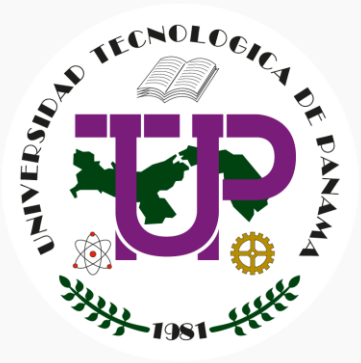


Universidad Tecnológica de Panamá
Facultad de Ingeniería Eléctrica
Maestría en Ingeniería Eléctrica

Semana 2

Introducción a Machine Learning

Ejemplo (Pima Diabetes Dataset)

Variable	Muestra 1	Muestra 2
Pregnancies	1	5
Glucose	189	166
Blood Pressure	60	72
Skin Thickness	23	19
Insulin	846	175
BMI	30.1	25.8
Diabetes Pedigree	0.398	0.587
Age	59	51

Ejemplo

Distancia Euclidiana

Modelo matemático:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Aplicación paso a paso:

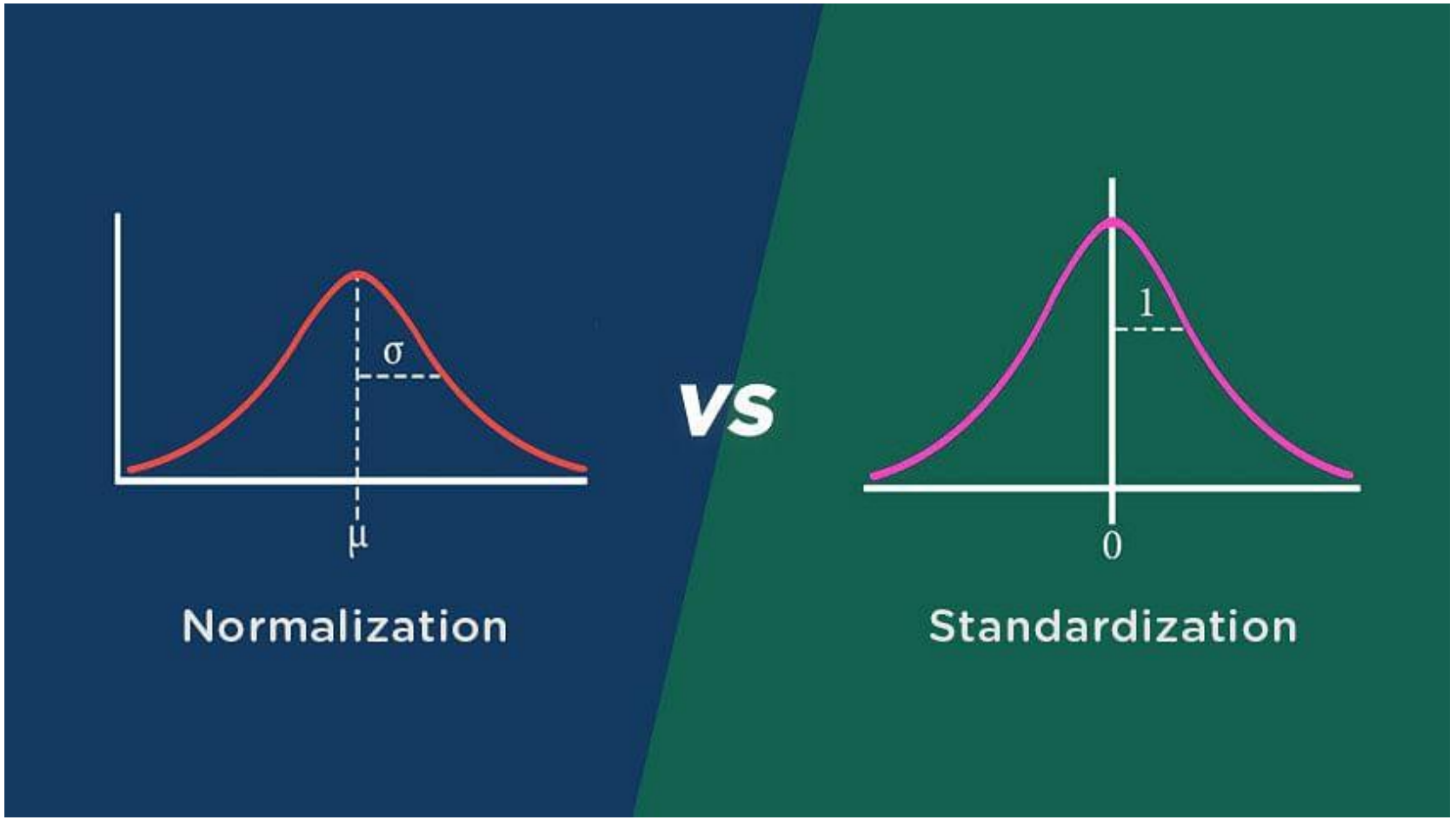
$$d(\text{Muestra 1}, \text{Muestra 2}) = \sqrt{(1-5)^2 + (189-166)^2 + (60-72)^2 + (23-19)^2 +}$$

$$+ (846-175)^2 + (30.1-25.8)^2 + (0.398-0.587)^2 + (59-51)^2$$

$$= \sqrt{16 + 529 + 144 + 16 + 44944 + 18.49 + 0.035721 + 64}$$

$$= \sqrt{45741.525721} \approx 213.98$$

Técnicas de pre-procesamiento de datos



Normalización

La **normalización de datos** es una técnica de **preprocesamiento** que consiste en transformar los valores de una o más variables para que se encuentren dentro de una **escala común**, sin distorsionar sus relaciones originales.

Definición formal de Normalización

Normalización es el proceso de **reescalar los datos a un rango específico**, típicamente entre **0 y 1**, usando una transformación matemática.

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Donde:

- x es un valor original.
- $x_{\min}$ y $x_{\max}$ son el valor mínimo y máximo de la variable.
- x_{norm} es el valor reescalado (normalizado) entre 0 y 1.

¿Por qué se normalizan los datos?

- Evita que variables con escalas grandes dominen el modelo.
- Mejora el rendimiento de algoritmos sensibles a la escala, como:
  - K-Nearest Neighbors (KNN)
  - Redes neuronales
  - SVMs
  - Regresión logística (cuando hay regularización)
- Acelera la convergencia de los algoritmos de optimización.

Definición formal de Estandarización

Estandarización es el proceso de **transformar los datos para que tengan media 0 y desviación estándar 1**, utilizando una transformación basada en la distribución de los datos. Esta técnica es útil para comparar variables que originalmente tienen escalas diferentes.

$$x_{\text{std}} = \frac{x - \mu}{\sigma}$$

Donde:

- x es un valor original.
- μ es la media de la variable.
- σ es la desviación estándar de la variable.
- x_{std} es el valor **estandarizado**, es decir, un valor reescalado que indica cuántas desviaciones estándar se aleja x de la media.

¿Por qué se estandarizan los datos?

- Permite comparar variables con escalas y unidades diferentes en un marco común.
- Mejora el comportamiento de modelos que asumen distribución normal o centrada, como:
  - Análisis de Componentes Principales (PCA)
  - Regresión lineal y logística con regularización
  - SVMs con kernel
  - K-Means clustering
- Reduce la influencia de variables con alta varianza sobre los modelos.
- Facilita la interpretación en términos de desviaciones estándar.
- Acelera la convergencia de algoritmos de optimización, como el descenso de gradiente.

Comparación de normalización vs estandarización

Acceder al ejemplo

Correlación de Pearson

- La correlación de Pearson es un coeficiente estadístico que cuantifica el grado de relación lineal entre dos variables numéricas.
- Se utiliza para determinar si existe una asociación y qué tan fuerte es esa asociación.

Correlación de Pearson

- Permite comprender y analizar las relaciones entre variables numéricas, facilitando la identificación de patrones o tendencias en los datos.
- La correlación de Pearson mide tanto la fuerza (qué tan estrechamente están relacionadas las variables) como la dirección (si la relación es positiva o negativa) de la relación lineal entre dos variables cuantitativas.

Correlación de Pearson

- Es útil en diversas áreas como la ciencia, economía y salud para explorar asociaciones entre variables y apoyar la toma de decisiones basada en datos.
- Un valor alto (cercano a 1 o -1) indica una relación lineal fuerte, mientras que un valor cercano a 0 sugiere poca o ninguna relación lineal.

Modelo Matemático de la Correlación de Pearson

La **correlación de Pearson** se expresa mediante el siguiente modelo matemático:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Donde:

- r: Coeficiente de correlación de Pearson.
- x_i, y_i: Valores individuales de las variables X e Y.
- $\bar{x}, \bar{y}$: Promedios de las variables X e Y.

Interpretación del Modelo

- **Numerador:** Representa la covarianza entre las dos variables, que mide cómo varían juntas.
- **Denominador:** Es el producto de las desviaciones estándar de las dos variables, que normaliza la covarianza para que el coeficiente esté en el rango de -1 a 1.

Valores de la Correlación de Pearson

Su valor varía de -1 a 1
1 indica una correlación positiva perfecta.
-1 indica una correlación negativa perfecta.
0 indica que no hay correlación lineal.

Nota Importante

Es importante tener en cuenta que **una alta correlación no implica causalidad**.

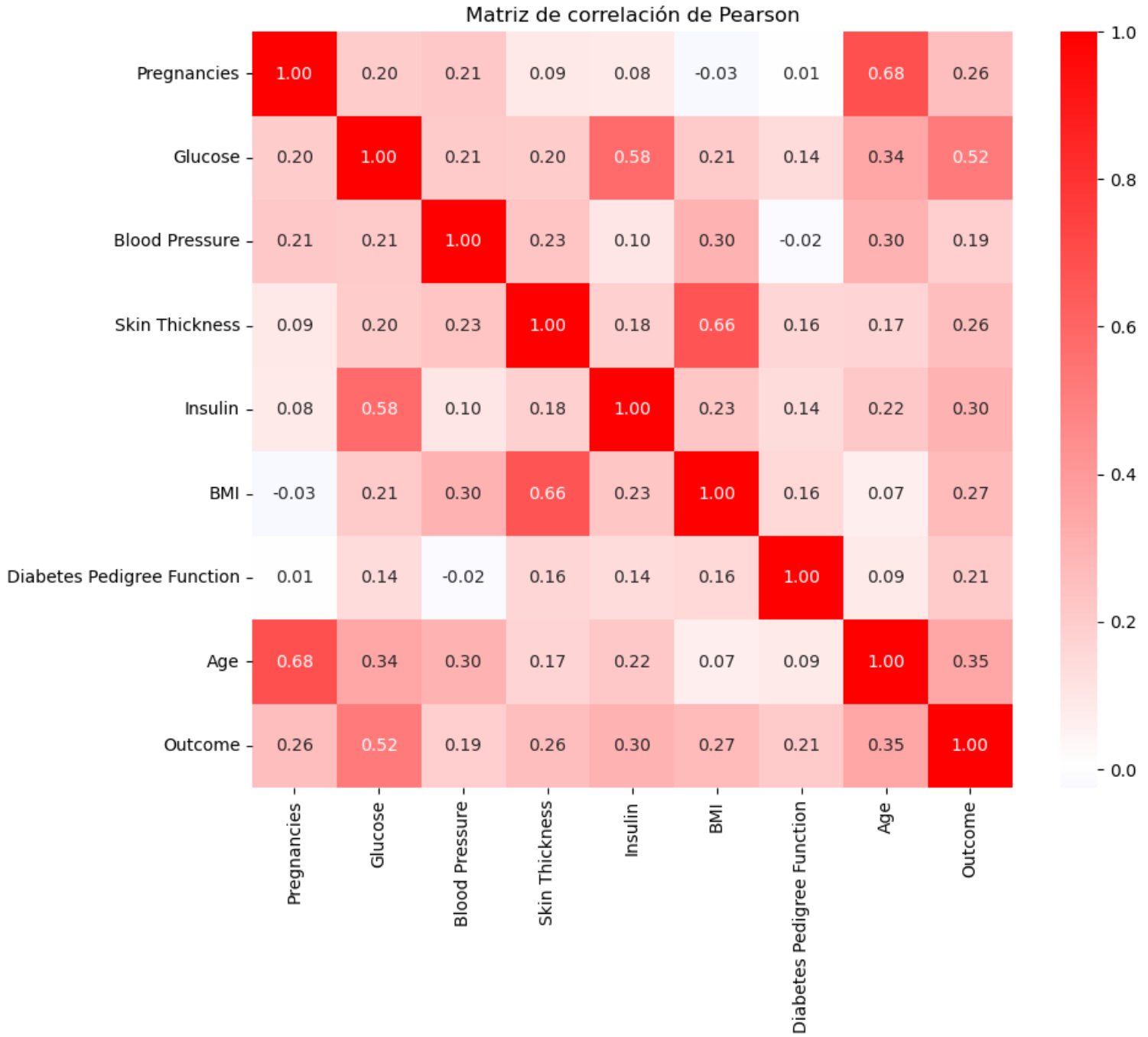
Ejemplo

Las ventas de helados y los incidentes de ahogamiento pueden tener una alta correlación, pero comer helado no causa ahogamientos.

- Este es un ejemplo de una correlación espuria, donde ambas variables están influenciadas por un tercer factor (como el clima cálido) en lugar de que una cause a la otra.
- Siempre se debe considerar la posibilidad de variables confusas. Se debe recordar que la correlación por sí sola no puede establecer una relación de causa y efecto.

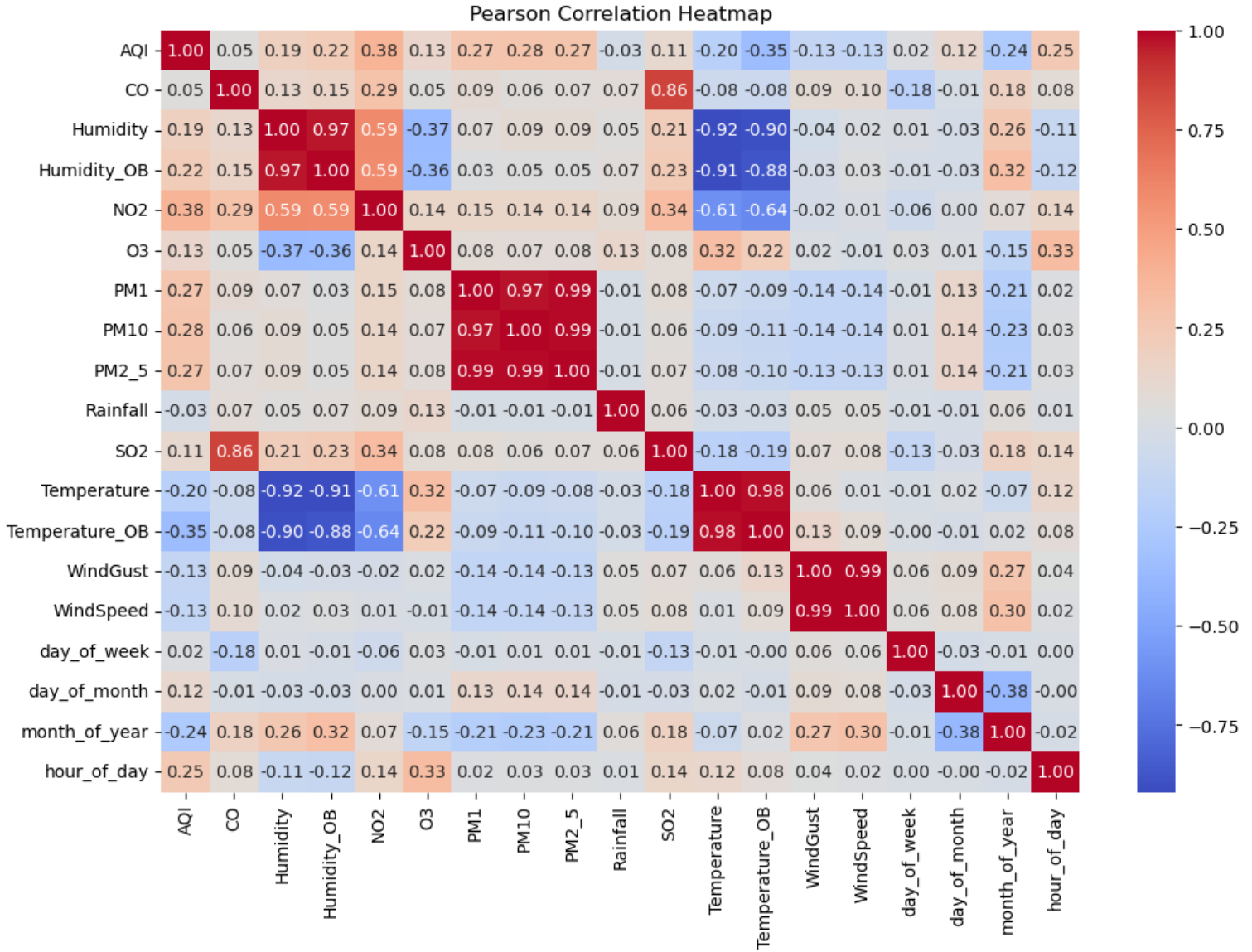
Ejemplo

Pima Indian Diabetes Dataset



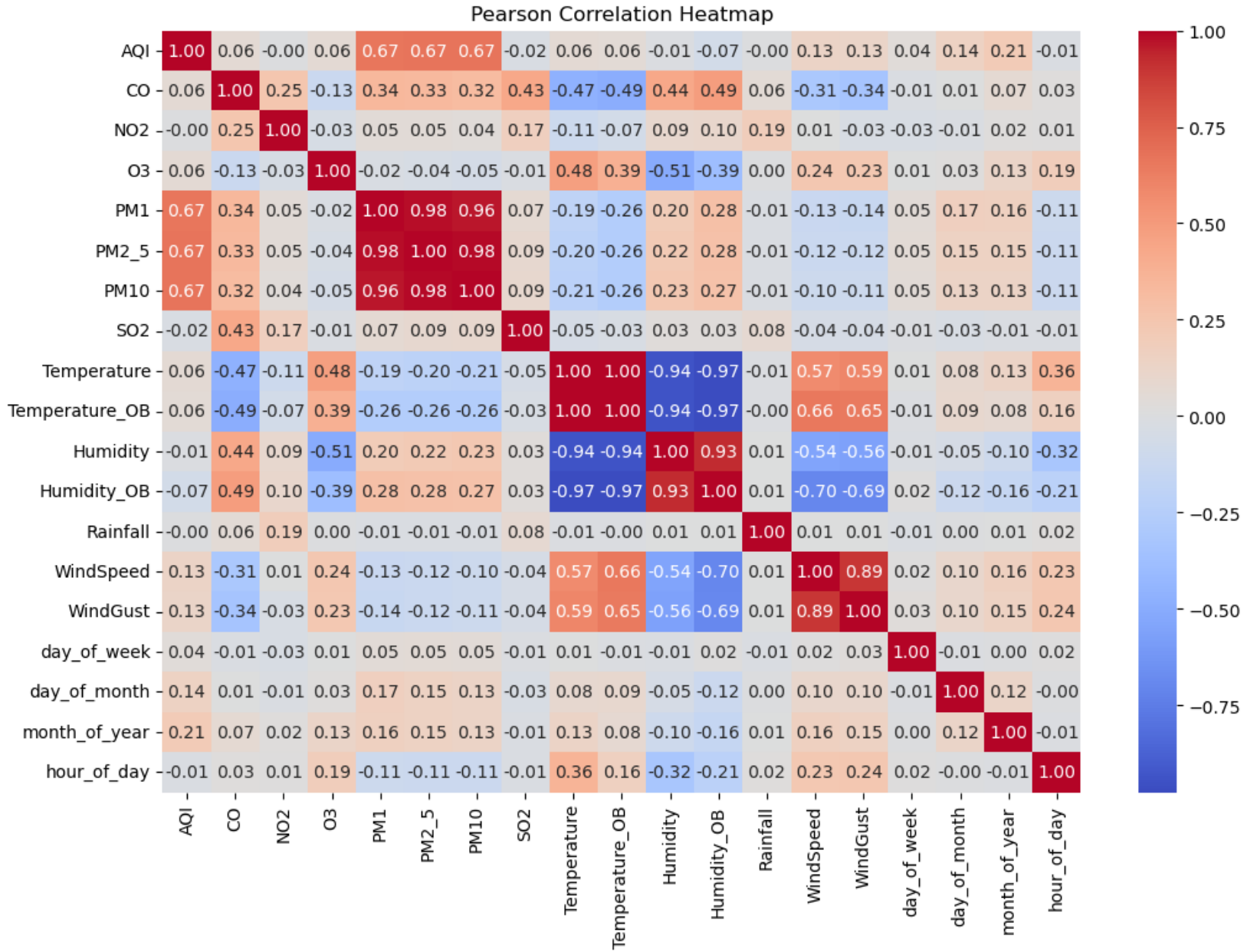
Ejemplo

Calidad de Aire en PTY



Ejemplo

Calidad de Aire en Azuero



Introducción a modelos lineales

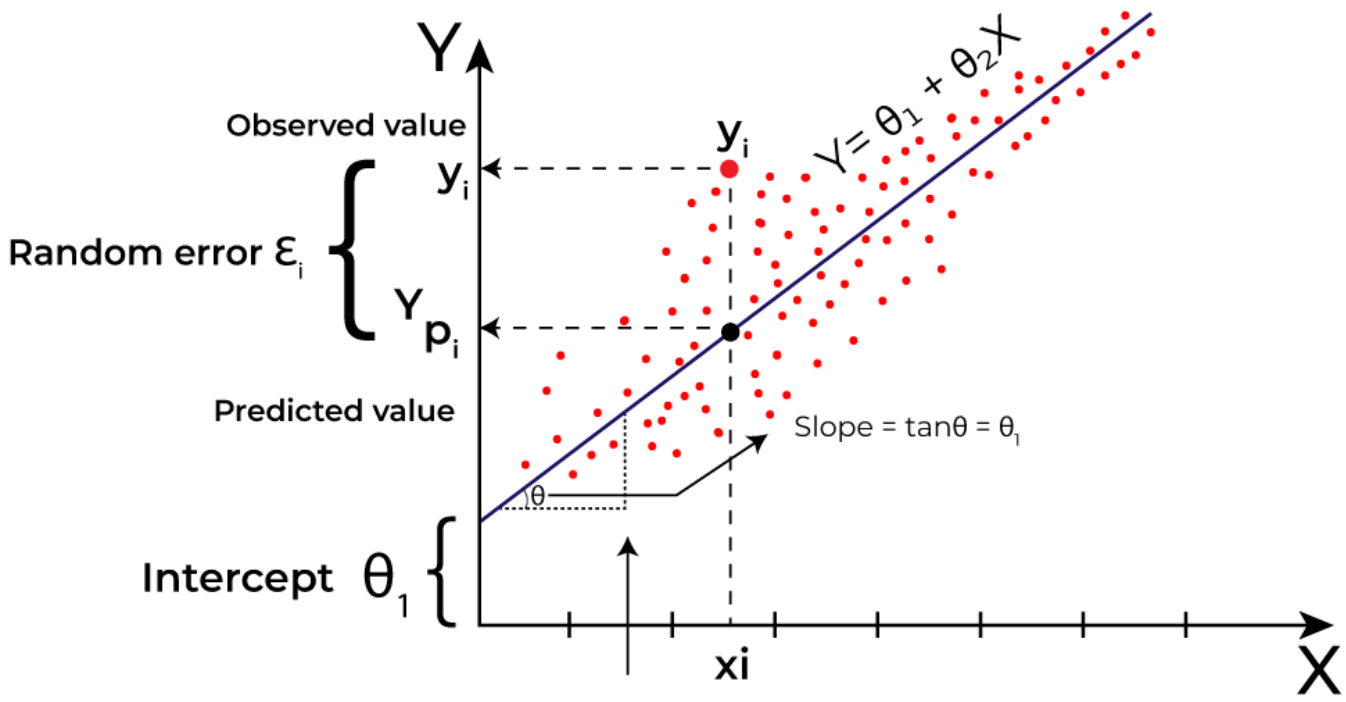
Los modelos lineales son una clase fundamental de modelos estadísticos que asumen una relación lineal entre las variables independientes y la variable dependiente.

Son ampliamente utilizados debido a su simplicidad, interpretabilidad y eficacia en una variedad de aplicaciones.

Regresión lineal y sus extensiones

- La regresión lineal es un método estadístico utilizado para modelar la relación entre una variable dependiente continua y una o más variables independientes.

- Sus extensiones incluyen técnicas como la regresión polinómica y la regresión múltiple, que permiten capturar relaciones más complejas.



Modelo matematico de la regresión lineal

El modelo matemático de la **regresión lineal múltiple** se expresa de la siguiente manera:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

Donde:

- y: Variable dependiente.
- β_0: Intercepto.
- $\beta_1, \beta_2, \ldots, \beta_n$: Coeficientes de las variables independientes.
- $x_1, x_2, \ldots, x_n$: Variables independientes.
- ϵ: Término de error.

Aplicaciones: estimación de relación entre variables

- La estimación de la relación entre variables es un proceso fundamental en el análisis de datos y la estadística. Consiste en identificar y cuantificar cómo una o más variables independientes están asociadas con una variable dependiente. Este análisis permite comprender patrones, realizar predicciones y tomar decisiones informadas basadas en datos.

- Se utiliza en una amplia variedad de campos para explorar relaciones entre factores y evaluar su impacto. Además, proporciona una base para desarrollar modelos predictivos y optimizar procesos en diferentes contextos.

Ejemplo de Linear Regression

Acceder al ejemplo

Regresión Logística

- La regresión logística es un modelo estadístico utilizado para predecir la probabilidad de una variable dependiente categórica basada en una o más variables independientes.
- A diferencia de la regresión lineal, que se utiliza para variables dependientes continuas, la regresión logística se aplica principalmente a problemas de clasificación, como la clasificación binaria (por ejemplo, sí/no, verdadero/falso).
- El modelo utiliza la función sigmoide para transformar los valores predichos en probabilidades que se encuentran en el rango de 0 a 1. Esto permite interpretar los resultados como probabilidades de pertenencia a una clase específica.

Modelo Matemático de la Regresión Logística

El modelo matemático de la **regresión logística** es:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

Donde:

- $P(y = 1 \mid x)$: Probabilidad de que la variable dependiente sea 1 dado x.
- β_0: Intercepto.
- $\beta_1, \beta_2, \ldots, \beta_n$: Coeficientes de las variables independientes.
- $x_1, x_2, \ldots, x_n$: Variables independientes.

Características Principales

- **Probabilidades**: El modelo predice probabilidades, lo que permite establecer umbrales para clasificar las observaciones.
- **Linealidad en el Logit**: Aunque la relación entre las variables independientes y la probabilidad no es lineal, el logit es lineal.
- **Función de Costo**: Utiliza la entropía cruzada como función de costo para optimizar los parámetros del modelo.

¿Qué es el *logit*?

El **logit** es la transformación matemática utilizada en la regresión logística para modelar la relación entre las variables independientes y la probabilidad de un resultado binario.

Representa el logaritmo del *odds* (razón de probabilidades) y se define como:

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

Donde:

- P: Probabilidad de que ocurra un evento (por ejemplo, $P(y = 1 \mid x)$).
- $1 - P$: Probabilidad de que no ocurra el evento.

Propiedades del Logit

- Linealidad: El logit permite que la regresión logística modele una relación lineal entre las variables independientes y el logit de la probabilidad.

- El uso del logit es fundamental en la regresión logística, ya que facilita la interpretación de los coeficientes del modelo y permite trabajar con probabilidades en un marco lineal.

Limitaciones

- **Linealidad en el Logit:** Asume que el logit de la probabilidad es lineal con respecto a las variables independientes.
- **Multicolinealidad:** La presencia de alta correlación entre variables independientes puede afectar la estabilidad del modelo.
- **No captura relaciones no lineales complejas:** Para relaciones más complejas, se requieren modelos más avanzados como redes neuronales o árboles de decisión.

Técnicas de regularización (L1, L2)

En el contexto de machine learning, las técnicas de regularización son métodos para prevenir el overfitting, que ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento y pierde capacidad de generalización en datos nuevos.

Estas técnicas añaden un término de penalización a la función de costo del modelo, restringiendo los valores de los coeficientes y mejorando la robustez del modelo.

Regularización L1 (Lasso)

La **regularización L1** agrega una penalización proporcional a la **suma de los valores absolutos de los coeficientes**. Su función de costo se define como:

$$J(\theta) = \text{Error del modelo} + \lambda \sum_{i=1}^{n} |\theta_i|$$

Donde:

- λ: Parámetro de regularización que controla la fuerza de la penalización.
- θ_i: Coeficientes del modelo.

Regularización L2 (Ridge)

La **regularización L2** agrega una penalización proporcional a la **suma de los cuadrados de los coeficientes**.

Su función de costo se define como:

$$J(\theta) = \text{Error del modelo} + \lambda \sum_{i=1}^{n} \theta_i^2$$

Donde:

- λ: Parámetro de regularización que controla la fuerza de la penalización.
- θ_i: Coeficientes del modelo.

Comparación entre L1 y L2

L1: Selección de características (coeficientes exactos en 0), útil para modelos interpretables.

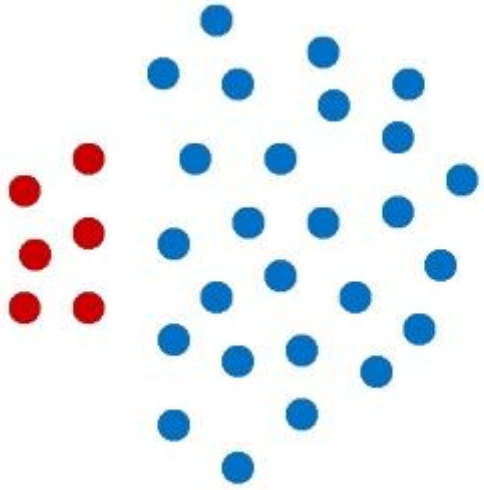
L2: Distribuye la penalización entre todas las características, reduciendo el impacto de cada una.

Aplicaciones

- **Regresión lineal y logística:** Para mejorar la generalización del modelo y evitar el overfitting.
- **Modelos de alta dimensionalidad:** Donde el número de características es mayor que el número de observaciones.
- **Selección de características:** Especialmente con L1, para identificar las variables más relevantes.

Balanceo de Clases

El balanceo de clases es muy importante en problemas de clasificación para garantizar que el modelo no esté sesgado hacia la clase mayoritaria.



Importancia del Balanceo

En problemas de clasificación, un desbalance significativo entre clases puede llevar a un modelo que favorezca la clase mayoritaria.

Esto puede resultar en un desempeño deficiente en la predicción de la clase minoritaria.

Consecuencias del desbalance de clases

- Reducción en la capacidad del modelo para generalizar.
- Métricas como la precisión pueden ser engañosas, ya que un modelo puede predecir siempre la clase mayoritaria y obtener una alta precisión.