Minería de datos

Hugo Andrés Dorado

Contenido

- Definiciones en minería de datos
- Pre procesamiento
- Tipo de aprendizaje
- Modelos supervisados
- Tipos de error
- Partición del conjunto de datos
- Medidas de desempeño
- Optimización de parámetros
- Interpretación de modelos

Conceptos en minería de datos

- **Características, variables de entrada (Features):** Variables medidas sobre las observaciones que se asocian luego a un variable salida.

- **Variable de salida:** Variable a explicar de interés.

- **Función costo:** Es una función que permite aproximar un conjunto de variables de entrada para generar una respuesta aproximada según la variable de salida.

$$Y = f(X) + \varepsilon$$

Preprocesamiento

- Dumificación de variables

Row Number	Direction
1	North
2	North-West
3	South
4	East
5	North-West

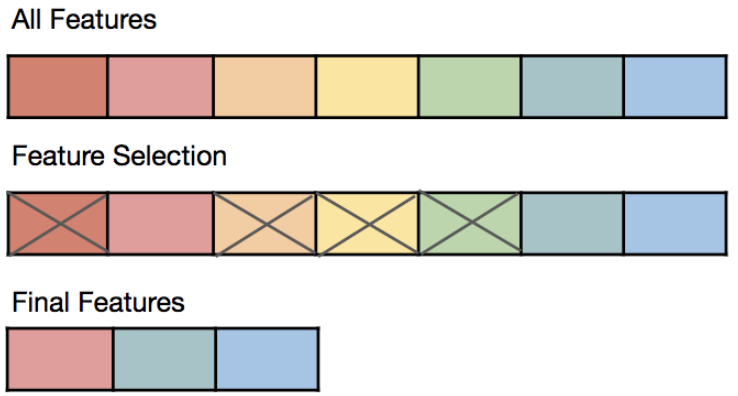
Row Number	Direction_N	Direction_S	Direction_W	Direction_E	Direction_NW
1	1	0	0	0	0
2	0	0	0	0	1
3	0	1	0	0	0
4	0	0	0	1	0
5	0	0	0	0	1

- Normalización de datos

Normalization Formula

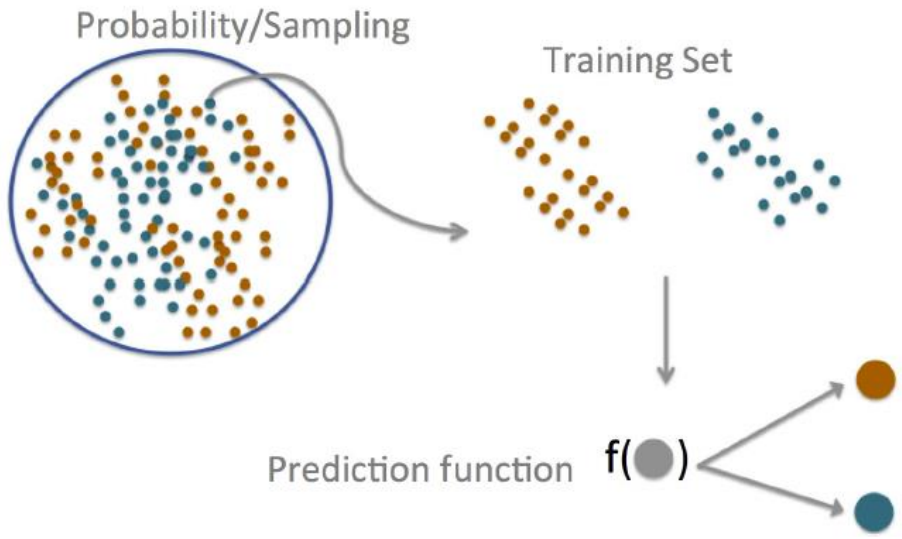
$$X_{normalized} = \frac{(X - X_{minimum})}{(X_{maximum} - X_{minimum})}$$

- Selección de atributos



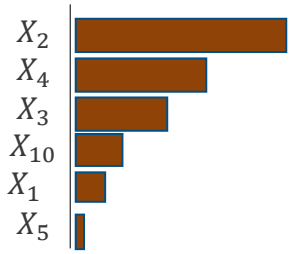
Tipo de aprendizaje

- Supervisado

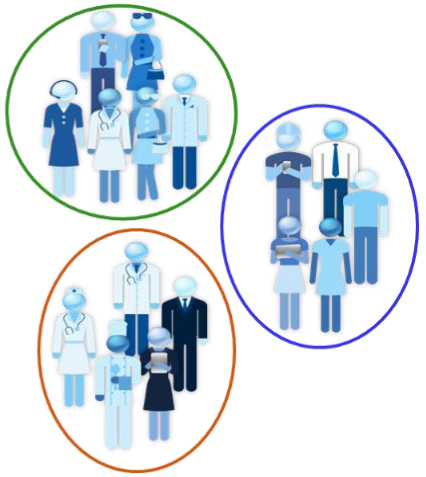


- Clasificación.
- Regresión.

Ranking de importancia de variables



- No supervisado



- Análisis cluster.
- PCA.
- TSE

Modelos supervisados

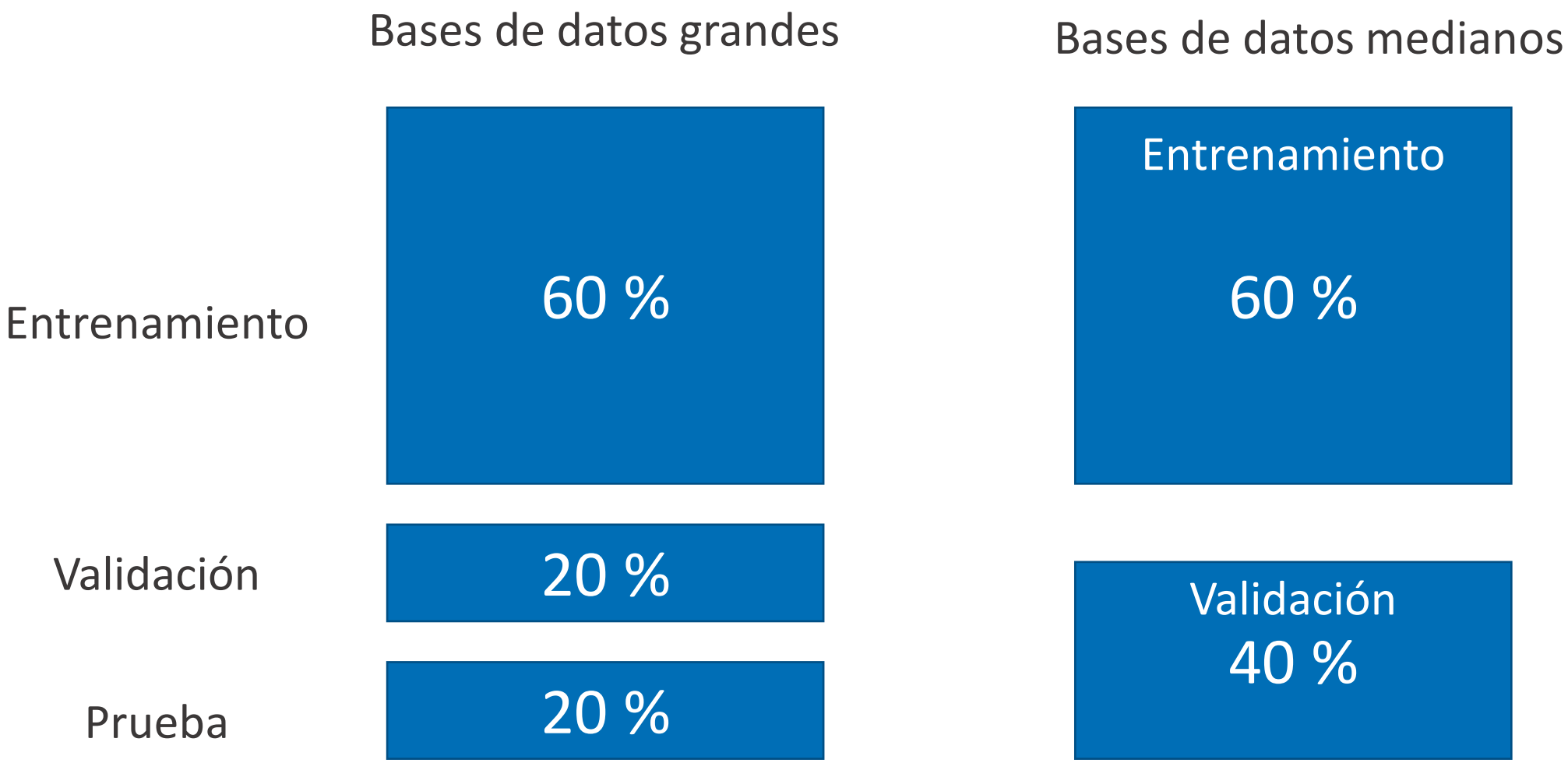
- **Pregunta**:¿Que mails son spam?,¿Que zonas son bosque?,¿Que clientes serán morosos?
- **Entrada de datos:** conjuntos de e-mail, Imágenes satelitales, información de clientes. (Ya clasificados)
- **Variables de entrada:** Frecuencia de ciertas palabras, índices espectrales por color, variables seleccionadas
- **Algoritmo:** Redes neuronales artificiales, suppor vector machine, J46
- **Parámetros:** (Tasa de decaimiento, neuronas ocultas) ,(costo), (umbral de confianza)
- **Evaluación.** (Precisión, exactitud, concordancia)

Tipos de error.

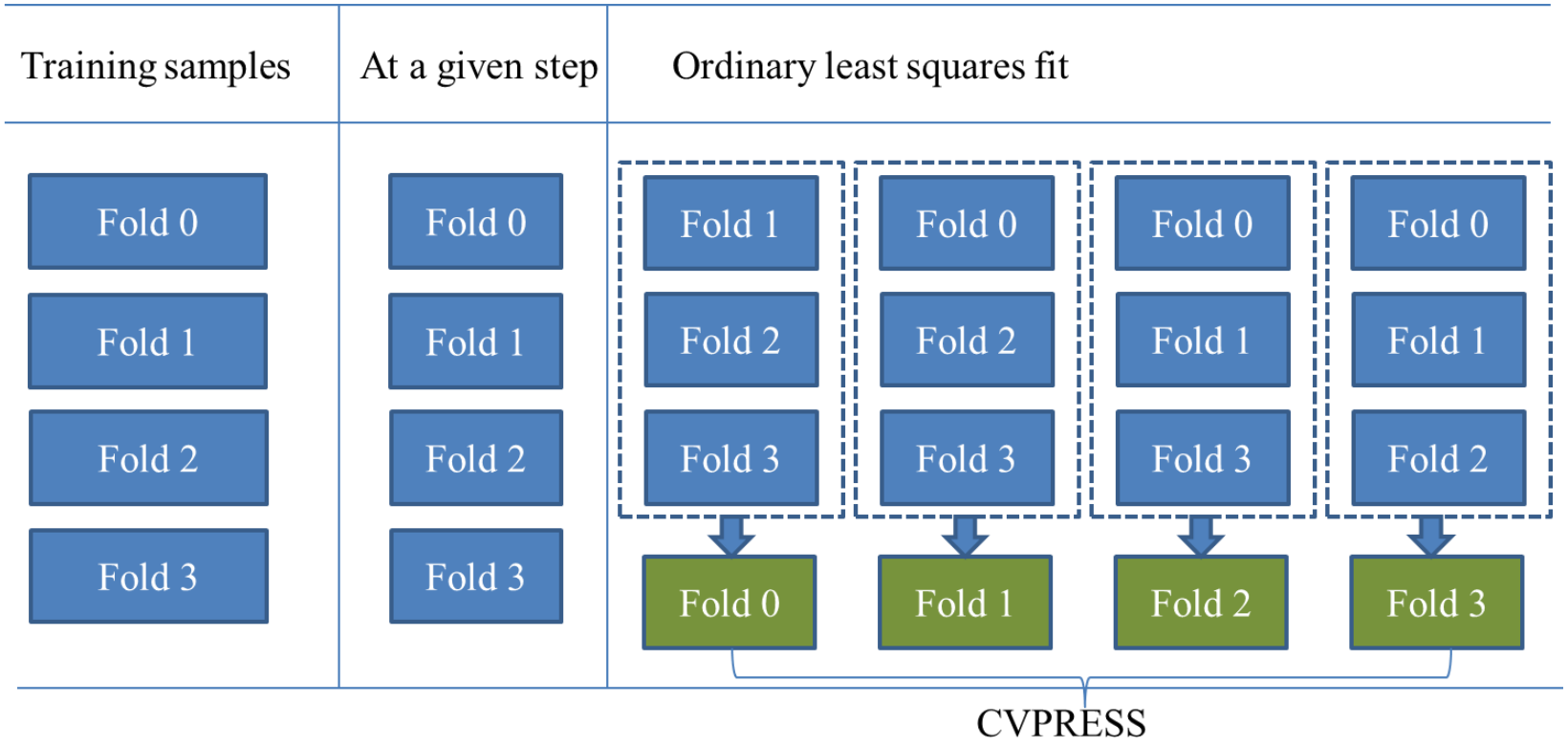
- **Error dentro de la muestra:** La tasa de error que se obtiene en los mismos datos para construir el modelo.

- **Error fuera de la muestra:** La tasa de error que se obtiene al traer nuevos datos no mostrados, también conocido como error de generalización.

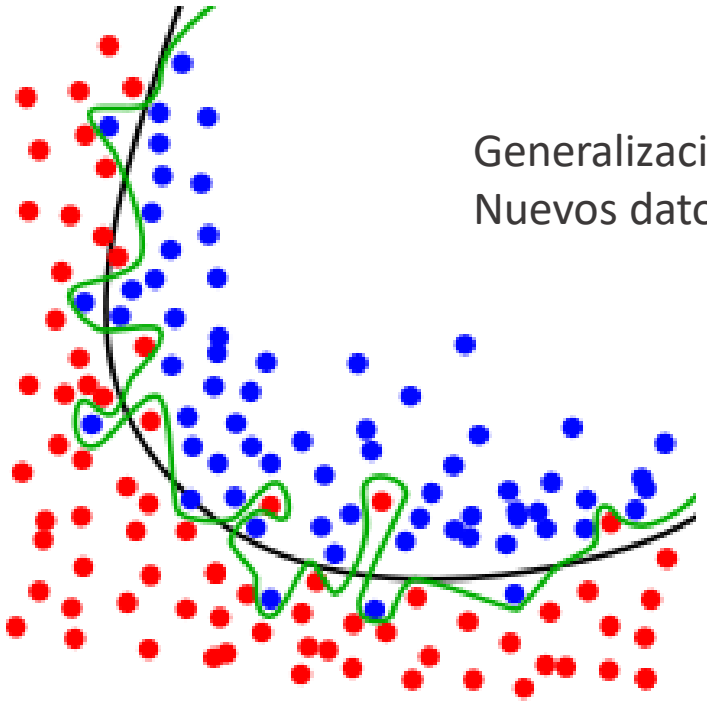
Partición del conjunto de datos



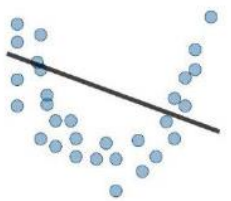
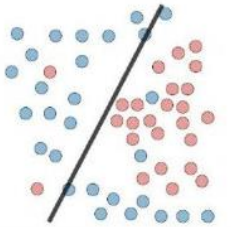
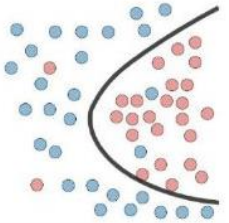
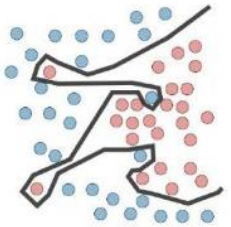
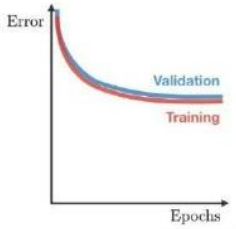
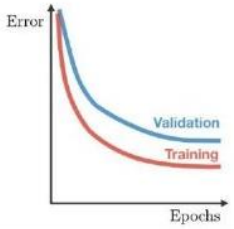
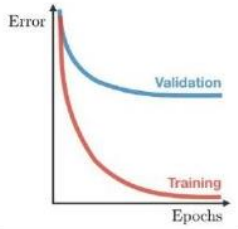
Partición del conjunto de datos - K - fold



Overfitting – sobre parametrización

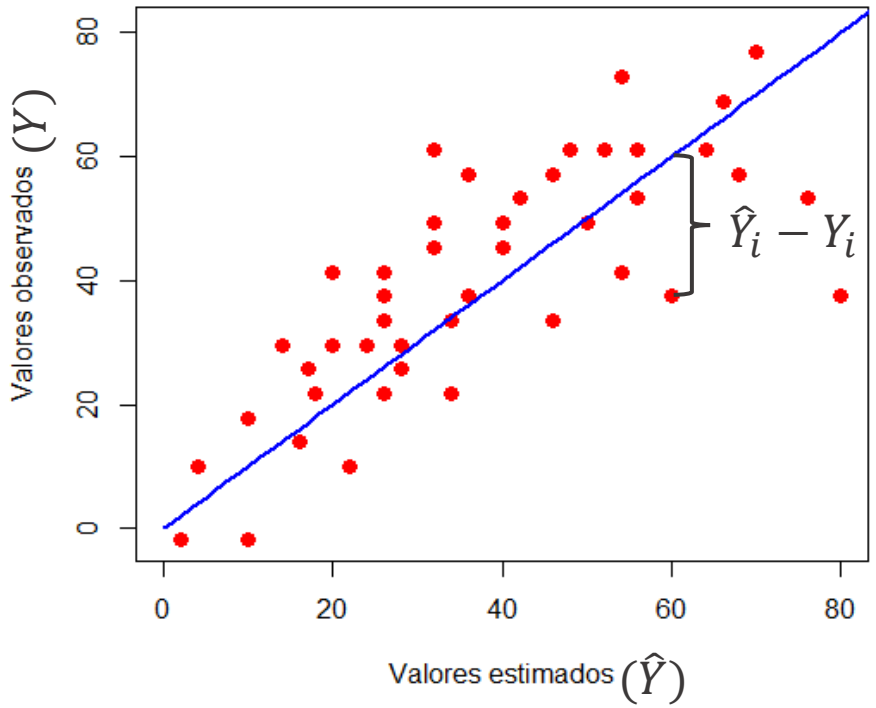


Generalización
Nuevos datos

	Underfitting	Just right	Overfitting
Symptoms	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Regression			
Classification			
Deep learning			
Remedies	- Complexify model - Add more features - Train longer		- Regularize - Get more data

Medidas de desempeño

Modelos de regresión



Raíz de errores cuadráticos medios

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n}(\hat{Y}_i - Y_i)^2}{n}}$$

n= número de observaciones

Coeficiente de determinación

$$R^2 = \rho^2_{\hat{y}y}$$

Un valor entre [0,1]

Medidas de desempeño

Modelos de clasificación

		Realidad	
Predicción		Ocurre	No ocurre
	Ocurre	A	B
	No ocurre	C	D

Accuracy /Exactitud

$$\frac{A+D}{A+B+C+D} = \frac{Casos\ acertados}{Casos\ posibles} \quad [0,1]$$

$$kappa = \frac{P_r(\alpha) + P_r(e)}{1 - P_r(e)} \quad [-1,1]$$

$P_r(\alpha)$ = Acuerdo relativo
$P_r(e)$ = Probabilidad por azar

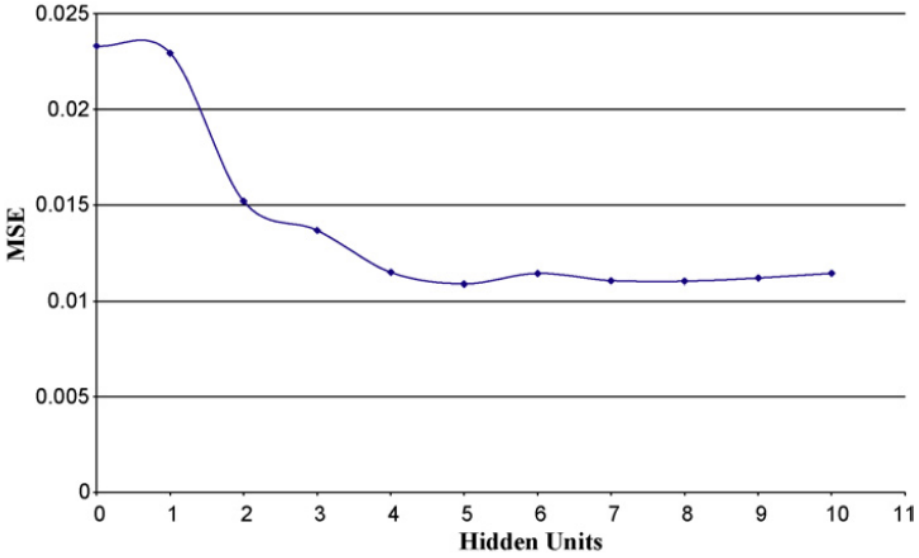
Valoración del Índice Kappa	
Valor de k	**Fuerza de la concordancia**
< 0.20	Pobre
0.21 – 0.40	Débil
0.41 – 0.60	Moderada
0.61 – 0.80	Buena
0.81 – 1.00	Muy buena

Optimización de parámetros

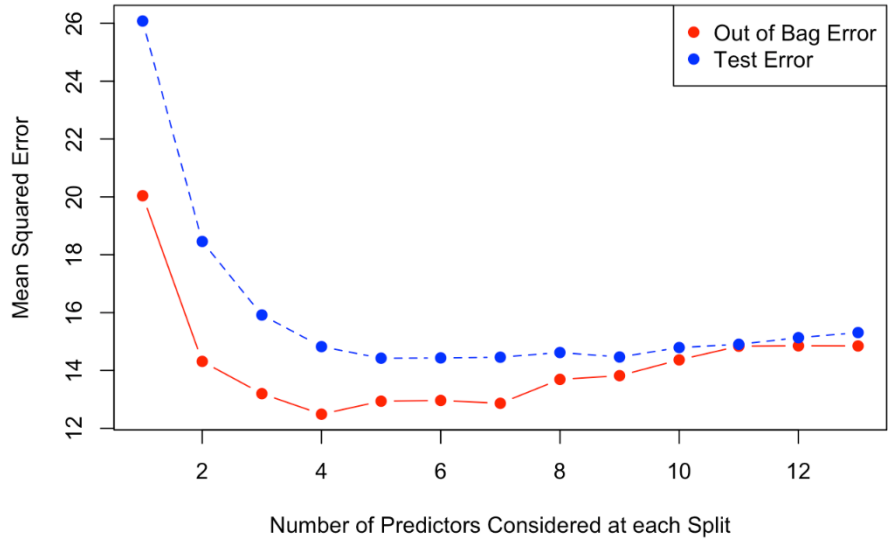
Model	*method* Value	Type	Libraries	Tuning Parameters
Neural Network	nnet	Classification, Regression	nnet	size, decay
Random Forest	rf	Classification, Regression	randomForest	mtry
CART	rpart	Classification, Regression	rpart	cp

http://topepo.github.io/caret/index.html

nnet (size)

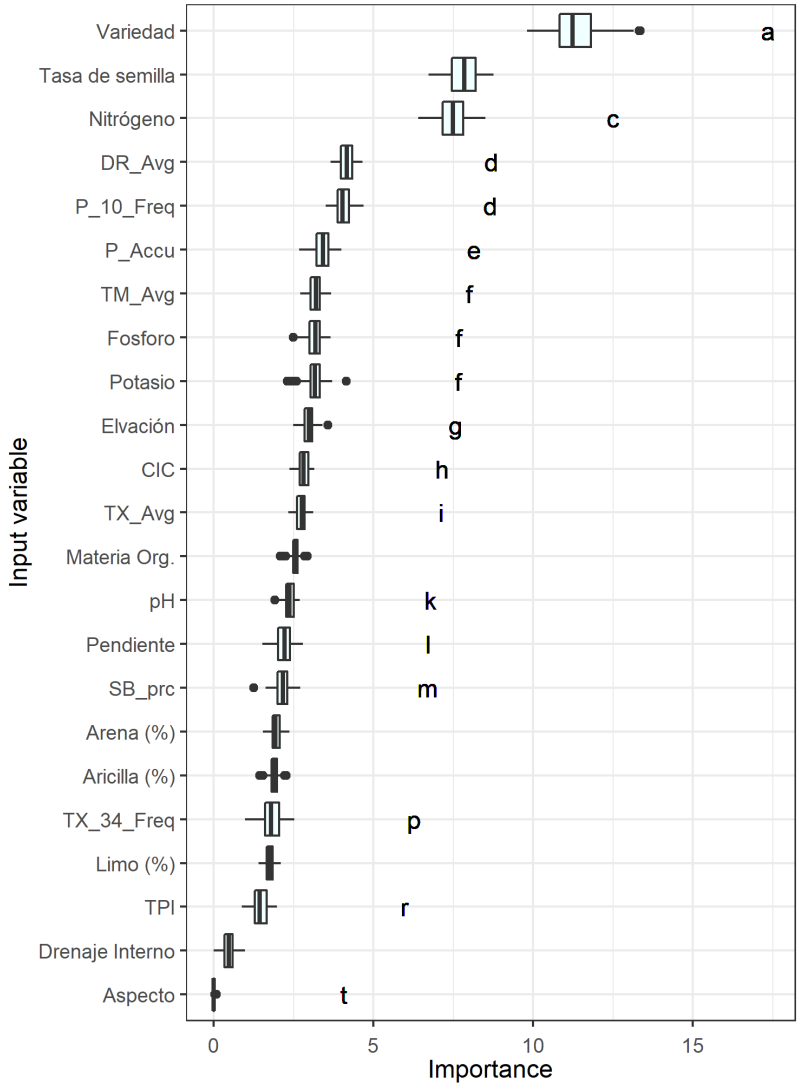


rf (mtry)



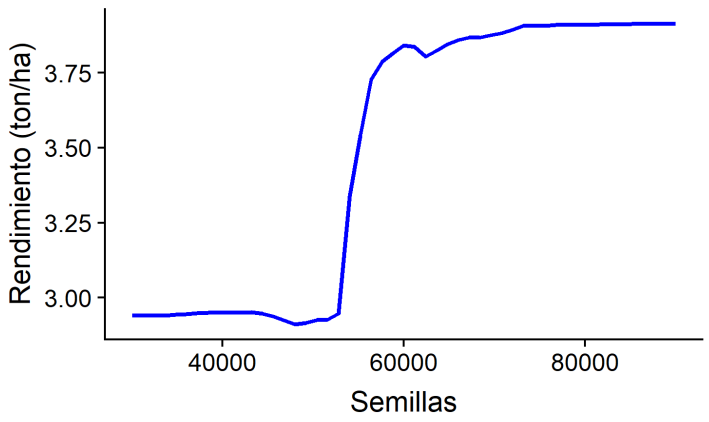
Interpretación de modelos

Importancia de variables

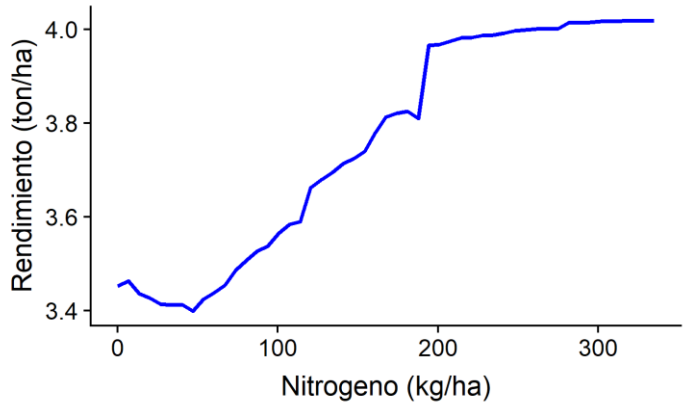


Dependencias parciales

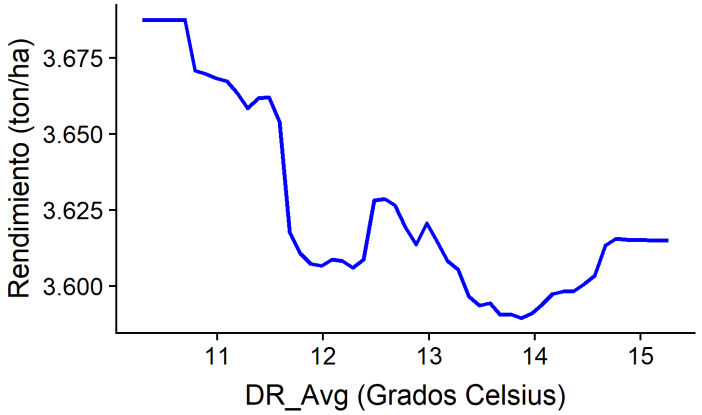
Alrededor de 60.000 semillas por HA



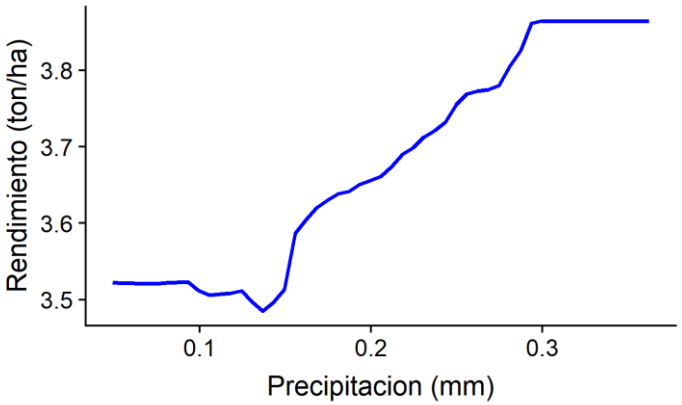
Una cantidad óptima de 180 kg/ha de nitrógeno



Un rango diurno de menos de 11.5^{O}C



Por lo menos 30% de días con lluvia



El mejor método de aprendizaje de máquina

Interpretable

Simple

Preciso

Rápido

Escalable

https://www.coursera.org/specializations/jhu-data-science

Puntos claves

- La minería de datos ofrece una gama variada de herramientas para explotar el conocimiento que pueda tener la información registrada en bases de datos.

- Para llevar a cabo un proyecto de minería de datos se debe seguir un cuidadoso procesamiento, análisis y evaluación de los datos.

- El mejor método de aprendizaje de máquina debe tener característica que le permiten tener un balance entre interpretación, complejidad, efectividad, velocidad y escalabilidad.

Gracias!

Hugo Andrés Dorado

Investigador Alianza Bioversity CIAT

h.a.dorado@cgiar.org