COLOMBIA POTENCIA DE LA VIDA

TIC

# UNIVERSIDAD SERGIO ARBOLEDA

Análisis de Datos
Nivel Integrador

TALENTO TECH

UNIVERSIDAD SERGIO ARBOLEDA

# PREPARACIÓN SEMANA 13

# REGRESIÓN AVANZADA

La regresión avanzada surge como una extensión de la regresión lineal ordinaria cuando nos enfrentamos a problemas más complejos o desafiantes. La regresión lineal ordinaria asume que la relación entre las variables independientes y la variable dependiente es lineal, lo cual puede ser limitante en situaciones donde la relación es más compleja o presenta características especiales.

**Necesidad de Regresión Avanzada:**

1. **Superación de Limitaciones Lineales:** En ocasiones, la relación entre las variables puede ser no lineal. Los modelos de regresión avanzada permiten capturar patrones más complejos y no lineales en los datos.
2. **Manejo de Multicolinealidad:** Cuando las variables independientes están altamente correlacionadas, la regresión lineal ordinaria puede verse afectada por la multicolinealidad. Los métodos de regresión avanzada, como Ridge y LASSO, ofrecen soluciones a este problema.
3. **Regularización para Evitar Overfitting:** En presencia de un gran número de características, los modelos de regresión pueden volverse propensos al sobreajuste (overfitting). Métodos como Ridge y LASSO introducen términos de regularización que ayudan a prevenir el sobreajuste.

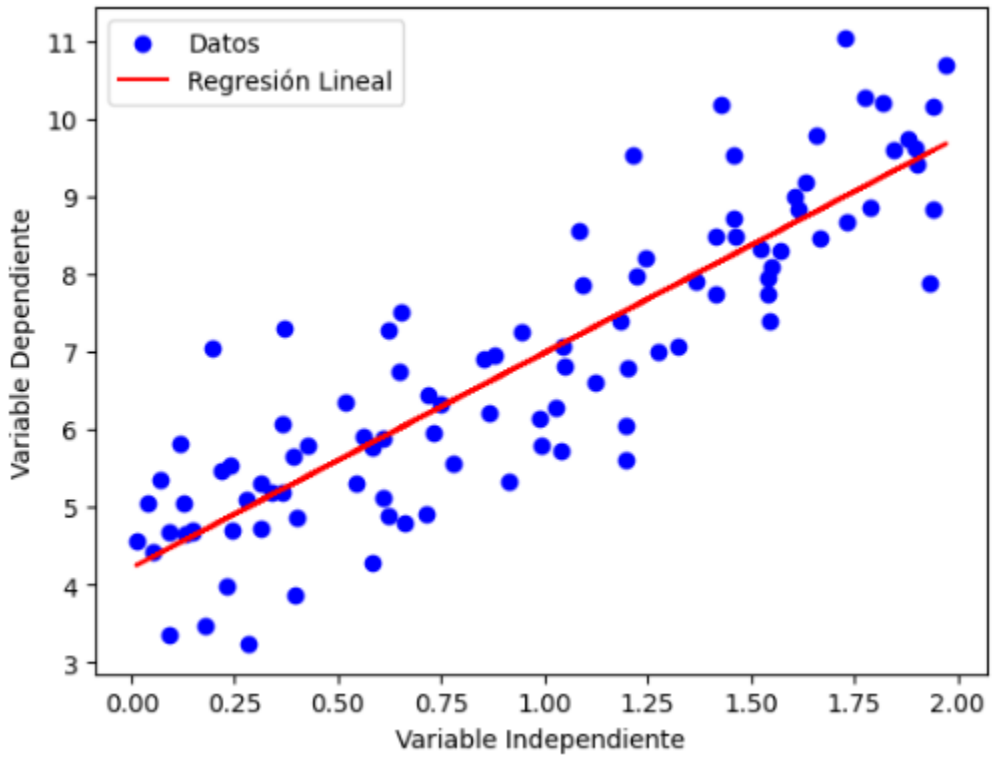
# REGRESIÓN AVANZADA

**Breve Revisión de la Regresión Lineal Ordinaria:** La regresión lineal ordinaria busca encontrar la línea (o hiperplano) que mejor se ajuste a los datos minimizando la suma de los cuadrados de las diferencias entre las observaciones reales y las predicciones del modelo.

En este ejemplo, la línea roja representa la regresión lineal ordinaria que busca minimizar la distancia vertical entre los puntos de datos y la línea. La regresión avanzada, como Ridge y LASSO, se introduce para abordar limitaciones y mejorar la capacidad del modelo para adaptarse a situaciones más complejas.

El código de los ejemplos proporcionados acá, podemos encontrarlo en el enlace: https://acortar.link/tx0ztf

# REGRESIÓN AVANZADA



# CONCEPTOS BÁSICOS DE REGULARIZACIÓN

La regularización s una técnica utilizada en modelos de regresión para evitar el sobreajuste (overfitting) al penalizar la complejidad del modelo. Los métodos de regularización introducen términos adicionales en la función de pérdida, que castigan a los modelos por tener coeficientes muy grandes. Dos enfoques comunes de regularización en regresión son Ridge (regresión Ridge) y LASSO (Least Absolute Shrinkage and Selection Operator).

- En el contexto de la regresión, la regularización agrega un término de penalización a la función de pérdida del modelo. Este término de penalización está relacionado con la magnitud de los coeficientes del modelo.
- La idea principal es evitar que los coeficientes tomen valores extremadamente grandes, lo que podría llevar a un sobreajuste del modelo.

# CONCEPTOS BÁSICOS DE REGULARIZACIÓN

**Penalización en Ridge y LASSO:**

1. **Ridge (Regresión Ridge):**
   - Ridge introduce una penalización L2, que es proporcional al cuadrado de los valores absolutos de los coeficientes. La función de pérdida para Ridge es la suma de los errores cuadráticos ordinarios y el término de penalización L2.

$$J(\theta) = \text{MSE}(\theta) + \alpha \sum_{i=1}^{n} \theta_i^2$$

   - $\alpha$ es el parámetro de regularización, que controla la fuerza de la penalización.

# CONCEPTOS BÁSICOS DE REGULARIZACIÓN

**Penalización en Ridge y LASSO:**

**2. LASSO (Least Absolute Shrinkage and Selection Operator):**

- LASSO introduce una penalización L1, que es proporcional al valor absoluto de los coeficientes. La función de pérdida para LASSO es la suma de los errores cuadráticos ordinarios y el término de penalización L1.

$$J(\theta) = \text{MSE}(\theta) + \alpha \sum_{i=1}^{n} |\theta_i|$$

- • *α es nuevamente el parámetro de regularización.*

# CONCEPTOS BÁSICOS DE REGULARIZACIÓN

**Ventajas y Desventajas de Ridge y LASSO:**

1. **Ridge:**
    - **Ventajas:**
        - Útil cuando hay multicolinealidad en los datos.
        - Conserva todas las características, incluso si algunas tienen una contribución pequeña.
    - **Desventajas:**
        - No realiza selección de características; todos los coeficientes son diferentes de cero.

# CONCEPTOS BÁSICOS DE REGULARIZACIÓN

2. **LASSO:**
    - **Ventajas:**
        - Realiza selección de características; algunos coeficientes pueden ser exactamente cero.
        - Útil para conjuntos de datos con muchas características irrelevantes.
    - **Desventajas:**
        - Puede tener dificultades con multicolinealidad.
        - No conserva todas las características.

Tanto Ridge y LASSO son técnicas valiosas para controlar la complejidad del modelo y mejorar la generalización en casos donde la regresión lineal ordinaria podría ser insuficiente. La elección entre Ridge y LASSO dependerá del problema específico y de las características de los datos.

# IMPLEMENTACIÓN PRÁCTICA

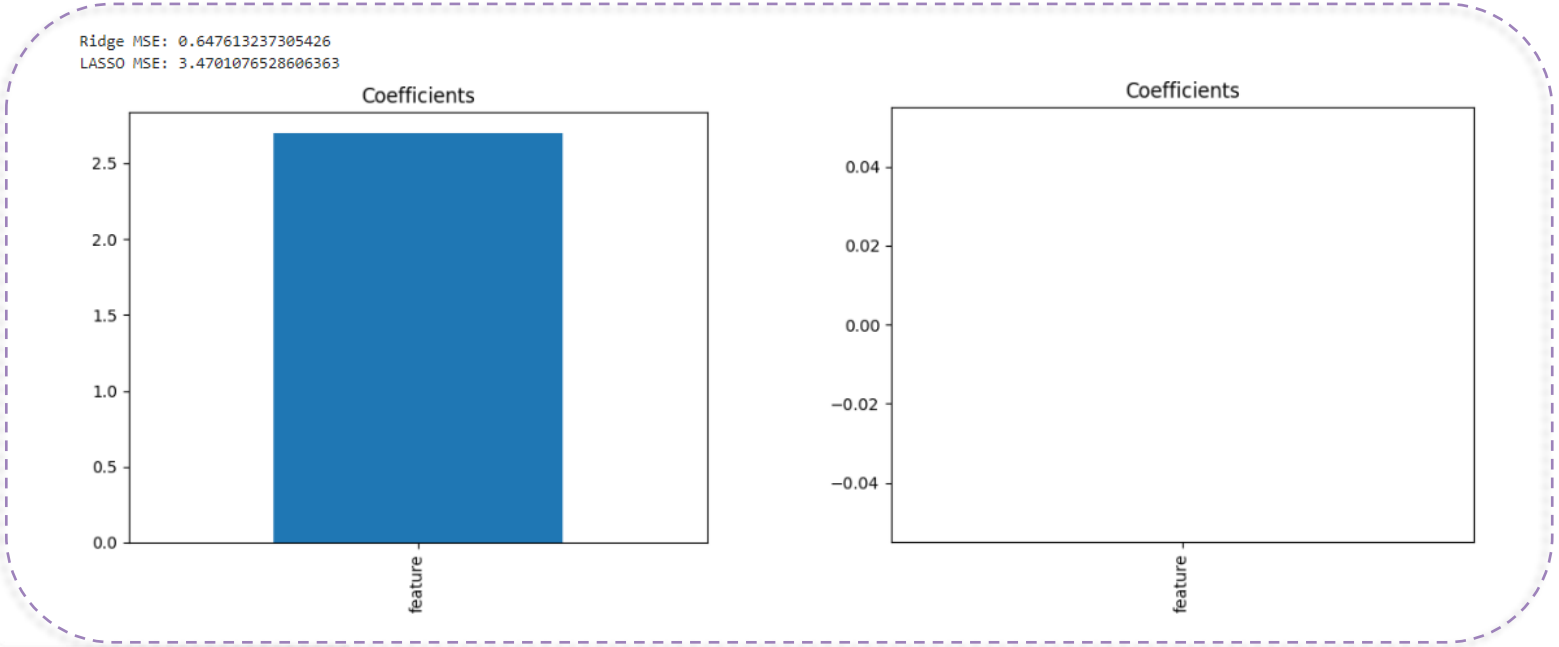
Veamos paso a paso

Primero, importamos las bibliotecas necesarias y generamos datos simulados para trabajar.

Luego, implementamos modelos de Ridge y LASSO y ajustamos los hiperparámetros.

Podemos evaluar el rendimiento de los modelos utilizando métricas como el error cuadrático medio (MSE).

Finalmente, podemos visualizar cómo los modelos afectan los coeficientes.



# SELECCIÓN DE HIPERPARÁMETROS EN RIDGE Y LASSO

La selección adecuada de hiperparámetros, como el parámetro de regularización ($\alpha$) en Ridge y LASSO, es crucial para obtener modelos de regresión robustos y generalizables. Aquí, discutiremos métodos comunes para encontrar el mejor valor de alpha, incluyendo la validación cruzada.

1. **Búsqueda Manual:**
    - Seleccionar varios valores de alpha manualmente.
    - Entrenar modelos con cada valor de alpha.
    - Evaluar el rendimiento del modelo en un conjunto de validación.
    - Elegir el alpha que proporciona el mejor rendimiento.

```
alphas = [0.1, 1.0, 10.0]
for alpha in alphas:
    ridge_model = Ridge(alpha=alpha)
    ridge_model.fit(X_train, y_train)
    mse = evaluate_model(ridge_model, X_val, y_val)
    print(f'Ridge MSE for alpha={alpha}: {mse}')
```

# SELECCIÓN DE HIPERPARÁMETROS EN RIDGE Y LASSO

2. **Búsqueda Grid:**
   - Definir un rango de valores para alpha.
   - Realizar una búsqueda exhaustiva (grid search) entrenando modelos con cada combinación de valores de hiperparámetros.
   - Evaluar el rendimiento de cada modelo y seleccionar el mejor.

```python
from sklearn.model_selection import GridSearchCV

param_grid = {'alpha': [0.1, 1.0, 10.0]}
ridge_model = Ridge()
grid_search = GridSearchCV(ridge_model, param_grid, cv=5, scoring='neg_mean_squared_error')
grid_search.fit(X_train, y_train)

best_alpha = grid_search.best_params_['alpha']
print(f'Best alpha for Ridge: {best_alpha}')
```

# SELECCIÓN DE HIPERPARÁMETROS EN RIDGE Y LASSO

3. **Validación Cruzada (Cross-validation):**
   - Dividir el conjunto de entrenamiento en varios pliegues (folds).
   - Entrenar el modelo en k-1 pliegues y evaluar en el pliegue restante, repitiendo esto k veces.
   - Calcular el rendimiento promedio del modelo para cada valor de alpha.

```python
from sklearn.model_selection import cross_val_score

alphas = [0.1, 1.0, 10.0]
for alpha in alphas:
    ridge_model = Ridge(alpha=alpha)
    scores = cross_val_score(ridge_model, X_train, y_train, scoring='neg_mean_squared_error', cv=5)
    mse_mean = -scores.mean()
    print(f'Cross-validated Ridge MSE for alpha={alpha}: {mse_mean}')
```

# MÁQUINAS DE SOPORTE VECTORIAL (SVM)

Las Máquinas de Soporte Vectorial (SVM) son un poderoso método de aprendizaje supervisado utilizado para la clasificación y la regresión. Enfocándonos en la clasificación, veamos los fundamentos de SVM y cómo encuentra el hiperplano óptimo.

1. **Hiperplano y Márgenes:**
   - El hiperplano es una superficie de decisión que separa las clases en el espacio de características.
   - Los márgenes son las distancias entre el hiperplano y los puntos más cercanos de cada clase. SVM busca maximizar estos márgenes.
2. **Vectores de Soporte:**
   - Son los puntos de datos más cercanos al hiperplano y son cruciales para determinar la posición y orientación del hiperplano.
   - SVM se centra en estos vectores de soporte para optimizar la separación entre clases.

# MÁQUINAS DE SOPORTE VECTORIAL (SVM)

3. **Función de Decisión y Margen:**
   - La función de decisión asigna puntos a una clase en función de su posición con respecto al hiperplano.
   - El margen es la distancia entre la función de decisión y el hiperplano.
4. **Optimización del Hiperplano:**
   - SVM busca encontrar el hiperplano que maximice los márgenes.
   - La función de optimización involucra minimizar la norma del vector de pesos (coeficientes del hiperplano) sujeto a la condición de que todos los puntos estén del lado correcto del hiperplano.

# MÁQUINAS DE SOPORTE VECTORIAL (SVM)

A continuación, un ejemplo práctico de cómo implementar SVM para clasificación en Python con scikit-learn:

```python
from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score

# Cargar un conjunto de datos de ejemplo (por ejemplo, Iris)
iris = datasets.load_iris()
X_train, X_test, y_train, y_test = train_test_split(iris.data, iris.target, test_size=0.2, random_state=42)

# Crear un clasificador SVM
svm_classifier = SVC(kernel='linear')
svm_classifier.fit(X_train, y_train)

# Predecir en el conjunto de prueba
y_pred = svm_classifier.predict(X_test)

# Evaluar la precisión del modelo
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy of SVM: {accuracy}')
```

```
Accuracy of SVM: 1.0
```

# COMPARACIÓN DE MODELOS

Al comprar los modelos SMV, Árboles de decisión y bosques aleatorios es crucial entender las fortalezas y debilidades de diferentes modelos para elegir el más adecuado para un escenario particular. Aquí, compararemos SVM, árboles de decisión y bosques aleatorios:

1. **Support Vector Machines (SVM):**
    - *Fortalezas:*
        - Efectivo en espacios de alta dimensión y especialmente útil cuando el número de dimensiones es mayor que el número de muestras.
        - Puede manejar eficientemente conjuntos de datos no lineales mediante el uso de kernels.
        - Buen rendimiento cuando hay una clara separación entre las clases.
    - *Debilidades:*
        - Puede ser computacionalmente costoso en conjuntos de datos grandes.
        - Sensible a la elección del kernel y a los hiperparámetros.

# COMPARACIÓN DE MODELOS

3. **Bosques Aleatorios:**
    - *Fortalezas:*
        - Reducción del sobreajuste al combinar múltiples árboles de decisión.
        - Manejo eficaz de conjuntos de datos grandes y de alta dimensión.
        - Proporciona importancia de características.
    - *Debilidades:*
        - Menos interpretable que un solo árbol de decisión.
        - Puede ser computacionalmente costoso en comparación con un solo árbol.

# COMPARACIÓN DE MODELOS

**Criterios para la Selección del Modelo:**

- **Tamaño del Conjunto de Datos:**
  - SVM puede ser eficiente en conjuntos de datos pequeños a medianos.
  - Bosques aleatorios son adecuados para conjuntos de datos grandes y complejos.
  - Árboles de decisión son rápidos y efectivos en conjuntos de datos pequeños.
- **Interpretación:**
  - Si la interpretación del modelo es crucial, los árboles de decisión proporcionan una visión clara.
  - Si la interpretación es menos importante y se prioriza el rendimiento, SVM o bosques aleatorios podrían ser opciones.

# COMPARACIÓN DE MODELOS

- **Relación entre Características y Resultados:**
  - SVM es efectivo cuando la relación entre características y resultados es compleja y no lineal.
  - Árboles de decisión son útiles cuando la relación es más simple y fácilmente interpretable.
  - Bosques aleatorios son una buena elección para equilibrar complejidad y rendimiento.

En última instancia, la elección entre SVM, árboles de decisión y bosques aleatorios dependerá del problema específico y de las características de los datos. Experimentar con diferentes modelos y evaluar su rendimiento en conjuntos de validación o mediante técnicas de validación cruzada es fundamental para tomar decisiones informadas.

# ¿Preguntas?