MACHINE LEARNING

TIPOS DE VARIABLES:

- FEATURES(Características): Son variables independientes o de entrada que se utilizan para hacer predicciones. También se les conoce como atributos.
- LABELS(Etiquetas): Esta es la variable dependiente o de salida que el modelo está tratando de predecir.

El *feature engineering* (ingeniería de características) es un paso fundamental en el aprendizaje automático, ya que implica seleccionar y transformar las características relevantes para mejorar el rendimiento del modelo. Este proceso puede incluir técnicas como normalización, codificación *one-hot*, escalado, reducción de dimensionalidad o la creación de nuevas características a partir de las ya existentes.

La relación entre las etiquetas y las características es clave en el aprendizaje supervisado. El modelo aprende de los datos etiquetados proporcionados, extrayendo patrones y relaciones entre las características y las etiquetas correspondientes. Analizando estas relaciones, el modelo puede hacer predicciones sobre datos nuevos no vistos.

Comprender bien esta relación permite extraer patrones significativos de los datos, entrenar modelos de manera efectiva y realizar predicciones precisas sobre datos nuevos. Por lo tanto, una ingeniería de características cuidadosa es esencial para garantizar que el modelo capture la información relevante y se generalice bien.

*ESCALA DE MEDIDAS:

Es el conjunto de los posibles valores que una cierta variable puede tomar. Esta escala determina qué tipos de operaciones matemáticas y estadísticas se pueden aplicar a datos.

En el contexto de Machine Learning, las escalas de medida son cruciales por varias razones:

1. **Selección de Algoritmos**: Algunos algoritmos de ML requieren escalas específicas para funcionar correctamente.
2. **Preprocesamiento de datos:** Normalizar o estandarizar las características es una práctica común en ML para asegurar que todas las características estén a la misma escala. Esto mejora la convergencia del modelo durante el entrenamiento y evita que las características con mayores escalas dominen el proceso de optimización.

3. **Interpretación de resultados:** Las métricas de evaluación del modelo dependen de las escalas de medición. Una escala incorrecta puede hacer interpretaciones erróneas de la efectividad del modelo.
4. **Visualización y Análisis Exploratorio de Datos:** Escalar adecuadamente las variables permite una interpretación más clara de los patrones y las correlaciones presentes en los datos.

*MEDIDAS DE TENDENCIA CENTRAL Y DISPERSION

Estas medidas son fundamentales en ML al permitir comprender la distribución y la variabilidad de los datos, facilitando la toma de decisiones informadas y la construcción de modelos predictivos precisos.

1. **Medidas de Tendencia Central (Media, Mediana y Moda)**: Ayudan a entender la distribución de los datos y pueden ser utilizadas para la imputación de valores faltantes. Aquí encontrarás las definiciones y algunos ejemplos: **->** **¿Qué son las medidas de tendencia central y para qué sirven?**

    - **Promedio o media**

        Este dato es ampliamente usado en estadística. Es la cantidad que **se obtiene al sumar todos los datos de un conjunto de valores para posteriormente dividir la cifra obtenida entre la cantidad de valores analizados**. El resultado se expresa en la misma unidad que los datos originales: metros, litros, gramos, horas, etc.

        A la hora de utilizar esta medida de tendencia central en un análisis es necesario tener en cuenta que considera todas las puntuaciones proporcionadas por las variables, por lo que cuando hay valores extremos no ofrece una visión real de la muestra.

        Ejemplo: Para obtener la media del conjunto de números 3, 4, 7, 8, 10 se deben sumar todas las cifras 2+4+7+8+9=30. El resultado hay que dividirlo entre 5, que corresponde al número de valores registrados 30/5=6. La media es 6.

    - **Mediana**

        Es el dato estadístico que ocupa la **posición central en un conjunto de datos cuando estos se organizan en orden de magnitud**, dejando la misma cantidad de valores a un lado y al otro**.**

- **Mediana para datos impares.**

  Obtener la mediana con una cantidad de datos impares es muy sencillo. En primer lugar, todas las cifras se deben ordenar de forma ascendente antes de localizar el centro del conjunto. La mediana será el número que se encuentre exactamente en el medio, de tal forma que el número de datos ubicados a la derecha y a la izquierda de la mediana será exactamente igual.

  Ejemplo: En el conjunto de datos ya ordenado de los números 1, 3, 5, 8, 10, 13, 15, la mediana será 8, puesto que divide el conjunto en dos partes iguales.

- **Mediana para datos pares.**

  En este caso el dato es un poco más laborioso de obtener. Una vez más, es necesario ordenar los datos de menor a mayor y tomar en consideración los dos datos que quedan en el centro del conjunto. La mediana se obtiene al sacar promedio de los dos valores centrales.

  Ejemplo: En el conjunto de números 1, 3, 6, 8, 9, 11, se toman los valores centrales 6 y 8 para hacer el cálculo. El resultado se obtiene con la siguiente operación (6+8)/2=7. La mediana en este ejercicio es igual a 7.

- ## Moda

  La moda es **la variable que más se repite en un conjunto de datos** o muestra poblacional. Una muestra puede presentar más de una moda. No hay una forma específica para obtener esta información, solamente hay que verificar cuál es el resultado que más se repite.

  Ejemplo: Si se busca saber cuál es color favorito en un grupo de diez alumnos, se requiere preguntar esta información a cada estudiante. Si cuatro niños responden azul, dos dicen rosa, dos contestan verde y el último dije amarillo, la moda será azul. Este es el dato que más se repite.

¿Para qué sirven las medidas de tendencia central?

Las medidas de tendencia central tienen distintos usos, entre ellos:

- **Resumir la información**.
- **Conocer el elemento promedio o típico** de un grupo.

- **Comparar e interpretar los resultados** obtenidos al analizar una colección de valores observados.
- **Estudiar el comportamiento de una misma variable** en distintas ocasiones.
- **Comparar los resultados** con otros grupos estadísticos o poblacionales
- **Ordenar los datos sistemáticamente**.
- **Aportar credibilidad** a una información, ya que arrojan promedios o sesgos en los datos reunidos.

2. **Medidas de dispersión (Rango, Desviación Estándar, Varianza, Coeficiente de variación):** Fundamentales en machine learning para entender la distribución de los datos, identificar outliers, seleccionar características relevantes y preprocesar los datos de manera efectiva antes de aplicar modelos predictivos.

Puntos clave

- Las medidas de dispersión ofrecen un valor numérico que indica el grado de variabilidad de una variable.
- El rango, la varianza, la desviación típica y el coeficiente de variación son las medidas de dispersión más conocidas.
- El rango muestra la diferencia entre el valor máximo y mínimo de una muestra o población.

Medidas de dispersión: Explicación sencilla

En otras palabras, las medidas de dispersión son números que indican si una variable se mueve mucho, poco, más o menos que otra. La razón de ser de este tipo de medidas es conocer de manera resumida una característica de la variable estudiada.

En este sentido, deben acompañar a las medidas de tendencia central. Juntas, ofrecen información de un sólo vistazo que luego podremos utilizar para comparar y, si fuera preciso, tomar decisiones.

Principales medidas de dispersión

Las medidas de dispersión más conocidas son: el rango, la varianza, la desviación típica y el coeficiente de variación (no confundir con coeficiente de determinación). A continuación veremos estas cuatro medidas.

Rango

El rango es un valor numérico que indica la diferencia entre el valor máximo y el mínimo de una población o muestra estadística. Su fórmula es:

$R = Máx_x - Mín_x$

Donde:

- **R →** Es el rango.
- **Máx →** Es el valor máximo de la muestra o población.
- **Mín →** Es el valor mínimo de la muestra o población estadística.
- **x →** Es la variable sobre la que se pretende calcular esta medida.

Rango (estadística): Qué es, fórmula y ejemplos

El rango es un valor numérico que indica la diferencia entre el valor máximo y el mínimo... ver más

Varianza

La varianza es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media. Formalmente se calcula como la suma de los residuos al cuadrado divididos entre el total de observaciones. Su fórmula es la siguiente:

$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$$

- **X →** Variable sobre la que se pretenden calcular la varianza

- x_i → Observación número i de la variable X. i puede tomará valores entre 1 y n.
- **N** → Número de observaciones.
- $\bar{x}$ → Es la media de la variable X.

Varianza: Qué es, fórmula y ejemplos

La varianza es una medida de dispersión que representa la variabilidad de una serie de datos respecto... ver más

Desviación típica

La desviación típica es otra medida que ofrece información de la dispersión respecto a la media. Su cálculo es exactamente el mismo que la varianza, pero realizando la raíz cuadrada de su resultado. Es decir, la desviación típica es la raíz cuadrada de la varianza.

$$\sigma = \sqrt{\frac{\sum_i^N (X_i - \bar{X})^2}{N}}$$

- **X** → Variable sobre la que se pretenden calcular la varianza
- x_i → Observación número i de la variable X. i puede tomará valores entre 1 y n.
- **N** → Número de observaciones.
- $\bar{x}$ → Es la media de la variable X.

Medidas de dispersión

Las medidas de dispersión tratan, a través del cálculo de diferentes fórmulas, de arrojar un valor numérico... ver más

Coeficiente de variación

Su cálculo se obtiene de dividir la desviación típica entre el valor absoluto de la media del conjunto y por lo general se expresa en porcentaje para su mejor comprensión.

$$CV = \frac{\sigma_x}{|\bar{X}|}$$

- **X** → Variable sobre la que se pretenden calcular la varianza
- **σ_x** → Desviación típica de la variable X.
- **$|\bar{x}|$** → Es la media de la variable X en valor absoluto con $\bar{x} \neq 0$

Coeficiente de variación: Qué es, usos y ejemplos

El coeficiente de variación, también denominado como coeficiente de variación de Pearson, es una medida estadística que... ver más

A continuación se muestra una imagen que resume las fórmulas anteriores:

MEDIDAS DE DISPERSIÓN

VARIANZA	DESVIACIÓN ESTÁNDAR
$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$	$\sigma = \sqrt{\frac{\sum_1^N (x_i - \bar{X})^2}{N}}$

- **X →** Variable sobre la que se pretenden calcular la varianza.
- **x_i →** Observación número i de la variable X. i puede tomará valores entre 1 y n.
- **N →** Número de observaciones.
- **$\bar{x}$ →** Es la media de la variable X.

RANGO ESTADÍSTICO	COEFICIENTE DE VARIACIÓN
$R = Máx_x - Mín_x$	$CV = \frac{\sigma_x}{\lvert\bar{X}\rvert}$
• **R →** Es el rango. • **Máx →** Es el valor máximo de la muestra o población. • **Mín →** Es el valor mínimo de la muestra o población estadística. • **x →** Es la variable sobre la que se pretende calcular esta medida.	• **X →** Variable sobre la que se pretenden calcular la varianza. • **σ_x →** Desviación típica de la variable X. • **$\lvert\bar{x}\rvert$ →** Es la media de la variable X en valor absoluto con $\bar{x} \neq 0$.

A efectos comparativos, es importante indicar que debemos comparar siempre variables con las mismas unidades de medida. Por ejemplo, no tendría mucho sentido decir que la variabilidad del producto interior bruto (PIB) es mayor que la de la venta de helados. Se podría indicar, pero comparar euros con número de helados no tiene sentido. Por tanto, siempre mejor comparar variables con la misma unidad de medida.

Lo mismo ocurre con las medidas de dispersión. Si lo que se quiere es comparar dos variables, es preferible hacerlo con las mismas medidas de dispersión para cada una de ellas y preferiblemente en la misma unidad.