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 Regresión No Paramétrica

A nuestro recordado Maestro

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

TALLER DE ESPECIALIZACIÓN "STATISTICAL SCIENCE INTRODUCTION"

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

André Omar Chávez Panduro
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)

Data Scientist — Interbank

Customer Intelligence Analyst — Claro

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

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

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

TALLER DE ESPECIALIZACIÓN "STATISTICAL SCIENCE INTRODUCTION"

« En medio de la dificultad, reside el nacimiento de la **OPORTUNIDAD**»

TALLER DE ESPECIALIZACIÓN "STATISTICAL SCIENCE INTRODUCTION"

AGENDA

- Introducción
- Formas Paramétricas
- Formas No Paramétricas

INTRODUCCION

La teoría de los **métodos no paramétricos** desarrolla procedimientos de inferencia estadística, que no realizan una suposición explícita con respecto a la forma funcional de la distribución de probabilidad de las observaciones de la muestra.

Si bien en la **estadística no paramétrica** también aparecen modelos y parámetros, ellos están definidos de una manera más general que en su contrapartida paramétrica.

MODELO

$$Y_i = r(X_i) + \varepsilon_i, \quad i = 1, \cdots, n$$

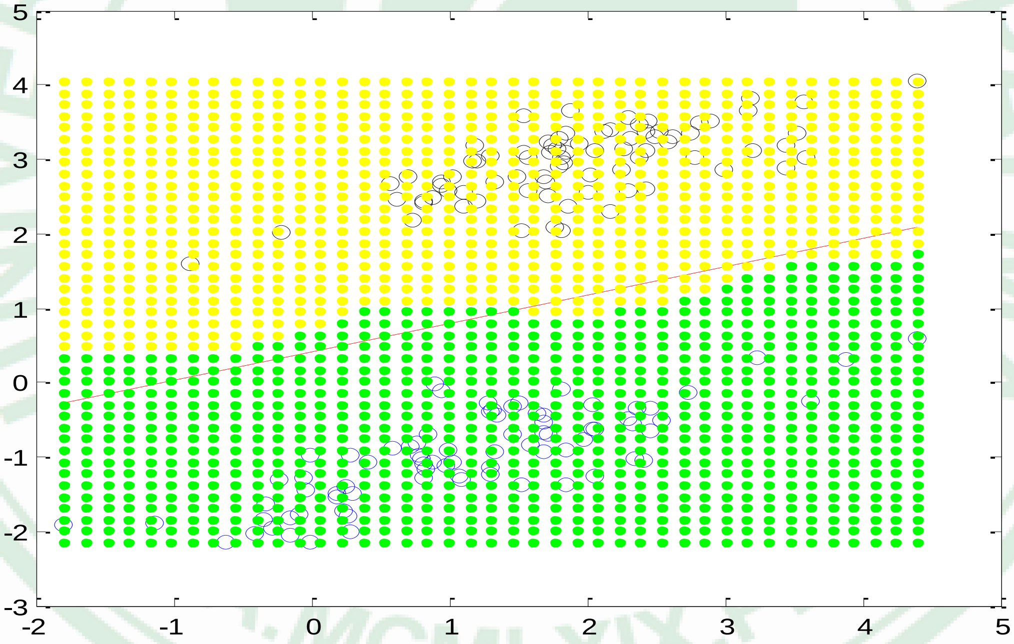
FORMAS PARAMÉTRICAS

$$p = \frac{e^Z}{1+e^Z} = \frac{1}{1+e^{-Z}},$$

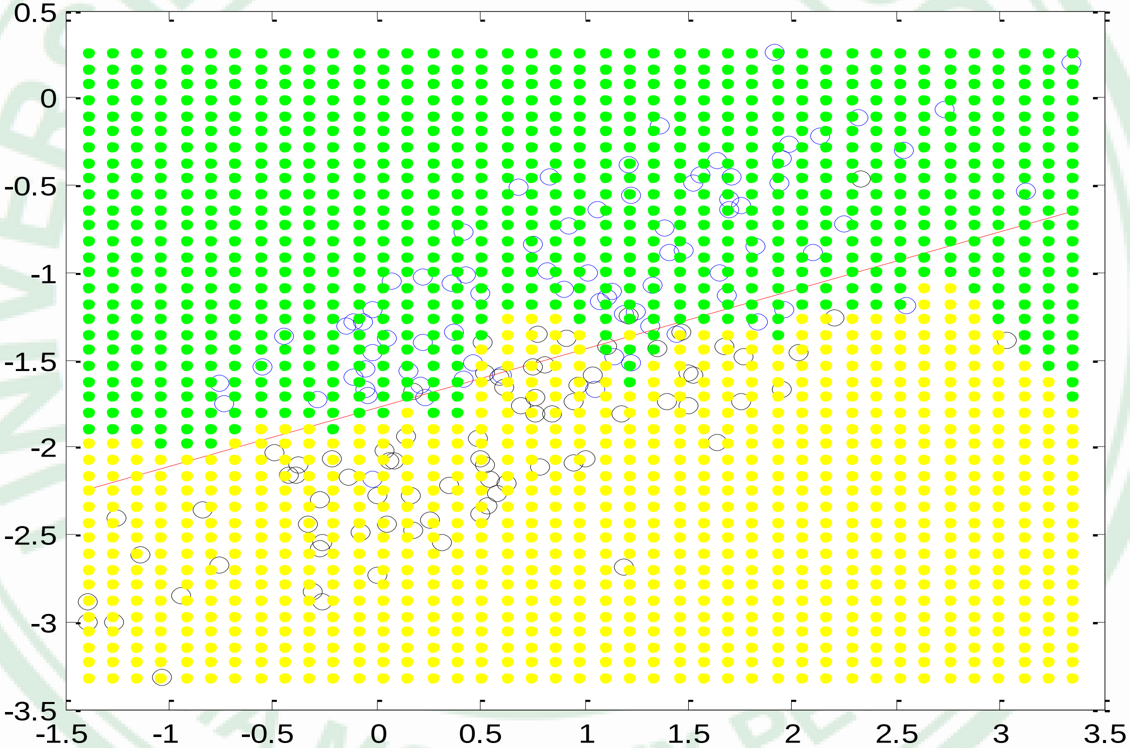
siendo

$$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k$$

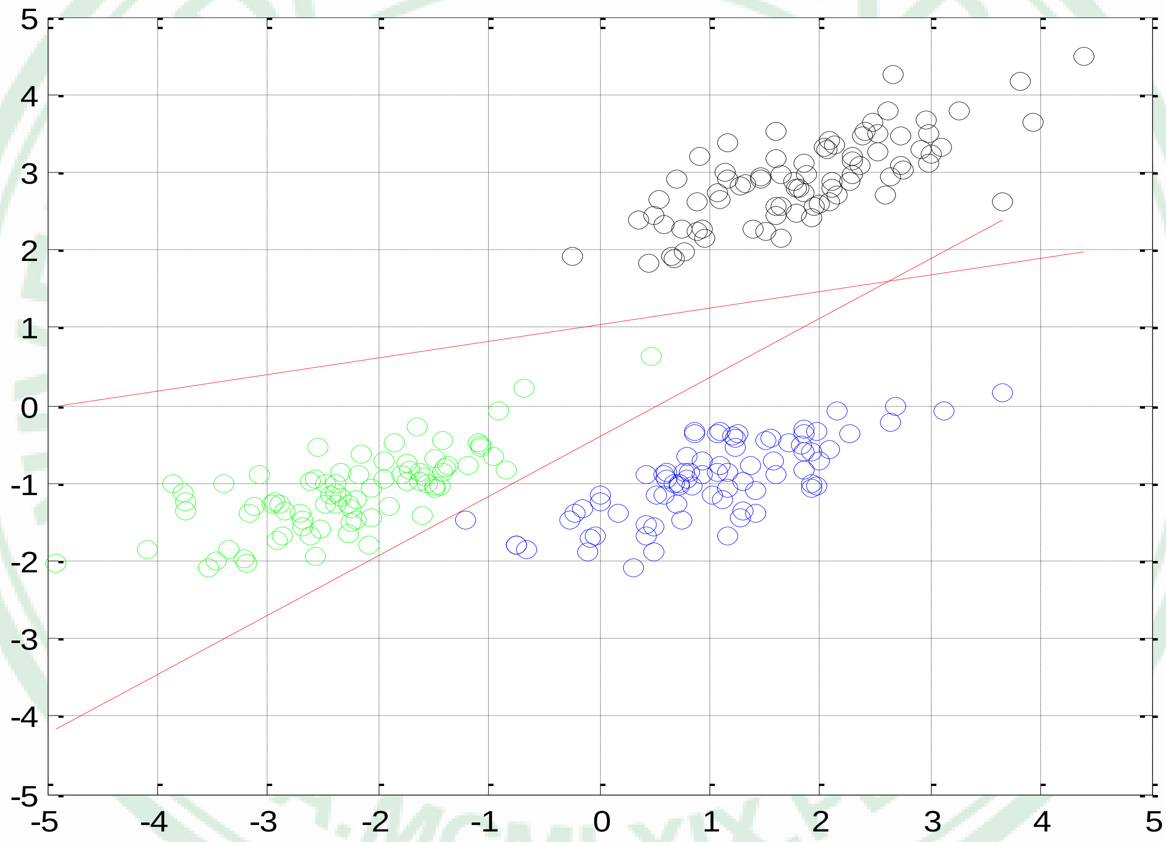
CLASIFICACIÓN MEDIANTE LOS VECINOS MÁS CERCANOS



CLASIFICACIÓN MEDIANTE LOS VECINOS MÁS CERCANOS



FUNCIONES DISCRIMINANTES



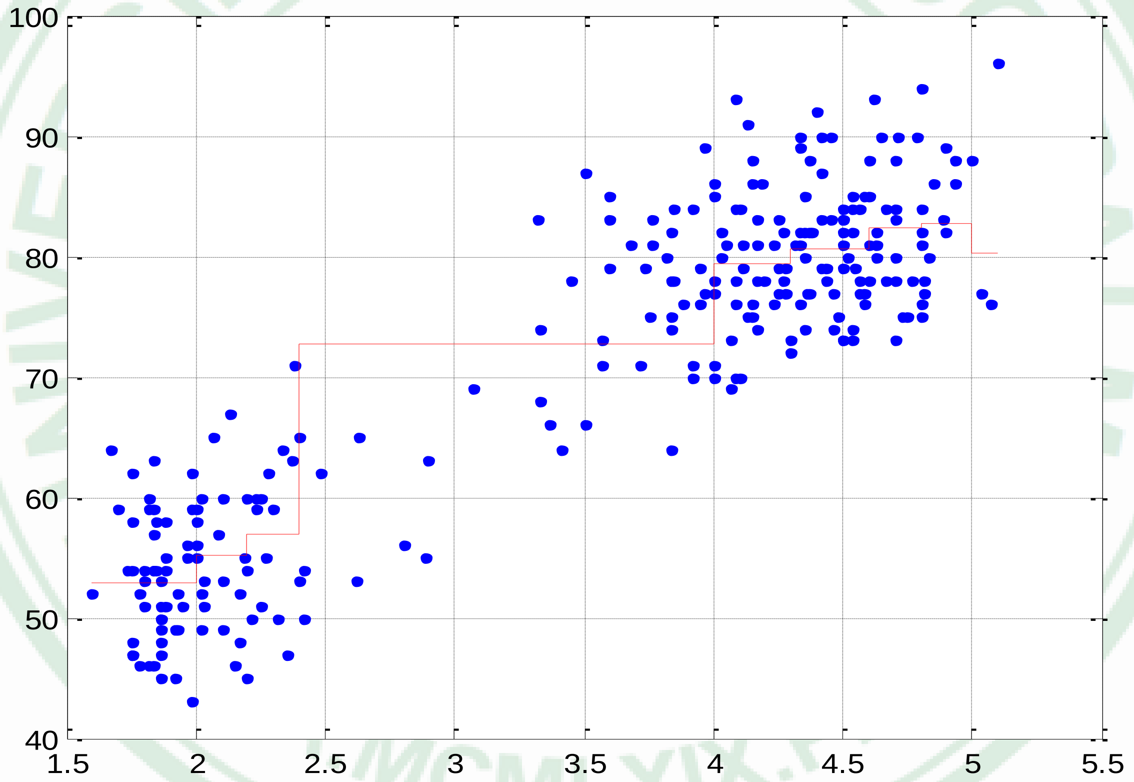
TALLER DE ESPECIALIZACIÓN "STATISTICAL SCIENCE INTRODUCTION"

FORMAS NO PARAMÉTRICAS

REGRESORGRAMA

- En el **Regresorgrama** se divide el intervalo de los valores de la variable predictora en varios subintervalos (usualmente 5). La amplitud de los subintervalos se elige de tal manera que haya aproximadamente igual número de datos en cada uno de ellos.
- Luego se promedia los valores de la variable de respuesta en cada subintervalo. Esto determina varios segmentos de línea que al unirse forma el **regresorgrama**. Lo malo de este estimador es que no es suave porque hay saltos en cada punto de corte.

REGRESORGRAMA



RUNNING

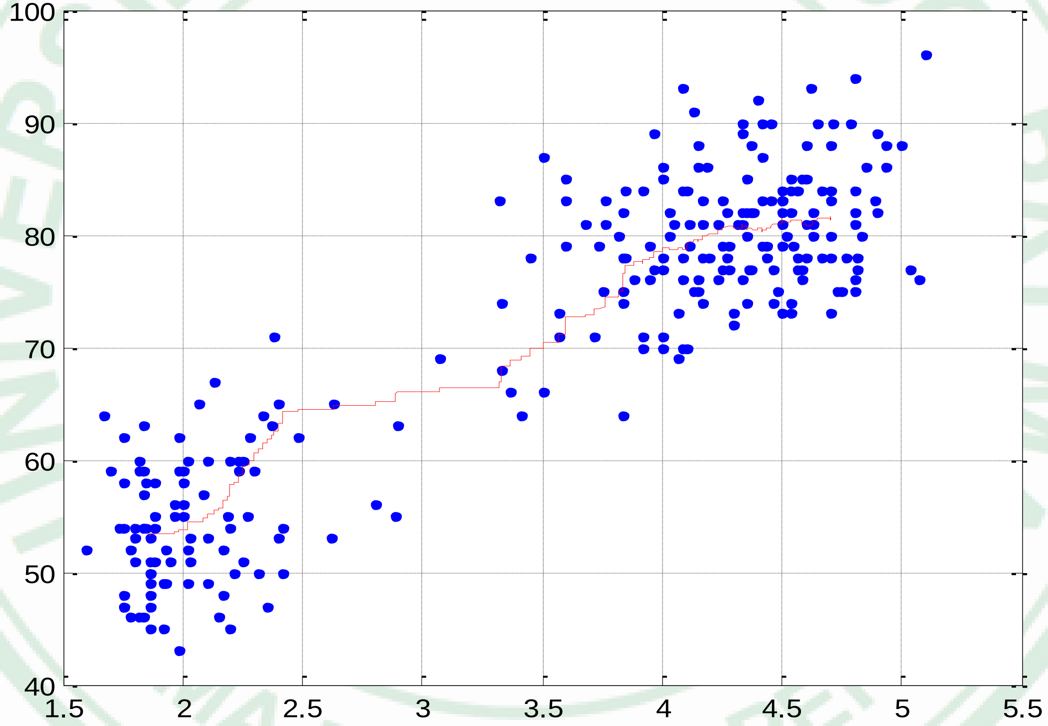
Primero, para cada valor de xi se define una vecindad simétrica N (xi) que contenga a dicho punto. La simetría esta en el sentido que el número de puntos k es el mismo tanto a la derecha como a la izquierda del punto dado, en los extremos esto no se puede lograr pero se trata de estar lo más cerca posible.

$$N(x_i) = \{\max(i-k,1),\ldots, i-1, i, i+1,\ldots, \min(i+k, n)\}.$$

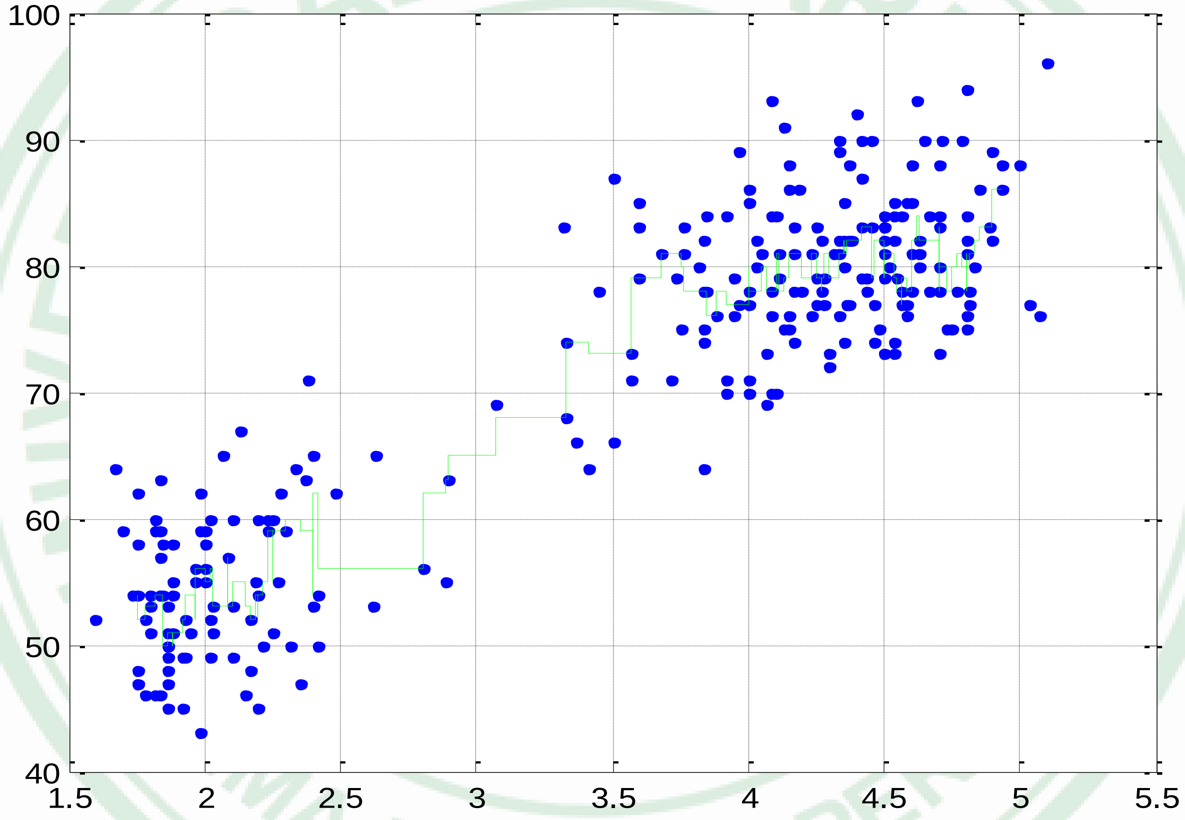
Luego se calcula el suavizador por running means, running medians ó running lines en el punto xi.

Método	Suavizador $s(x_i)$
Running Means	$s(x_i) =$ promedio de las y's en $N(x_i)$
Running Medians	$s(x_i) =$ mediana de las y's en $N(x_i)$
Running Lines	$s(x_i) =$ valor estimado de la regresión mínimo cuadrática para $x = x_i$ que se obtiene usando los puntos (x_j, y_j) con x_j que cae en $N(x_i)$

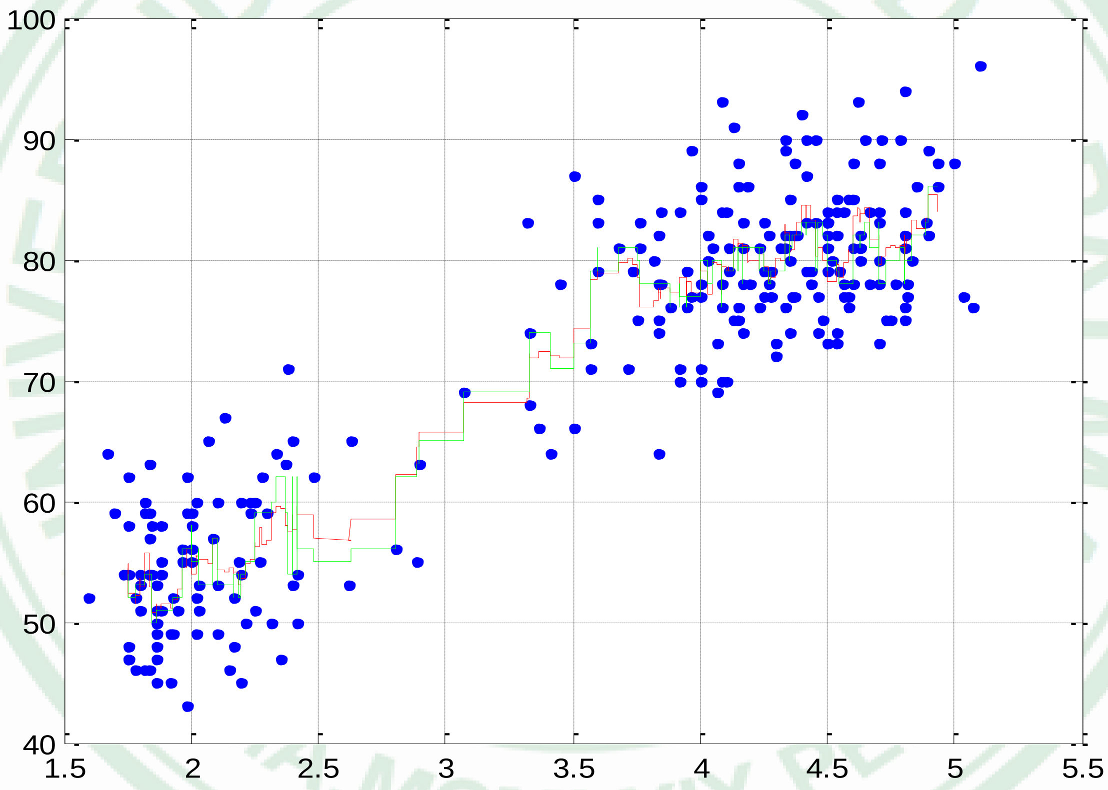
RUNNING MEANS



RUNNING MEDIANS



RUNNING MEANS Y MEDIANS



TALLER DE ESPECIALIZACIÓN "STATISTICAL SCIENCE INTRODUCTION"

¡Gracias!

TALLER DE ESPECIALIZACIÓN “STATISTICAL SCIENCE INTRODUCTION”