



International Center for Tropical Agriculture
Since 1967 Science to cultivate change

Fuente de datos, procesamiento datos faltantes y modelos de clasificación

Marzo, 2019

Juan Camilo Rivera
j.c.rivera@cgiar.org

Hugo Dorado
h.a.dorado@cgiar.org



Fuentes de información

Datos abiertos

WorldClim

Descripción:

Es un conjunto de capas de variables de clima con resolución cerca de un kilómetro. 1970 - 2000

Variables para la version 2.0:

- Temperatura maxima, minima y promedio.
- Precipitación.
- Radiación Solar
- Velocidad del viento
- Presión de vapor de agua.



Variables Bioclimaticas:

BIO1 = Annual Mean Temperature

BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))

BIO3 = Isothermality (BIO2/BIO7) (* 100)

BIO4 = Temperature Seasonality (standard deviation *100)

BIO5 = Max Temperature of Warmest Month

BIO6 = Min Temperature of Coldest Month

BIO7 = Temperature Annual Range (BIO5-BIO6)

BIO8 = Mean Temperature of Wettest Quarter

BIO9 = Mean Temperature of Driest Quarter

BIO10 = Mean Temperature of Warmest Quarter

BIO11 = Mean Temperature of Coldest Quarter

BIO12 = Annual Precipitation

BIO13 = Precipitation of Wettest Month

BIO14 = Precipitation of Driest Month

BIO15 = Precipitation Seasonality (Coefficient of Variation)

BIO16 = Precipitation of Wettest Quarter

BIO17 = Precipitation of Driest Quarter

BIO18 = Precipitation of Warmest Quarter

BIO19 = Precipitation of Coldest Quarter

Pagina web:

<http://worldclim.org/version2>

NOAA

National Oceanic Atmospheric Administration

- **Descripción:**

Creación de NCEI (National Centers for Environmental Information) la unión de cinco centros de información de oceanografía, clima y geofísica.

- **Página web:**

<https://www.ncdc.noaa.gov/cdo-web/datatools/findstation>





- **Descripción:**

Es una organización que pertenece a la University East Anglia que ayuda a los científicos a estudiar más a fondo los problemas de cambio de climático.

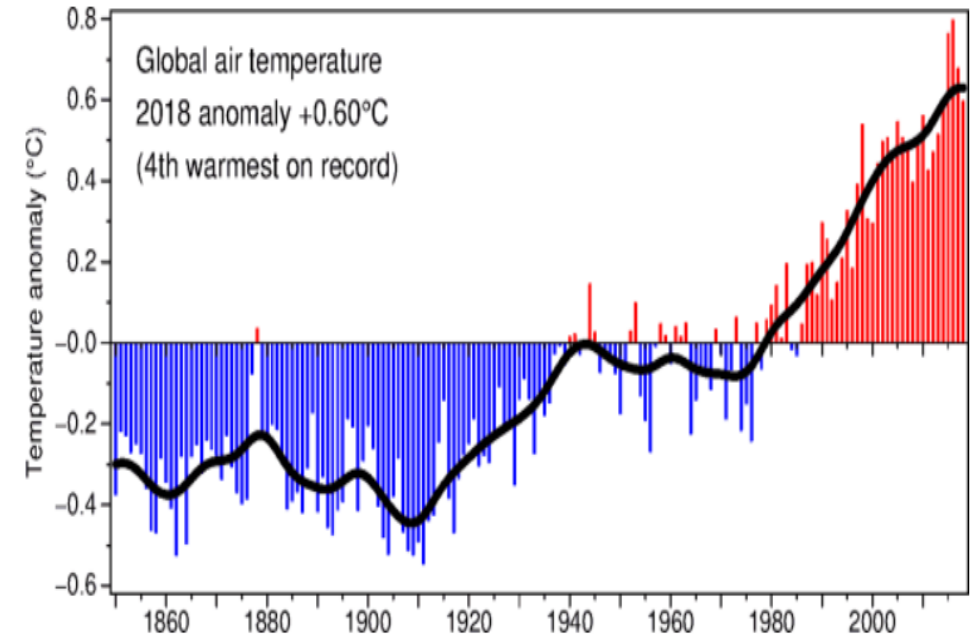
- **Tipo de formato:**

<https://crudata.uea.ac.uk/cru/data/temperature/#file>
or.

- **Página web:**

<http://www.cru.uea.ac.uk/>

<http://fabiolexcastrosig.blogspot.com/>



SOILDGRID

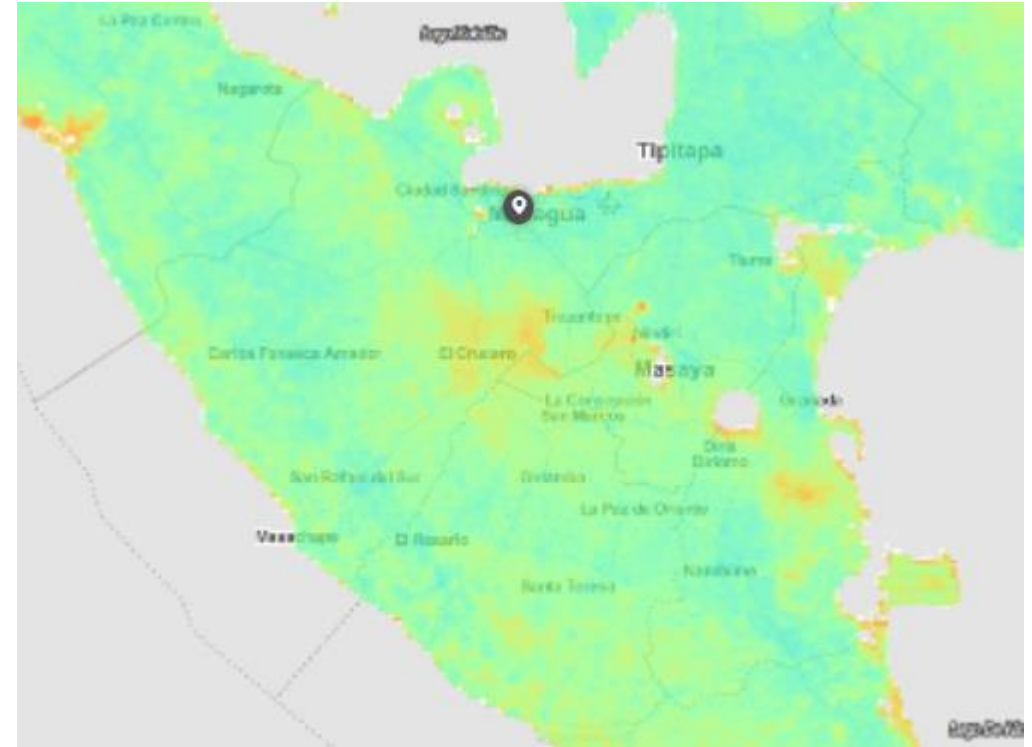


- **Descripción:**

Es un sistema automatizado de suelos basado en una compilación en datos de perfiles de suelo y sensores remotos de datos.

Pagina web:

- https://soilgrids.org/#!/?layer=ORCDRC_Msl3_250m&vector=1



CHIRPS



- **Descripción:**

Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS es una base datos de 30 años de precipitaciones a nivel global.

Pagina web:

- <http://chg.geog.ucsb.edu/data/chirps/>

Forma descargarlo

<http://fabiolexcastro.sig.blogspot.com/2016/07/descarga-de-automatizada-de-archivos.html>

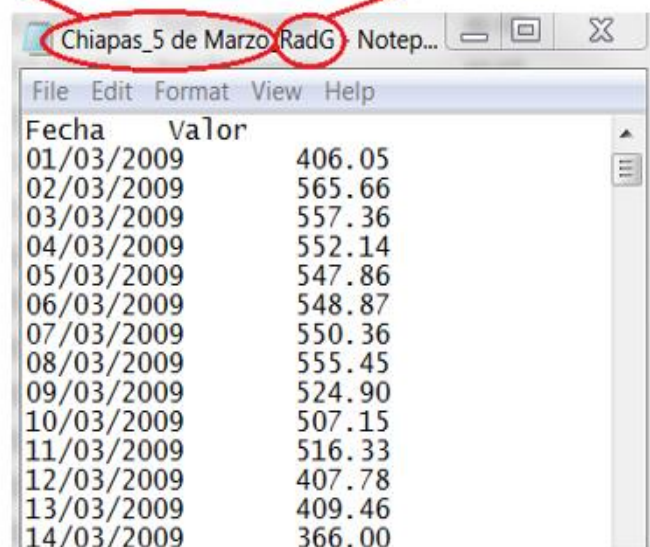


Archivos planos

Tipos

Nombre estación

Variable climatologica



Fecha	Valor
01/03/2009	406.05
02/03/2009	565.66
03/03/2009	557.36
04/03/2009	552.14
05/03/2009	547.86
06/03/2009	548.87
07/03/2009	550.36
08/03/2009	555.45
09/03/2009	524.90
10/03/2009	507.15
11/03/2009	516.33
12/03/2009	407.78
13/03/2009	409.46
14/03/2009	366.00



	A	B	C	D	E	F
1	DATE	ESOL	RAIN	RHUM	TMAX	TMIM
.557	4/5/2009	412.8747	0	70.99139	36	24.3016
.558	4/6/2009	513.9043	0	75.20833	34.8	24.9
.559	4/7/2009	396.5338	0	73.85714	34.1	25.6
.560	4/8/2009	397.8491	0	74.09524	33.9	25.4
.561	4/9/2009	448.4498	0	76.82609	34.6	24.9
.562	4/10/2009	481.8188	0	66.20671	39	24.8
.563	4/11/2009	448.1053	0	73.66386	35.9	25.4

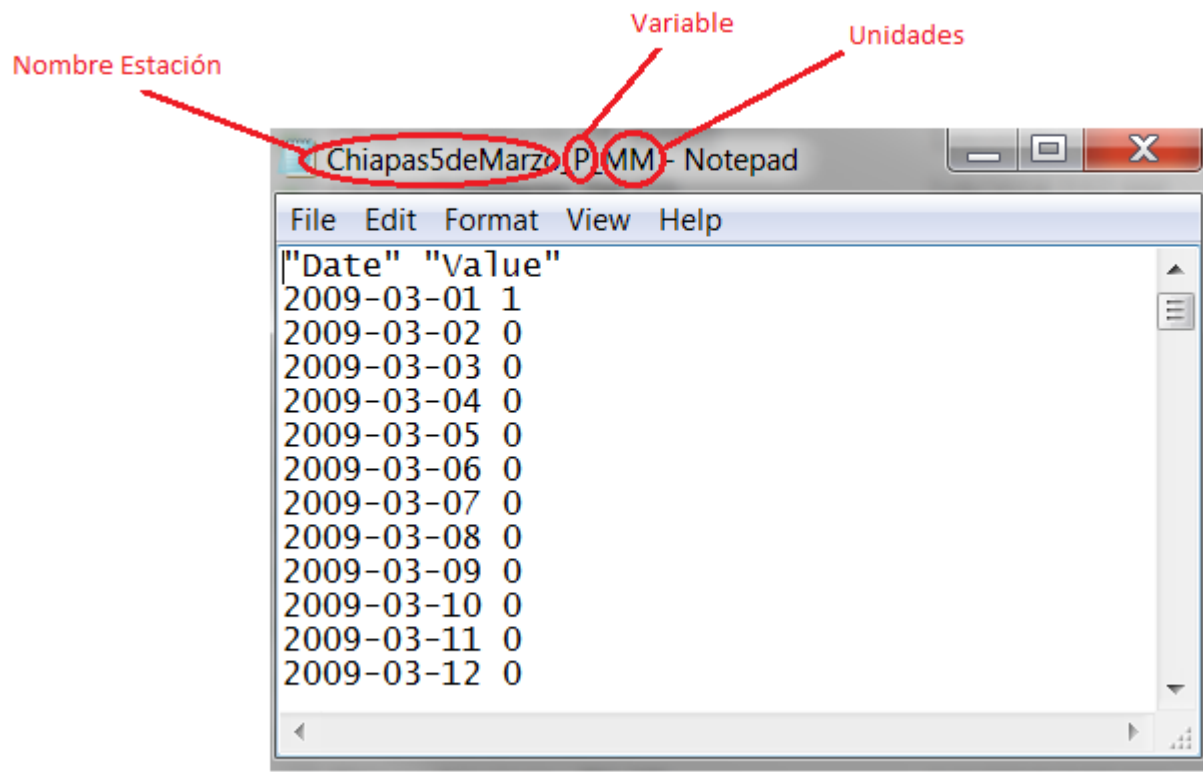
Variables y unidades

Abreviación	Significado (Ingles)	Significado (español)
TX	Maximum temperature	Temperatura máxima
TM	Minimum temperature	Temperatura mínima
P	Precipitation	Precipitación
RH	Relative humidity	Humedad relative
SR	Solar radiation	Radiación solar

Unidades

Abreviacion	Unidad de Medida
CD	Grados Celisus
FD	Grados Fahrenheit
MM	Mililitros
NE	Número entre 0 y 100
CCM2	Calorias por centimetro cuadrado
MJM2	Mega Julio por metro cuadrado
WAM2	Watts por metro cuadrado

Formato único



Llenado de faltantes

Date	value
19800101	NA
19800102	NA
19800103	NA
19800104	NA
19800105	NA
19800106	NA
19800107	NA
19800108	NA
19800109	NA
19800110	NA
19800111	35.2
19800112	NA
19800113	NA
19800114	36.2
19800115	35.2
19800116	NA

Vector Autoregresivo Regresión (VAR)

$$x_t = A_1 \cdot x_{t-1} + \dots + A_p \cdot x_{t-p} + u_t$$

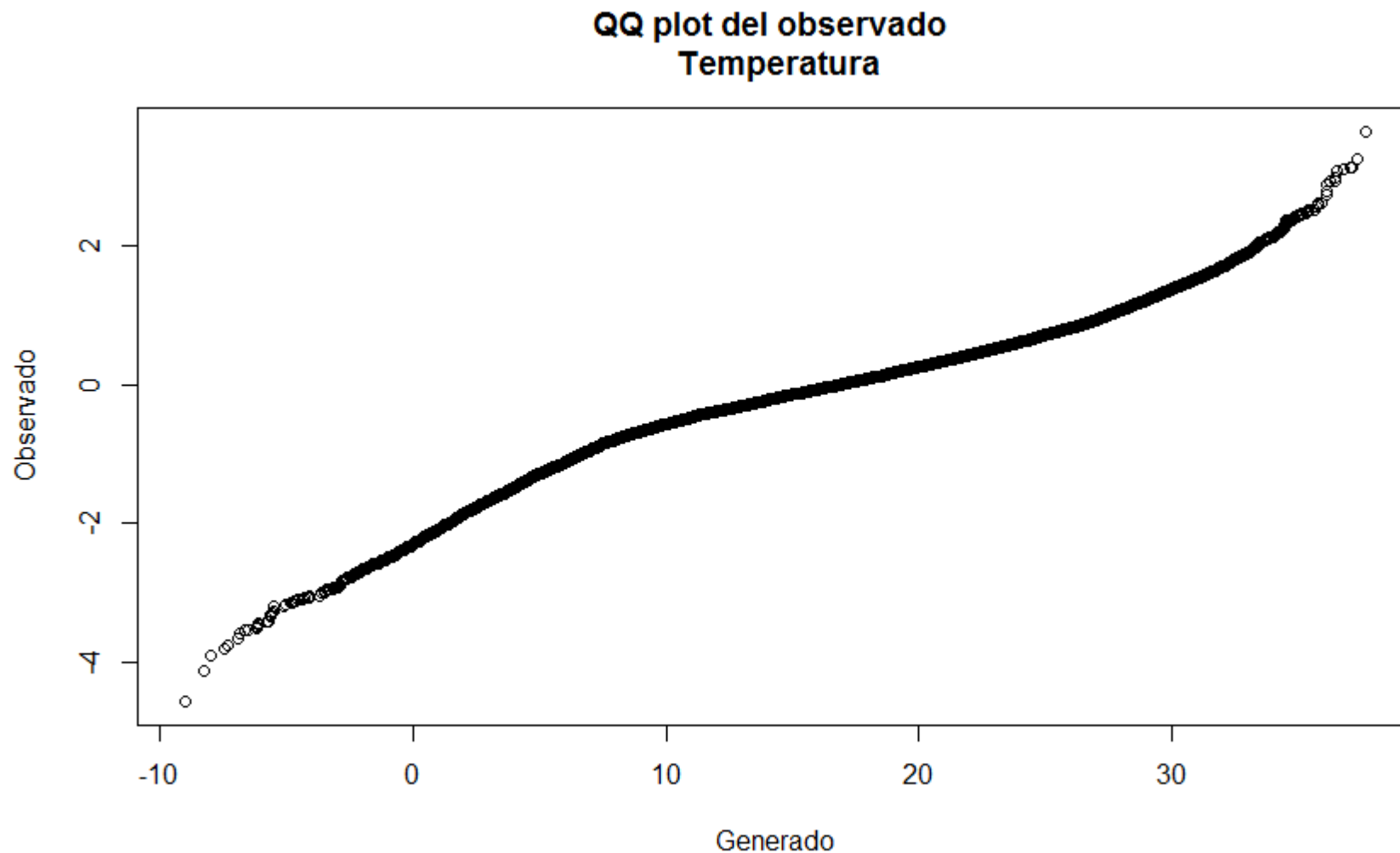
x_t = Vector de dimension K, conjunto de variables de clima.

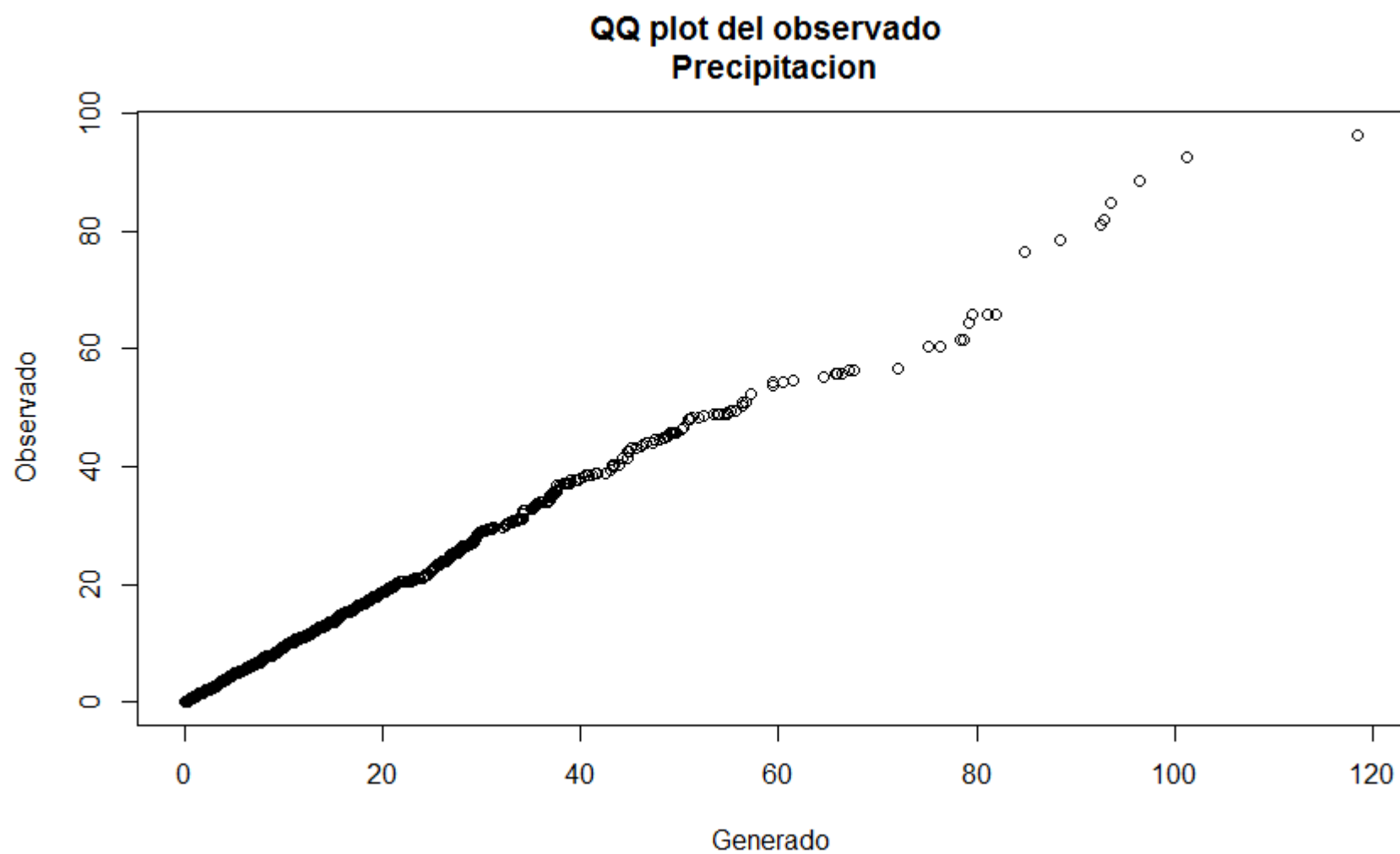
A_i = Es el coeficiente de la matriz K x K

u_t = Es un proceso estocastico de dimension K

- Modelo estocastico usado para capturar la relación lineal entre multiple series de tiempo.
- Es una generalización de los modelos AR modelos autoregresivos.

Ejemplo de Rmwagen





TAREAS EN MINERIA DE DATOS

- **AGRUPACION**

Consiste en dividir un conjunto de instancias de un dominio dado, descrito por un número de atributos discretos o de valor continuo, en un conjunto de grupos (clústeres) basándose en la similitud entre las instancias

- **REGRESIÓN**

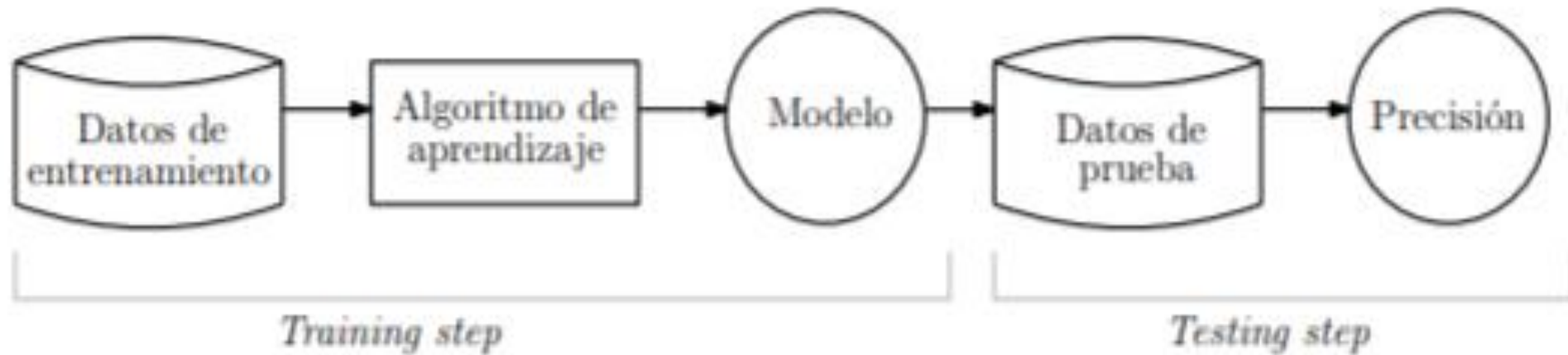
Los modelos de regresión predicen valores numéricos en lugar de etiquetas de clase discretas

- **CLASIFICACIÓN**

Asignar instancias de un dominio dado, descritas por un conjunto de atributos discretos o de valor continuo, a un conjunto de clases.

APRENDIZAJE SUPERVISADO

Regresión y clasificación



k-fold cross validation

Una partición aleatoria del conjunto de datos en k conjuntos del mismo tamaño, usando $k - 1$ conjuntos para entrenar el modelo y el conjunto restante para evaluarlo, repitiendo el proceso k veces y promediando el error estimado.



Evaluación de Resultados

- CLASIFICACIÓN

Matriz de confusión

		Clase predicha	
		P	N
Clase verdadera	P	TP	FN
	N	FP	TN

Verdadero positivo (*True Positive*, TP): número de clasificaciones correctas en la clase positiva (*P*).

Verdadero negativo (*True Negative*, TN): número de clasificaciones correctas en la clase negativa (*N*).

Falso negativo (*False Negative*, FN): número de clasificaciones incorrectas de clase positiva clasificada como negativa.

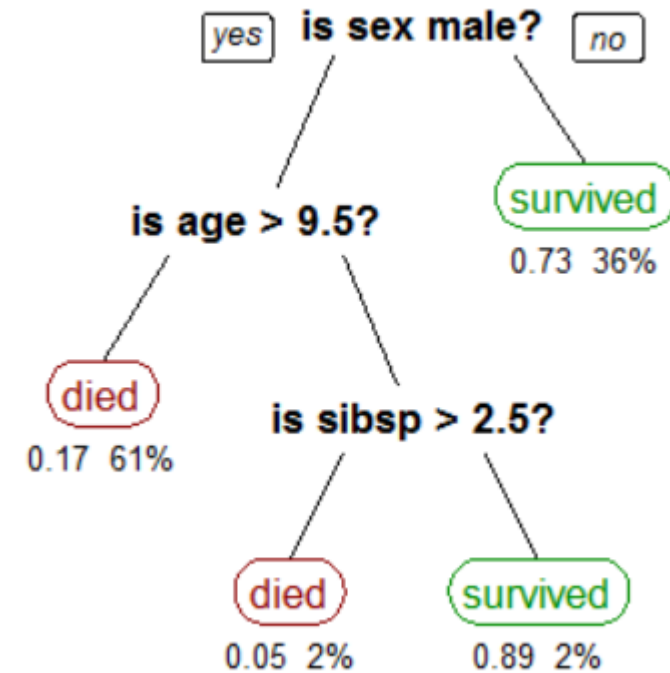
Falso positivo (*False Positive*, FP): número de clasificaciones incorrectas de clase negativa clasificada como positiva.

Random Forest

Método de **regresión** y **clasificación**.

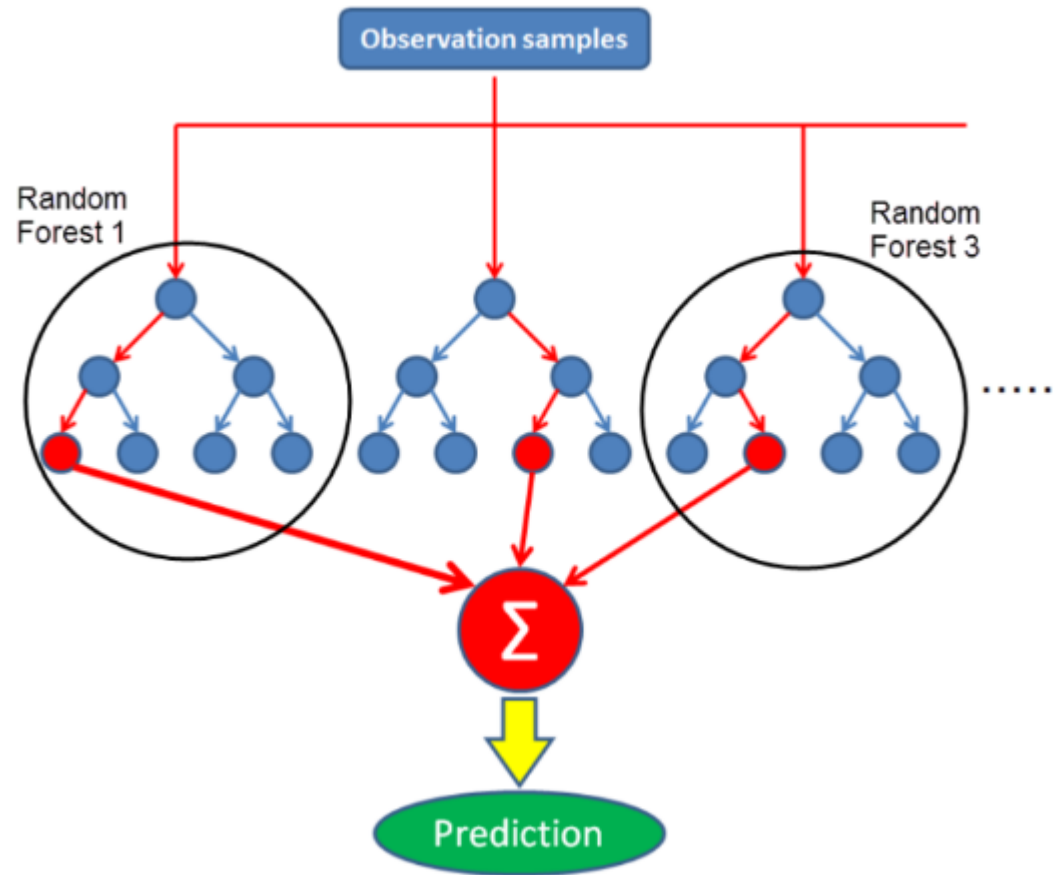
Esta basado en **CART** (Árboles de decision).

- Cada nodo corresponde a una variable de entrada.
- Ramas son los posibles valores que puede tomar la variable.
- Las ramas inferiores muestran la variable a predecir.



Estructura Random Forest

- Cadena de Árboles aleatorios
- No correlacionados
- Combinados usando nodo optimización
- El resultado más votado será el ganador



Thank you!



WE'RE PROUD TO
HAVE CELEBRATED 50 YEARS
OF AGRICULTURAL RESEARCH
FOR DEVELOPMENT

International Center for Tropical Agriculture - CIAT

Headquarters and Regional Office
for South America and the Caribbean

+57 2 445 0000

Km 17 Recta Cali-Palmira
A.A. 6713, Cali, Colombia

✉ ciat.cgiar.org

🌐 ciat.cgiar.org



CIAT is a CGIAR Research Center