UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Tecnologías de la información

Extracción de Conocimiento en Bases de Datos

Análisis Supervisado

IDGS91N

PROFESOR:
Enrique Mascote

Alumno:
Emanuel Chavira

28 de Noviembre de 2025

Introducción

En la actualidad, muchas organizaciones utilizan **aprendizaje supervisado** para predecir comportamientos y tomar decisiones informadas. Los algoritmos de aprendizaje supervisado se dividen en dos grandes categorías: modelos de **regresión**, que estiman valores continuos (como precios o ventas), y modelos de **clasificación**, que predicen categorías discretas (por ejemplo, si un cliente comprará o no). El objetivo de este documento es investigar algunos de los algoritmos de regresión y clasificación más utilizados, comprender su funcionamiento, conocer sus métricas de evaluación y estudiar sus fortalezas y limitaciones. Posteriormente se presenta un caso práctico que ilustra la elección e implementación de un algoritmo en un problema real.

Investigación de algoritmos

Algoritmos de regresión

Regresión lineal

Qué resuelve. La regresión lineal es uno de los algoritmos más sencillos de aprendizaje supervisado. Su objetivo es modelar una relación lineal entre una o más variables independientes y una variable dependiente continua. El modelo ajusta una recta (en el caso univariante) o un hiperplano (en el caso multivariante) que minimiza la suma de los errores cuadrados entre las predicciones y los valores reales.

Principio de funcionamiento. Supone que existe una relación lineal entre las variables predictoras y la respuesta, representada por la ecuación ($y = m\,x + b$). Para encontrar la recta de mejor ajuste se utiliza el método de *mínimos cuadrados*, que minimiza la suma de los residuos (diferencia entre valor real y valor predicho) al cuadrado【884934342811455†L150-L178】. El algoritmo estima los coeficientes (m) e (b) que optimizan esta función de coste.

Métricas de evaluación. Las métricas más empleadas en regresión lineal incluyen:

- **Error absoluto medio (MAE):** promedia el valor absoluto de las diferencias entre los valores reales y las predicciones. Su ventaja es que tiene la misma unidad que la variable dependiente.
- **Error cuadrático medio (MSE):** calcula el promedio de los cuadrados de los errores. Es una métrica diferenciable que penaliza de forma más severa los errores grandes, aunque su unidad es el cuadrado de la variable de salida.

- **Raíz del error cuadrático medio (RMSE):** es la raíz cuadrada del MSE, de modo que la métrica vuelve a tener la misma unidad que la variable predicha.
- **Coeficiente de determinación ((R^2)):** indica qué proporción de la varianza de la variable dependiente es explicada por el modelo; un (R^2) cercano a 1 sugiere un buen ajuste.

Fortalezas. La regresión lineal es un algoritmo fácil de comprender e implementar. Sus coeficientes pueden interpretarse como el cambio en la variable dependiente asociado a una unidad de cambio en la variable independiente. Además, se entrena rápidamente y suele emplearse como *modelo de referencia* para comparar con métodos más complejos.

Limitaciones. Asume que la relación entre las variables es lineal; si esta condición no se cumple, el modelo presenta un desempeño deficiente. Es sensible a la multicolinealidad entre variables predictoras y puede sufrir tanto *sobreajuste* como *subajuste*. Su capacidad para capturar relaciones complejas es limitada.

Regresión de bosques aleatorios (Random Forest)

Qué resuelve. El algoritmo de *Random Forest* es una técnica de **aprendizaje por conjunto** que combina numerosos árboles de decisión para mejorar la capacidad predictiva. Puede utilizarse tanto en problemas de clasificación como de regresión. En el caso de regresión, el modelo construye varios árboles a partir de distintas muestras de bootstrap y promedia sus predicciones para estimar un valor continuo.

Principio de funcionamiento. Random Forest genera diversas muestras de entrenamiento seleccionando aleatoriamente observaciones con reemplazo (bootstrap). Para cada muestra construye un árbol de decisión; en cada nodo se evalúa un subconjunto aleatorio de características para determinar la mejor división, lo que introduce diversidad entre los árboles. Una vez entrenados los árboles, las predicciones se combinan mediante voto mayoritario (clasificación) o promedio (regresión).

Métricas de evaluación. Al ser un modelo de regresión, se utilizan las mismas métricas que en la regresión lineal: MAE, MSE, RMSE y (R^2). Estas métricas cuantifican el error medio, el error cuadrático medio, su raíz y la proporción de varianza explicada, respectivamente.

Fortalezas. Random Forest ofrece elevada precisión al reducir la varianza de cada árbol individual mediante el promedio de múltiples árboles. Es robusto frente al ruido y a los valores atípicos, no asume una forma funcional específica de los datos (es no paramétrico) y permite estimar la importancia de las características. Puede manejar valores faltantes y variables numéricas o categóricas sin necesidad de escalado.

Limitaciones. Su mayor desventaja es el coste computacional: entrenar muchos árboles y promediarlos requiere considerable tiempo y memoria, y las predicciones pueden ser lentas. El modelo es menos interpretable que un árbol individual porque la combinación de árboles funciona como una "caja negra". También puede sobreajustar si se construyen demasiados árboles o si estos son muy profundos.

Algoritmos de clasificación

Regresión logística

Qué resuelve. La regresión logística es un modelo estadístico utilizado para **clasificación binaria**. Predice la probabilidad de que ocurra un evento (clase 1) frente a la alternativa (clase 0) a partir de variables independientes. Utiliza una función logística (sigmoide) para transformar la combinación lineal de las características en una probabilidad.

Principio de funcionamiento. El modelo calcula la probabilidad (P(Y=1) =). La regresión logística se estima mediante **máxima verosimilitud**, buscando los parámetros () que maximizan la probabilidad de observar los datos dados los parámetros. Para asignar clases, se compara la probabilidad con un umbral (p. ej., 0,5) y se clasifica en la clase correspondiente. Existen extensiones para casos multiclase (regresión logística multinomial y ordinal).

Métricas de evaluación. Las métricas de clasificación más comunes son:

- **Precisión global (accuracy):** proporción de predicciones correctas entre todas las predicciones. Puede ser engañosa en conjuntos desequilibrados.
- **Precisión (precision):** mide cuántas de las predicciones positivas son realmente positivas.
- **Exhaustividad o sensibilidad (recall):** indica qué proporción de casos positivos fue detectada por el modelo.
- **Puntaje F1:** media armónica entre precisión y exhaustividad; equilibra ambas métricas.

También pueden utilizarse métricas como AUC-ROC o log-loss para problemas binarios o multiclase.

Fortalezas. La regresión logística es fácil de implementar, interpretar y entrenar. Sus coeficientes permiten conocer la dirección y magnitud de la asociación entre las características y la probabilidad de pertenecer a una clase. No requiere supuestos de distribución sobre las clases y se puede extender a problemas multiclase. El modelo suele ser rápido y sirve como base para modelos más complejos.

Limitaciones. Supone una relación lineal entre las variables explicativas y el logaritmo de las probabilidades (*log-odds*). Si esta relación no existe o si las clases no son linealmente separables, el

modelo pierde rendimiento. No puede capturar relaciones no lineales sin añadir transformaciones o funciones de base. Además, es propenso al sobreajuste cuando el número de predictores supera al número de observaciones y es sensible a la multicolinealidad.

Árbol de decisión (clasificación)

Qué resuelve. Un árbol de decisión clasifica instancias dividiendo recursivamente el espacio de características en regiones más pequeñas en función de preguntas binarias. Cada nodo interno representa una decisión basada en una característica, y cada hoja representa una clase. Los árboles de decisión pueden utilizarse para clasificación (categorías) o regresión (valores continuos).

Principio de funcionamiento. El proceso comienza en un nodo raíz que representa todo el conjunto de datos. En cada nodo se selecciona la característica y el umbral que mejor dividen los datos de acuerdo con un criterio como la **impureza de Gini** o la **entropía**. Las ramas conducen a nuevos nodos donde el proceso se repite hasta que se alcanza un criterio de parada (por ejemplo, todas las instancias pertenecen a la misma clase o se ha alcanzado una profundidad máxima). Las hojas contienen la clase predicha.

Métricas de evaluación. Al igual que en la regresión logística, los árboles de decisión se evalúan mediante precisión, precisión por clase, recall y F1. Para comparar modelos también se utilizan la matriz de confusión o el área bajo la curva ROC.

Fortalezas. Los árboles de decisión son fáciles de entender y visualizar: muestran de forma explícita las reglas de decisión. Son versátiles; pueden manejar datos numéricos y categóricos y no requieren escalado de variables. Capturan relaciones no lineales y pueden tratar valores faltantes mediante diversas estrategias.

Limitaciones. Su principal desventaja es la tendencia a **sobreajustar**: árboles muy profundos memorizan el conjunto de entrenamiento y pierden capacidad de generalización. Además, son inestables (pequeños cambios en los datos generan árboles muy diferentes). Pueden sesgarse hacia variables con muchos niveles y no siempre capturan interacciones complejas. La construcción y poda de árboles grandes puede resultar costosa computacionalmente.

Caso de estudio y justificación

Definición del problema

Para ilustrar el uso de un algoritmo de clasificación, se planteó predecir si un tumor es **maligno** o **benigno** a partir de mediciones obtenidas en imágenes digitales de biopsias. El objetivo es apoyar a los médicos en el diagnóstico temprano de cáncer de mama. El conjunto de datos **Breast Cancer Wisconsin**, disponible en la biblioteca scikit-learn, contiene 569 observaciones y 30 características numéricas, como el radio medio, la textura y la simetría de la célula.

Justificación del algoritmo elegido

El problema es binario (maligno vs. benigno) y las características tienen relaciones aproximadamente lineales con la probabilidad de malignidad. Por ello se seleccionó la **regresión logística**. Este algoritmo permite estimar probabilidades de manera directa e interpretar el impacto de cada característica en la probabilidad de cáncer. Además, su entrenamiento es rápido y sus resultados sirven como punto de referencia para comparar con algoritmos más complejos.

Diseño e implementación del modelo

Variables y estructura de datos

Las variables de entrada consisten en 30 mediciones obtenidas a partir de la imagen de la biopsia, entre ellas el radio, la textura, el perímetro, el área y la compacidad de la célula. La variable objetivo es binaria (1 = maligno, 0 = benigno). Los datos se dividen en conjunto de **entrenamiento** (80 %) y **prueba** (20 %) utilizando muestreo estratificado para preservar la proporción de clases.

La regresión logística requiere normalizar o estandarizar las características para evitar que los coeficientes estén sesgados por diferencias de escala. Por ello se aplicó un **escalador estándar** (*StandardScaler*) antes de entrenar el modelo. El flujo general se muestra en la Figura 1.

1. **Carga y exploración de datos.** Se obtiene el conjunto de datos desde sklearn.datasets.
2. **Preprocesamiento.** Se divide en entrenamiento y prueba y se estandarizan las variables.
3. **Entrenamiento.** Se ajusta un modelo de regresión logística con regularización por defecto (solucionador *lbfgs* y límite de 10 000 iteraciones).
4. **Predicción y evaluación.** Se calculan las métricas accuracy, precision, recall y F1 sobre el conjunto de prueba.

Código de implementación (Python)

El siguiente fragmento de código implementa el flujo descrito usando scikit-learn. Incluye la preparación de los datos, el entrenamiento del modelo y el cálculo de las métricas.

```
import pandas as pd
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# Cargar el conjunto de datos
data = load_breast_cancer()
X = pd.DataFrame(data.data, columns=data.feature_names)
y = pd.Series(data.target)

# División entrenamiento/prueba con estratificación
X_train, X_test, y_train, y_test = train_test_split(
   X, y, test_size=0.20, random_state=42, stratify=y)

# Escalado de características
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Entrenamiento de la regresión logística
log_reg = LogisticRegression(max_iter=10000, solver='lbfgs')
log_reg.fit(X_train_scaled, y_train)

# Predicciones y métricas
y_pred = log_reg.predict(X_test_scaled)
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
print(f"Accuracy: {accuracy:.3f}")
print(f"Precision: {precision:.3f}")
print(f"Recall: {recall:.3f}")
print(f"F1 Score: {f1:.3f}")
```

Resultados y evaluación

El modelo de regresión logística entrenado obtuvo los siguientes valores en el conjunto de prueba:

Métrica	Valor
Accuracy	0.982
Precision	0.986
Recall	0.986
F1 Score	0.986

Los resultados indican que el modelo clasifica correctamente el 98,2 % de los casos. La precisión y el recall son elevados (≈ 0,986), lo que significa que el modelo detecta la mayoría de los tumores malignos con pocos falsos positivos. El puntaje F1 confirma el equilibrio entre precisión y recall. La matriz de confusión muestra una sola clasificación errónea para cada clase (41 benignos y 71 malignos predichos correctamente), lo que respalda la robustez del modelo.

Análisis y posibles mejoras

Aunque el desempeño es alto, es posible explorar mejoras:

- **Regularización y ajuste de hiperparámetros.** Ajustar el parámetro de regularización (C) mediante validación cruzada para evitar sobreajuste o bajoajuste.
- **Selección de características.** Evaluar la importancia de las variables y eliminar atributos redundantes para simplificar el modelo y reducir la multicolinealidad.
- **Modelos alternativos.** Probar algoritmos más complejos como bosques aleatorios o máquinas de soporte vectorial, que pueden capturar relaciones no lineales, y comparar sus métricas.

Conclusiones y recomendaciones

Este estudio comparó algoritmos de regresión y clasificación destacados en aprendizaje supervisado. La **regresión lineal** constituye una herramienta básica para aproximar relaciones lineales y sirve como referencia; no obstante, presenta limitaciones cuando las relaciones son no lineales o existen colinealidades. El **bosque aleatorio** supera algunas de estas limitaciones al agregar múltiples árboles, proporcionando mayor precisión y robustez a costa de interpretabilidad y requerimientos computacionales.

En clasificación se revisaron la **regresión logística** y los **árboles de decisión**. La regresión logística destaca por su simplicidad, rapidez e interpretabilidad, aunque requiere una relación lineal entre las variables explicativas y el log-odds y no maneja bien patrones no lineales. Los árboles de decisión son intuitivos y capturan relaciones no lineales, pero pueden sobreajustar y ser inestables.

El caso de estudio demostró que la regresión logística clasifica tumores con gran exactitud (F1 ≈ 0,986), lo que la hace adecuada como modelo inicial para problemas de diagnóstico médico. Se recomienda complementarla con técnicas de validación cruzada y selección de características para mejorar la generalización. Para problemas con relaciones complejas o conjuntos de datos grandes, se sugiere investigar modelos basados en árboles o redes neuronales, evaluando cuidadosamente el equilibrio entre precisión e interpretabilidad.

Referencias

1. Gupta, M. (2025). *Linear Regression in Machine Learning*. GeeksforGeeks. Recuperado de https://www.geeksforgeeks.org/machine-learning/ml-linear-regression/
2. Amiya Ranjan Rout. (2025). *Advantages and Disadvantages of Logistic Regression*. GeeksforGeeks. Recuperado de https://www.geeksforgeeks.org/data-science/advantages-and-disadvantages-of-logistic-regression/
3. Saloni 1297. (2025). *Decision Tree*. GeeksforGeeks. Recuperado de https://www.geeksforgeeks.org/machine-learning/decision-tree/
4. Analytics Vidhya. (2021). *Know the Best Evaluation Metrics for Your Regression Model*. Recuperado de https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model
5. GeeksforGeeks. (2025). *Evaluation Metrics in Machine Learning*. Recuperado de https://www.geeksforgeeks.org/machine-learning/metrics-for-machine-learning-model/
6. Ravinder Kamat. (2025). *What are the Advantages and Disadvantages of Random Forest?*. GeeksforGeeks. Recuperado de https://www.geeksforgeeks.org/machine-learning/what-are-the-advantages-and-disadvantages-of-random-forest