Universidad **Ricardo Palma**

RECTORADO

PROGRAMA DE ESPECIALIZACIÓN EN CIENCIA DE DATOS

Formamos seres humanos para una cultura de paz

TALLER DE ESTADÍSTICA PARA LA CIENCIA DE DATOS

R + Python

MÓDULO 5 : ANALISIS MULTIVARIADO II

Análisis de Reresión Logística Binaria

A nuestro recordado Maestro

Dr. Erwin Kraenau Espinal, Presidente de la Comisión de Creación de la Maestría en Ciencia de los Datos

TALLER DE ESTADÍSTICA PARA CIENCIA DE DATOS

EXPOSITORES

José Antonio Cárdenas Garro
UNMSM
MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricardo Palma

Predictive Modelling Specialist — Scotiabank

Portfolio and Consumption Analyst — Scotiabank

Data Analyst — ONP (Oficina de Normalización Previsional)

Correo : josecardenasgarro@gmail.com
LinkedIn : www.linkedin.com/in/jos%C3%A9-antonio-c%C3%A1rdenas-garro-599266b0

André Omar Chávez Panduro
UNMSM
MSc in Data Science Candidate
Promotion "Erwin Kraenau Espinal"
Universidad Ricardo Palma

Data Scientist — Interbank

Customer Intelligence Analyst — Claro

Data Analyst — INEI (Instituto Nacional de Estadística e Informática)

Correo : andrecp38@gmail.com / 09140205@unmsm.edu.pe
LinkedIn : www.linkedin.com/in/andré-chávez-a90078b9

« Divide las dificultades que examinas en tantas partes como sea posible , para su mejor solución»

Agenda

- Introducción a los Problemas de Clasificación o Aprendizaje Supervisado.
- Regresión Logística Binaria.

CLASIFICACIÓN: DEFINICIÓN

- Dada una colección de registros (Conjunto de Entrenamiento) cada registro contiene un conjunto de variables (atributos) denominado x, con una variable (atributo) adicional que es la clase denominada y.

- El objetivo de la ***clasificación*** es encontrar un modelo (una función) para predecir la clase a la que pertenecería cada registro, esta asignación una clase se debe hacer con la mayor precisión posible.

- Un conjunto de prueba (tabla de testing) se utiliza para determinar la precisión del modelo. Por lo general, el conjunto de datos dado se divide en dos conjuntos al azar de el de entrenamiento y el de prueba.

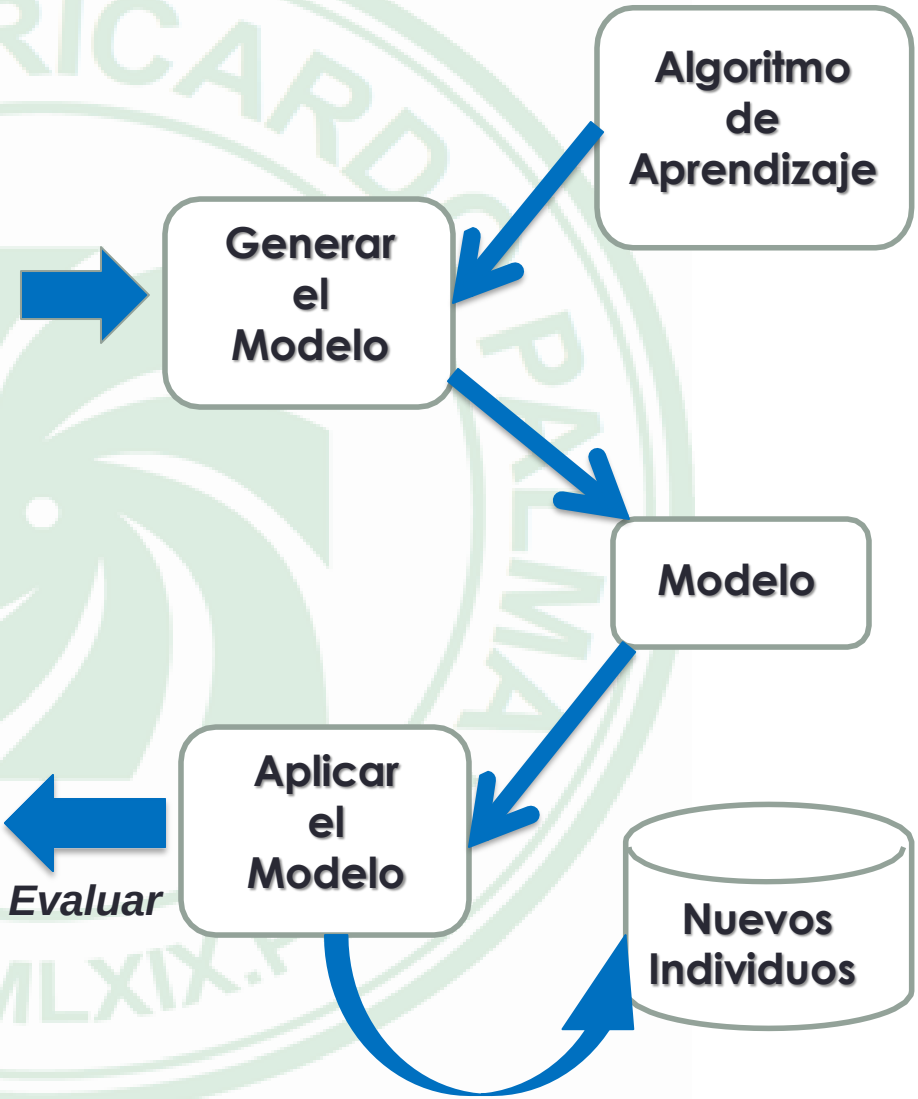
MODELO GENERAL DE LOS MÉTODOS DE CLASIFICACIÓN

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
1	SI	SOLTERO	S/ 1,000	NO
2	SI	CASADO	S/ 5,000	NO
3	NO	CASADO	S/ 3,500	SI
4	SI	VIUDO	S/ 4,500	NO
5	NO	SOLTERO	S/ 2,000	NO
6	NO	SOLTERO	S/ 1,500	SI

Tabla de Aprendizaje

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
7	SI	SOLTERO	S/ 4,000	NO
8	SI	CASADO	S/ 5,500	NO
9	NO	CASADO	S/ 6,500	SI

Tabla de Testing

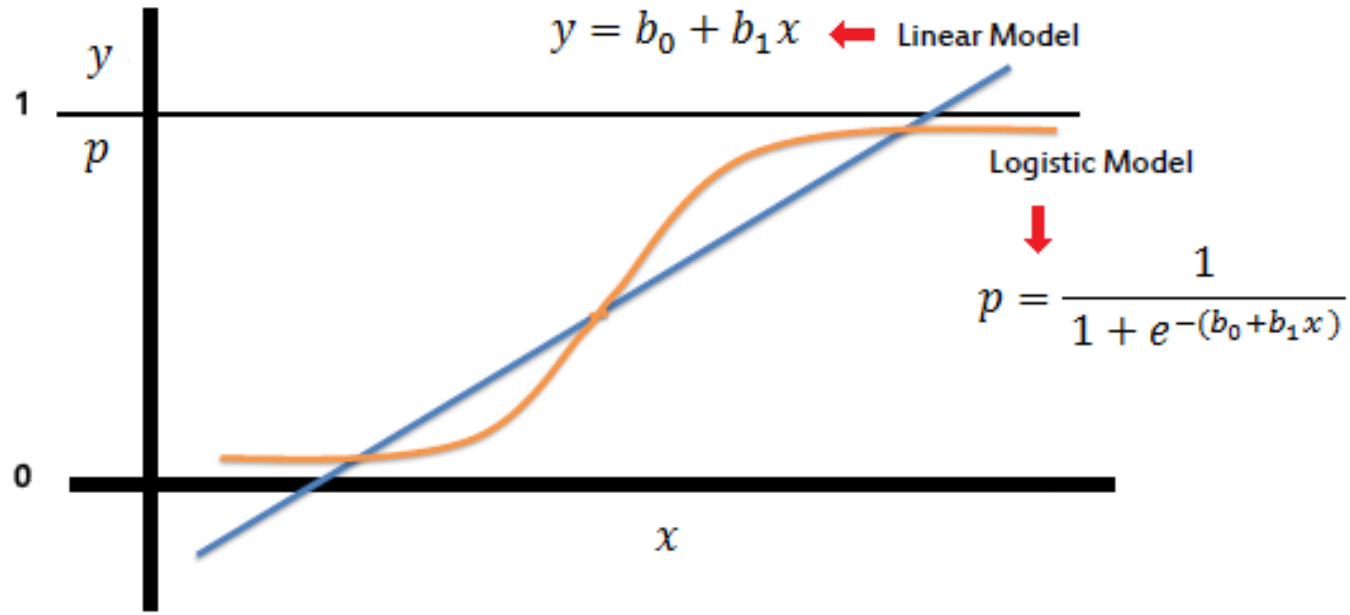


DEFINICIÓN DE CLASIFICACIÓN

- Dada una base de datos $D = \{t_1, t_2, \dots, t_n\}$ de tuplas o registros (individuos) y un conjunto de clases $C = \{C_1, C_2, \dots, C_m\}$, el **problema de la clasificación** es encontrar una función $f: D \rightarrow C$ tal que cada t_i es asignada una clase C_j.

- $f: D \rightarrow C$ podría ser una Red Neuronal, un Árbol de Decisión, un modelo basado en Análisis Discriminante, o una Red Beyesiana.

Regresión Logística



DEFINICIÓN

- Es un modelo predictivo supervisado.
- La regresión logística es un modelo de elección discreta en el que la variable dependiente es cualitativa.
- Es flexible en cuanto a la naturaleza de las variables explicativas, pues éstas pueden ser de escala y categóricas.

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

- Para este modelo se considera que la variable respuesta, es una variable dicotómica que tomas dos valores.
- Para estos modelos dicotómicos, las dos categorías deben de ser mutuamente excluyentes.
- La variable respuesta se puede expresar de la siguiente forma:

$$Y_i = \begin{cases} 1, \Pr ob(Y_i = 1) = P_i \\ 0, \Pr ob(Y_i = 0) = 1 - P_i \end{cases}$$

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

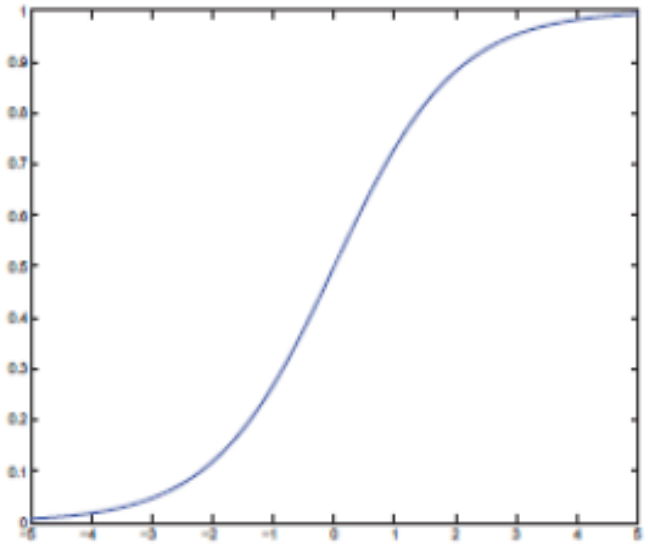
Ejemplo:

La variable Morosidad toma los siguientes valores:

- "1" si el cliente es moroso.
- "0" si el cliente es no moroso.

¿Es dicotómica?
¿Es cualitativa?
¿Es mutuamente excluyente?

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO



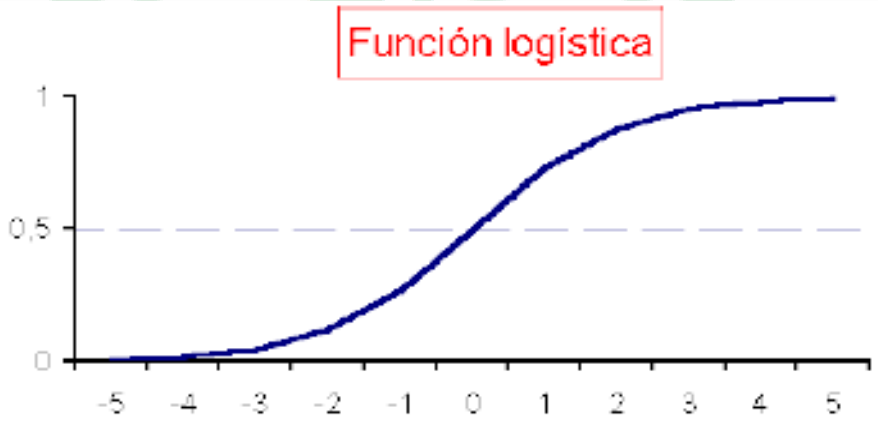
Se basa en la función logística:

$$f(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+\frac{1}{e^{z}}} = \frac{e^{z}}{1+e^{z}}$$

Está acotada entre 0 y 1:

$$\lim_{z\to-\infty} f(z) = 0, \quad \lim_{z\to\infty} f(z) = 1.$$

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO



La representación matemática del modelo es la siguiente:

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

z_i : Variable dependiente del modelo: "Moroso" y " No Moroso"

p_i : Probabilidad de que el cliente sea "Moroso"

β_i : Coeficientes del modelo (parámetros a estimar)

x_i : Variables explicativas del modelo

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

Odds Ratio

Es la razón entre la probabilidad de que se produzca un suceso y la probabilidad de que no se produzca ese suceso.

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

Indica cuánto más probable es ser un cliente "Moroso" que "No Moroso"

¡Gracias!