


Inteligencia Artificial con Python y scikit-learn

[Install](#) [User Guide](#) [API](#) [Examples](#) [Community](#) [More](#)

scikit-learn

Machine Learning in Python

[Getting Started](#) [Release Highlights for 1.6](#)

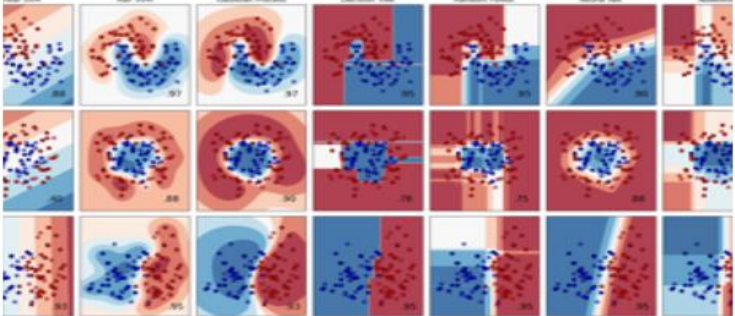
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: [Gradient boosting](#), [nearest neighbors](#), [random forest](#), [logistic regression](#), and [more...](#)

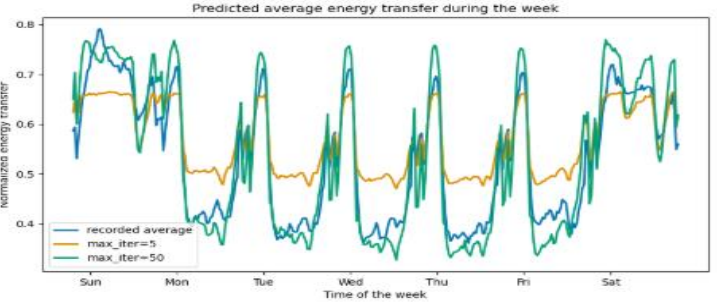


Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, stock prices.

Algorithms: [Gradient boosting](#), [nearest neighbors](#), [random forest](#), [ridge](#), and [more...](#)




Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, grouping experiment outcomes.

Algorithms: [k-Means](#), [HDBSCAN](#), [hierarchical clustering](#), and [more...](#)





Aprendizaje automático aplicado

Evaluación del modelo



Generalización, sobreajuste y subajuste

Generalización. La capacidad del modelo para generalizar es su habilidad para hacer predicciones precisas en datos nuevos. Un modelo bien probado debe demostrar que no solo memoriza los datos de entrenamiento, sino que entiende los patrones subyacentes.

- **Sobreajuste (Overfitting):** El modelo aprende demasiado bien los detalles y ruido de los datos de entrenamiento, lo que lleva a un mal desempeño en el conjunto de prueba.
- **Subajuste (Underfitting):** El modelo no captura suficientemente la relación entre los datos, llevando a un bajo desempeño en ambos conjuntos.



Métricas de Evaluación

Las métricas elegidas para probar el modelo dependen del tipo de problema que estás resolviendo:

Para problemas de clasificación:

- **Precisión (Accuracy):** Porcentaje de predicciones correctas.
- **Precisión y Recall:** Útiles en casos desbalanceados.
- **F1-Score:** Promedio ponderado de precisión y recall.
- **ROC-AUC:** Medida del rendimiento general en clasificación binaria.

Métricas de Evaluación

Para problemas de regresión:

- **Error Absoluto Medio (MAE):** Promedio del valor absoluto de los errores.
- **Error Cuadrático Medio (MSE):** Promedio del cuadrado de los errores.
- **Coeficiente de determinación (R^2):** Cuánto de la variabilidad de los datos es explicada por el modelo.

Para problemas de clustering:

- **Silhouette Score:** Evalúa la separación entre los grupos.
- **Homogeneidad:** Mide la consistencia dentro de un cluster.



Errores y Diagnóstico

Durante la prueba, es común analizar errores para identificar áreas de mejora:

- ¿Qué tipos de datos predice mal el modelo?
- ¿Son los errores sistemáticos (bias) o aleatorios (variance)?
- ¿El modelo maneja bien datos raros o casos extremos?

Validación Cruzada

A veces, para una mejor evaluación, se usa **validación cruzada**, dividiendo los datos en múltiples subconjuntos y entrenando/pruebas múltiples veces para obtener una medida promedio de desempeño.

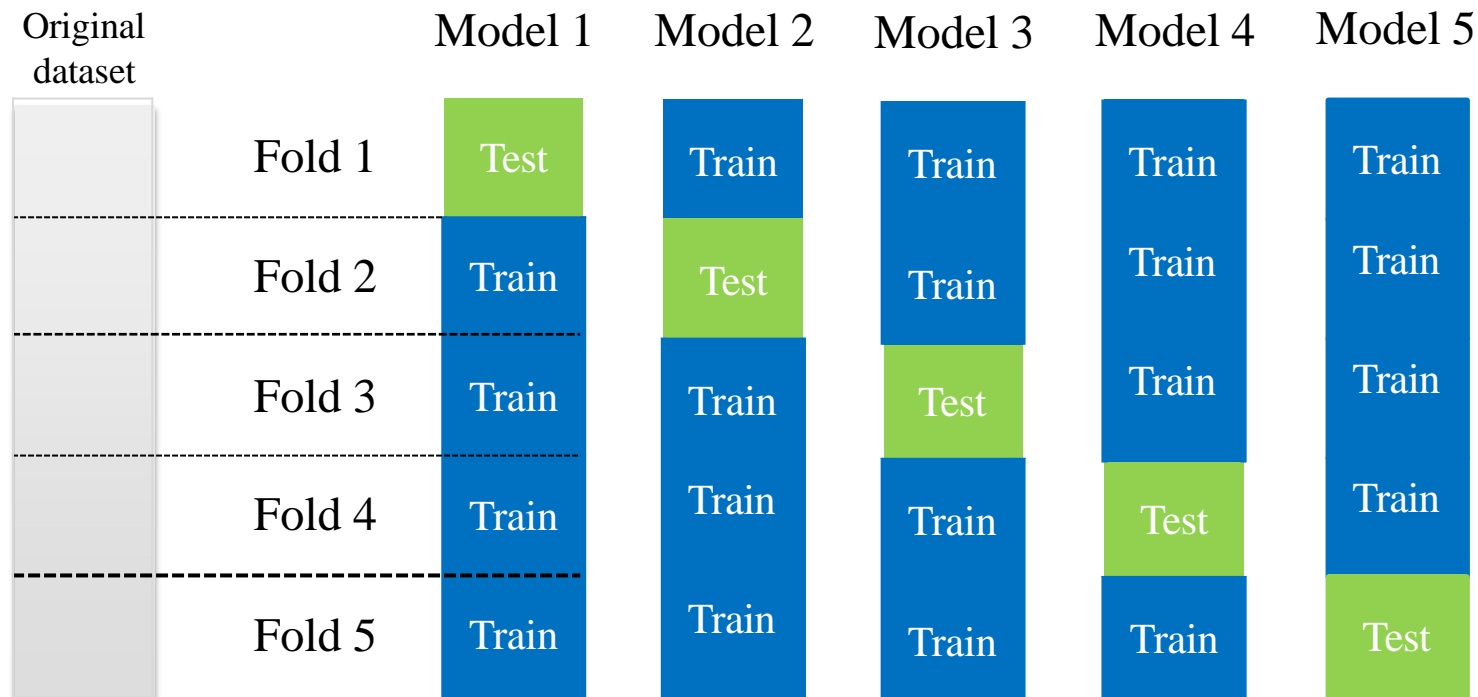
- **Utiliza varias divisiones de prueba de entrenamiento, no solo una sola**
- **Cada división se utiliza para entrenar y evaluar un modelo independiente**
- **¿Por qué es mejor?**
 - *La puntuación de precisión de un método de aprendizaje supervisado puede variar, dependiendo de las muestras que terminen en el conjunto de entrenamiento.*
 - *El uso de varias divisiones de prueba de tren proporciona estimaciones más estables y confiables sobre el rendimiento promedio del clasificador.*
 - *Los resultados se promedian en varios conjuntos de entrenamiento diferentes en lugar de basarse en un único modelo entrenado en un conjunto de entrenamiento determinado.*

random_state	Test set accuracy
0	1.00
1	0.93
5	0.93
7	0.67
10	0.87

Precisión del clasificador k-NN ($k = 5$) en el conjunto de prueba de datos de fruta para diferentes valores de random_state en train_test_split.

Ejemplo de validación cruzada (5 veces)

- El dataset se divide en varias partes iguales llamadas **folds**.
- Cada fold se utiliza alternativamente como un conjunto de prueba, mientras que los restantes se emplean como conjunto de entrenamiento.
- Este proceso se repite tantas veces como folds se hayan definido.



Evaluación del modelo

<u>True</u> negative	TN	FP
	FN	TP
<u>True</u> positive		
	<u>Predicted</u> negative	<u>Predicted</u> positive

Label 1 = positive class
(class of interest)

Label 0 = negative class
(everything else)

TP = true positive
FP = false positive (Type I
error)
TN = true negative
FN = false negative (Type II
error)

Evaluación del modelo

- Una vez que un modelo es entrenado, la **evaluación** del mismo **proporciona retroalimentación crítica sobre las características de rendimiento del modelo entrenado.**
- Ayuda a comprender **qué instancias de datos se están clasificando** o prediciendo incorrectamente.
- Lo que a su vez podría sugerir mejores características o **refinamientos al modelo de aprendizaje** en la fase de refinamiento de características y modelos.
- Considerar que pueden existir **aspectos posibles de la evaluación del rendimiento del modelo que están más allá de la precisión** promedio que pueden ser críticos de medir.

Evaluación del modelo

- Por ejemplo, en una aplicación de salud que utiliza un **clasificador para detectar tumores en una imagen médica**, es posible que desee que el clasificador se equivoque por precaución.
- Y marcar cualquier cosa que incluso tenga una pequeña posibilidad de ser canceroso.
- Incluso si esto significa a veces clasificar incorrectamente el tejido sano como enfermo.



Confusion matrix

- Calcula la matriz de confusión para **evaluar la precisión de una clasificación**.
- Por definición, una matriz de confusión C es tal que C_{ij} es igual al número de **observaciones que se sabe que están en grupo i y se prevé que estarán en grupo j** .
- Por lo tanto, en la clasificación binaria, el recuento de verdaderos negativos es C_{00} , falsos negativos es C_{10} , verdaderos positivos es C_{11} , y falsos positivos es C_{01} ,

`sklearn.metrics.confusion_matrix`

```
sklearn.metrics.confusion_matrix(y_true, y_pred, *, labels=None, sample_weight=None, normalize=None) [source]
```

Confusion matrix

`sklearn.metrics.confusion_matrix`

```
sklearn.metrics.confusion_matrix(y_true, y_pred, *, labels=None, sample_weight=None, normalize=None) [source]
```

Binary Prediction Outcomes

<u>True</u> negative	TN	FP
	FN	TP
<u>True</u> positive		
	<u>Predicted</u> negative	<u>Predicted</u> positive

Label 1 = positive class
(class of interest)

Label 0 = negative class
(everything else)

TP = true positive
FP = false positive (Type I
error)
TN = true negative
FN = false negative (Type II
error)

Confusion matrix

- La **diagonal principal** contiene **la suma de todas las predicciones correctas**.
- La **otra diagonal refleja los errores del clasificador**: los falsos positivos y los falsos negativos.

	N (modelo)	S (modelo)
n (real)	Negativos Reales	Falsos Positivos
p (real)	Falsos Negativos	Positivos reales

Confusion matrix

Matriz de confusión		Estimado por el modelo			
		Negativo (N)	Positivo (P)		
Real	Negativo	a: (TN)	b: (FP)		
	Positivo	c: (FN)	d: (TP)	Precisión ("precision") Porcentaje predicciones positivas correctas:	$d/(b+d)$
		Sensibilidad, exhaustividad ("Recall") Porcentaje casos positivos detectados	Especificidad ("Specificity") Porcentaje casos negativos detectados	Exactitud ("accuracy") Porcentaje de predicciones correctas (No sirve en datasets poco equilibrados)	
		$d/(d+c)$	$a/(a+b)$	$(a+d)/(a+b+c+d)$	

Referencias

Python Intermedio

<https://python-intermedio.readthedocs.io/es/latest/>

Pandas_Cheat_Sheet.

https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf

NearestNeighborsClassification

https://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html

Confusionmatrix

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html