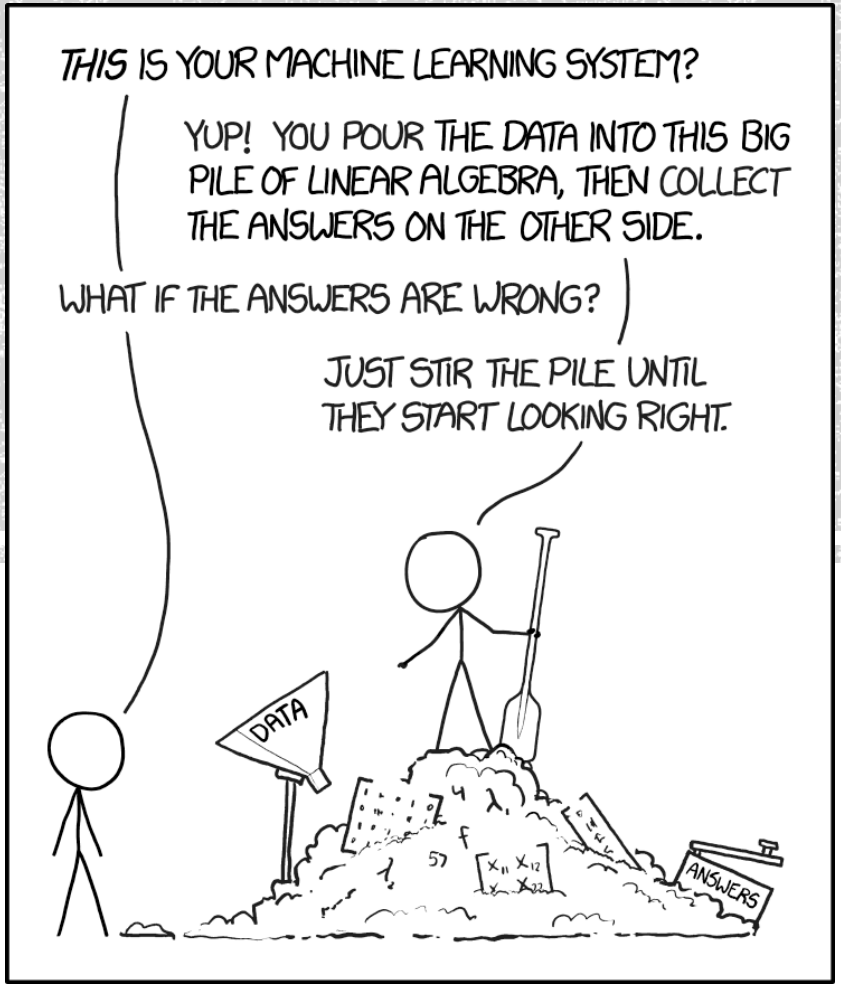
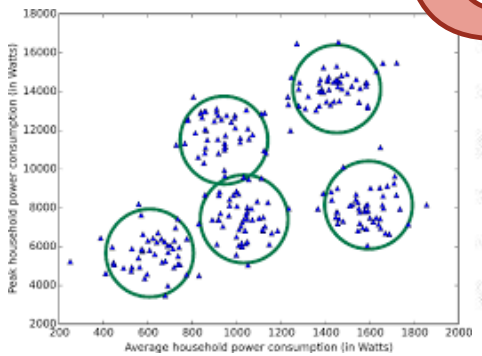
APRENDIZAJE AUTOMÁTICO

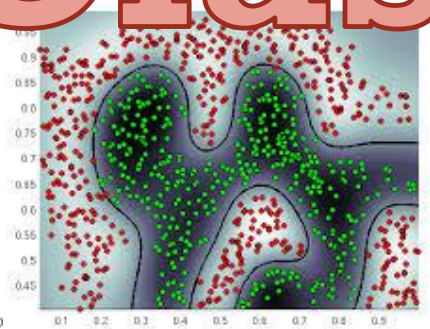


AGENDA

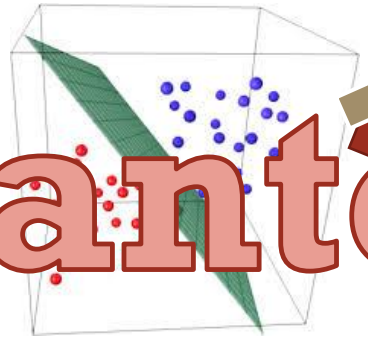
Aprendizaje
automático



Aprendizaje
no supervisado

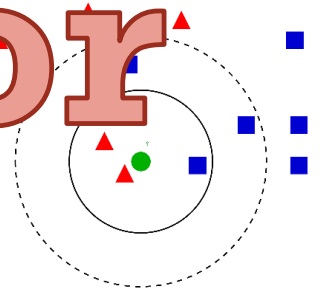


Aprendizaje
supervisado



Clasificación

Métricas de
Evaluación de la
clasificación

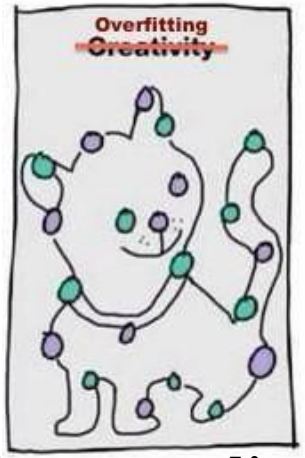


KNN

Clase anterior

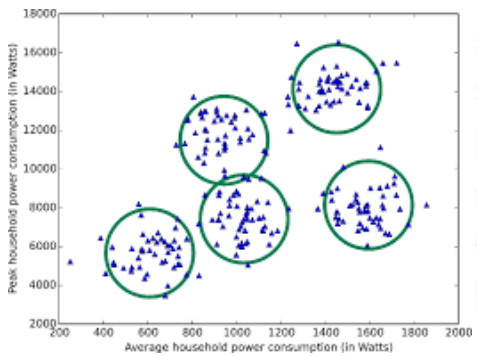
AGENDA

Protocolos

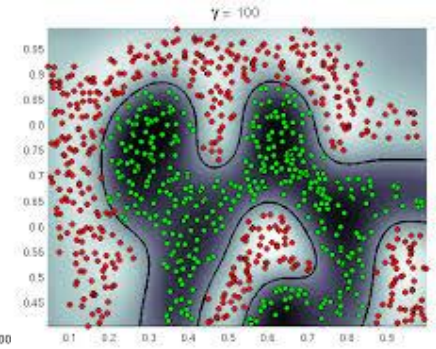


Sobre aprendizaje (Overfitting)

Aprendizaje automático

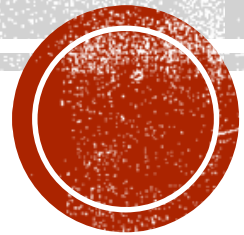


Aprendizaje no supervisado



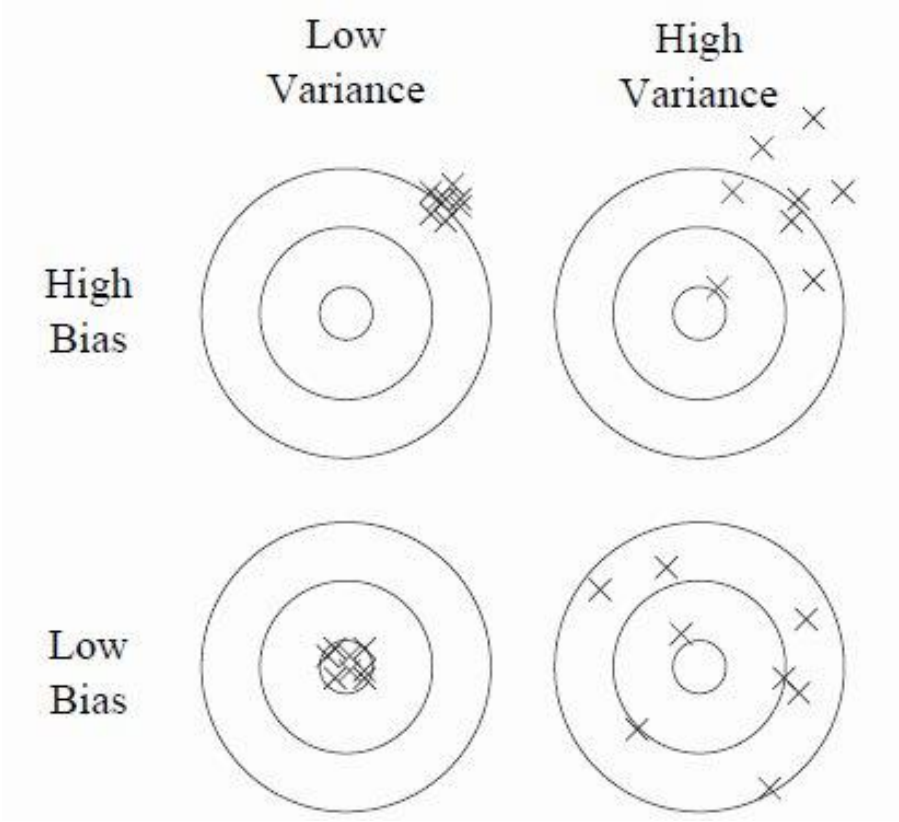
Aprendizaje supervisado

PROTOCOLOS DE EVALUACIÓN



SESGO/ VARIANZA

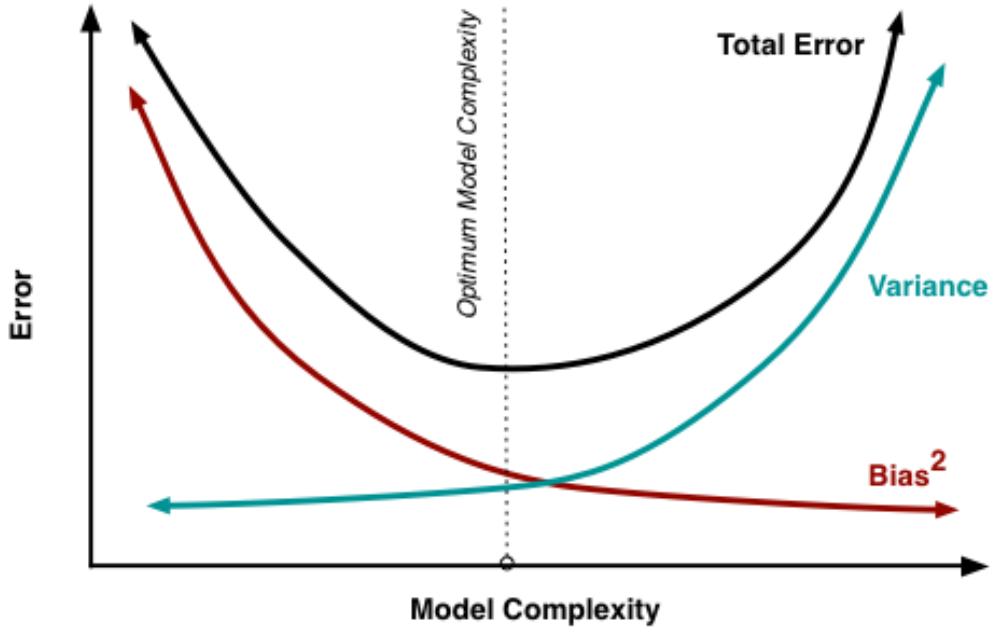
- **Sesgo** (bias): que tan lejos está el modelo de la verdad
- **Varianza**: Qué tanto varían los datos de la predicción para una misma instancia



Domingo, 2012

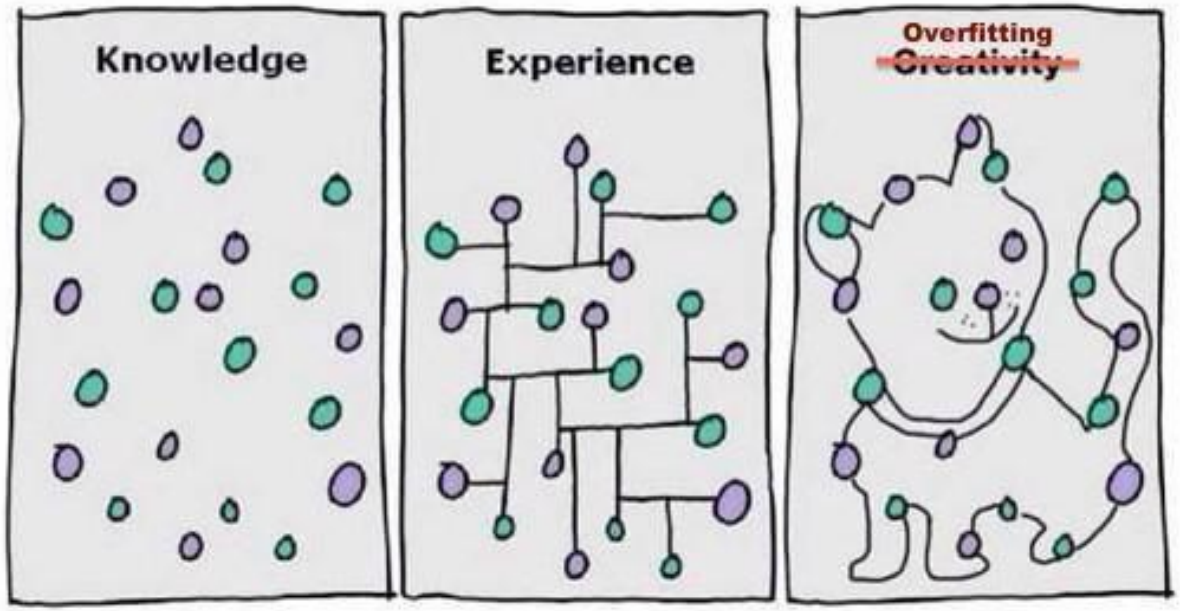
SESGO/ VARIANZA

- Ambos son fuente de error
- Se debe determinar un **compromiso** entre ambos tipos de error
- Parámetros de los modelos controlan la complejidad



http://scott.fortmann-roe.com/docs/BiasVariance.html

SOBRE APRENDIZAJE (OVERFITTING)

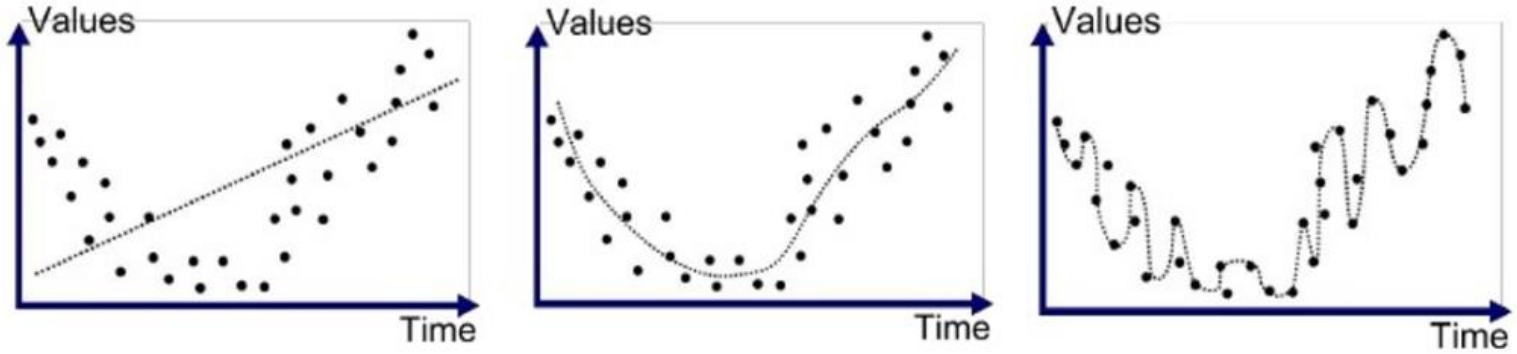


http://blog.algotrading101.com/design-theories/what-is-curve-fitting-overfitting-in-trading/

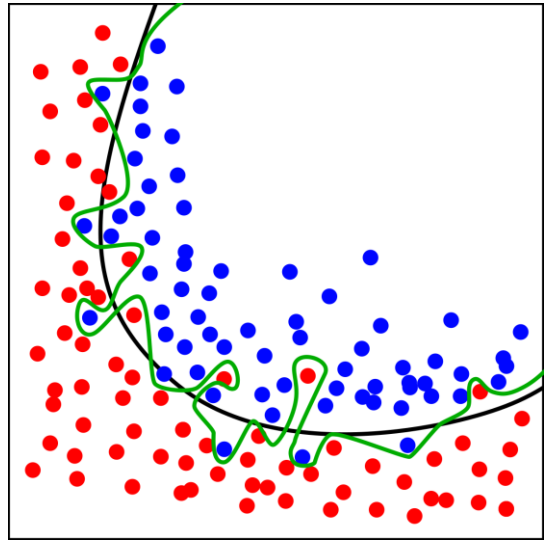
- **Sobre aprendizaje**: Los modelos aprenden a describir los errores aleatorios o el "ruido" del conjunto de entrenamiento.
- Ocurre cuando un modelo se vuelve excesivamente **complejo**

SOBRE APRENDIZAJE (OVERFITTING)

Regresión



Clasificación



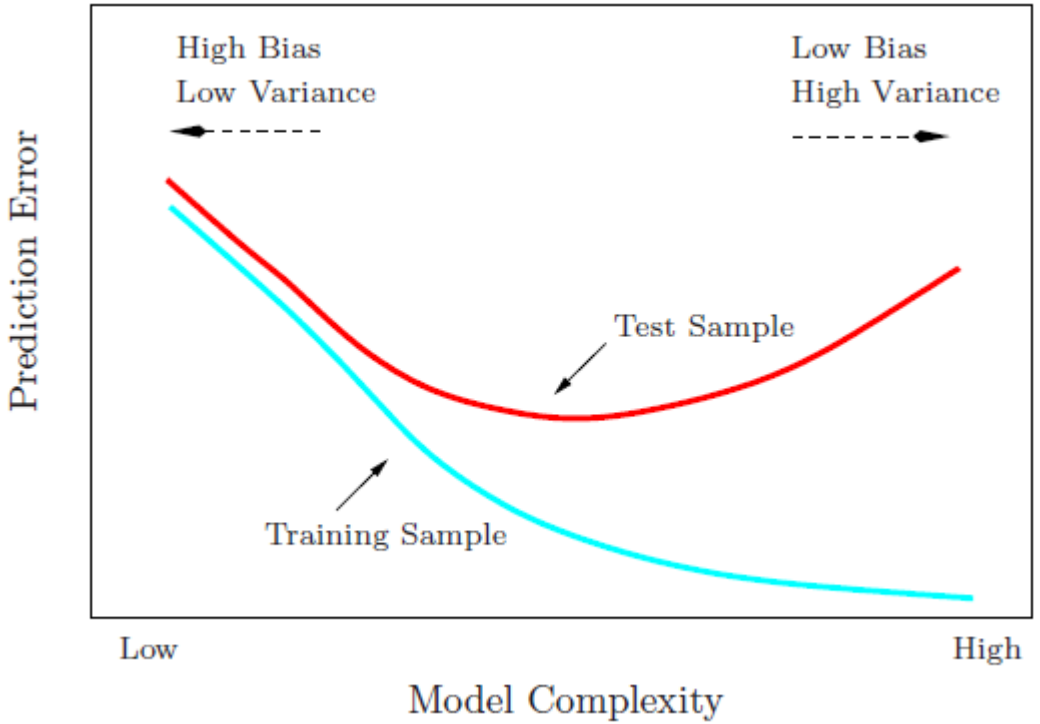
¿Cómo es el sesgo y la varianza de estos modelos?

- La **complejidad** de un modelo debe ajustarse de tal manera que permita la **generalización**, al utilizarse con datos que no haya conocido durante el proceso de entrenamiento

https://en.wikipedia.org/wiki/Overfitting

SOBRE APRENDIZAJE (OVERFITTING)

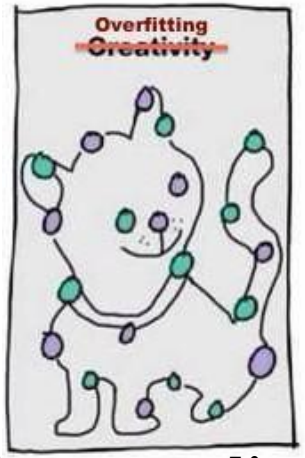
- Los modelos tienden a ajustarse al conjunto de datos usado para su aprendizaje → el **error de entrenamiento** es un mal estimador
- Queremos encontrar la complejidad del modelo que nos permita minimizar el **error de test**



https://onlinecourses.science.psu.edu/stat857/node/160

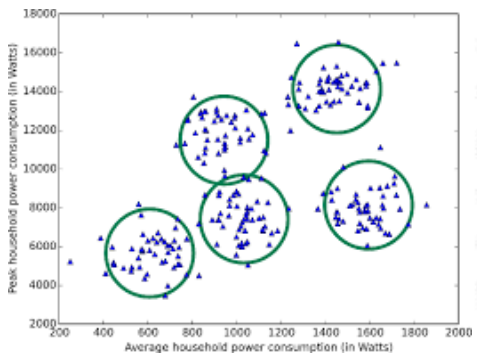
AGENDA

Protocolos

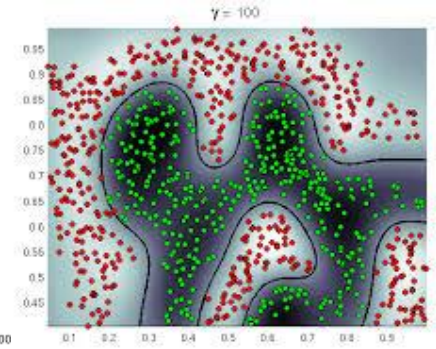


Sobre aprendizaje (Overfitting)

Aprendizaje automático

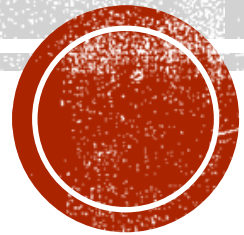


Aprendizaje no supervisado



Aprendizaje supervisado

PROTOCOLOS DE EVALUACIÓN

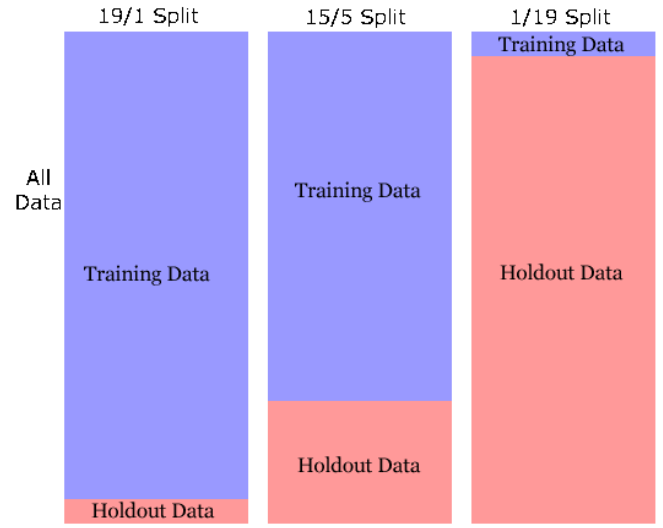


PROTOCOLOS DE EVALUACIÓN

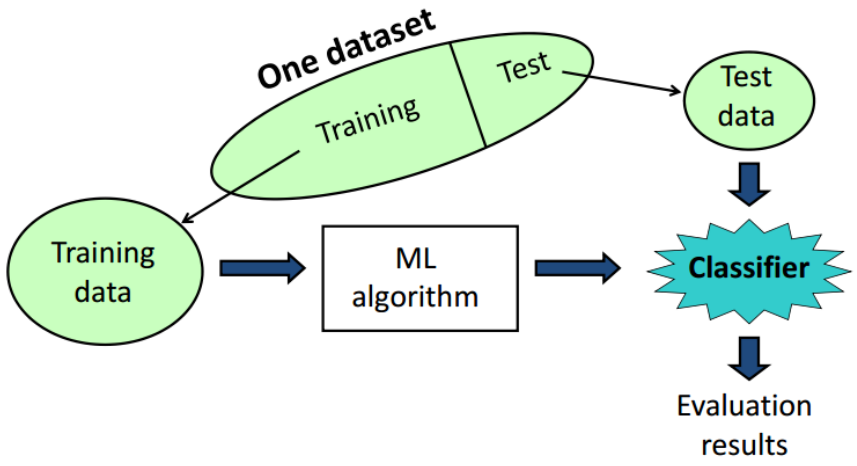
- Aplican para aprendizaje supervisado en general (tanto para clasificación como para regresión.
- Evaluar cual sería la capacidad de **generalización** del modelo a datos nuevos
- Diferenciar entre el **error de entrenamiento** y el **error de test.** Evitar el sesgo causado por la **subestimación del error** al evaluar con el mismo set de entrenamiento.
- Permitir establecer un compromiso entre sesgo y varianza, luchando contra el **sobre aprendizaje**, en busca de un modelo con buenas **capacidades predictivas**

PROTOCOLOS DE EVALUACIÓN

- **Holdout**: particionar el conjunto de datos en 2:
  - **Conjunto de entrenamiento**: con el que se aprende el algoritmo de clasificación
  - **Conjunto de validación o test**: separa al comienzo del procedimiento y no se considera en el aprendizaje
  - **Aleatoriedad** del particionamiento
  - **Compromiso**: entre mas datos mejor el aprendizaje, entre mas datos mejor la evaluación
- **Repeated holdout**: repetir el procedimiento y agregar las métricas de evaluación



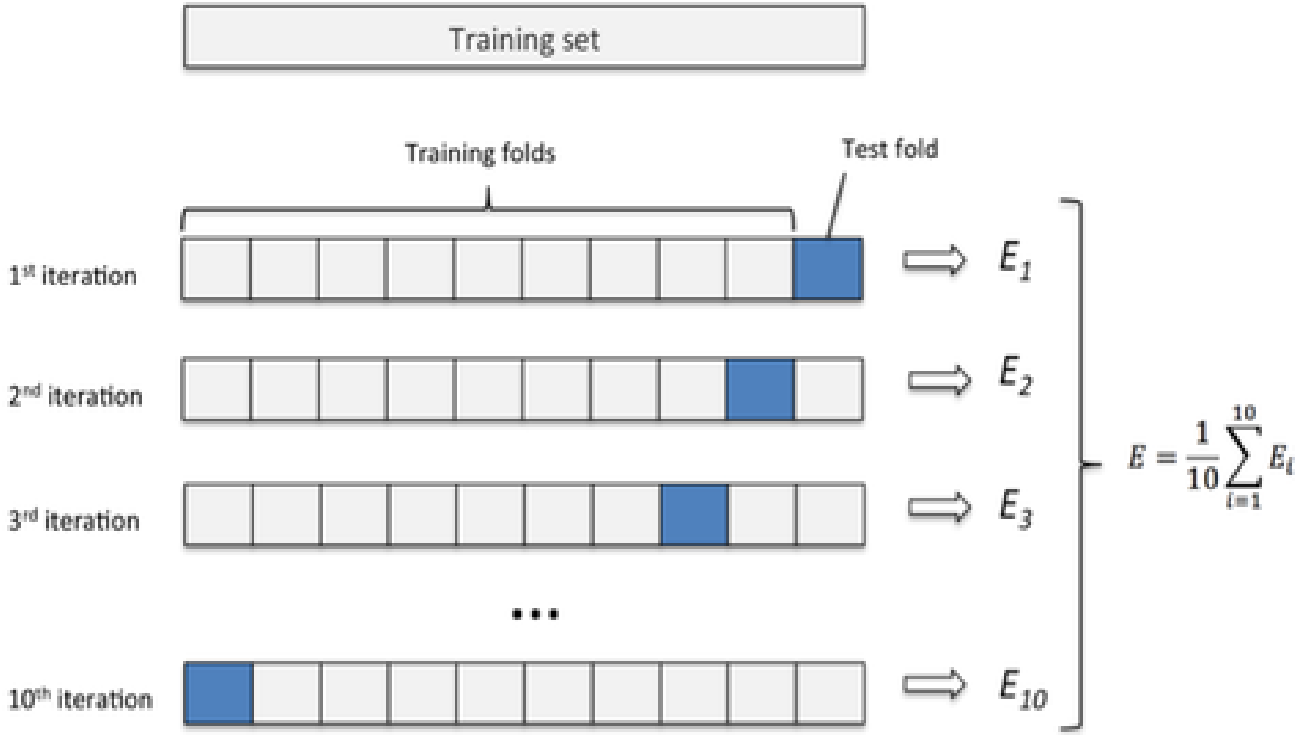
https://webdocs.cs.ualberta.ca/~aixplore/learning/DecisionTrees/InterArticle/6-DecisionTree.html



Ian Witten, Weka MOOC

PROTOCOLOS DE EVALUACIÓN

- **K-fold cross-validation**:
  - Particionar el set de datos en K conjuntos disyuntos del mismo tamaño
  - K-1 partes se usan para entrenamiento, 1 parte se usa para el test
  - Se repite el proceso K veces
  - Se agregan las métricas de evaluación



Sebastian Raschka, 2015

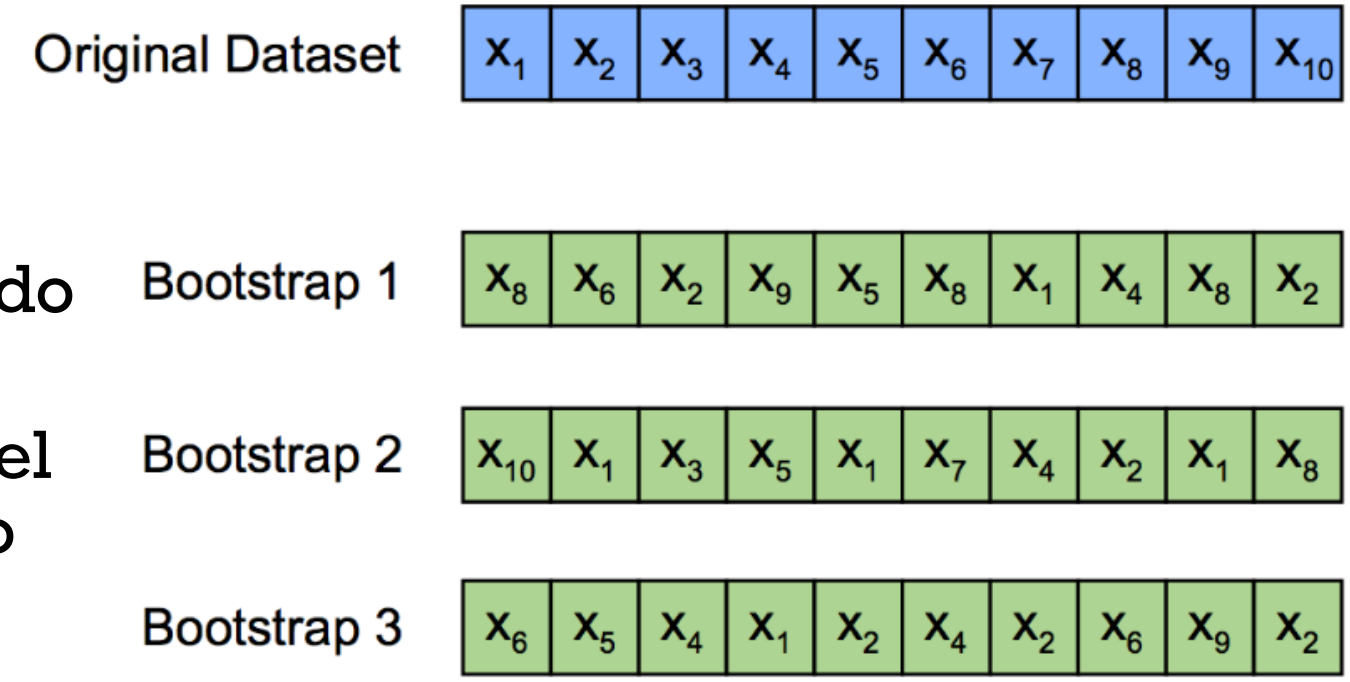
PROTOCOLOS DE EVALUACIÓN

- **K-fold cross-validation,**
  Escogencia del K:
  - Permite balancear entre sesgo y varianza
  - **LOOCV** (Leave One Out Cross-Validation): partes de tamaño 1
  - Por defecto se estima que los mejores resultados se obtienen con un valor de K entre 5 y 10

Sebastian Raschka, 2015

PROTOCOLOS DE EVALUACIÓN

- **Bootstrapping**:
  - Consideración de varios conjuntos de entrenamiento/test utilizando muestreo con remplazo
  - Por lo general muestreos del mismo tamaño del conjunto original



Sebastian Raschka, 2015

TALLER DE CLASIFICACIÓN CON KNN

- Dataset: Iris
- Evaluar los diferentes protocolos y establecer un valor de K, así como un intervalo de confianza para la exactitud de la predicción.