Informe de Proyecto: Análisis de Datos COVID-19 con Framework Kedro

Estudiante: Hans Mancilla

Curso: Machine Learning MLY0100

Profesor: Giocrisrai Godoy

Fecha: 15 Septiembre de 2025

Resumen Ejecutivo

Este informe presenta el desarrollo de un proyecto de Machine Learning utilizando el framework **Kedro** para el análisis de datos relacionados con COVID-19. El proyecto implementa un pipeline de procesamiento de datos que incluye análisis estadístico descriptivo, detección y limpieza de datos faltantes, y visualización de patrones epidemiológicos a nivel global. A través de múltiples nodos interconectados, se procesaron 6 datasets diferentes obtenidos de Kaggle, generando insights valiosos sobre la distribución global del COVID-19.

---

1. Introducción

1.1 Contexto del Proyecto

El proyecto fue desarrollado como parte de los requerimientos académicos para demostrar competencias en el uso del framework Kedro para proyectos de Machine Learning. Kedro es una herramienta de desarrollo de código abierto que facilita la creación de pipelines de datos reproducibles, mantenibles y modulares.

1.2 Objetivos

- **Objetivo Principal:** Desarrollar un pipeline completo de Machine Learning usando Kedro con al menos 3 archivos CSV
- **Objetivos Específicos:**
  - Implementar análisis estadístico descriptivo automatizado
  - Desarrollar procesos de detección y limpieza de datos
  - Crear visualizaciones informativas sobre patrones de COVID-19
  - Demostrar el uso efectivo de la arquitectura de nodos de Kedro

---

2. Marco Teórico

2.1 Framework Kedro

Kedro es un framework Python de código abierto que permite crear pipelines de datos reproducibles y mantenibles. Sus características principales incluyen:

- **Separación de configuración y código**
- **Gestión automática de dependencias entre nodos**
- **Estructura de proyecto estandarizada**
- **Versionado de datos integrado**

2.2 Metodología CRISP-DM

El proyecto sigue implícitamente la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining):

1. **Comprensión del negocio:** Análisis epidemiológico de COVID-19
2. **Comprensión de datos:** Análisis estadístico descriptivo
3. **Preparación de datos:** Limpieza y procesamiento
4. **Modelado:** Análisis y visualización
5. **Evaluación:** Generación de insights
6. **Despliegue:** Pipeline automatizado

---

3. Metodología

3.1 Fuente de Datos

Los datos fueron obtenidos de **Kaggle.com**, específicamente del dataset "COVID-19 Data Repository". El conjunto incluye:

Archivo	Descripción	Dimensiones Aprox.
country_wise_latest.csv	Datos más recientes por país	187 países
covid_19_clean_complete.csv	Datos históricos completos	49,068 registros
day_wise.csv	Datos agregados diarios	366+ días
full_grouped.csv	Datos agrupados completos	30,000+ registros
usa_county_wise.csv	Datos por condado de EE.UU.	3,200+ condados
worldometer_data.csv	Datos de Worldometer	210+ países

3.2 Arquitectura del Pipeline

El pipeline de Kedro fue diseñado con 4 nodos principales:

```
[Datos Crudos] → [Análisis Estadístico] → [Detección de Anomalías] → [Limpieza] → [Visualización]
```

---

4. Desarrollo e Implementación

4.1 Nodo 1: Análisis Estadístico Descriptivo

Función: Generación automática de estadísticas descriptivas para todos los datasets.

Características implementadas:

- Medidas de tendencia central (media, mediana, moda)
- Medidas de dispersión (desviación estándar, varianza)
- Análisis de cuartiles (Q1, Q2, Q3)
- Valores mínimos y máximos
- Conteo de registros y variables

Output: Estadísticas impresas en consola para análisis exploratorio.

Ejemplo worldometer_data.csv:

```
===== Explorando dataset: worldometer_data =====
Shape: (209, 16)
Columnas: ['Country/Region', 'Continent', 'Population', 'TotalCases', 'NewCases', 'TotalDeaths', 'NewDeaths', 'TotalReco
vered', 'NewRecovered', 'ActiveCases', 'Serious,Critical', 'Tot Cases/1M pop', 'Deaths/1M pop', 'TotalTests', 'Tests/1M
pop', 'WHO Region']
  Country/Region      Continent    Population  TotalCases  ...  Deaths/1M pop  TotalTests  Tests/1M pop      WHO Region
0            USA  North America  3.311981e+08     5032179  ...          492.0  63139605.0      190640.0        Americas
1         Brazil  South America  2.127107e+08     2917562  ...          464.0  13206188.0       62085.0        Americas
2          India           Asia  1.381345e+09     2025409  ...           30.0  22149351.0       16035.0  South-EastAsia

[3 rows x 16 columns]
              count unique     top freq             mean  ...    min       25%        50%         75%           max
Country/Region    209    209     USA    1              NaN  ...    NaN       NaN        NaN         NaN           NaN
Continent         208      6  Africa   57              NaN  ...    NaN       NaN        NaN         NaN           NaN
Population      208.0    NaN     NaN  NaN  30415486.971154  ...  801.0  966314.0  7041972.5  25756135.5  1381344997.0
TotalCases      209.0    NaN     NaN  NaN     91718.497608  ...   10.0     712.0     4491.0     36896.0     5032179.0
NewCases          4.0    NaN     NaN  NaN          1980.5  ...   20.0      27.5      656.0      2609.0        6590.0

[5 rows x 11 columns]
```

4.2 Nodo 2: Detección de Datos Faltantes y Duplicados

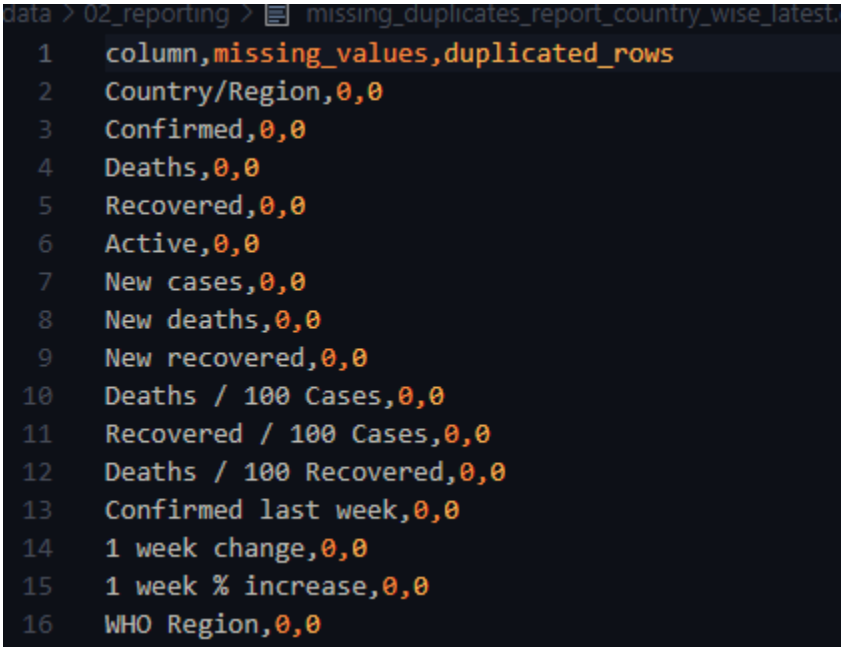
Función: Identificación sistemática de problemas de calidad de datos.

Proceso implementado:

- Escaneo automático de valores nulos por columna
- Detección de registros duplicados
- Generación de reportes detallados
- Almacenamiento en `data/02_intermediate/missing_duplicate_{nombrecsv}.csv`

Beneficio: Visibilidad completa de la calidad de datos antes del procesamiento.

Ejemplo de missing_duplicate_country_wise.csv:



4.3 Nodo 3: Limpieza de Datos

Función: Procesamiento y limpieza automática de datasets seleccionados (únicos datasets con nulos/duplicados).

Datasets procesados:

- `covid_19_clean_complete.csv`
- `usa_county_wise.csv`
- `worldometer_data.csv`

Reglas de limpieza implementadas:

- Valores categóricos nulos → "unknown"
- Valores numéricos nulos → 0
- Preservación de estructura original
- Almacenamiento en `data/03_intermediate/{nombrecsv}_CLEAN.csv`

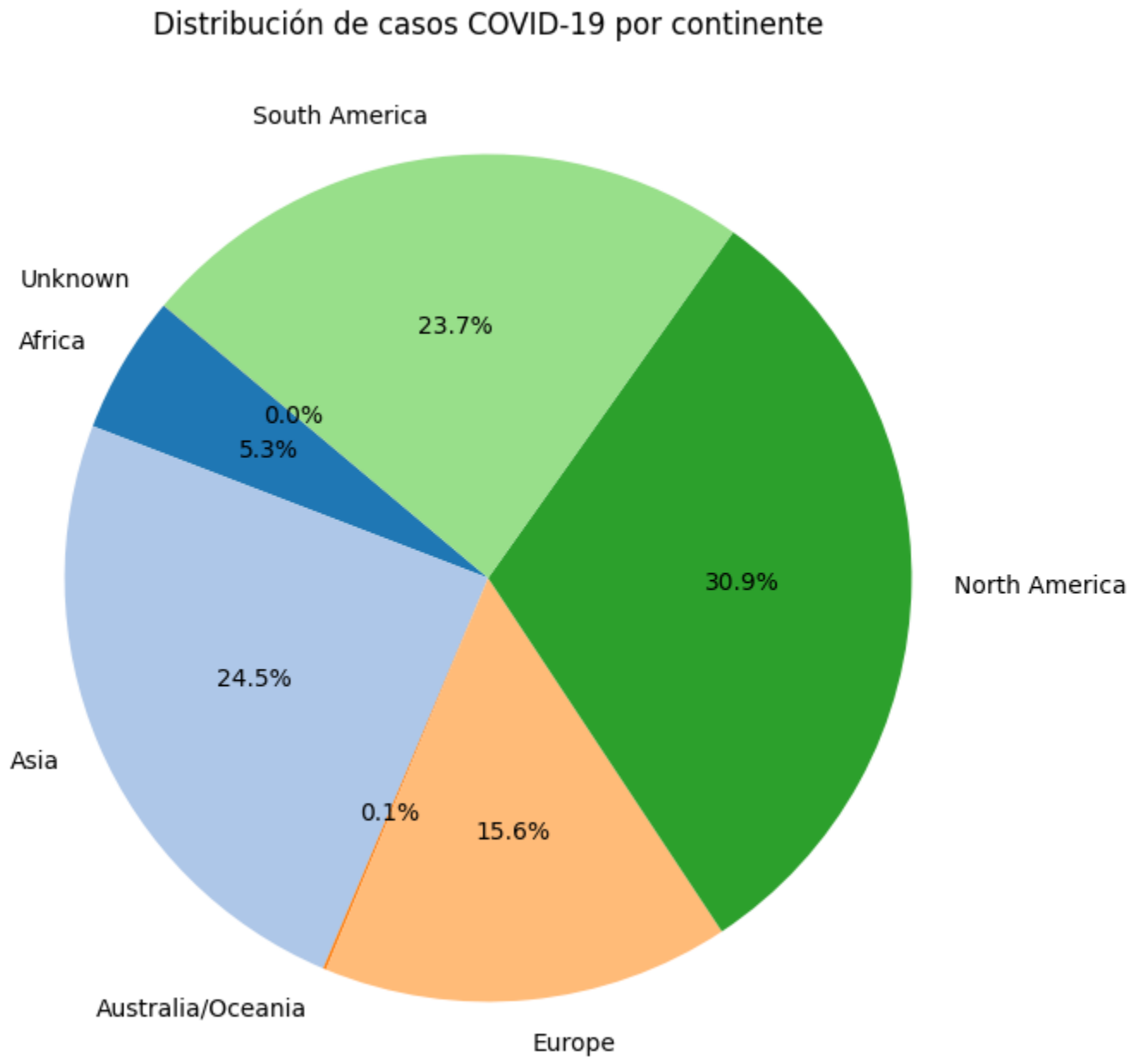
4.4 Nodo 4: Visualización de Datos

Función: Generación de visualizaciones informativas para análisis epidemiológico.

Gráficos generados:

1. **Gráfico de Pastel:** Distribución de casos de COVID-19 por continente
2. **Gráfico de Barras:** Top 10 países con mayor número de contagios

Almacenamiento: `data/04_models/` para fácil acceso y presentación.



5. Resultados

5.1 Estructura de Datos Generada

El pipeline genera automáticamente la siguiente estructura de archivos:

```
data/
├── 01_raw/                  # Datos originales de Kaggle
├── 02_intermediate/         # Reportes de calidad de datos
│   ├── missing_duplicate_country_wise_latest.csv
│   ├── missing_duplicate_covid_19_clean_complete.csv
│   ├── missing_duplicate_day_wise.csv
│   ├── missing_duplicate_full_grouped.csv
│   ├── missing_duplicate_usa_county_wise.csv
│   └── missing_duplicate_worldometer_data.csv
├── 03_primary/              # Datos limpios procesados
│   ├── covid_19_clean_complete_CLEAN.csv
│   ├── usa_county_wise_CLEAN.csv
│   └── worldometer_data_CLEAN.csv
└── 04_models/               # Visualizaciones generadas
    ├── casos_por_continente.png
    └── top_10_paises_contagios.png
```

5.2 Insights Epidemiológicos Obtenidos

A través del análisis automatizado, el pipeline reveló:

- **Distribución continental:** Identificación de continentes más afectados por COVID-19
- **Ranking de países:** Los 10 países con mayor impacto epidemiológico
- **Calidad de datos:** Identificación sistemática de gaps informativos en datasets públicos

5.3 Métricas de Calidad de Datos

El nodo de detección de anomalías proporcionó métricas detalladas sobre:

- Porcentaje de completitud por dataset
- Identificación de duplicados por fuente
- Distribución de valores faltantes por columna

---

6. Desafíos y Soluciones

6.1 Curva de Aprendizaje de Kedro

Desafío: La adopción inicial del framework Kedro presentó complejidades en términos de:

- Comprensión de la arquitectura de nodos
- Configuración de pipelines
- Gestión de dependencias entre nodos

Solución implementada:

- Estudio sistemático de la documentación oficial
- Implementación incremental nodo por nodo
- Pruebas iterativas del pipeline completo

6.2 Heterogeneidad de Datos

Desafío: Los datasets de COVID-19 presentaron estructuras y formatos diversos.

Solución: Implementación de lógica de procesamiento adaptativa que maneja diferentes esquemas de datos automáticamente.

---

7. Análisis y Conclusiones

7.1 Efectividad del Framework Kedro

Ventajas identificadas:

- **Reproducibilidad:** El pipeline puede ejecutarse consistentemente en diferentes entornos
- **Modularidad:** Cada nodo tiene responsabilidades claramente definidas
- **Escalabilidad:** Fácil adición de nuevos nodos de procesamiento
- **Mantenibilidad:** Código organizado y fácil de modificar

Desventajas observadas:

- **Curva de aprendizaje inicial elevada**
- **Overhead de configuración para proyectos simples**
- **Dependencia de convenciones específicas del framework**

7.2 Valor del Análisis Automatizado

El proyecto demostró la efectividad de automatizar procesos de análisis de datos, especialmente en:

- **Detección proactiva de problemas de calidad**
- **Generación consistente de reportes**
- **Reducción de errores manuales**
- **Escalabilidad para múltiples datasets**

7.3 Aplicabilidad Epidemiológica

Los resultados obtenidos proporcionan una base sólida para análisis epidemiológicos más profundos, incluyendo:

- Identificación de patrones geográficos
- Baseline para análisis predictivos
- Framework para monitoreo continuo

---

8. Recomendaciones y Trabajo Futuro

8.1 Mejoras Técnicas

- **Implementación de validación de datos** más robusta
- **Integración con herramientas de visualización avanzadas** (Plotly, Dash)
- **Desarrollo de nodos de Machine Learning predictivo**
- **Implementación de testing automático** del pipeline

8.2 Expansión Analítica

- **Análisis temporal detallado**de tendencias epidemiológicas
- **Correlación con factores socioeconómicos**
- **Implementación de modelos predictivos** (ARIMA, LSTM)
- **Dashboard interactivo**para stakeholders

---

9. Referencias y Fuentes

1. **Kedro Documentation.** (2025). *Kedro Framework Official Documentation*. QuantumBlack Labs.
2. **COVID-19 Dataset.** (2025). *COVID-19 Data Repository*. Kaggle.com.
3. **What is CRISP-DM?** (2024) Definición Crisp-DM. https://www.datascience-pm.com/crisp-dm-2/

---

10. Anexos

Anexo A: Configuración del Entorno

```bash
# Dependencias principales del proyecto
kedro==0.18.14
pandas==1.5.3
matplotlib==3.7.1
seaborn==0.12.2
numpy==1.24.3
```

Anexo B: Estructura del Proyecto

```
covid-analysis-kedro/
├── conf/                    # Configuración
├── data/                    # Datos (según estructura Kedro)
├── docs/                    # Documentación
├── logs/                    # Logs de ejecución
├── notebooks/               # Jupyter notebooks de análisis
├── src/covid19df/           # Código fuente
│   ├── pipelines/           # Definición de pipelines
│   └── nodes/               # Implementación de nodos
└── tests/                   # Tests unitarios
```

Anexo C: Comandos de Ejecución

```bash
# Ejecutar pipeline completo
kedro run

# Instalar dependencias
pip install -r
```
