



Técnicas de Machine Learning para

Clasificación (Fundamentos y aplicaciones)

M.Sc. Angelo Jonathan Diaz Soto

2025

Data Science – Business Intelligence – Big Data – Machine Learning – Artificial Intelligence – Innovation and
(+51) 976 760 803
Technology www.datayanalytics.com info@datayanalytics.com

Contenido





- Aplicaciones
- Modelos de Clasificación
- Métricas
de
Evaluación
Curva
ROC ■


Métodos de validación cruzada

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Aplicaciones Sanitario:

Detección de enfermedades.



Predicción de aparición de enfermedades. Análisis  de la actividad postural.

- ▣ Predicción de estancia hospitalaria.
- ▣ Análisis de señales cerebrales. Clasificación de
- ▣ secuencias de ADN.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Aplicaciones



Retail:

- ▣ Estimación de la demanda. Fijación de precios.
- ▣ Predicción del comportamiento de los compradores. Segmentación de ▣ clientes.
 - ▣ Búsqueda de clientes basándose en comportamientos en las redes sociales, - interacciones en la web Optimización



de la usabilidad

Web/Móvil.

- Optimización de las horas que maximizan el impacto en redes sociales de una campaña de marketing.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com



Aplicaciones



Logística:

- Predecir de fallos en equipos tecnológicos.
- Aplicación en data analytics a partir de sensores. □

Mantenimiento predictivo en aeronáutica.

- ▣ Análisis de telemetría en coches.
- ▣ Predicción de retrasos de aviones.
- ▣ Predecir el tráfico urbano.
- ▣ Vehículos autónomos.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

fraude en las transacciones electrónicas. ▣

Predicción de riesgos financieros.

- ▣ Predicción de recesión.
- ▣ Fijación de precios de productos bancarios.
- ▣ Segmentación de clientes.

Aplicaciones

Financiero:

- ▣ Detección de



(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Aplicaciones

Energético:





- ❑ Estimación de demanda energética Predicción del clima.

Seguridad:

- ❑ Detectar intrusiones en una red de comunicaciones de datos.
- ❑ Detección de objetos.
- ❑ Sistemas Anti-spam.
- ❑ Detectando software malicioso.

RRHH:

- ❑ Análisis de empleados más rentables.

Algoritmos de Aprendizaje Supervisado



■ Árboles de decisión

- Classification de Naïve Bayes
- Regresión Logística
- Análisis Discriminante (lineal y cuadrático)
- Support Vector Machines (SVM)
- Random Forest
- Métodos “Ensemble”
(Conjuntos de clasificadores)
- Vecinos más cercanos (KNNs)
- Redes Neuronales Artificiales





Métricas de Evaluación

$$\hat{Y}_i = 1 \quad \hat{Y}_i = 0$$



Valor estimado \hat{Y}_i / Valor real Y_i
 $Y_i = 0$ $Y_i = 1$

$P_{11}P_{12}$

$P_{21}P_{22}$

Donde P_{11} y P_{22} corresponderá a predicciones correctas (valores 0 bien

predichos en el primer caso y valores 1 bien predichos en el segundo caso), mientras que P_{12} y P_{21} corresponderá a predicciones erróneas (valores 1 mal predichos en el primer caso y valores 0 mal predichos en el segundo caso).

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Métricas de Evaluación




A partir de estos valores se pueden definir los índices que aparecen en el siguiente cuadro:



☐ **Tasa de errores:** Cociente entre las predicciones correctas y el total de predicciones.

$$P_{11} + P_{12} + P_{21} + P_{22}$$

 ² **Tasa de aciertos:** Cociente entre las predicciones incorrectas y el total de predicciones.

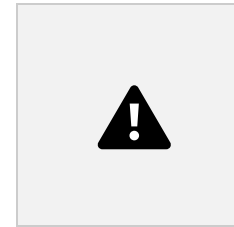



$$P_{11} + P_{12} + P_{21} + P_{22}$$



(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Métricas de Evaluación



 ³ **Especificidad:** Proporción entre la frecuencia valores cero correctos y el total de valores cero observados.



$$P_{11} + P_{21}$$

- 4 **Sensibilidad:** Proporción entre los valores uno correctos predichos por el algoritmo y el total de elementos que son realmente 1.



$$P_{21} + P_{22}$$

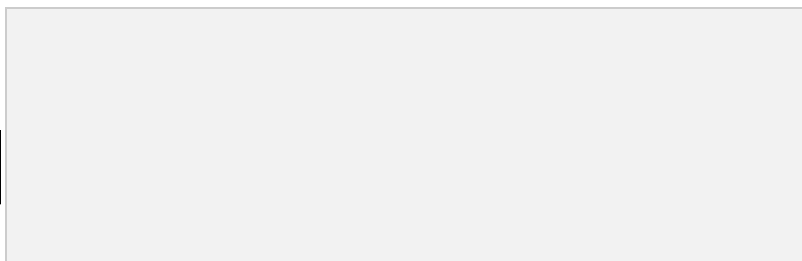
(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Métricas de Evaluación



- 5 **Tasa de falsos ceros:**

Proporción entre la frecuencia de valores cero incorrectos y el total de valores cero observados.



$$P_{11} + P_{21}$$

- 6 **Tasa de falsos unos:** Proporción entre la frecuencia de valores uno incorrectos y el total de valores uno observados.

$$P_{12} + P_{22}$$

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Métricas de Evaluación



- Un método para evaluar clasificadores alternativo a la métrica expuesta es la curva ROC (Receiver Operating Characteristic).
- La curva ROC es una representación gráfica del

rendimiento del clasificador que muestra la distribución de las fracciones de verdaderos positivos y de falsos positivos.

- La fracción de verdaderos positivos se conoce como **sensibilidad**, sería la probabilidad de clasificar correctamente a un individuo cuyo estado real sea definido como positivo.
- La **especificidad** es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea clasificado como negativo. Esto es igual a restar uno de la fracción de falsos positivos.
- La curva ROC también es conocida como la representación de sensibilidad frente a (1-especificidad).

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Curva ROC

- En definitiva, se considera un **modelo inútil**, cuando la curva ROC recorre la diagonal positiva del gráfico.



- ❑ En tanto que en un **test perfecto**, la curva ROC recorre los bordes izquierdo y superior del gráfico.
- ❑ La curva ROC permite comparar modelos a través del área bajo su curva.



Métodos de validación



- Los métodos de validación, también conocidos como

resampling, son estrategias que permiten estimar la capacidad predictiva de los modelos cuando se aplican a nuevas observaciones, haciendo uso únicamente de los datos de entrenamiento.

- La idea en la que se basan todos ellos es la siguiente: el modelo se ajusta empleando un subconjunto de observaciones del conjunto de **entrenamiento** y se evalúa

(calcular una métrica que mide cómo de bueno es el modelo, por ejemplo, accuracy) con las observaciones

restantes.

- Este proceso se repite múltiples veces y los resultados se agregan y

promedian. Gracias a las repeticiones, se compensan las posibles desviaciones que puedan surgir por el reparto aleatorio de las observaciones.

- ❑ La diferencia entre métodos suele ser la forma en la que se generan los subconjuntos de entrenamiento/validación.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

k-Fold-Cross-Validation (CV)

- ❑ Las observaciones de entrenamiento se reparten en **k folds** (conjuntos) del mismo tamaño. El modelo se ajusta con todas las observaciones excepto las del primer fold y se evalúa

prediciendo las observaciones del fold que ha quedado excluido, obteniendo así la primera métrica.

- ❑ El proceso **se repite k veces**, excluyendo un fold distinto en cada iteración. Al final, se generan k valores de la métrica, que se agregan



(normalmente con la media y la desviación típica)
generando la estimación final de validación.



(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

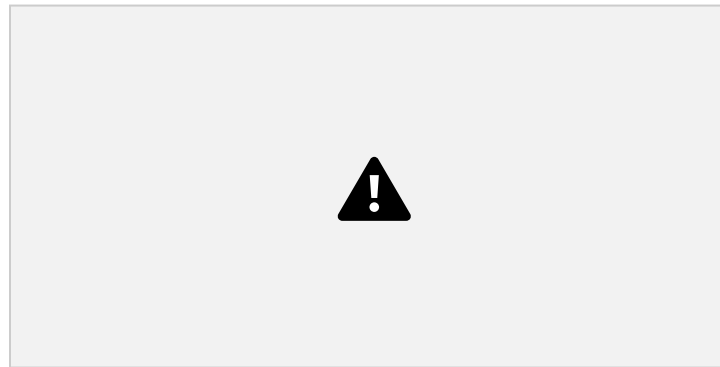
Leave-One-Out Cross-Validation (LOOCV)



- LOOCV es un caso especial de k-Fold-Cross-Validation en el que el número k de folds es igual al número de

observaciones disponibles en el conjunto de entrenamiento.

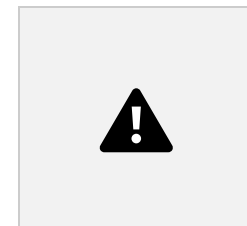
- El modelo se ajusta cada vez con todas las observaciones excepto una, que se emplea para evaluar el modelo. Este método supone un **coste computacional muy elevado**, el modelo se ajusta tantas veces como observaciones de entrenamiento, por lo que, en la práctica, no suele compensar emplearlo.



(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Repeated k-Fold-Cross-Validation (repeated CV)

- Es exactamente igual al método k-Fold-Cross-Validation pero



repitiendo el proceso completo n veces.

- Por ejemplo, 10-Fold-Cross-Validation con 5 repeticiones implica a un total de 50 iteraciones ajuste-validación, pero no equivale a un 50-Fold-Cross-Validation.



Leave-Group-Out Cross-Validation (LGOCV)



□ LGOCV,
train/test
Monte Carlo
simplemente
generar



□ La
a cada

también conocido como repeated
splits o
Cross-Validation, consiste
en
múltiples divisiones aleatorias
entrenamiento-test
(solo dos conjuntos por repetición).
proporción de observaciones que va
conjunto se

determina de antemano, 80%-20% suele dar buenos
resultados.

- Este método, aunque más simple de implementar que CV,
requiere de muchas repeticiones (>50) para generar
estimaciones estables.

Bootstrapping



■ Una muestra bootstrap es una muestra obtenida a partir de la muestra original por muestreo aleatorio con

reposición, y del mismo tamaño que la muestra original. ■ Muestreo aleatorio con reposición

(resampling with replacement)

significa que, después de que una observación sea extraída, se vuelve a poner a disposición para las siguientes extracciones.

■ Como resultado de este tipo de



muestreo, algunas observaciones aparecerán múltiples veces en la muestra bootstrap y otras ninguna.

- ❑ Las observaciones no seleccionadas reciben el nombre de out-of-bag (OOB).
- ❑ Por cada iteración de bootstrapping se genera una nueva muestra bootstrap, se ajusta el modelo con ella y se evalúa con las observaciones out-of-bag.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Bootstrapping



❑

muestra del mismo tamaño que la muestra
muestreo aleatorio con reposición.

❑ Ajustar el
modelo
empleando



la nueva muestra generada en el

2

paso 1.

Calcular el error del modelo empleando aquellas observaciones

3

de la muestra original que no se han incluido en la nueva muestra. A este error se le conoce como error de validación.

Repetir el proceso n veces y calcular la media de los n errores

4

de validación.

Finalmente, y tras las n repeticiones, se ajusta el modelo final

5

empleando todas las observaciones de entrenamiento originales.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com

Recomendaciones para la elección del método de validación





- Si el tamaño de la muestra es pequeño, se recomienda emplear **Repeated k-Fold-Cross-Validation**, ya que consigue un buen equilibrio bias-varianza y, dado que no son muchas observaciones, el coste computacional no es excesivo.

- Si el objetivo principal es **comparar modelos** más que obtener una estimación precisa de las métricas, se recomienda

bootstrapping ya que tiene menos varianza.

- Si el tamaño muestral es muy grande, la diferencia entre métodos se reduce y toma más importancia la eficiencia computacional. En estos casos, **10-Fold-Cross-Validation** simple es suficiente.



Otras métricas de validación

Existe una gran variedad de métricas que permiten evaluar la capacidad predictiva de un algoritmo. La idoneidad de cada una **depende completamente del problema en cuestión**, y su correcta elección dependerá del entendimiento del problema al que se enfrenta. A continuación, se describen algunas de las más utilizadas.



Accuracy es

el porcentaje de observaciones
correctamente
clasificadas respecto al total de predicciones.

- **Kappa** o Cohen's Kappa es el valor de accuracy normalizado respecto del porcentaje de acierto esperado por azar. A diferencia de accuracy, cuyo rango de valores puede ser $[0, 1]$, el de kappa es $[-1, 1]$.

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com



Referencias Bibliográficas

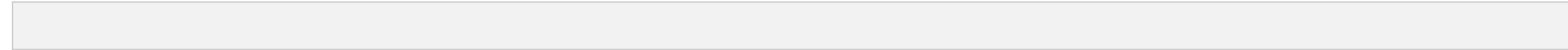
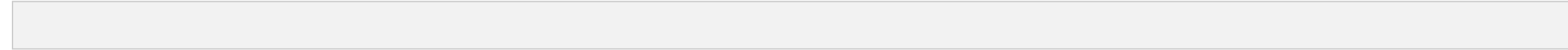


- An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics).
- Introduction to Machine Learning with Python: A Guide for Data Scientists.

- ❑ An Introduction to Statistical Learning by James, Gareth et al.
- ❑ Applied Predictive Modeling by Max Kuhn and Kjell Johnson.
- ❑ <https://www.aprendemachinelearning.com/aplicaciones-del-machine-learning/>
- ❑ <https://aprendeia.com/todo-sobre-aprendizaje-supervisado-en-machine-learning/>

(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com





(+51) 976 760 803 www.datayanalytics.com info@datayanalytics.com