la probabilidad de ocurrencia de un suceso particular.

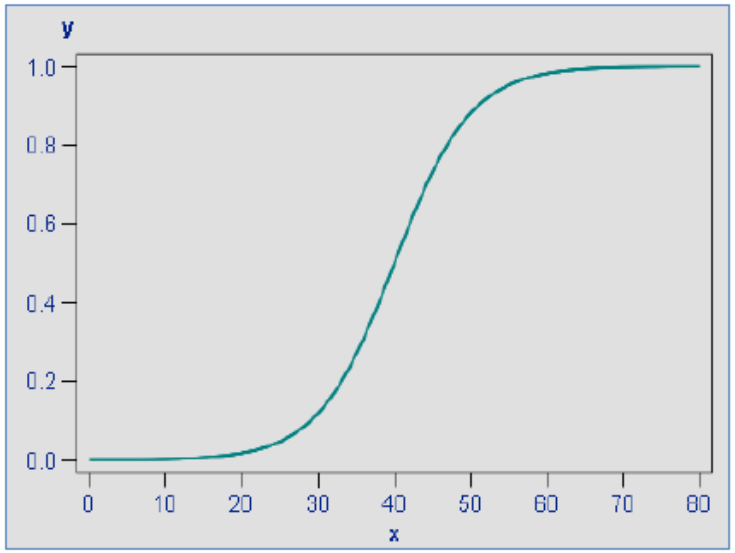
Modelo de Regresión Logìstica

Variable de Salida es Cualitativa:

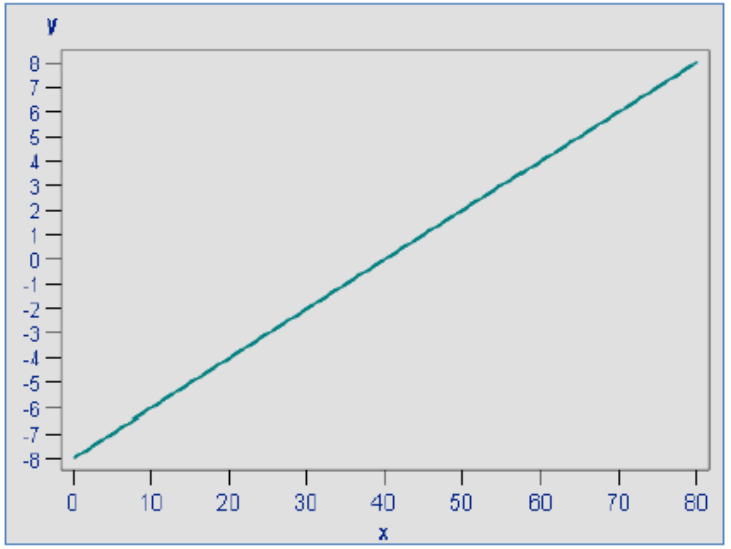
- 1=si
- 0=no

Para la linealización del modelo se realiza una transformación para la variable de respuesta.

Curva Logística: $Y = p_i$



Transformación Lineal: $Y = \log \frac{p_i}{1 - p_i}$



$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_i x_i)}}$$

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_i x_i$$

Regresión Logística

- A diferencia de la Regresión Múltiple, en este modelo las variables explicativas pueden ser categóricas y no necesitan la condición de tener una distribución conjunta normal multivariada y adicionalmente las variables explicativas pueden ser categóricas.

La representación matemática del modelo es la siguiente:

$$g_i = \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

g_i : Variable dependiente del modelo: "Bueno" y "Malo".

p_i : Probabilidad de que el cliente sea bueno en los próximos 6 meses.

β_i : Coeficientes del modelo (Parámetros a estimar).

x_i : Variables explicativas del modelo.

Regresión Logística

- A diferencia de la Regresión Múltiple, en este modelo las variables explicativas pueden ser categóricas y no necesitan la condición de tener una distribución conjunta normal multivariada y adicionalmente las variables explicativas pueden ser categóricas.

La representación matemática del modelo es la siguiente:

$$g_i = \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

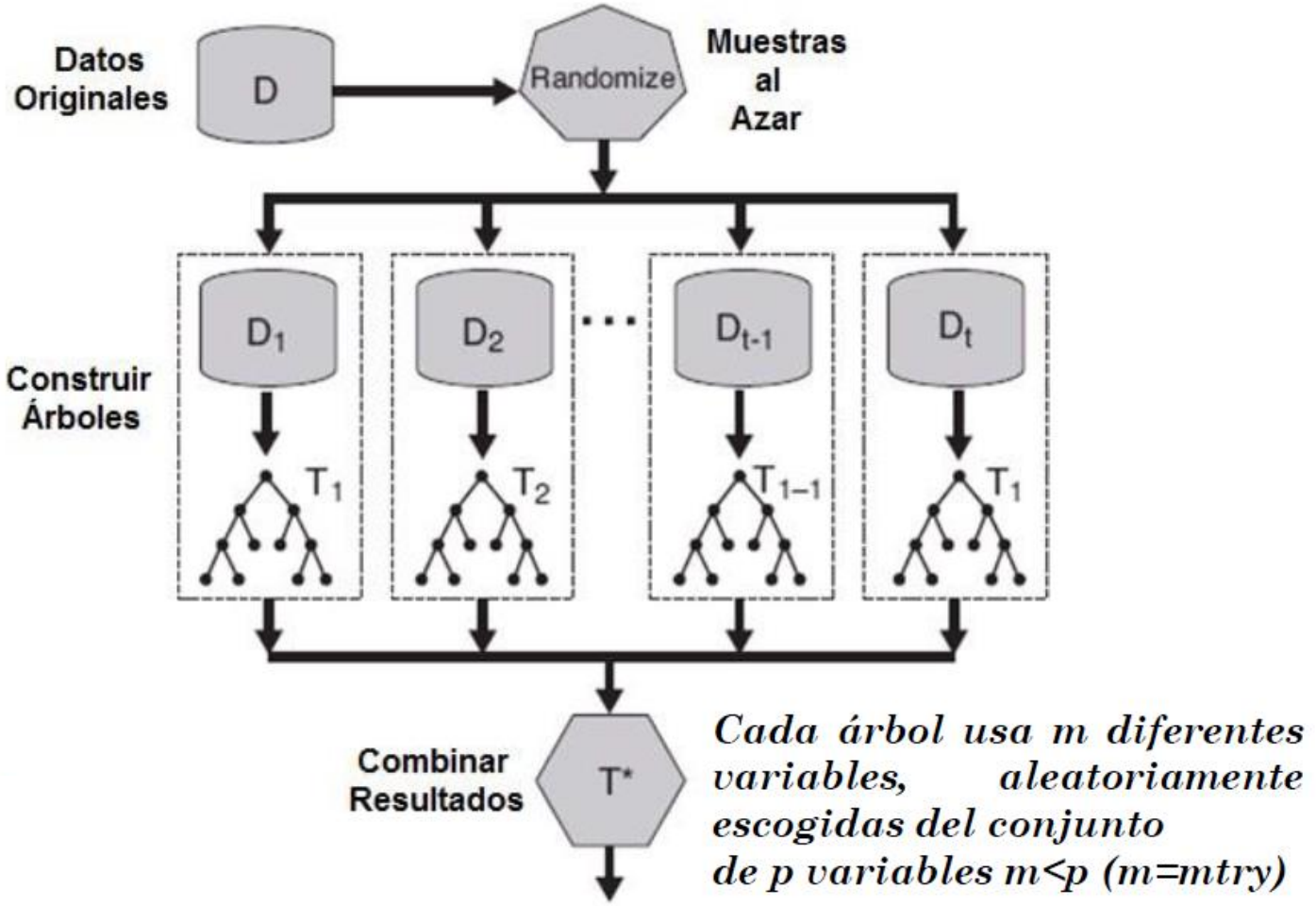
g_i : Variable dependiente del modelo: "Bueno" y "Malo".

p_i : Probabilidad de que el cliente sea bueno en los próximos 6 meses.

β_i : Coeficientes del modelo (Parámetros a estimar).

x_i : Variables explicativas del modelo.

2.Random Forest



Cada árbol usa m diferentes variables, aleatoriamente escogidas del conjunto de p variables m<p (m=mtry)

3.Cross-Validation

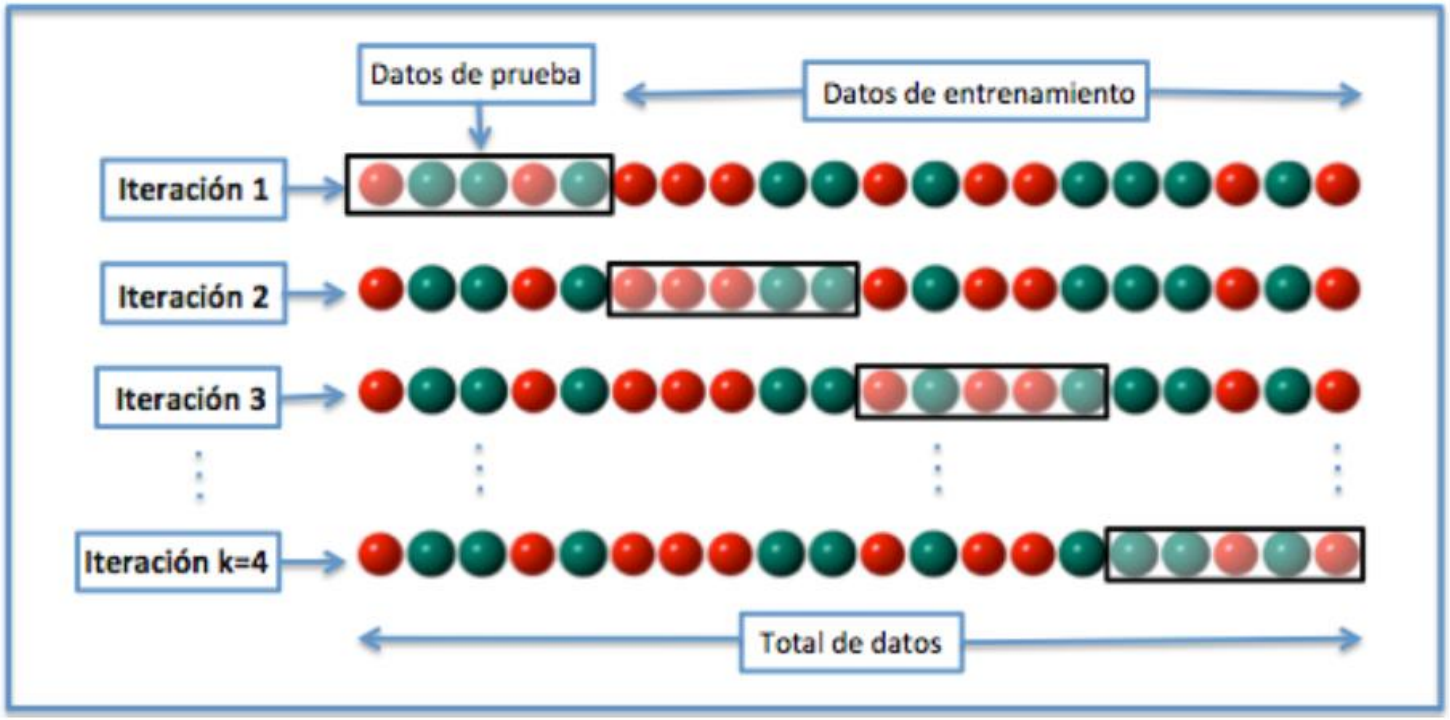
Cross-Validation o Validación Cruzada es un método que permite probar el rendimiento de un modelo predictivo de Machine Learning. Este método brinda una mejor comprensión del rendimiento del modelo en todo el conjunto de datos en lugar de una sola división de prueba/entrenamiento.

El proceso sigue estos pasos:

1. El número de iteraciones está definido, por defecto es 5.
2. El conjunto de datos se divide de acuerdo con estas iteraciones, donde cada uno tiene un conjunto único de datos de prueba.
3. Un modelo es entrenado y probado para cada iteración.
4. Cada iteración devuelve una métrica para sus datos de prueba.
5. La media y la desviación estándar de estas métricas se pueden calcular para proporcionar una única métrica para el proceso.

Cross-Validation

Prueba el modelo en múltiples iteraciones de un conjunto de datos:



K grupos → K iteraciones