**Dataset original (con NaN)**

| **Fila** | **Age** | **Fare** | **SibSp** |
| --- | --- | --- | --- |
| 0 | 22.0 | 7.25 | 1 |
| 1 | **NaN** | 71.28 | 1 |
| 2 | 24.0 | 8.05 | 0 |
| 3 | 35.0 | **NaN** | 1 |
| 4 | 28.0 | 8.46 | 0 |

**1. KNNImputer (k=2)**

El algoritmo calcula distancias entre filas **usando las columnas disponibles**.  
Para cada NaN busca los **2 vecinos más cercanos** y promedia sus valores.

**Caso A → Fila 1 (Age = NaN)**

* Usamos variables sin NaN: Fare=71.28, SibSp=1.
* Comparamos con otras filas (distancia euclídea simplificada):
* Distancia con Fila 0 (Age=22, Fare=7.25, SibSp=1) → muy grande por la diferencia en Fare.
* Distancia con Fila 3 (Age=35, SibSp=1) → muy cercana porque ambos tienen SibSp=1.
* Distancia con Fila 4 (Age=28, SibSp=0) → algo más lejana pero válida.

*Vecinos elegidos ≈ Fila 3 (Age=35) y Fila 4 (Age=28).* Promedio:

**Caso B → Fila 3 (Fare = NaN)**

* Usamos Age=35, SibSp=1.
* Comparamos con otras filas con Fare conocido:
* Fila 1 (Fare=71.28, SibSp=1) → muy cercana.
* Fila 4 (Fare=8.46, SibSp=0) → también cercana.

Vecinos elegidos ≈ Fila 1 y Fila 4.  
 Promedio:

Resultado KNN:

* Age imputado = **31.5**
* Fare imputado = **39.87**

**2. IterativeImputer (MICE)**

Aquí ya no se buscan vecinos, sino que se ajustan **modelos de regresión**.  
El proceso es iterativo:

1. Se inicializan NaN con medias aproximadas.
2. Luego, para cada variable con NaN, se entrena un modelo de regresión usando las demás variables como predictores.
3. Se predicen los valores faltantes y se repite varias veces hasta converger.

**Caso A → Fila 1 (Age = NaN)**

* Modelo:
* Observamos que Fare=71.28 es muy alto (clase alta).
* El modelo aprende que pasajeros con tarifas altas tendían a ser mayores.  
   Predice Age ≈ 50.13

**Caso B → Fila 3 (Fare = NaN)**

* Modelo:
* Observamos que Age=35 y SibSp=1 son valores comunes en pasajeros de clase media.
* El modelo aprende que tarifas altas suelen acompañar a edades diferentes, pero aquí predice algo moderado.
* Predice Fare ≈ 33.81

Resultado MICE:

* Age imputado = **50.13**
* Fare imputado = **33.81**

**Comparación:**

| **Método** | **Age imputado** | **Fare imputado** |
| --- | --- | --- |
| **KNNImputer** | 31.5 | 39.87 |
| **MICE** | 50.13 | 33.81 |

Conclusión:

* **KNN** → busca “parecidos” y promedia → valores más conservadores.
* **MICE** → usa regresiones multivariadas → valores más extremos pero más coherentes con correlaciones globales.

**Tips para elegir la imputación correcta**

**🔹 1. Media (mean)**

Úsala cuando:

* La variable es **numérica**.
* Los datos siguen una **distribución aproximadamente normal** (sin muchos outliers).  
  Problema: sensible a valores extremos.

Ejemplo: temperatura corporal en °C.

**🔹 2. Mediana (median)**

Úsala cuando:

* La variable es **numérica**.
* Hay **outliers** o asimetría en los datos.  
  ⚠️ Problema: pierde un poco de variabilidad.

Ejemplo: ingresos anuales (suelen tener valores muy altos que distorsionan la media).

**🔹 3. Moda (most\_frequent)**

Úsala cuando:

* La variable es **categórica** (sexo, estado civil, ciudad, etc.).
* O para numéricas discretas con pocos valores repetidos.  
  Problema: puede sobre-representar la categoría más común.

Ejemplo: columna “Embarked” en Titanic.

**4. Constante (constant)**

Úsala cuando:

* Quieres marcar explícitamente los valores faltantes.
* Prefieres poner "Desconocido", 0, 999, etc.  
  Problema: introduces un valor que no pertenece realmente a la distribución.

Ejemplo: “No aplica” en una encuesta.

**5. KNNImputer**

Úsalo cuando:

* Quieres usar la **similitud entre registros**.
* Tus variables están relacionadas de forma local (vecinos similares → valores parecidos).  
  Problema: costoso en datasets muy grandes, y sensible a la escala (conviene normalizar primero).

Ejemplo: imputar calificaciones de alumnos basándote en alumnos con notas similares.

**🔹 6. IterativeImputer (MICE)**

Úsalo cuando:

* Hay **correlación fuerte entre variables**.
* Quieres un método más **sofisticado y realista**.  
  Problema: más lento, y puede generar estimaciones “extremas” si el modelo aprende correlaciones raras.

Ejemplo: imputar valores médicos (glucosa, presión, colesterol) donde las variables están relacionadas fisiológicamente.

**Resumen Express:**

* **Media** → rápido, normal y sin outliers.
* **Mediana** → rápido, con outliers.
* **Moda** → categóricas.
* **Constante** → marcar faltantes explícitamente.
* **KNN** → cuando crees que “los parecidos se parecen en todo”.
* **MICE** → cuando hay correlaciones entre variables y quieres imputación predictiva.

**Tips según tipos de variables**

1. **Numéricas continuas** (edad, salario, temperatura):
   * Pequeños NaN → mean o median son rápidos.
   * Con outliers → mejor median.
   * Con correlación con otras variables → MICE.
2. **Numéricas discretas** (número de hijos, calificaciones enteras):
   * Usar median o most\_frequent para mantener coherencia.
   * Si hay patrones con otras variables → KNN o MICE.
3. **Categóricas** (sexo, ciudad, estado civil):
   * Usar most\_frequent.
   * O constant para poner “Desconocido” y no distorsionar la moda.

**📌 Tips según tamaño del dataset**

**🔹 Dataset pequeño (pocas filas, pocas columnas)**

* Métodos complejos como **MICE** pueden sobreajustar o dar estimaciones inestables (muy poca info).
* Mejor usar **mediana, moda o constante**.

👉 Ejemplo: 100 registros de una encuesta.

**🔹 Dataset mediano (miles de filas, <100 columnas)**

* Aquí ya se puede usar **KNN** (si normalizas las variables).
* **MICE** también es útil si hay correlaciones claras.
* Media/mediana siguen siendo válidos si los faltantes son pocos.

👉 Ejemplo: Titanic (~890 filas, 12 variables).

**🔹 Dataset grande (decenas de miles o millones de filas, muchas columnas)**

* **Cuidado con KNN y MICE** → son muy costosos en memoria y tiempo.
* Lo práctico es usar **estrategias simples**: media, mediana, moda, constante.
* Para algo más avanzado: usar **modelos predictivos especializados** (árboles, XGBoost) para imputar, en lugar de KNN/MICE.

👉 Ejemplo: Big Data de ventas de un supermercado (millones de registros).

**📌 Tips según proporción de valores faltantes**

* Si faltan **<5% de los datos** en una columna → cualquier método simple (media, mediana, moda).
* Si faltan **entre 5–30%** → mejor usar KNN o MICE (capturan relaciones).
* Si faltan **>30–40%** → puede ser mejor eliminar la variable, porque la imputación ya mete demasiado “ruido”.

| **Tipo de variable / Tamaño del dataset** | **Pequeño (≤1k filas, pocas columnas)** | **Mediano (1k–100k filas, <100 columnas)** | **Grande (>100k filas, muchas columnas)** |
| --- | --- | --- | --- |
| **Numéricas continuas** (edad, salario, temperatura) | - Media o mediana simples.- Mejor **mediana** si hay outliers.- Evitar MICE (pocos datos). | - Media/mediana si pocos NaN.- **KNN** si hay patrones locales.- **MICE** si hay correlación fuerte. | - Media/mediana (rápidos).- Evitar KNN/MICE (muy pesados). |
| **Numéricas discretas** (hijos, calificaciones) | - Mediana o moda.- Constante si quieres diferenciar “sin info”. | - Moda si pocos NaN.- KNN/MICE si hay dependencia con otras columnas. | - Moda o constante.- Evitar métodos costosos. |
| **Categóricas** (sexo, ciudad, estado civil) | - Moda.- Constante = “Desconocido”. | - Moda si pocos NaN.- Constante si quieres preservar la categoría “NA”. | - Moda (rápido).- Constante (“Desconocido”). |
| **Cuando hay mucha correlación entre variables** | - Pocos datos: usar **mediana/moda**. | - Ideal para **MICE** (predicciones multivariadas). | - MICE no escala bien → preferir medianas o modelos predictivos simples. |
| **Proporción de NaN** | <5% → cualquier método simple. >30% → eliminar variable. | 5–30% → **KNN o MICE** útiles. <5% → simple. | Igual que mediano, pero más pragmático: <5% simple, >30% eliminar. |