Portada apuntes IA

Tabla de contenido

[Regresión 3](#_Toc183981737)

[Tabla: Modelos de Regresión y Hiperparámetros con Explicación 3](#_Toc183981738)

[Métricas para Modelos de Regresión 6](#_Toc183981739)

[Clasificación 7](#_Toc183981740)

[Tabla: Modelos de Clasificación, Hiperparámetros y Métricas 7](#_Toc183981741)

[Metricas 8](#_Toc183981742)

[Optimización de Hiperparámetros 9](#_Toc183981743)

[1. Búsqueda Exhaustiva (Grid Search) 9](#_Toc183981744)

[2. Búsqueda Aleatoria (Randomized Search) 9](#_Toc183981745)

# Regresión

## **Tabla: Modelos de Regresión y Hiperparámetros con Explicación**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modelo** | **Hiperparámetro** | **Rango de Valores** | **Valor por Defecto** | **Explicación del Hiperparámetro** |
| **Regresión Lineal** | fit\_intercept | True, False | True | Indica si se debe calcular el intercepto del modelo. Si es False, no se añade un término constante. |
|  | normalize | True, False | False | Determina si las características deben ser normalizadas antes de ajustar el modelo. Si True, se normalizan. |
| modelo = LinearRegression(fit\_intercept=True, normalize=False) | | | | |
| **Regresión Polinómica** | degree | 1 a n (entero) | 2 | Grado del polinomio. A mayor grado, el modelo se ajusta más a los datos, pero puede haber sobreajuste. |
|  | interaction\_only | True, False | False | Si es True, solo se consideran interacciones entre las características, sin términos cuadráticos. |
|  | include\_bias | True, False | True | Si es True, se incluye un término de sesgo en el modelo. Si es False, no se incluye. |
| modelo\_polinomico = make\_pipeline( PolynomialFeatures(degree=2, interaction\_only=False, include\_bias=True), LinearRegression() | | | | |
| **Regresión Ridge** | alpha | 0.0001 a 1000 | 1.0 | Término de regularización. A mayor valor de alpha, más fuerte es la regularización, lo que reduce la magnitud de los coeficientes. |
|  | fit\_intercept | True, False | True | Si es True, el modelo ajusta un intercepto (sesgo). Si es False, no se ajusta ningún término constante. |
|  | normalize | True, False | False | Si es True, las características se normalizan antes de ajustar el modelo. |
|  | solver | 'auto', 'svd', 'cholesky', 'lsqr', 'saga' | 'auto' | Método utilizado para resolver la ecuación. auto selecciona el mejor según las características del modelo. |
|  | max\_iter | Entero positivo | None | Número máximo de iteraciones para los métodos iterativos, como 'saga'. Define el número máximo de intentos de ajuste. |
| modelo\_ridge = Ridge(alpha=1.0, fit\_intercept=True, normalize=False, solver='auto', max\_iter=None) | | | | |
| **Regresión Lasso** | alpha | 0.0001 a 1000 | 1.0 | Término de regularización. Similar a Ridge, controla la magnitud de los coeficientes, pero en este caso puede reducir algunos coeficientes a cero. |
|  | fit\_intercept | True, False | True | Si es True, se ajusta un término constante. Si es False, no se incluye el intercepto. |
|  | normalize | True, False | False | Si es True, las características se normalizan antes de ajustar el modelo. |
|  | max\_iter | Entero positivo | 1000 | Número máximo de iteraciones para el ajuste del modelo. Si el valor de max\_iter es alcanzado antes de converger, se detiene. |
|  | selection | 'cyclic', 'random' | 'cyclic' | Método de selección de características en el modelo. 'cyclic' usa todas las características de forma cíclica, mientras que 'random' selecciona aleatoriamente. |
| modelo\_lasso = Lasso(alpha=1.0, fit\_intercept=True, normalize=False, max\_iter=1000, selection='cyclic') | | | | |
| **ElasticNet** | alpha | 0.0001 a 1000 | 1.0 | Término de regularización, similar al de Ridge y Lasso. Controla el grado de penalización de los coeficientes. |
|  | l1\_ratio | 0 a 1 (proporción entre Lasso y Ridge) | 0.5 | Controla la mezcla entre las regularizaciones L1 (Lasso) y L2 (Ridge). Un valor de 1 se comporta como Lasso, mientras que 0 es igual a Ridge. |
|  | fit\_intercept | True, False | True | Si es True, el modelo ajusta un intercepto (sesgo). Si es False, no se ajusta un término constante. |
|  | normalize | True, False | False | Si es True, las características se normalizan antes de ajustar el modelo. |
|  | max\_iter | Entero positivo | 1000 | Número máximo de iteraciones para el ajuste del modelo. |
| modelo\_elasticnet = ElasticNet(alpha=1.0, l1\_ratio=0.5, fit\_intercept=True, normalize=False, max\_iter=1000) | | | | |
| **SGDRegressor** | loss | 'squared\_error', 'huber', 'epsilon\_insensitive', 'squared\_epsilon\_insensitive' | 'squared\_error' | Define la función de pérdida que se utiliza para el ajuste del modelo. 'squared\_error' es para regresión estándar, mientras que 'huber' y 'epsilon\_insensitive' son útiles para datos con valores atípicos. |
|  | alpha | 0.0001 a 1000 | 0.0001 | Término de regularización. Controla la magnitud de los coeficientes del modelo. |
|  | fit\_intercept | True, False | True | Si es True, el modelo ajusta un término constante. Si es False, no se ajusta ningún intercepto. |
|  | max\_iter | Entero positivo | 1000 | Número máximo de iteraciones para el ajuste del modelo. |
| modelo\_sgd = SGDRegressor(loss='squared\_error', alpha=0.0001, fit\_intercept=True, max\_iter=1000) | | | | |
| **LinearSVC** | C | 0.0001 a 1000 | 1.0 | Controla la regularización. Un valor más alto de C reduce la regularización y ajusta el modelo más estrechamente a los datos. |
|  | loss | 'hinge', 'squared\_hinge' | 'hinge' | Función de pérdida que se utiliza. 'hinge' es para clasificación estándar, y 'squared\_hinge' es para pérdida cuadrática. |
|  | max\_iter | Entero positivo | 1000 | Número máximo de iteraciones para el ajuste del modelo. |
| modelo\_svc = LinearSVC(C=1.0, loss='hinge', max\_iter=1000) | | | | |
| **Árboles de Decisión** | max\_depth | None, 1 a n (entero) | None | Profundidad máxima del árbol. Si es None, se expande hasta que cada hoja contenga menos de min\_samples\_split muestras. |
|  | min\_samples\_split | 2 a n (entero) | 2 | Número mínimo de muestras necesarias para dividir un nodo. |
|  | min\_samples\_leaf | 1 a n (entero) | 1 | Número mínimo de muestras necesarias para estar en una hoja del árbol. |
|  | criterion | 'gini', 'entropy' | 'gini' | Función utilizada para medir la calidad de la división en los nodos del árbol. 'gini' es para el índice de Gini, y 'entropy' para la entropía de la información. |
| modelo\_arbol = DecisionTreeClassifier(max\_depth=None, min\_samples\_split=2, min\_samples\_leaf=1, criterion='gini') | | | | |

## **Métricas para Modelos de Regresión**

|  |  |  |
| --- | --- | --- |
| **Métrica** | **Explicación** | **Código Python** |
| **MSE** | El **Error Cuadrático Medio (MSE)** mide el promedio de los errores al cuadrar las diferencias entre las predicciones y los valores reales. Penaliza más los grandes errores. | from sklearn.metrics import mean\_squared\_error mse = mean\_squared\_error(y\_train, y\_pred) |
| **R²** | El **Coeficiente de Determinación (R²)** mide qué tan bien el modelo explica la variabilidad de la variable dependiente. Un valor cercano a 1 indica que el modelo es bueno. | from sklearn.metrics import r2\_score r2 = r2\_score(y\_train, y\_pred) |
| **RMSE** | La **Raíz del Error Cuadrático Medio (RMSE)** es la raíz cuadrada del MSE. Proporciona una medida de error en las mismas unidades que los datos. | import numpy as np rmse = np.sqrt(mean\_squared\_error(y\_train, y\_pred)) |
| **MAE** | El **Error Absoluto Medio (MAE)** calcula el promedio de las diferencias absolutas entre las predicciones y los valores reales. Es más robusto frente a los valores atípicos que el MSE. | from sklearn.metrics import mean\_absolute\_error mae = mean\_absolute\_error(y\_train, y\_pred) |

# Clasificación

## **Tabla: Modelos de Clasificación, Hiperparámetros y Métricas**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modelo** | **Cuándo Usarlo** | **Hiperparámetros** | **Métricas y Cuándo Usarlas** |
| **Regresión Logística** (LogisticRegression) | Problemas binarios/multiclase con datos linealmente separables o casi lineales. | - **penalty**: 'l1', 'l2', 'elasticnet', 'none' (Default: 'l2'). - **C**: Regularización (float > 0, Default: 1.0). - **solver**: 'lbfgs', 'liblinear', 'sag', 'saga' (Default depende del penalty). - **max\_iter**: Máx. iteraciones (int, Default: 100). - **multi\_class**: 'auto', 'ovr', 'multinomial' (Default: 'auto'). | - **Accuracy**: Usar en problemas balanceados para medir qué tan frecuentemente el modelo predice correctamente. - **ROC-AUC**: Ideal para problemas desbalanceados y medir la capacidad de distinguir entre clases. - **Precision/Recall**: Útil si los falsos positivos o negativos tienen un costo mayor. |
| **Árbol de Decisión** (DecisionTreeClassifier) | Relaciones no lineales en datos, interpretabilidad visual. | - **criterion**: 'gini', 'entropy' (Default: 'gini'). - **max\_depth**: Profundidad máxima (int > 0 o None, Default: None). - **min\_samples\_split**: Muestras mínimas para dividir (int > 1 o float, Default: 2). - **min\_samples\_leaf**: Muestras mínimas por hoja (int o float, Default: 1). - **max\_features**: 'auto', 'sqrt', 'log2', None (Default: None). | - **Accuracy**: Medir desempeño en problemas balanceados. - **F1-Score**: Útil para problemas desbalanceados ya que balancea precisión y exhaustividad. - **Confusión**: Para observar distribuciones de errores entre clases. - **Precision**: Si los falsos positivos son críticos (ej., diagnóstico médico). |
| **Bosques Aleatorios** (RandomForestClassifier) | Datos no lineales o de alta dimensionalidad. Modelo robusto frente al sobreajuste. | - **n\_estimators**: Nº de árboles (int > 0, Default: 100). - **criterion**: 'gini', 'entropy', 'log\_loss' (Default: 'gini'). - **max\_depth**: ProfundidaHistorial del chat  ChatGPT  ChatGPT  Explorar GPT  Hoy  Algoritmos ML Regresión Clasificación  7 días anteriores  SCP transferencia archivo remoto  Imputación con SimpleImputer  MNIST Ensemble Voting  Ensemble learning y precisión  Dimensionalidad de MNIST PCA  Curva ROC con probabilidades  Hiperparámetros de entrenamiento  HDFS Connection Refused Error  30 días anteriores  Ryzen 9 5900X vs i7 7700K  MongoDB Clúster Shardeado  Matriz de confusión  Verificar nodos Hadoop  JupyterLab Node.js Issue  Función de recompensa DeepRacer  Ver colecciones en mongosh  Docker Compose Replica Set  Fixing RF Input Error  Ayuda con código  Visualización de dígitos  Parametros incompatibles LogisticRegression  Hadoop Streaming en EMR  Log Transformation Error Fix  PPK to PEM Conversion  Leer CSV desde HDFS  Recompensa velocidad AWS DeepRacer  Push rejected solution  Push rejected solution  Tamaño y tiempo Parquet  Leer letras en vez de palabras  Ejercicios de distribución normal  Uso de verbose en ColumnTransformer  Ejemplo de consulta Nor  octubre  Hadoop Sudoku Puzzle Command  Evitar bloqueo en Ubuntu  Verificar instalación de MongoDB  Data Preprocessing with Pipelines  Valores únicos en DataFrame  Error de archivo no encontrado  Soporte MongoDB i386 Debian  Hadoop SSH Permission Issue  Probabilidades peso recién nacidos  Ajustes en pipeline Titanic  Error en ColumnTransformer  mayo  Mejoras en la Base  Crear Bean JavaEE Simple  Autoincrementación de Clave Primaria  Recuperación de datos GET  Cambiar a un plan superior  Más acceso a los mejores modelos  ChatGPT  Compartir  User  Tú dijiste:  Hazme un resumen sobre que algoritmos hay que utilizar en regresion y en clasificacion en machine learning que parametros principales hay que controlar  ChatGPT Plus  ChatGPT  En machine learning, la elección del algoritmo y el control de parámetros clave son fundamentales para obtener buenos resultados en tareas de regresión y clasificación. Aquí tienes un resumen organizado por tipo de tarea:  Regresión  La regresión se utiliza para predecir valores continuos. Los algoritmos más comunes incluyen:  1. Algoritmos más usados:  Regresión Lineal:  Supone una relación lineal entre las variables.  Parámetros clave:  Learning rate (si se usa en contextos de optimización).  Regularización (L1: Lasso, L2: Ridge).  Regresión Polinómica:  Extiende la regresión lineal con términos polinómicos.  Parámetros clave:  Grado del polinomio.  Regularización para evitar sobreajuste.  Árboles de Decisión:  Divide los datos en intervalos basándose en reglas.  Parámetros clave:  Profundidad máxima del árbol.  Número mínimo de muestras por hoja (min\_samples\_leaf).  Bosques Aleatorios (Random Forests):  Combina múltiples árboles para mejorar la precisión.  Parámetros clave:  Número de árboles.  Profundidad máxima.  Número de características seleccionadas en cada división.  Gradient Boosting (Ej.: XGBoost, LightGBM, CatBoost):  Mejora iterativamente las predicciones del modelo.  Parámetros clave:  Número de árboles.  Tasa de aprendizaje (learning rate).  Profundidad máxima de los árboles.  Support Vector Regression (SVR):  Modela la regresión mediante un margen que minimiza errores.  Parámetros clave:  C (penalización por error).  Epsilon (margen de tolerancia).  Kernel (lineal, RBF, etc.).  Redes Neuronales:  Modelos basados en múltiples capas y no linealidades.  Parámetros clave:  Arquitectura (número de capas y neuronas).  Learning rate.  Función de activación.  Clasificación  La clasificación se utiliza para predecir categorías discretas. Los algoritmos más comunes incluyen:  1. Algoritmos más usados:  Regresión Logística:  Modelo lineal para probabilidades.  Parámetros clave:  Regularización (L1, L2).  Solvers para optimización (liblinear, sag, lbfgs).  Árboles de Decisión:  Divide datos en reglas jerárquicas.  Parámetros clave:  Profundidad máxima.  min\_samples\_split y min\_samples\_leaf.  Bosques Aleatorios:  Reduce el sobreajuste de los árboles individuales.  Parámetros clave:  Número de árboles.  Profundidad máxima.  max\_features.  Gradient Boosting (Ej.: XGBoost, LightGBM, CatBoost):  Similar a regresión, pero optimizado para clasificación.  Parámetros clave:  Tasa de aprendizaje (learning rate).  Número de árboles.  Profundidad máxima.  Support Vector Machines (SVM):  Encuentra un hiperplano óptimo para la separación de clases.  Parámetros clave:  Kernel (lineal, RBF, etc.).  C (penalización por errores).  Gamma (para kernels no lineales).  K-Nearest Neighbors (KNN):  Clasifica según los vecinos más cercanos.  Parámetros clave:  Número de vecinos (k).  Métrica de distancia (Euclidiana, Manhattan).  Redes Neuronales:  Aplicables a problemas complejos.  Parámetros clave:  Arquitectura de la red.  Learning rate.  Épocas y tamaño del batch.  Naive Bayes:  Basado en probabilidades condicionales.  Parámetros clave:  Distribución asumida (Gaussiana, Bernoulli, Multinomial).  Parámetros generales importantes a controlar:  Hiperparámetros de regularización:  Evitan sobreajuste (ej.: alpha, lambda).  Tasa de aprendizaje:  En algoritmos iterativos, controla el tamaño de los pasos.  Cantidad de datos:  Tamaño del conjunto de entrenamiento.  Balance de clases (para clasificación):  Métodos como pesos de clase o sobremuestreo/infra-muestreo.  Selección de características:  Reducir la dimensionalidad mejora la interpretabilidad.  Validación cruzada:  Divide los datos para evaluar el modelo con diferentes subconjuntos.  Métrica de evaluación:  Determina cómo medir el éxito (RMSE, MAE, accuracy, AUC-ROC, etc.).  Este enfoque combinado de algoritmos y ajustes permite abordar problemas con mayor precisión y robustez.  Tú dijiste:  for i in range(2,5,1) y luego quiero que meta un valor de i en una variable que sea num\_(el valor de i)  ChatGPT Plus  ChatGPT  Para lograr lo que describes, necesitas un bucle for que itere por los valores de i y cree dinámicamente variables cuyo nombre incluya el valor de i. Esto puede hacerse utilizando un diccionario (que es más seguro y eficiente que manejar variables dinámicamente).  Aquí tienes un ejemplo de cómo hacerlo en Python:  # Diccionario para almacenar las variables dinámicas  variables = {}  # Bucle para iterar de 2 a 4 (inclusive, paso 1)  for i in range(2, 5, 1):  # Crear el nombre de la variable dinámica  variable\_name = f"num\_{i}"  # Asignar un valor a la variable  variables[variable\_name] = i  # Imprimir las variables dinámicas almacenadas en el diccionario  for name, value in variables.items():  print(f"{name}: {value}")  En este caso:  El nombre de la variable se construye como un string usando f"num\_{i}".  Usamos un diccionario para asociar el nombre dinámico con su valor.  Al final, puedes acceder a las variables dinámicas con variables["num\_2"], variables["num\_3"], etc.  Esto es mucho más manejable y seguro que tratar de crear nombres de variables directamente en el espacio de nombres global o local.  Tú dijiste:  en que caso hay que utilizar micro macro o otros  ChatGPT Plus  ChatGPT  En machine learning, micro, macro, y otras estrategias como weighted para calcular métricas (por ejemplo, precision, recall, y F1-score) son esenciales en problemas de clasificación con múltiples clases o clases desbalanceadas. Aquí te explico cuándo usar cada una y por qué:  1. Micro-average  Cómo funciona: Calcula métricas considerando las predicciones de todas las clases juntas, como si fueran un solo problema binario. Suma los verdaderos positivos, falsos positivos, y falsos negativos para todas las clases antes de calcular las métricas.  Uso recomendado:  Cuando todas las clases tienen igual importancia, sin importar el balance de datos.  Para problemas de clasificación multi-clase o multi-etiqueta en los que deseas evaluar el rendimiento general del modelo sin destacar diferencias entre clases.  Ventajas:  Da más peso a las clases frecuentes porque acumula los conteos globalmente.  Es útil cuando quieres optimizar el rendimiento general en lugar de enfocarte en clases específicas.  Ejemplo de uso:  Clasificar imágenes donde cada clase tiene muchas muestras (p. ej., clasificación de dígitos del 0 al 9 en MNIST).  2. Macro-average  Cómo funciona: Calcula métricas para cada clase por separado y luego toma el promedio de esas métricas, dándole igual peso a cada clase.  Uso recomendado:  Cuando quieres evaluar el rendimiento en todas las clases de manera equilibrada, independientemente de su tamaño o frecuencia.  Útil si tienes clases desbalanceadas y deseas que cada clase tenga la misma importancia.  Ventajas:  Permite detectar si el modelo tiene bajo rendimiento en clases minoritarias.  No se sesga por clases más grandes o frecuentes.  Ejemplo de uso:  Problemas médicos donde las clases minoritarias (como enfermedades raras) son tan importantes como las mayoritarias (como no tener la enfermedad).  3. Weighted-average  Cómo funciona: Calcula métricas para cada clase y pondera su contribución al promedio según el número de muestras de cada clase.  Uso recomendado:  Cuando tienes un conjunto de datos desbalanceado, pero quieres tener en cuenta la proporción de cada clase.  Útil si las clases más frecuentes deben tener mayor peso en la evaluación general.  Ventajas:  Considera tanto el rendimiento general como la distribución de datos entre clases.  Es una alternativa intermedia entre micro y macro cuando hay desbalance.  Ejemplo de uso:  Evaluar un modelo en un dataset desbalanceado para predecir clientes leales frente a no leales, donde los clientes leales son minoría pero económicamente relevantes.  4. Otras estrategias  Per-class (clase por clase):  Métricas calculadas por separado para cada clase.  Útil cuando quieres examinar específicamente el rendimiento en clases clave.  Subset accuracy (exact match en problemas multi-etiqueta):  Evalúa si el conjunto completo de predicciones coincide exactamente con las etiquetas reales.  Muy estricto, se usa poco en la práctica, excepto para aplicaciones de sistemas críticos.  Resumen de cuándo usar cada una:  Estrategia ¿Clases balanceadas? Igual importancia para todas las clases Clases desbalanceadas y peso proporcional Foco en clases minoritarias  Micro-average ✅ ✅ ❌ ❌  Macro-average ✅/❌ ✅ ❌ ✅  Weighted ❌ ❌ ✅ ✅  Per-class ✅/❌ ✅ (por clase específica) N/A N/A  En general, elegir la estrategia adecuada depende de la distribución de datos y los objetivos del proyecto.  ChatGPT puede cometer errores. Considera verificar la información importante.  ?d máxima (int > 0 o None, Default: None). - **min\_samples\_split**: Divisiones mínimas (int > 1 o float, Default: 2). - **max\_features**: 'sqrt', 'log2', None (Default: 'sqrt'). | - **ROC-AUC**: Recomendado para problemas desbalanceados. - **F1-Score**: Evaluar el equilibrio entre precisión y recall en datos desbalanceados. - **Precision/Recall**: Dependiendo del costo de errores. - **Confusión**: Para entender errores entre clases específicas. |
| **Máquinas de Soporte Vectorial** (SVC) | Datos no linealmente separables; ideal para baja/moderada dimensionalidad. | - **C**: Regularización (float > 0, Default: 1.0). - **kernel**: 'linear', 'poly', 'rbf', 'sigmoid', 'precomputed' (Default: 'rbf'). - **degree**: Grado (si kernel='poly') (int > 0, Default: 3). - **gamma**: 'scale', 'auto', float > 0 (Default: 'scale'). - **probability**: Calcular probabilidades (True o False, Default: False). | - **ROC-AUC**: Medir capacidad del modelo para distinguir entre clases en problemas desbalanceados. - **F1-Score**: Si el dataset está desbalanceado. - **Precision**: Cuando falsos positivos tienen más impacto. - **Confusión**: Análisis detallado de predicciones erróneas. |
| **K-Vecinos Más Cercanos** (KNeighborsClassifier) | Problemas pequeños con datos bien distribuidos; no asume distribución específica de los datos. | - **n\_neighbors**: Nº de vecinos (int > 0, Default: 5). - **weights**: 'uniform', 'distance' o función personalizada (Default: 'uniform'). - **metric**: 'minkowski', 'euclidean', 'manhattan' (Default: 'minkowski'). - **p**: Parámetro para Minkowski (int > 0, Default: 2). | - **Accuracy**: Evaluar modelos en problemas balanceados. - **Confusión**: Explorar qué errores son más comunes. - **F1-Score**: Si el dataset tiene clases desbalanceadas. - **Precision/Recall**: Dependiendo de los costos de errores en el problema. |

## Metricas

|  |  |  |
| --- | --- | --- |
| **Métrica** | **Código en Python** | **Descripción** |
| **Accuracy** | clf.fit(X\_train, y\_train)  y\_pred = clf.predict(X\_test  accuracy = accuracy\_score(y\_test, y\_pred) | Proporción de predicciones correctas respecto al total. Ideal para problemas balanceados. |
| **Precision** | precision = precision\_score(y\_test, y\_pred) | Proporción de verdaderos positivos entre todas las predicciones positivas. Útil si los falsos positivos son costosos. |
| **Recall (Sensibilidad)** | recall = recall\_score(y\_test, y\_pred) | Proporción de verdaderos positivos correctamente identificados. Importante si los falsos negativos son más costosos. |
| **F1-Score** | f1 = f1\_score(y\_test, y\_pred) | Promedio armónico de precisión y sensibilidad. Recomendado para problemas desbalanceados. |
| **ROC-AUC** | roc\_auc = roc\_auc\_score(y\_test, clf.predict\_proba(X\_test)[:,1]) | Medida evaluando una capacidad efectiva para distinguir con área bajo la curva perfecta |

## **Optimización de Hiperparámetros**

### **1. Búsqueda Exhaustiva (Grid Search)**

La búsqueda exhaustiva explora todas las combinaciones posibles de los hiperparámetros dentro de los rangos especificados.

* **Código:**

python

Copiar código

from sklearn.model\_selection import GridSearchCV

# Ejemplo con Random Forest

param\_grid = {

'n\_estimators': [50, 100, 200],

'max\_depth': [None, 10, 20],

'min\_samples\_split': [2, 5, 10]

}

grid\_search = GridSearchCV(estimator=RandomForestClassifier(), param\_grid=param\_grid, cv=5, scoring='accuracy')

grid\_search.fit(X\_train, y\_train)

best\_params = grid\_search.best\_params\_

print("Mejores Hiperparámetros:", best\_params)

* **Ventajas**: Encuentra la combinación óptima de hiperparámetros dentro de los rangos definidos.
* **Desventajas**: Costosa en tiempo si los rangos de hiperparámetros son amplios.

### **2. Búsqueda Aleatoria (Randomized Search)**

La búsqueda aleatoria selecciona combinaciones al azar dentro de los rangos definidos, explorando menos combinaciones pero potencialmente encontrando una configuración adecuada más rápido.

* **Código:**

python

Copiar código

from sklearn.model\_selection import RandomizedSearchCV

from scipy.stats import randint

# Ejemplo con Random Forest

param\_dist = {

'n\_estimators': randint(50, 200),

'max\_depth': [None, 10, 20, 30],

'min\_samples\_split': randint(2, 10)

}

random\_search = RandomizedSearchCV(estimator=RandomForestClassifier(), param\_distributions=param\_dist, n\_iter=50, cv=5, scoring='accuracy')

random\_search.fit(X\_train, y\_train)

best\_params = random\_search.best\_params\_

print("Mejores Hiperparámetros:", best\_params)

* **Ventajas**: Más rápido que la búsqueda exhaustiva, especialmente en hiperparámetros con muchos valores posibles.
* **Desventajas**: Puede no explorar todas las configuraciones óptimas.

# Ensamblaje

## **Modelos de Ensamblaje y Hiperparámetros con Explicación**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Modelo** | **Hiperparámetro** | **Rango de Valores** | **Valor por Defecto** | **Explicación del Hiperparámetro** |
| **Bagging** | n\_estimators | 1 a n (entero positivo) | 10 | Número de clasificadores o estimadores base a entrenar en paralelo. Un valor mayor mejora el rendimiento hasta cierto punto. |
|  | max\_samples | 1 a n (fracción o número entero de muestras) | 1.0 | Fracción de muestras utilizadas para entrenar cada clasificador. Un valor menor reduce el sobreajuste pero también puede reducir la precisión. |
|  | max\_features | 1 a n (fracción o número entero de características) | 1.0Historial del chat  ChatGPT  ChatGPT  Explorar GPT  Hoy  Algoritmos ML Regresión Clasificación  7 días anteriores  SCP transferencia archivo remoto  Imputación con SimpleImputer  MNIST Ensemble Voting  Ensemble learning y precisión  Dimensionalidad de MNIST PCA  Curva ROC con probabilidades  Hiperparámetros de entrenamiento  HDFS Connection Refused Error  30 días anteriores  Ryzen 9 5900X vs i7 7700K  MongoDB Clúster Shardeado  Matriz de confusión  Verificar nodos Hadoop  JupyterLab Node.js Issue  Función de recompensa DeepRacer  Ver colecciones en mongosh  Docker Compose Replica Set  Fixing RF Input Error  Ayuda con código  Visualización de dígitos  Parametros incompatibles LogisticRegression  Hadoop Streaming en EMR  Log Transformation Error Fix  PPK to PEM Conversion  Leer CSV desde HDFS  Recompensa velocidad AWS DeepRacer  Push rejected solution  Push rejected solution  Tamaño y tiempo Parquet  Leer letras en vez de palabras  Ejercicios de distribución normal  Uso de verbose en ColumnTransformer  Ejemplo de consulta Nor  octubre  Hadoop Sudoku Puzzle Command  Evitar bloqueo en Ubuntu  Verificar instalación de MongoDB  Data Preprocessing with Pipelines  Valores únicos en DataFrame  Error de archivo no encontrado  Soporte MongoDB i386 Debian  Hadoop SSH Permission Issue  Probabilidades peso recién nacidos  Ajustes en pipeline Titanic  Error en ColumnTransformer  mayo  Mejoras en la Base  Crear Bean JavaEE Simple  Autoincrementación de Clave Primaria  Recuperación de datos GET  Cambiar a un plan superior  Más acceso a los mejores modelos  ChatGPT  Compartir  User  Tú dijiste:  Hazme un resumen sobre que algoritmos hay que utilizar en regresion y en clasificacion en machine learning que parametros principales hay que controlar  ChatGPT Plus  ChatGPT  En machine learning, la elección del algoritmo y el control de parámetros clave son fundamentales para obtener buenos resultados en tareas de regresión y clasificación. Aquí tienes un resumen organizado por tipo de tarea:  Regresión  La regresión se utiliza para predecir valores continuos. Los algoritmos más comunes incluyen:  1. Algoritmos más usados:  Regresión Lineal:  Supone una relación lineal entre las variables.  Parámetros clave:  Learning rate (si se usa en contextos de optimización).  Regularización (L1: Lasso, L2: Ridge).  Regresión Polinómica:  Extiende la regresión lineal con términos polinómicos.  Parámetros clave:  Grado del polinomio.  Regularización para evitar sobreajuste.  Árboles de Decisión:  Divide los datos en intervalos basándose en reglas.  Parámetros clave:  Profundidad máxima del árbol.  Número mínimo de muestras por hoja (min\_samples\_leaf).  Bosques Aleatorios (Random Forests):  Combina múltiples árboles para mejorar la precisión.  Parámetros clave:  Número de árboles.  Profundidad máxima.  Número de características seleccionadas en cada división.  Gradient Boosting (Ej.: XGBoost, LightGBM, CatBoost):  Mejora iterativamente las predicciones del modelo.  Parámetros clave:  Número de árboles.  Tasa de aprendizaje (learning rate).  Profundidad máxima de los árboles.  Support Vector Regression (SVR):  Modela la regresión mediante un margen que minimiza errores.  Parámetros clave:  C (penalización por error).  Epsilon (margen de tolerancia).  Kernel (lineal, RBF, etc.).  Redes Neuronales:  Modelos basados en múltiples capas y no linealidades.  Parámetros clave:  Arquitectura (número de capas y neuronas).  Learning rate.  Función de activación.  Clasificación  La clasificación se utiliza para predecir categorías discretas. Los algoritmos más comunes incluyen:  1. Algoritmos más usados:  Regresión Logística:  Modelo lineal para probabilidades.  Parámetros clave:  Regularización (L1, L2).  Solvers para optimización (liblinear, sag, lbfgs).  Árboles de Decisión:  Divide datos en reglas jerárquicas.  Parámetros clave:  Profundidad máxima.  min\_samples\_split y min\_samples\_leaf.  Bosques Aleatorios:  Reduce el sobreajuste de los árboles individuales.  Parámetros clave:  Número de árboles.  Profundidad máxima.  max\_features.  Gradient Boosting (Ej.: XGBoost, LightGBM, CatBoost):  Similar a regresión, pero optimizado para clasificación.  Parámetros clave:  Tasa de aprendizaje (learning rate).  Número de árboles.  Profundidad máxima.  Support Vector Machines (SVM):  Encuentra un hiperplano óptimo para la separación de clases.  Parámetros clave:  Kernel (lineal, RBF, etc.).  C (penalización por errores).  Gamma (para kernels no lineales).  K-Nearest Neighbors (KNN):  Clasifica según los vecinos más cercanos.  Parámetros clave:  Número de vecinos (k).  Métrica de distancia (Euclidiana, Manhattan).  Redes Neuronales:  Aplicables a problemas complejos.  Parámetros clave:  Arquitectura de la red.  Learning rate.  Épocas y tamaño del batch.  Naive Bayes:  Basado en probabilidades condicionales.  Parámetros clave:  Distribución asumida (Gaussiana, Bernoulli, Multinomial).  Parámetros generales importantes a controlar:  Hiperparámetros de regularización:  Evitan sobreajuste (ej.: alpha, lambda).  Tasa de aprendizaje:  En algoritmos iterativos, controla el tamaño de los pasos.  Cantidad de datos:  Tamaño del conjunto de entrenamiento.  Balance de clases (para clasificación):  Métodos como pesos de clase o sobremuestreo/infra-muestreo.  Selección de características:  Reducir la dimensionalidad mejora la interpretabilidad.  Validación cruzada:  Divide los datos para evaluar el modelo con diferentes subconjuntos.  Métrica de evaluación:  Determina cómo medir el éxito (RMSE, MAE, accuracy, AUC-ROC, etc.).  Este enfoque combinado de algoritmos y ajustes permite abordar problemas con mayor precisión y robustez.  Tú dijiste:  for i in range(2,5,1) y luego quiero que meta un valor de i en una variable que sea num\_(el valor de i)  ChatGPT Plus  ChatGPT  Para lograr lo que describes, necesitas un bucle for que itere por los valores de i y cree dinámicamente variables cuyo nombre incluya el valor de i. Esto puede hacerse utilizando un diccionario (que es más seguro y eficiente que manejar variables dinámicamente).  Aquí tienes un ejemplo de cómo hacerlo en Python:  # Diccionario para almacenar las variables dinámicas  variables = {}  # Bucle para iterar de 2 a 4 (inclusive, paso 1)  for i in range(2, 5, 1):  # Crear el nombre de la variable dinámica  variable\_name = f"num\_{i}"  # Asignar un valor a la variable  variables[variable\_name] = i  # Imprimir las variables dinámicas almacenadas en el diccionario  for name, value in variables.items():  print(f"{name}: {value}")  En este caso:  El nombre de la variable se construye como un string usando f"num\_{i}".  Usamos un diccionario para asociar el nombre dinámico con su valor.  Al final, puedes acceder a las variables dinámicas con variables["num\_2"], variables["num\_3"], etc.  Esto es mucho más manejable y seguro que tratar de crear nombres de variables directamente en el espacio de nombres global o local.  Tú dijiste:  en que caso hay que utilizar micro macro o otros  ChatGPT Plus  ChatGPT  En machine learning, micro, macro, y otras estrategias como weighted para calcular métricas (por ejemplo, precision, recall, y F1-score) son esenciales en problemas de clasificación con múltiples clases o clases desbalanceadas. Aquí te explico cuándo usar cada una y por qué:  1. Micro-average  Cómo funciona: Calcula métricas considerando las predicciones de todas las clases juntas, como si fueran un solo problema binario. Suma los verdaderos positivos, falsos positivos, y falsos negativos para todas las clases antes de calcular las métricas.  Uso recomendado:  Cuando todas las clases tienen igual importancia, sin importar el balance de datos.  Para problemas de clasificación multi-clase o multi-etiqueta en los que deseas evaluar el rendimiento general del modelo sin destacar diferencias entre clases.  Ventajas:  Da más peso a las clases frecuentes porque acumula los conteos globalmente.  Es útil cuando quieres optimizar el rendimiento general en lugar de enfocarte en clases específicas.  Ejemplo de uso:  Clasificar imágenes donde cada clase tiene muchas muestras (p. ej., clasificación de dígitos del 0 al 9 en MNIST).  2. Macro-average  Cómo funciona: Calcula métricas para cada clase por separado y luego toma el promedio de esas métricas, dándole igual peso a cada clase.  Uso recomendado:  Cuando quieres evaluar el rendimiento en todas las clases de manera equilibrada, independientemente de su tamaño o frecuencia.  Útil si tienes clases desbalanceadas y deseas que cada clase tenga la misma importancia.  Ventajas:  Permite detectar si el modelo tiene bajo rendimiento en clases minoritarias.  No se sesga por clases más grandes o frecuentes.  Ejemplo de uso:  Problemas médicos donde las clases minoritarias (como enfermedades raras) son tan importantes como las mayoritarias (como no tener la enfermedad).  3. Weighted-average  Cómo funciona: Calcula métricas para cada clase y pondera su contribución al promedio según el número de muestras de cada clase.  Uso recomendado:  Cuando tienes un conjunto de datos desbalanceado, pero quieres tener en cuenta la proporción de cada clase.  Útil si las clases más frecuentes deben tener mayor peso en la evaluación general.  Ventajas:  Considera tanto el rendimiento general como la distribución de datos entre clases.  Es una alternativa intermedia entre micro y macro cuando hay desbalance.  Ejemplo de uso:  Evaluar un modelo en un dataset desbalanceado para predecir clientes leales frente a no leales, donde los clientes leales son minoría pero económicamente relevantes.  4. Otras estrategias  Per-class (clase por clase):  Métricas calculadas por separado para cada clase.  Útil cuando quieres examinar específicamente el rendimiento en clases clave.  Subset accuracy (exact match en problemas multi-etiqueta):  Evalúa si el conjunto completo de predicciones coincide exactamente con las etiquetas reales.  Muy estricto, se usa poco en la práctica, excepto para aplicaciones de sistemas críticos.  Resumen de cuándo usar cada una:  Estrategia ¿Clases balanceadas? Igual importancia para todas las clases Clases desbalanceadas y peso proporcional Foco en clases minoritarias  Micro-average ✅ ✅ ❌ ❌  Macro-average ✅/❌ ✅ ❌ ✅  Weighted ❌ ❌ ✅ ✅  Per-class ✅/❌ ✅ (por clase específica) N/A N/A  En general, elegir la estrategia adecuada depende de la distribución de datos y los objetivos del proyecto.  ChatGPT puede cometer errores. Considera verificar la información importante.  ? | Número de características utilizadas por cada estimador base. Puede ser un número o una fracción de las características disponibles. |
|  | bootstrap | True, False | True | Si se utiliza muestreo con reemplazo (True) o sin reemplazo (False) para crear las muestras de entrenamiento. |
| from sklearn.ensemble import BaggingClassifier  from sklearn.tree import DecisionTreeClassifier  import numpy as np  # Datos de ejemplo  X = np.array([[1, 2], [2, 3], [3, 4], [4, 5]])  y = np.array([0, 1, 0, 1])  # Crear el modelo BaggingClassifier con los hiperparámetros especificados  modelo\_bagging = BaggingClassifier(  base\_estimator=DecisionTreeClassifier(),  n\_estimators=10, # Número de clasificadores base  max\_samples=1.0, # Fracción de muestras para entrenar cada clasificador  max\_features=1.0, # Número de características utilizadas por cada estimador base  bootstrap=True # Muestreo con reemplazo  )  # Ajustar el modelo  modelo\_bagging.fit(X, y)  # Hacer predicciones  predicciones = modelo\_bagging.predict(X)  print("Predicciones:", predicciones) | | | | |
| **Random Forest** | n\_estimators | 1 a n (entero positivo) | 100 | Número de árboles de decisión en el bosque. Un valor más alto puede mejorar el rendimiento pero aumenta el tiempo de computación. |
|  | max\_depth | None, 1 a n (entero) | None | Profundidad máxima de los árboles. Si es None, se expanden hasta que cada hoja contenga menos de min\_samples\_split muestras. |
|  | min\_samples\_split | 2 a n (entero) | 2 | Número mínimo de muestras necesarias para dividir un nodo. Controla el tamaño de los árboles. |
|  | min\_samples\_leaf | 1 a n (entero) | 1 | Número mínimo de muestras necesarias para estar en una hoja del árbol. |
|  | max\_features | 1 a n (fracción o número entero de características) | sqrt | Número máximo de características para dividir cada nodo. 'sqrt' es el valor predeterminado, que selecciona la raíz cuadrada de las características. |
| from sklearn.ensemble import RandomForestClassifier  import numpy as np  # Datos de ejemplo  X = np.array([[1, 2], [2, 3], [3, 4], [4, 5]])  y = np.array([0, 1, 0, 1])  # Crear el modelo RandomForestClassifier con los hiperparámetros especificados  modelo\_random\_forest = RandomForestClassifier(  n\_estimators=100, # Número de árboles en el bosque  max\_depth=None, # Profundidad máxima de los árboles  min\_samples\_split=2, # Número mínimo de muestras para dividir un nodo  min\_samples\_leaf=1, # Número mínimo de muestras en una hoja del árbol  max\_features='sqrt' # Número máximo de características para dividir cada nodo  )  # Ajustar el modelo  modelo\_random\_forest.fit(X, y)  # Hacer predicciones  predicciones = modelo\_random\_forest.predict(X)  print("Importancia de características:", modelo\_random\_forest.feature\_importances\_)  print("Predicciones:", predicciones) | | | | |
| **AdaBoost** | n\_estimators | 1 a n (entero positivo) | 50 | Número de estimadores base a entrenar. Un valor mayor puede mejorar el rendimiento, pero también puede aumentar el riesgo de sobreajuste. |
|  | learning\_rate | 0 a 1 (flotante) | 1.0 | Tasa de aprendizaje que ajusta la contribución de cada estimador al modelo final. Un valor pequeño puede mejorar el rendimiento, pero aumentar el tiempo de entrenamiento. |
|  | algorithm | 'SAMME', 'SAMME.R' | 'SAMME.R' | Algoritmo utilizado para el entrenamiento. 'SAMME.R' es más rápido y eficaz, mientras que 'SAMME' es más clásico. |
| from sklearn.ensemble import AdaBoostClassifier  from sklearn.tree import DecisionTreeClassifier  import numpy as np  # Datos de ejemplo  X = np.array([[1, 2], [2, 3], [3, 4], [4, 5]])  y = np.array([0, 1, 0, 1])  # Crear el modelo AdaBoostClassifier con los hiperparámetros especificados  modelo\_adaboost = AdaBoostClassifier(  base\_estimator=DecisionTreeClassifier(),  n\_estimators=50, # Número de estimadores base  learning\_rate=1.0, # Tasa de aprendizaje  algorithm='SAMME.R' # Algoritmo utilizado para el entrenamiento  )  # Ajustar el modelo  modelo\_adaboost.fit(X, y)  # Hacer predicciones  predicciones = modelo\_adaboost.predict(X)  print("Predicciones:", predicciones) | | | | |
| **Gradient Boosting** | n\_estimators | 1 a n (entero positivo) | 100 | Número de árboles a entrenar. Un número mayor generalmente mejora el rendimiento, pero puede aumentar el riesgo de sobreajuste. |
|  | learning\_rate | 0.0 a 1.0 (flotante) | 0.1 | Tasa de aprendizaje que controla el impacto de cada estimador. Un valor pequeño permite que el modelo se entrene durante más iteraciones sin sobreajustar. |
|  | max\_depth | 1 a n (entero positivo) | 3 | Profundidad máxima de cada árbol. El valor predeterminado ayuda a evitar el sobreajuste en la mayoría de los casos. |
|  | min\_samples\_split | 2 a n (entero) | 2 | Número mínimo de muestras necesarias para dividir un nodo. Esto controla el crecimiento de los árboles. |
| from sklearn.ensemble import GradientBoostingClassifier  import numpy as np  # Datos de ejemplo  X = np.array([[1, 2], [2, 3], [3, 4], [4, 5]])  y = np.array([0, 1, 0, 1])  # Crear el modelo GradientBoostingClassifier con los hiperparámetros especificados  modelo\_gb = GradientBoostingClassifier(  n\_estimators=100, # Número de árboles a entrenar  learning\_rate=0.1, # Tasa de aprendizaje  max\_depth=3, # Profundidad máxima de cada árbol  min\_samples\_split=2 # Número mínimo de muestras para dividir un nodo  )  # Ajustar el modelo  modelo\_gb.fit(X, y)  # Hacer predicciones  predicciones = modelo\_gb.predict(X)  print("Importancia de características:", modelo\_gb.feature\_importances\_)  print("Predicciones:", predicciones) | | | | |
| **XGBoost** | n\_estimators | 1 a n (entero positivo) | 100 | Número de árboles de decisión a entrenar. Un número más alto puede mejorar la precisión, pero aumenta el tiempo de entrenamiento. |
|  | learning\_rate | 0.0 a 1.0 (flotante) | 0.3 | Tasa de aprendizaje que controla el impacto de cada árbol. Un valor pequeño reduce el sobreajuste y mejora la generalización. |
|  | max\_depth | 1 a n (entero positivo) | 6 | Profundidad máxima de los árboles. Un valor mayor permite modelos más complejos, pero puede causar sobreajuste. |
|  | subsample | 0.5 a 1.0 (flotante) | 1.0 | Fracción de muestras utilizada para entrenar cada árbol. Si es menor que 1.0, reduce el riesgo de sobreajuste. |
|  | colsample\_bytree | 0.5 a 1.0 (flotante) | 1.0 | Fracción de características utilizadas para entrenar cada árbol. Reducir este valor ayuda a mejorar la generalización. |
| from xgboost import XGBClassifier  import numpy as np  # Datos de ejemplo  X = np.array([[1, 2], [2, 3], [3, 4], [4, 5]])  y = np.array([0, 1, 0, 1])  # Crear el modelo XGBClassifier con los hiperparámetros especificados  modelo\_xgb = XGBClassifier(  n\_estimators=100, # Número de árboles de decisión  learning\_rate=0.3, # Tasa de aprendizaje  max\_depth=6, # Profundidad máxima de los árboles  subsample=1.0, # Fracción de muestras utilizada  colsample\_bytree=1.0 # Fracción de características utilizada  )  # Ajustar el modelo  modelo\_xgb.fit(X, y)  # Hacer predicciones  predicciones = modelo\_xgb.predict(X)  print("Importancia de características:", modelo\_xgb.feature\_importances\_)  print("Predicciones:", predicciones) | | | | |
| **LightGBM** | n\_estimators | 1 a n (entero positivo) | 100 | Número de árboles de decisión. |
|  | learning\_rate | 0.0 a 1.0 (flotante) | 0.1 | Tasa de aprendizaje que ajusta la contribución de cada árbol al modelo final. |
|  | max\_depth | -1, 1 a n (entero positivo) | -1 | Profundidad máxima de los árboles. Si es -1, no hay límite. |
|  | num\_leaves | 31 a 2^max\_depth (entero) | 31 | Número de hojas por árbol. Un valor mayor permite árboles más complejos. |
| from lightgbm import LGBMClassifier  import numpy as np  # Datos de ejemplo  X = np.array([[1, 2], [2, 3], [3, 4], [4, 5]])  y = np.array([0, 1, 0, 1])  # Crear el modelo LGBMClassifier con los hiperparámetros especificados  modelo\_lgbm = LGBMClassifier(  n\_estimators=100, # Número de árboles de decisión  learning\_rate=0.1, # Tasa de aprendizaje  max\_depth=-1, # Profundidad máxima de los árboles  num\_leaves=31 # Número de hojas por árbol  )  # Ajustar el modelo  modelo\_lgbm.fit(X, y)  # Hacer predicciones  predicciones = modelo\_lgbm.predict(X)  print("Importancia de características:", modelo\_lgbm.feature\_importances\_)  print("Predicciones:", predicciones) | | | | |
| **CatBoost** | iterations | 1 a n (entero positivo) | 1000 | Número de árboles a entrenar. |
|  | learning\_rate | 0.0 a 1.0 (flotante) | 0.03 | Tasa de aprendizaje. |
|  | depth | 1 a 16 (entero positivo) | 6 | Profundidad máxima de los árboles. A mayor profundidad, más complejo es el modelo, pero mayor riesgo de sobreajuste. |
|  | l2\_leaf\_reg | 0 a 100 (flotante) | 3.0 | Término de regularización L2 que controla el sobreajuste. |
|  | border\_count | 32 a 255 (entero positivo) | 128 | Número de valores discretos para representar las características continuas. |
| # Datos de ejemplo  X = np.array([[1, 2], [2, 3], [3, 4], [4, 5]])  y = np.array([0, 1, 0, 1])  # Crear el modelo CatBoostClassifier con los hiperparámetros especificados  modelo\_catboost = CatBoostClassifier(  iterations=1000, # Número de árboles a entrenar  learning\_rate=0.03, # Tasa de aprendizaje  depth=6, # Profundidad máxima de los árboles  l2\_leaf\_reg=3.0, # Término de regularización L2  border\_count=128 # Número de valores discretos para características continuas  )  # Ajustar el modelo (nota: verbose=0 para suprimir salida durante el ajuste)  modelo\_catboost.fit(X, y, verbose=0)  # Hacer predicciones  predicciones = modelo\_catboost.predict(X)  print("Predicciones:", predicciones) | | | | |
| **StackingClassifier** | estimators | List of tuples (modelo, nombre) | None | Lista de modelos base (estimadores). Cada elemento de la lista es una tupla con el nombre del modelo y el estimador correspondiente. |
|  | final\_estimator | Clasificador (puede ser cualquier estimador) | LogisticRegression() | Clasificador final que se utiliza para combinar las predicciones de los modelos base. Por defecto es una regresión logística. |
|  | cv | None, int (entero positivo o Cross-Validation Generator) | 5 | Número de divisiones para la validación cruzada utilizada para entrenar los modelos base. |
|  | stack\_method | 'auto', 'predict\_proba', 'predict' | 'auto' | Método para combinar las predicciones de los modelos base. 'auto' elige automáticamente el mejor. |
|  | passthrough | True, False | False | Si es True, las predicciones de los estimadores base se pasan directamente al final\_estimator además de las predicciones combinadas. |
|  | n\_jobs | -1 a n (entero) | None | Número de núcleos a utilizar para el entrenamiento paralelo de los modelos base. Si es -1, se usan todos los núcleos. |
| # Datos de ejemplo  X = np.array([[1, 2], [2, 3], [3, 4], [4, 5]])  y = np.array([0, 1, 0, 1])  # Crear los estimadores base  estimadores = [  ('decision\_tree', DecisionTreeClassifier()),  ('svm', SVC(probability=True))  ]  # Crear el modelo StackingClassifier con los hiperparámetros especificados  modelo\_stacking = StackingClassifier(  estimators=estimadores,  final\_estimator=LogisticRegression(), # Clasificador final  cv=5, # Validación cruzada con 5 divisiones  stack\_method='auto', # Método para combinar las predicciones  passthrough=False, # No pasar las predicciones de los estimadores base al final\_estimator  n\_jobs=-1 # Usar todos los núcleos disponibles para el entrenamiento  )  # Ajustar el modelo  modelo\_stacking.fit(X, y)  # Hacer predicciones  predicciones = modelo\_stacking.predict(X)  print("Predicciones:", predicciones) | | | | |

# Reducción de dimensionalidad

|  |  |  |
| --- | --- | --- |
| **Elemento** | **Descripción** | **Código en Python** |
| **PCA (Principal Component Analysis)** | Técnica de reducción de dimensionalidad que transforma las características originales en un conjunto de componentes principales, que explican la mayor parte de la varianza. Es útil cuando se tienen muchas variables correlacionadas. | from sklearn.decomposition import PCA  pca = PCA(n\_components=2)  X\_reduced = pca.fit\_transform(X\_train) |
| **n\_components** | Número de componentes principales a retener en la reducción de la dimensionalidad. Si se establece un valor entre 0 y 1, se seleccionan suficientes componentes para explicar esa proporción de la varianza. | PCA(n\_components=0.95) (explica el 95% de la varianza) |
| **svd\_solver** | Algoritmo utilizado para calcular los componentes principales. Los valores posibles son 'auto', 'full', 'arpack' y 'randomized'. Dependiendo de los datos, algunos algoritmos pueden ser más rápidos o más precisos. | PCA(svd\_solver='randomized') |
| **whiten** | Si es True, los componentes principales serán reescalados (whitened) para tener varianza unitaria. Esto puede ser útil para algunos modelos de aprendizaje automático que requieren variables con varianza similar. | PCA(whiten=True) |
| **copy** | Si es True, se hace una copia de los datos originales al realizar la transformación. Si es False, la transformación se realiza en los datos originales (puede ahorrar memoria). | PCA(copy=True) |
| **random\_state** | Semilla para asegurar la reproducibilidad de la descomposición si se usa el svd\_solver='randomized'. Especifica la semilla del generador de números aleatorios. | PCA(random\_state=42) |
| **mean\_centering** | Centra los datos (resta la media de cada característica) antes de aplicar PCA. Es la práctica estándar en la mayoría de los casos. Este parámetro se aplica de manera predeterminada. | PCA() (automáticamente centra los datos) |

## Historial del chat

[ChatGPT](https://chatgpt.com/)

[Explorar GPT](https://chatgpt.com/gpts)

### Hoy

1. [Algoritmos ML Regresión Clasificación](https://chatgpt.com/c/674dbf48-42fc-8003-bb38-02471dfe51e7)

### 7 días anteriores

1. [SCP transferencia archivo remoto](https://chatgpt.com/c/6748c479-373c-8003-bf7c-057350cfab3c)
2. [Imputación con SimpleImputer](https://chatgpt.com/c/6748bc7f-d2e8-8003-be19-0352ea1d16c8)
3. [MNIST Ensemble Voting](https://chatgpt.com/c/67489627-cc18-8003-8778-ec4642c0e4b7)
4. [Ensemble learning y precisión](https://chatgpt.com/c/67489032-aa6c-8003-811e-dc6682fa2b6f)
5. [Dimensionalidad de MNIST PCA](https://chatgpt.com/c/67487c2f-5f68-8003-87f0-a7bcb95f386f)
6. [Curva ROC con probabilidades](https://chatgpt.com/c/6745d88c-4b00-8003-8cfb-127085019f1b)
7. [Hiperparámetros de entrenamiento](https://chatgpt.com/c/67472a51-d5a4-8003-8e1b-5e5c94a89a07)
8. [HDFS Connection Refused Error](https://chatgpt.com/c/671a8fee-4f74-8003-a5d9-a14c63f09905)

### 30 días anteriores

1. [Ryzen 9 5900X vs i7 7700K](https://chatgpt.com/c/673f8844-a3c4-8003-87ec-7fbb63e36dc3)
2. [MongoDB Clúster Shardeado](https://chatgpt.com/c/673f7159-f0ac-8003-a040-7c1f204f9374)
3. [Matriz de confusión](https://chatgpt.com/c/673ca96c-7ba8-8003-8203-bbdc6809b885)
4. [Verificar nodos Hadoop](https://chatgpt.com/c/673e3169-f4b4-8003-86aa-9b59e7ea9101)
5. [JupyterLab Node.js Issue](https://chatgpt.com/c/673e003f-ec8c-8003-b7ba-1a689e1b7d5c)
6. [Función de recompensa DeepRacer](https://chatgpt.com/c/672247b0-f564-8003-80de-57490a807376)
7. [Ver colecciones en mongosh](https://chatgpt.com/c/673ce6f9-c994-8003-b714-b7b0f249c1e9)
8. [Docker Compose Replica Set](https://chatgpt.com/c/673b60fc-e87c-8003-b89a-1bc4f735cd80)
9. [Fixing RF Input Error](https://chatgpt.com/c/673ccae6-9640-8003-9359-89887be6c269)
10. [Ayuda con código](https://chatgpt.com/c/673cba68-1a74-8003-a14e-39d05777d611)
11. [Visualización de dígitos](https://chatgpt.com/c/673cacc6-99b0-8003-9a26-b2c897b441e6)
12. [Parametros incompatibles LogisticRegression](https://chatgpt.com/c/673b8909-42dc-8003-b21a-aba66756144d)
13. [Hadoop Streaming en EMR](https://chatgpt.com/c/673636fc-92b8-8003-b374-9e4a56051a43)
14. [Log Transformation Error Fix](https://chatgpt.com/c/673606ea-0680-8003-a51d-2aed8db098b2)
15. [PPK to PEM Conversion](https://chatgpt.com/c/6734fa3e-049c-8003-8d33-7b7b262dd25c)
16. [Leer CSV desde HDFS](https://chatgpt.com/c/6734ebf0-21f8-8003-b3a6-dcae270ab380)
17. [Recompensa velocidad AWS DeepRacer](https://chatgpt.com/c/6734b844-3430-8003-9730-823019b7618e)
18. [Push rejected solution](https://chatgpt.com/c/67321858-ac88-8003-83b5-b87af9364931)
19. [Push rejected solution](https://chatgpt.com/c/672d188e-5138-8003-9c2a-e72aa1824412)
20. [Tamaño y tiempo Parquet](https://chatgpt.com/c/672d1258-f0b4-8003-9641-3d67c2695839)
21. [Leer letras en vez de palabras](https://chatgpt.com/c/672d03de-87dc-8003-b2fa-a4d0663168f8)
22. [Ejercicios de distribución normal](https://chatgpt.com/c/672ce6f9-6438-8003-b6e1-e3d6493a04d7)
23. [Uso de verbose en ColumnTransformer](https://chatgpt.com/c/6728f67c-09a0-8003-ac00-64203c86e82c)
24. [Ejemplo de consulta Nor](https://chatgpt.com/c/6728e192-391c-8003-bedd-7fd941f6c0be)

### octubre

1. [Hadoop Sudoku Puzzle Command](https://chatgpt.com/c/67226dd8-6fcc-8003-87f2-6c6fd546e846)
2. [Evitar bloqueo en Ubuntu](https://chatgpt.com/c/67226b6c-c400-8003-8bee-8312ac64ec73)
3. [Verificar instalación de MongoDB](https://chatgpt.com/c/67213405-a50c-8003-ad08-3c9669bfe31a)
4. [Data Preprocessing with Pipelines](https://chatgpt.com/c/671fcb9e-d370-8003-bf41-def39e352660)
5. [Valores únicos en DataFrame](https://chatgpt.com/c/671fbe85-7068-8003-b78f-3f33ecc557cc)
6. [Error de archivo no encontrado](https://chatgpt.com/c/671fb61c-42d4-8003-aa7e-0dec3e02d9f8)
7. [Soporte MongoDB i386 Debian](https://chatgpt.com/c/671f9ea7-8088-8003-a143-719faae7c57c)
8. [Hadoop SSH Permission Issue](https://chatgpt.com/c/67193362-7a50-8003-82bd-a48c6a5f0b7e)
9. [Probabilidades peso recién nacidos](https://chatgpt.com/c/6717f059-a10c-8003-aa4f-f8ddda6b7055)
10. [Ajustes en pipeline Titanic](https://chatgpt.com/c/6717d53e-8a0c-8003-8ecd-0899a4b1034a)
11. [Error en ColumnTransformer](https://chatgpt.com/c/6717d316-1338-8003-b2f1-2d120c3687d0)

### mayo

1. [Mejoras en la Base](https://chatgpt.com/c/ed4e6c28-3aff-4543-8227-da3df18a8951)
2. [Crear Bean JavaEE Simple](https://chatgpt.com/c/94c1d941-097f-4345-b69f-990964516da2)
3. [Autoincrementación de Clave Primaria](https://chatgpt.com/c/2dd27d0f-a3d1-46cd-890e-14534e9ff092)
4. [Recuperación de datos GET](https://chatgpt.com/c/d4d227ba-0087-4b0a-827f-71afda938ce6)

Cambiar a un plan superiorMás acceso a los mejores modelos

ChatGPT

Compartir



##### Tú dijiste:

Hazme un resumen sobre que algoritmos hay que utilizar en regresion y en clasificacion en machine learning que parametros principales hay que controlar

###### ChatGPT Plus

En machine learning, la elección del algoritmo y el control de parámetros clave son fundamentales para obtener buenos resultados en tareas de regresión y clasificación. Aquí tienes un resumen organizado por tipo de tarea:

### Regresión

La regresión se utiliza para predecir valores continuos. Los algoritmos más comunes incluyen:

#### 1. Algoritmos más usados:

* Regresión Lineal:
  + Supone una relación lineal entre las variables.
  + Parámetros clave:
    - Learning rate (si se usa en contextos de optimización).
    - Regularización (L1: Lasso, L2: Ridge).
* Regresión Polinómica:
  + Extiende la regresión lineal con términos polinómicos.
  + Parámetros clave:
    - Grado del polinomio.
    - Regularización para evitar sobreajuste.
* Árboles de Decisión:
  + Divide los datos en intervalos basándose en reglas.
  + Parámetros clave:
    - Profundidad máxima del árbol.
    - Número mínimo de muestras por hoja (min\_samples\_leaf).
* Bosques Aleatorios (Random Forests):
  + Combina múltiples árboles para mejorar la precisión.
  + Parámetros clave:
    - Número de árboles.
    - Profundidad máxima.
    - Número de características seleccionadas en cada división.
* Gradient Boosting (Ej.: XGBoost, LightGBM, CatBoost):
  + Mejora iterativamente las predicciones del modelo.
  + Parámetros clave:
    - Número de árboles.
    - Tasa de aprendizaje (learning rate).
    - Profundidad máxima de los árboles.
* Support Vector Regression (SVR):
  + Modela la regresión mediante un margen que minimiza errores.
  + Parámetros clave:
    - C (penalización por error).
    - Epsilon (margen de tolerancia).
    - Kernel (lineal, RBF, etc.).
* Redes Neuronales:
  + Modelos basados en múltiples capas y no linealidades.
  + Parámetros clave:
    - Arquitectura (número de capas y neuronas).
    - Learning rate.
    - Función de activación.

### Clasificación

La clasificación se utiliza para predecir categorías discretas. Los algoritmos más comunes incluyen:

#### 1. Algoritmos más usados:

* Regresión Logística:
  + Modelo lineal para probabilidades.
  + Parámetros clave:
    - Regularización (L1, L2).
    - Solvers para optimización (liblinear, sag, lbfgs).
* Árboles de Decisión:
  + Divide datos en reglas jerárquicas.
  + Parámetros clave:
    - Profundidad máxima.
    - min\_samples\_split y min\_samples\_leaf.
* Bosques Aleatorios:
  + Reduce el sobreajuste de los árboles individuales.
  + Parámetros clave:
    - Número de árboles.
    - Profundidad máxima.
    - max\_features.
* Gradient Boosting (Ej.: XGBoost, LightGBM, CatBoost):
  + Similar a regresión, pero optimizado para clasificación.
  + Parámetros clave:
    - Tasa de aprendizaje (learning rate).
    - Número de árboles.
    - Profundidad máxima.
* Support Vector Machines (SVM):
  + Encuentra un hiperplano óptimo para la separación de clases.
  + Parámetros clave:
    - Kernel (lineal, RBF, etc.).
    - C (penalización por errores).
    - Gamma (para kernels no lineales).
* K-Nearest Neighbors (KNN):
  + Clasifica según los vecinos más cercanos.
  + Parámetros clave:
    - Número de vecinos (k).
    - Métrica de distancia (Euclidiana, Manhattan).
* Redes Neuronales:
  + Aplicables a problemas complejos.
  + Parámetros clave:
    - Arquitectura de la red.
    - Learning rate.
    - Épocas y tamaño del batch.
* Naive Bayes:
  + Basado en probabilidades condicionales.
  + Parámetros clave:
    - Distribución asumida (Gaussiana, Bernoulli, Multinomial).

### Parámetros generales importantes a controlar:

1. Hiperparámetros de regularización:
   * Evitan sobreajuste (ej.: alpha, lambda).
2. Tasa de aprendizaje:
   * En algoritmos iterativos, controla el tamaño de los pasos.
3. Cantidad de datos:
   * Tamaño del conjunto de entrenamiento.
4. Balance de clases (para clasificación):
   * Métodos como pesos de clase o sobremuestreo/infra-muestreo.
5. Selección de características:
   * Reducir la dimensionalidad mejora la interpretabilidad.
6. Validación cruzada:
   * Divide los datos para evaluar el modelo con diferentes subconjuntos.
7. Métrica de evaluación:
   * Determina cómo medir el éxito (RMSE, MAE, accuracy, AUC-ROC, etc.).

Este enfoque combinado de algoritmos y ajustes permite abordar problemas con mayor precisión y robustez.

##### Tú dijiste:

for i in range(2,5,1) y luego quiero que meta un valor de i en una variable que sea num\_(el valor de i)

###### ChatGPT Plus

Para lograr lo que describes, necesitas un bucle for que itere por los valores de i y cree dinámicamente variables cuyo nombre incluya el valor de i. Esto puede hacerse utilizando un diccionario (que es más seguro y eficiente que manejar variables dinámicamente).

Aquí tienes un ejemplo de cómo hacerlo en Python:

python

Copiar código

# Diccionario para almacenar las variables dinámicas

variables = {}

# Bucle para iterar de 2 a 4 (inclusive, paso 1)

for i in range(2, 5, 1):

# Crear el nombre de la variable dinámica

variable\_name = f"num\_{i}"

# Asignar un valor a la variable

variables[variable\_name] = i

# Imprimir las variables dinámicas almacenadas en el diccionario

for name, value in variables.items():

print(f"{name}: {value}")

En este caso:

1. El nombre de la variable se construye como un string usando f"num\_{i}".
2. Usamos un diccionario para asociar el nombre dinámico con su valor.
3. Al final, puedes acceder a las variables dinámicas con variables["num\_2"], variables["num\_3"], etc.

Esto es mucho más manejable y seguro que tratar de crear nombres de variables directamente en el espacio de nombres global o local.

##### Tú dijiste:

en que caso hay que utilizar micro macro o otros

###### ChatGPT Plus

En machine learning, micro, macro, y otras estrategias como weighted para calcular métricas (por ejemplo, precision, recall, y F1-score) son esenciales en problemas de clasificación con múltiples clases o clases desbalanceadas. Aquí te explico cuándo usar cada una y por qué:

### 1. Micro-average

Cómo funciona: Calcula métricas considerando las predicciones de todas las clases juntas, como si fueran un solo problema binario. Suma los verdaderos positivos, falsos positivos, y falsos negativos para todas las clases antes de calcular las métricas.

* Uso recomendado:
  + Cuando todas las clases tienen igual importancia, sin importar el balance de datos.
  + Para problemas de clasificación multi-clase o multi-etiqueta en los que deseas evaluar el rendimiento general del modelo sin destacar diferencias entre clases.
* Ventajas:
  + Da más peso a las clases frecuentes porque acumula los conteos globalmente.
  + Es útil cuando quieres optimizar el rendimiento general en lugar de enfocarte en clases específicas.
* Ejemplo de uso:
  + Clasificar imágenes donde cada clase tiene muchas muestras (p. ej., clasificación de dígitos del 0 al 9 en MNIST).

### 2. Macro-average

Cómo funciona: Calcula métricas para cada clase por separado y luego toma el promedio de esas métricas, dándole igual peso a cada clase.

* Uso recomendado:
  + Cuando quieres evaluar el rendimiento en todas las clases de manera equilibrada, independientemente de su tamaño o frecuencia.
  + Útil si tienes clases desbalanceadas y deseas que cada clase tenga la misma importancia.
* Ventajas:
  + Permite detectar si el modelo tiene bajo rendimiento en clases minoritarias.
  + No se sesga por clases más grandes o frecuentes.
* Ejemplo de uso:
  + Problemas médicos donde las clases minoritarias (como enfermedades raras) son tan importantes como las mayoritarias (como no tener la enfermedad).

### 3. Weighted-average

Cómo funciona: Calcula métricas para cada clase y pondera su contribución al promedio según el número de muestras de cada clase.

* Uso recomendado:
  + Cuando tienes un conjunto de datos desbalanceado, pero quieres tener en cuenta la proporción de cada clase.
  + Útil si las clases más frecuentes deben tener mayor peso en la evaluación general.
* Ventajas:
  + Considera tanto el rendimiento general como la distribución de datos entre clases.
  + Es una alternativa intermedia entre micro y macro cuando hay desbalance.
* Ejemplo de uso:
  + Evaluar un modelo en un dataset desbalanceado para predecir clientes leales frente a no leales, donde los clientes leales son minoría pero económicamente relevantes.

### 4. Otras estrategias

* Per-class (clase por clase):
  + Métricas calculadas por separado para cada clase.
  + Útil cuando quieres examinar específicamente el rendimiento en clases clave.
* Subset accuracy (exact match en problemas multi-etiqueta):
  + Evalúa si el conjunto completo de predicciones coincide exactamente con las etiquetas reales.
  + Muy estricto, se usa poco en la práctica, excepto para aplicaciones de sistemas críticos.

### Resumen de cuándo usar cada una:

| Estrategia | ¿Clases balanceadas? | Igual importancia para todas las clases | Clases desbalanceadas y peso proporcional | Foco en clases minoritarias |
| --- | --- | --- | --- | --- |
| Micro-average | ✅ | ✅ | ❌ | ❌ |
| Macro-average | ✅/❌ | ✅ | ❌ | ✅ |
| Weighted | ❌ | ❌ | ✅ | ✅ |
| Per-class | ✅/❌ | ✅ (por clase específica) | N/A | N/A |

En general, elegir la estrategia adecuada depende de la distribución de datos y los objetivos del proyecto.

ChatGPT puede cometer errores. Considera verificar la información importante.