Participemos en un proyecto...


TITANIC

Compromiso...

-Jack, ¿de verdad me amas?

-Que se hunda el barco si miento

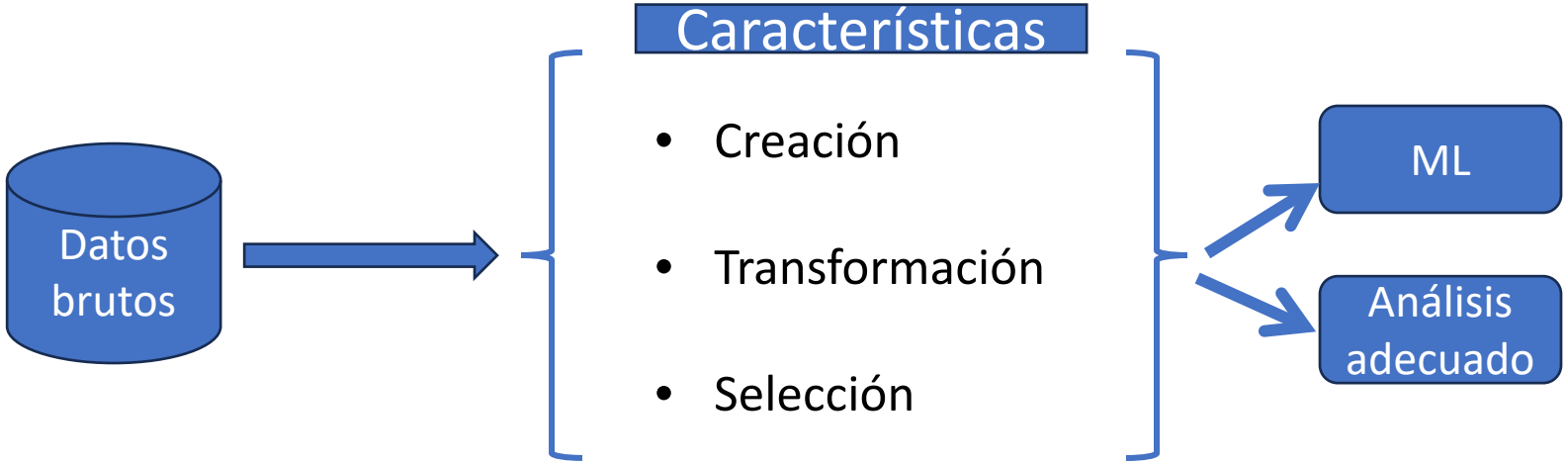
Riesgos...


HEY JACK
I'M PREGNANT

Ingeniería de Características

¿Qué es la ingenieria de caracrerísticas?

- La ingeniería de características es una de las etapas más importantes en la preparación de datos, ya que la calidad y relevancia de las características influyen en gran medida en la calidad de los resultados.



Es un proceso iterativo que implica experimentación y refinamiento a medida que se busca la mejor representación de los datos.

Clasificación de variables

Variables Cualitativas (Categóricas): Estas variables representan cualidades o características que no se pueden medir numéricamente, sino que se agrupan en categorías o clases.

Clasificación de variables

Variables Cualitativas (Categóricas): Estas variables representan cualidades o características que no se pueden medir numéricamente, sino que se agrupan en categorías o clases.

- **Nominales**: Las variables nominales representan categorías sin ningún orden inherente. Ejemplos son el género, la raza, el color, la nacionalidad, el estado civil, etc.

Clasificación de variables

Variables Cualitativas (Categóricas): Estas variables representan cualidades o características que no se pueden medir numéricamente, sino que se agrupan en categorías o clases.

- **Nominales**: Las variables nominales representan categorías sin ningún orden inherente. Ejemplos son el género, la raza, el color, la nacionalidad, el estado civil, etc.

- **Ordinales**: Las variables ordinales representan categorías con un orden inherente, pero las diferencias entre las categorías no son uniformes ni cuantificables. Ejemplos son la clasificación socioeconómica (baja, media, alta), la educación (primaria, secundaria, universitaria), etc.

Clasificación de variables

Variables Cuantitativas (Numéricas): Estas variables representan cantidades medibles y cuantificables. Se dividen en dos subtipos:

Clasificación de variables

Variables Cuantitativas (Numéricas): Estas variables representan cantidades medibles y cuantificables. Se dividen en dos subtipos:

- **Discretas**: Las variables discretas toman valores separados o contables y no pueden tomar valores intermedios. Ejemplos son la cantidad de hijos en una familia, el número de estudiantes en una clase, etc.

Clasificación de variables

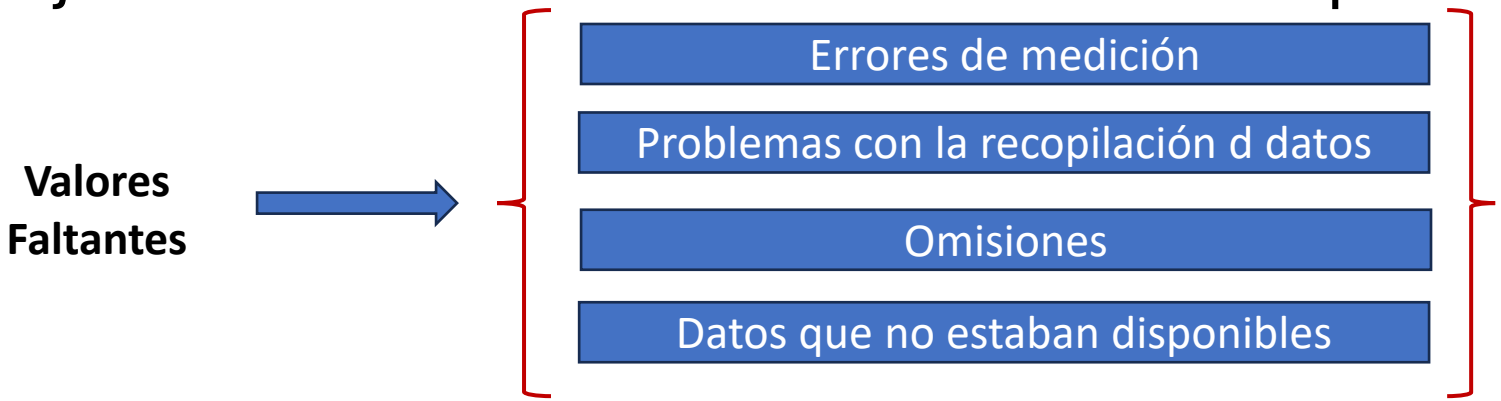
Variables Cuantitativas (Numéricas): Estas variables representan cantidades medibles y cuantificables. Se dividen en dos subtipos:

- **Discretas**: Las variables discretas toman valores separados o contables y no pueden tomar valores intermedios. Ejemplos son la cantidad de hijos en una familia, el número de estudiantes en una clase, etc.

- **Continuas**: Las variables continuas pueden tomar cualquier valor dentro de un rango específico y, en teoría, pueden tener infinitos valores posibles. Ejemplos son la altura, el peso, la temperatura, etc.

¿Qué es la imputación de datos?

Es el proceso de reemplazar o llenar valores faltantes en un conjunto de datos con valores estimados o predichos.



La imputación de datos es una parte importante del preprocesamiento de datos y es esencial para garantizar la integridad y la utilidad de los datos.

Técnicas para hacer imputación de datos

Imputación de media o mediana:

Los valores faltantes se reemplazan por la media (promedio) o la mediana de la característica correspondiente. Esta es una técnica simple y rápida, pero puede no ser adecuada si los valores faltantes están sesgados o siguen una distribución no normal.

Técnicas para hacer imputación de datos

Imputación de valores constantes:

Se reemplazan los valores faltantes por un valor constante predefinido, como cero o un valor específico que tenga sentido en el contexto. Esta técnica es útil cuando los valores faltantes tienen un significado especial.

Técnicas para hacer imputación de datos

Imputación de vecinos más cercanos (K-NN):

Los valores faltantes se estiman utilizando los valores de los vecinos más cercanos en el espacio de características. Esta técnica es especialmente útil en datos tabulares.

Técnicas para hacer imputación de datos

Imputación mediante modelos de aprendizaje automático:

Se utilizan algoritmos de aprendizaje automático para predecir los valores faltantes en función de las otras características. Ejemplos de modelos incluyen regresión, random forests o redes neuronales.

Técnicas para hacer imputación de datos

Imputación por interpolación o extrapolación:

Se utilizan técnicas de interpolación (para valores faltantes dentro del rango conocido) o extrapolación (para valores fuera del rango conocido) para estimar los valores faltantes.

Técnicas para hacer imputación de datos

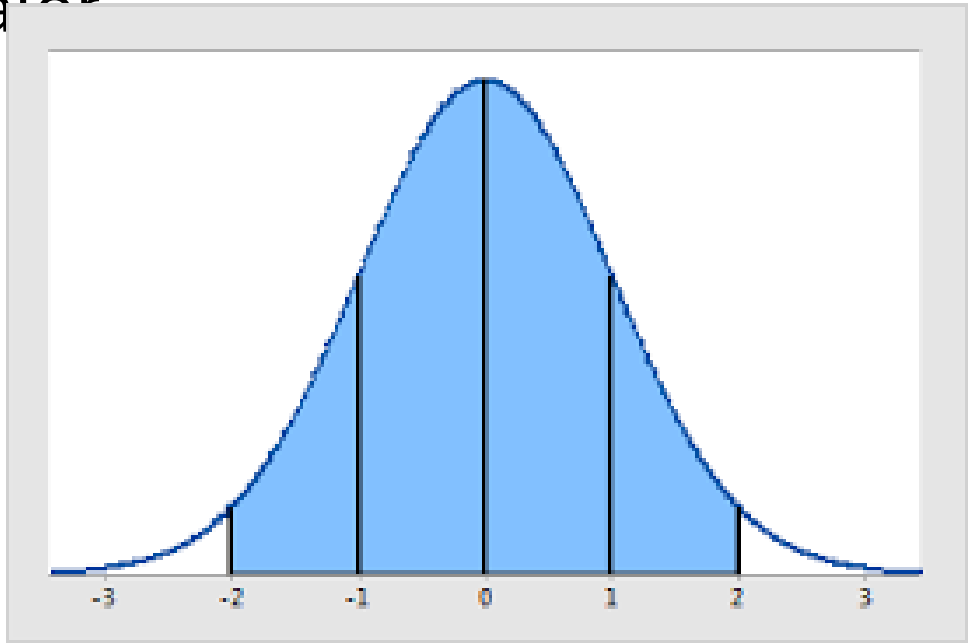
Imputación basada en reglas de negocio o conocimiento del dominio:

A veces, los valores faltantes se pueden estimar utilizando reglas específicas del dominio o conocimiento experto.

¿Qué es una distribución normal?

Una distribución normal, es una distribución continua que se caracteriza por su forma de campana simétrica.

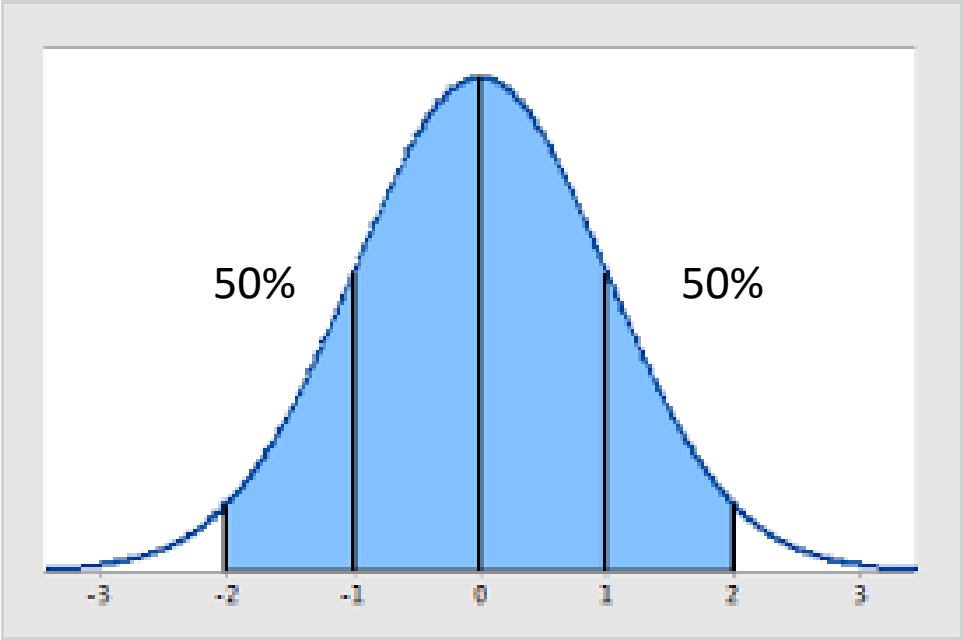
En una distribución normal, la mayoría de los datos se agrupan alrededor de un valor central y disminuyen a medida que nos alejamos de ese valor.



¿Qué es una distribución normal?

Simetría:

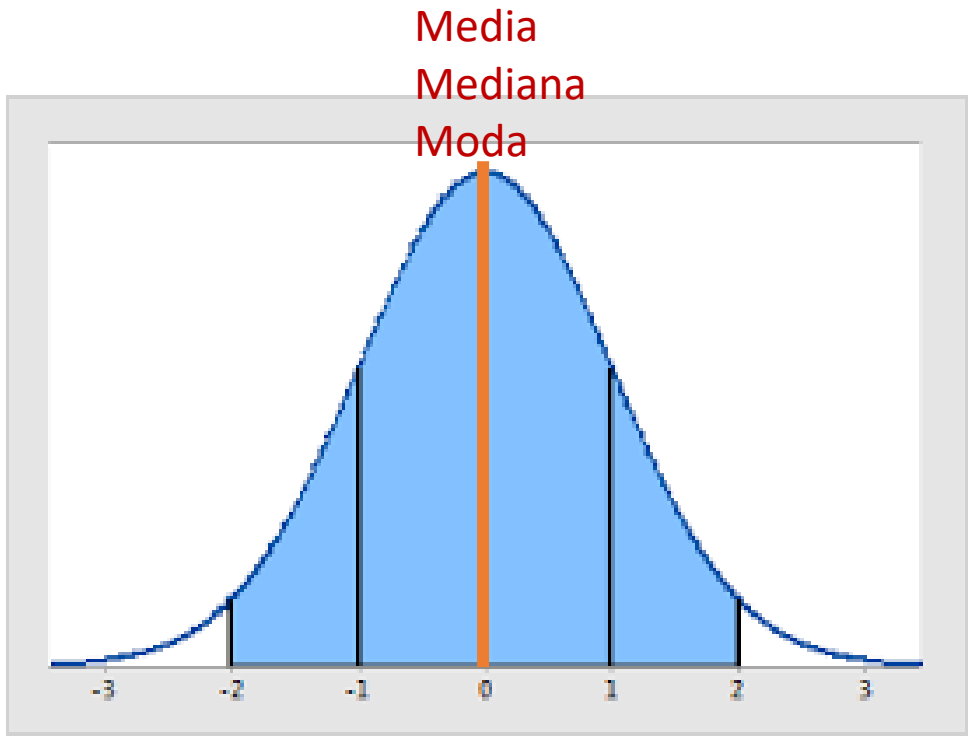
La distribución es simétrica alrededor de su media. Esto significa que la mitad de los datos se encuentran a la izquierda de la media y la otra mitad a la derecha.



¿Qué es una distribución normal?¶

Media, mediana y moda iguales:

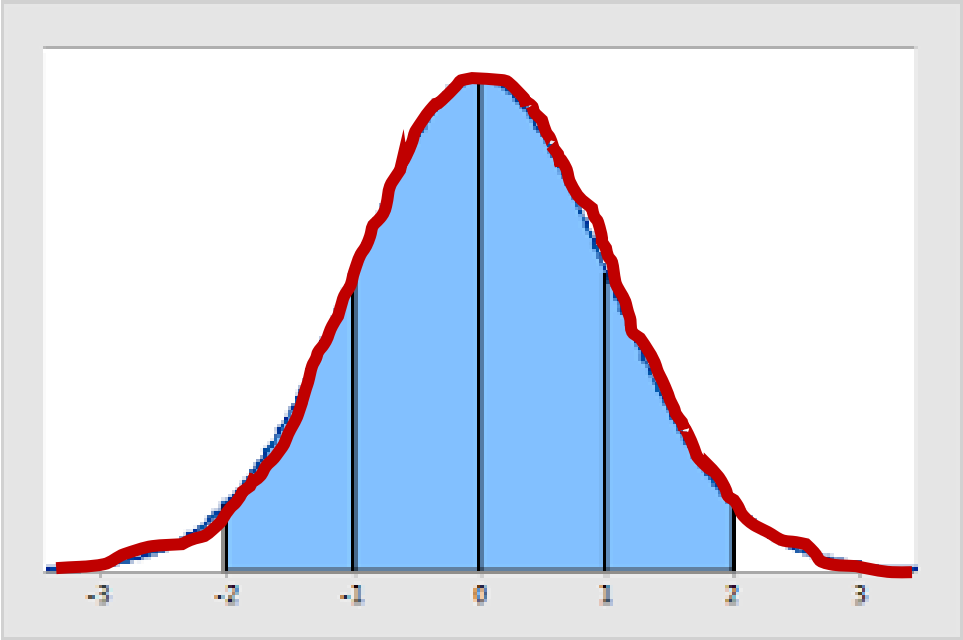
En una distribución normal, la media, la mediana y la moda son iguales y se ubican en el centro de la distribución.



¿Qué es una distribución normal?¶

Forma de campana:

La forma de la distribución se asemeja a una campana con una cola hacia ambos lados. A medida que te alejas de la media, la densidad de probabilidad disminuye gradualmente.



Transformación de los datos

Transformación:

Es un proceso en el análisis y preparación de datos que implica la modificación de los valores originales de las variables o características en un conjunto de datos para mejorar su calidad.

Transformación de los datos

Estandarización:

Consiste en ajustar las escalas de las variables para que tengan una media de 0 y una desviación estándar de 1.

Transformación de los datos

Normalización:

Consiste en ajustar las escalas de las variables para que estén en un rango específico [0,1].

Reducción de la dimensionalidad

Es un proceso en el análisis de datos y el aprendizaje automático que se utiliza para disminuir el número de características o variables en un conjunto de datos, manteniendo al mismo tiempo la mayor cantidad de información relevante.

Reducción de la dimensionalidad:¿Por qué?

Reducción de la complejidad:

Los conjuntos de datos con muchas características pueden volverse complejos y difíciles de manejar. La reducción de dimensionalidad puede simplificar los datos y hacer que sean más fáciles de entender y analizar.

Reducción de la dimensionalidad:¿Por qué?

Eliminación de características irrelevantes o redundantes:

No todas las características en un conjunto de datos contribuyen de manera significativa a la información. La reducción de dimensionalidad puede ayudar a identificar y eliminar características que no aportan valor.

Reducción de la dimensionalidad:¿Por qué?

Mejora de la eficiencia computacional:

Al reducir la cantidad de características, se puede acelerar el tiempo de entrenamiento y predicción de modelos de aprendizaje automático.

Reducción de la dimensionalidad:¿Por qué?

Visualización de datos:

La reducción de dimensionalidad puede ser útil para representar datos en un espacio de menor dimensión que permita su visualización y comprensión.

Reducción de la dimensionalidad:¿Cómo?

Selección de características:

En este enfoque, se selecciona un subconjunto de las características originales para conservar, descartando las demás. La selección de características se basa en diversos criterios, como la importancia de las características, la correlación con la variable objetivo o la eliminación de características altamente redundantes.

Reducción de la dimensionalidad:¿Cómo?

Extracción de características:

En este enfoque, se crean nuevas características que son combinaciones lineales de las características originales. Algunos métodos de extracción de características comunes incluyen Análisis de Componentes Principales (**PCA**) y T-Distributed Stochastic Neighbor Embedding (t-SNE).

Reducción de la dimensionalidad:¿Cómo?

Análisis de Componentes Principales (PCA):

El objetivo que persigue es capturar la mayor cantidad posible de la variabilidad en los datos en un conjunto reducido de variables (llamadas componentes principales), lo que facilita la interpretación y el análisis de los datos.

PCA:Pasos

Centrar los datos:

Para cada variable en el conjunto de datos, resta la media de esa variable a cada observación.

PCA:Pasos

Calcular la matriz de covarianza o correlación:

La matriz de covarianza se utiliza cuando las variables tienen diferentes unidades o escalas, mientras que la matriz de correlación se utiliza cuando las variables están en las mismas unidades y se quiere considerar también las relaciones lineales.

PCA:Pasos

Calcular los autovalores y autovectores:

Los autovalores y autovectores de la matriz de covarianza (o de la matriz de correlación) proporcionan información sobre la cantidad de variabilidad en los datos y en qué dirección se encuentra esa variabilidad. Los autovectores representan las direcciones de máxima variación, y los autovalores indican cuánta variabilidad hay en esas direcciones.

PCA:Pasos

Ordenar los autovalores y autovectores:

Ordena los autovalores de mayor a menor. Los autovectores correspondientes a los autovalores más grandes son los componentes principales más importantes y capturan la mayor cantidad de variabilidad en los datos.

PCA:Pasos

Seleccionar los componentes principales:

Elige los primeros k autovectores, donde k es el número de dimensiones que deseas reducir o el número de componentes principales que deseas conservar.

PCA:Pasos

Construir la matriz de transformación:

Utiliza los autovectores seleccionados como columnas para construir la matriz de transformación. Cada componente principal se representa como una combinación lineal de las variables originales.

PCA:Pasos

Proyectar los datos en el nuevo espacio:

Multiplica la matriz de datos centrados por la matriz de transformación para obtener el conjunto de datos transformado en el espacio de los componentes principales.

¿Qué aprendiste?

- Tienes una idea muy clara del potencial que ofrece la IA al medio empresarial

¿Qué aprendiste?

- Tienes una idea muy clara del potencial que ofrece la IA al medio empresarial
- Entiendes que la IA es un área muy extensa y conoces su alcance, limitantes y estado actual.

¿Qué aprendiste?

- Tienes una idea muy clara del potencial que ofrece la IA al medio empresarial
- Entiendes que la IA es un área muy extensa y conoces su alcance, limitantes y estado actual.
- Comprendes la lógica de programación que siguen los procesos computacionales asociados a los procesos ETL, feature engineering y representación de la información, entre otros.

¿Qué aprendiste?

- Tienes una idea muy clara del potencial que ofrece la IA al medio empresarial
- Entiendes que la IA es un área muy extensa y conoces su alcance, limitantes y estado actual.
- Comprendes la lógica de programación que siguen los procesos computacionales asociados a los procesos ETL, feature engineering y representación de la información, entre otros.
- Sabes interpretar resultados y proponer estrategias derivadas de tus interpretaciones.