🏥 Proyecto de Análisis de Datos de Salud

📋 Descripción del Proyecto

Este proyecto presenta un análisis completo de datos de salud con **100,000 registros ficticios** de pacientes, utilizando tecnologías modernas de análisis de datos como **Dask**, **Matplotlib**, **Seaborn** y **Plotly** para crear visualizaciones estáticas e interactivas.

🎯 Objetivos Cumplidos

✅ **Generación de Base de Datos Ficticia**: 100,000 registros realistas de salud
✅ **Análisis Exploratorio con Dask**: Procesamiento eficiente de grandes volúmenes de datos
✅ **Visualizaciones Estáticas**: Gráficos con Matplotlib y Seaborn
✅ **Visualizaciones Interactivas**: Dashboards con Plotly
✅ **Datos Estructurados**: Archivos CSV y JSON para reutilización

📊 Datos Generados

Base de Datos Principal

- **Archivo**: `data/health_data_100k.csv`
- **Registros**: 100,000 pacientes
- **Columnas**: 27 variables
- **Tamaño**: ~115 MB

Variables Incluidas

- **Demográficas**: Edad, género, provincia, fecha de nacimiento
- **Métricas de Salud**: IMC, presión arterial, frecuencia cardíaca, glucosa, colesterol, hemoglobina
- **Diagnósticos**: Diagnóstico principal, medicamento principal
- **Temporales**: Fechas de consultas, número de consultas
- **Económicas**: Costos de tratamiento, tipo de seguro médico
- **Estilo de Vida**: Actividad física, estado de fumador

🔍 Análisis Realizados

1. Estadísticas Descriptivas con Dask

- **Archivo**: `analysis/estadisticas_descriptivas.csv`
- Estadísticas completas de todas las variables numéricas
- Distribuciones categóricas
- Métricas de tendencia central y dispersión

2. Análisis Demográfico

- **Visualizaciones**: `analisis_demografico.png`, `demografico_interactivo.html`
- Distribución de edades por género
- Pirámide poblacional

- Distribución geográfica por provincias
- Análisis de seguros médicos

3. Análisis de Métricas de Salud

- **Visualizaciones**: `metricas_salud.png`, `metricas_3d.html`
- Matriz de correlaciones entre variables de salud
- Análisis de IMC por género
- Relaciones entre presión arterial, edad e IMC
- Visualización 3D interactiva

4. Análisis de Diagnósticos

- **Visualizaciones**: `diagnosticos_comunes.png`, `diagnosticos_sunburst.html`
- Top 15 diagnósticos más comunes
- Distribución por grupos de edad y género
- Análisis de costos por diagnóstico
- Número de consultas por diagnóstico

5. Análisis Temporal

- **Visualizaciones**: `tendencias_temporales.png`, `temporal_interactivo.html`
- Evolución de consultas por mes
- Análisis estacional
- Tendencias temporales por diagnóstico

📈 Visualizaciones Generadas

Gráficos Estáticos (PNG)

1. `analisis_demografico.png` - Análisis demográfico completo
2. `analisis_diagnosticos.png` - Análisis de diagnósticos por grupos
3. `diagnosticos_comunes.png` - Top diagnósticos más frecuentes
4. `metricas_salud.png` - Correlaciones y distribuciones de métricas
5. `tendencias_temporales.png` - Evolución temporal de consultas

Dashboards Interactivos (HTML)

1. `dashboard_completo.html` - **Dashboard principal con 9 visualizaciones**
2. `demografico_interactivo.html` - Análisis demográfico interactivo
3. `metricas_3d.html` - Visualización 3D de métricas de salud
4. `diagnosticos_sunburst.html` - Diagnósticos en formato sunburst
5. `temporal_interactivo.html` - Tendencias temporales interactivas

Gráficos Plotly Especializados (HTML)

1. `distribucion_edad_genero.html` - Distribución de edades por género
2. `correlaciones_heatmap.html` - Mapa de calor de correlaciones
3. `scatter_3d_metricas.html` - Scatter 3D de métricas principales
4. `treemap_costos.html` - Treemap de costos por diagnóstico y provincia
5. `violin_metricas.html` - Distribuciones violin por diagnóstico
6. `timeline_consultas.html` - Timeline de consultas médicas
7. `radar_perfiles.html` - Perfiles de salud por grupo de edad

📁 Estructura del Proyecto

```
health_analytics_project/
├── data/
│   ├── health_data_100k.csv           # Base de datos principal
│   ├── dashboard_data.json            # Datos para dashboard web
│   ├── edad_genero_data.json          # Datos edad/género
│   ├── correlaciones_data.json        # Matriz de correlaciones
│   ├── costos_data.json              # Datos de costos
│   ├── timeline_data.json            # Datos temporales
│   ├── radar_data.json               # Datos para gráfico radar
│   └── sunburst_data.json            # Datos para sunburst
├── analysis/
│   ├── estadisticas_descriptivas.csv  # Estadísticas con Dask
│   └── metricas_clave.json           # Métricas principales
├── visualizations/
│   ├── [5 archivos PNG]              # Gráficos estáticos
│   └── [13 archivos HTML]            # Dashboards interactivos
├── generate_health_data.py           # Generador de datos ficticios
├── health_analytics_dask.py          # Análisis principal con Dask
├── plotly_charts_generator.py        # Generador de gráficos Plotly
└── RESUMEN_PROYECTO.md              # Este documento
```

🔑 Métricas Clave del Dataset

- **Total de Pacientes**: 100,000
- **Edad Promedio**: 45.0 años
- **IMC Promedio**: 26.2
- **Costo Total de Tratamientos**: €75,003,901
- **Costo Promedio por Paciente**: €750
- **Total de Consultas**: 500,196
- **Consultas Promedio por Paciente**: 5.0
- **Diagnósticos Únicos**: 10
- **Provincias Cubiertas**: 24
- **Pacientes en Estado Crítico**: 1,140 (1.1%)
- **Porcentaje con Hipertensión**: 19.9%

🏆 Diagnósticos Más Comunes

1. **Hipertensión**: 19,934 pacientes (19.9%)
2. **Diabetes Tipo 2**: 14,836 pacientes (14.8%)
3. **Ansiedad**: 10,199 pacientes (10.2%)
4. **Obesidad**: 10,172 pacientes (10.2%)
5. **Artritis**: 9,704 pacientes (9.7%)

🛠️ Tecnologías Utilizadas

- **Python 3.11**: Lenguaje principal
- **Dask**: Procesamiento eficiente de grandes datasets
- **Pandas**: Manipulación de datos

- **NumPy**: Operaciones numéricas
- **Matplotlib**: Visualizaciones estáticas
- **Seaborn**: Visualizaciones estadísticas
- **Plotly**: Dashboards interactivos
- **Faker**: Generación de datos ficticios

🎨 Características Destacadas

Análisis con Dask

- Procesamiento eficiente de 100,000 registros
- Estadísticas descriptivas optimizadas
- Manejo de memoria eficiente

Visualizaciones Interactivas

- Gráficos 3D navegables
- Dashboards con múltiples vistas
- Filtros y zoom interactivos
- Tooltips informativos

Datos Realistas

- Correlaciones médicas realistas
- Distribuciones demográficas coherentes
- Costos de tratamiento variables
- Fechas y temporalidad consistente

🚀 Uso del Proyecto

Para Análisis de Datos

1. Cargar `health_data_100k.csv` en cualquier herramienta de análisis
2. Utilizar los archivos JSON para análisis específicos
3. Reproducir análisis ejecutando los scripts Python

Para Dashboards Web

1. Abrir cualquier archivo HTML en navegador
2. Utilizar `dashboard_completo.html` como punto de entrada
3. Integrar datos JSON en aplicaciones web personalizadas

Para Desarrollo

1. Modificar scripts para análisis adicionales
2. Extender visualizaciones con nuevas métricas
3. Adaptar generación de datos para otros casos de uso

📊 Insights Principales

1. **Correlación Edad-Salud**: Clara correlación entre edad y métricas como presión arterial y colesterol
2. **Distribución de Género**: Ligero predominio femenino (52% vs 48%)

3. **Costos Variables**: Los costos varían significativamente por diagnóstico y edad
4. **Patrones Temporales**: Variaciones estacionales en consultas médicas
5. **Riesgo Cardiovascular**: Identificación de grupos de alto riesgo

🔮 Posibles Extensiones

- **Machine Learning**: Modelos predictivos de riesgo
- **Dashboard Web Completo**: Aplicación web interactiva
- **Análisis Geoespacial**: Mapas de distribución de enfermedades
- **Análisis de Series Temporales**: Predicción de tendencias
- **Clustering**: Segmentación de pacientes por perfiles de salud

---

Proyecto completado exitosamente ✅
Fecha: Julio 2025
Autor: Yefferson josue rodriguez castillo