Domando Datos Reales con Pandas

Limpieza, Análisis y Ética en el Agro

Ingeniería de Software I

11 de enero de 2026

Curso de IA Aplicada

¿Por qué NumPy no es suficiente?

En la vida real, los datos son **sucios** y **heterogéneos**.

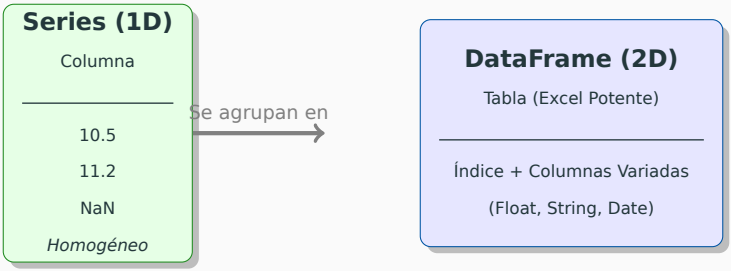
El problema de NumPy:

- Solo maneja números.
- No sabe qué es una fecha.
- Si falta un dato, todo falla.
- Acceso por índice opaco: `data[0, 4]`

La solución Pandas:

- Etiquetas: `data['Humedad']`.
- Series Temporales nativas.
- Manejo robusto de `NaN`.
- **SQL para Python.**

Anatomía: Series vs DataFrame



Ingesta de Datos (I/O)

No confíes en los valores por defecto. Define tus tipos.

```
import pandas as pd

# Carga robusta para Agro
df = pd.read_csv(
    'cosecha_2024.csv',
    sep=';',                    # Delimitador comun en Latam
    parse_dates=['fecha'],      # Interpretar tiempo
    na_values=['-', 'error'],   # Convertir basura a NaN
    index_col='fecha'           # Usar fecha como eje X
)

# Inspeccion rapida
print(df.info())
```

Tip Pro

Si una columna numérica tiene un solo texto, Pandas la volverá texto (`object`). Usa `.info()` siempre.

Navegación: .loc vs .iloc

La confusión #1 en entrevistas de trabajo.

Comando	Busca por...	Ejemplo
`.loc[]`	**Etiqueta** (Nombre)	`df.loc['2024-01', 'pH']`
`.iloc[]`	**Posición** (Índice)	`df.iloc[0, 4]`

```
# Filtrado Booleano (Mascara)
# "Dias calurosos con humedad alta"
alerta = df[ (df['temp'] > 30) & (df['hum'] > 80) ]
```

Limpieza y Valores Nulos

En el campo, los sensores fallan. ¿Qué hacemos con los huecos?

- **Borrar:** `dropna()` → Perdemos datos valiosos.
- **Promedio:** `fillna(mean)` → Aplana la curva.
- **Interpolación:** La opción correcta para series temporales.

```
# Relleno inteligente (conecta los puntos)
df['humedad'] = df['humedad'].interpolate(method='time')

# Sanear errores fisicos
# Humedad > 100% es imposible -> convertir a NaN
df.loc[df['humedad'] > 100, 'humedad'] = pd.NA
```

Análisis: GroupBy y Resample

GroupBy: Para categorías (Lotes, Variedades). **Resample:** Para tiempo (Días, Meses).

```
# ¿Cual lote produce mas?
rendimiento = df.groupby('id_lote')['kilos'].sum()

# Promedio mensual de clima
clima_mes = df.resample('M').agg({
    'temp': 'mean',
    'lluvia': 'sum'
})
```

Ética del Dato

Limpiar datos es **alterar la realidad**.

Preguntas Obligatorias

1. **Trazabilidad:** ¿Guardo el original `raw`?
2. **Sesgo:** Al rellenar con promedios, ¿oculto fallas sistémicas en zonas vulnerables?
3. **Transparencia:** ¿Está documentado mi proceso de limpieza?

Ejercicio Rápido: Calidad de Leche

Detectar vacas con problemas de salud (pH fuera de 6.6 - 8.6).

```
data = {
    'vaca': ['V1', 'V2', 'V3'],
    'ph':   [6.7,  6.2,  7.1],  # V2 acida, V3 alcalina
    'lts':  [20,   18,   15]
}
df = pd.DataFrame(data)

# 1. Filtrar anomalías
enfermas = df[ (df['ph'] < 6.6) | (df['ph'] > 6.8) ]

# 2. Calcular perdida economica (precio = 2000)
perdida = enfermas['lts'].sum() * 2000
print(f"Perdida hoy: ${perdida}")
```

Próximos Pasos

¿Listos para el reto semanal?

Descarguen `clima_corrupto.csv` del repositorio.

¡A programar!