**Guion de Video (Duración total: ~7 minutos)**

**1. Carátula y Bienvenida (Juan Carlos Garcés)**

|  |  |
| --- | --- |
| **Tiempo Estimado** | **0:00 – 0:30 (30 segundos)** |
| **Guion** | Buenos días/tardes. Somos el equipo integrado por Fernando Salazar, Edén Mariscal, Daniel Olivares, Adrián Alejandro Montiel y un servidor, Juan Carlos Garcés. Presentamos la **Fase 1: Manipulación y Preparación de Datos** de nuestro proyecto: **Implementación de MLOps para la Predicción de Niveles de Obesidad**. |
| **Visuales** | *Mostrar Diapositiva 1 (Carátula).* |

**2. Contexto del Problema (Fernando Salazar)**

|  |  |
| --- | --- |
| **Tiempo Estimado** | **0:30 – 1:15 (45 segundos)** |
| **Guion** | El proyecto aborda la obesidad, considerada una de las **principales causas de enfermedades crónicas no transmisibles** (OMS, 2023). La urgencia es clara: en Latinoamérica, **más del 60% de la población adulta presenta sobrepeso u obesidad** (Organización Panamericana de la Salud, 2022). Nuestro trabajo se basa en el *dataset* de estimación de niveles de obesidad, que incluye datos de individuos de México, Perú y Colombia, enfocados en hábitos alimenticios y condición física. |
| **Visuales** | *Mostrar Diapositiva 2 (Contexto). Mostrar gráficos de estadísticas globales/regionales si están disponibles en las fuentes.* |

**3. Objetivo y Enfoque MLOps (Edén Mariscal)**

|  |  |
| --- | --- |
| **Tiempo Estimado** | **1:15 – 2:00 (45 segundos)** |
| **Guion** | Nuestro objetivo es aplicar la metodología MLOps para construir un **modelo predictivo** que clasifique el nivel de obesidad de un individuo en una de las siete categorías posibles, como *Normal Weight* u *Obesity Type I, II, o III*. Buscamos que este modelo sirva para la salud preventiva. La clave es el enfoque MLOps: garantizar la **reproducibilidad y la trazabilidad** en cada etapa, desde el análisis de requerimientos (utilizando el ML Canvas) hasta la preparación final de los datos. |
| **Visuales** | *Mostrar Diapositiva 3 (Objetivo). Resaltar las palabras clave: MLOps, Predictivo, Trazabilidad.* |

**4. Actividades por Rol: Data Engineer (Adrián Alejandro Montiel Ramírez)**

|  |  |
| --- | --- |
| **Tiempo Estimado** | **2:00 – 3:00 (60 segundos)** |
| **Guion** | Como Data Engineer, mi enfoque fue asegurar la infraestructura inicial. Esto incluyó la configuración del proyecto y la **ingesta de datos** del archivo CSV. El dataset inicial constaba de **2,153 filas y 18 columnas**, todas con tipo de dato *object* y con presencia de valores nulos. La tarea crucial fue aplicar **DVC (Data Version Control)** para registrar las versiones tanto del dataset crudo como del limpio. Esto garantiza que los demás roles trabajen con versiones estables y trazables, permitiendo la reproducibilidad de los experimentos. |
| **Visuales** | *Mostrar Diapositiva 4 (Data Engineer). Mostrar captura de pantalla del código de importación de librerías y configuración de ruta raíz (sys.path).* |

**5. Actividades por Rol: Data Scientist (Daniel Olivares)**

|  |  |
| --- | --- |
| **Tiempo Estimado** | **3:00 – 4:00 (60 segundos)** |
| **Guion** | Mi rol se centró en el Análisis Exploratorio de Datos (EDA) y la limpieza profunda. Encontramos una fila duplicada y nulos en casi todas las 18 columnas, siendo la columna mixed\_type\_col la más afectada con **241 nulos**. Realizamos la estandarización de variables categóricas para manejar inconsistencias (como el uso de mayúsculas/minúsculas en NObeyesdad). Además, la eliminación de valores atípicos fue clave, donde aplicamos reglas de negocio y el método IQR, resultando en la eliminación de **168 filas, un 7.95%** del total inicial. Finalmente, como parte de la ingeniería de características, creamos el **Índice de Masa Corporal (IMC)**, una variable fundamental para la predicción. |
| **Visuales** | *Mostrar Diapositiva 5 (Data Scientist). Resaltar el porcentaje de datos eliminados (7.95%) y la creación del IMC.* |

**6. Actividades por Rol: ML Engineer (Edén Mariscal)**

|  |  |
| --- | --- |
| **Tiempo Estimado** | **4:00 – 5:00 (60 segundos)** |
| **Guion** | Como ML Engineer, me aseguré de que la preparación de datos fuera robusta para el modelado. Esto se logra mediante la construcción de **Pipelines de Scikit-learn**. Este enfoque garantiza un tratamiento de datos consistente y previene la fuga de datos (*data leakage*). Definimos un ColumnTransformer que aplica diferentes tratamientos: imputación por mediana y estandarización con StandardScaler para variables numéricas (incluyendo el nuevo IMC), y codificación OneHotEncoder e imputación por moda para las categóricas. Finalmente, dividimos el conjunto limpio en **1,556 muestras para entrenamiento y 390 para prueba**. |
| **Visuales** | *Mostrar Diapositiva 6 (ML Engineer). Mostrar una representación esquemática del Pipeline o el resumen del ColumnTransformer.* |

**7. Actividades por Rol: Software Engineer (Fernando Salazar)**

|  |  |
| --- | --- |
| **Tiempo Estimado** | **5:00 – 5:45 (45 segundos)** |
| **Guion** | Mi función fue dar soporte de código, asegurando la **modularización** y la organización. Configuramos la estructura de directorios, asegurando que las funciones como cargar\_dataframe o calcular\_imc fueran invocables desde la carpeta de código fuente (src). Esto facilita la colaboración y la escalabilidad. Además, fui responsable de la lógica para el **guardado de artefactos**, asegurando que los conjuntos de datos limpios y listos para el modelado se almacenaran correctamente con sus respectivos nombres de archivo, como obesity\_estimation\_cleaned.csv y obesity\_estimation\_model.csv. |
| **Visuales** | *Mostrar Diapositiva 7 (Software Engineer). Mostrar la estructura de carpetas o una referencia a la colaboración en GitHub.* |

**8. Síntesis MLOps: Trazabilidad (Adrián Alejandro Montiel Ramírez )**

|  |  |
| --- | --- |
| **Tiempo Estimado** | **5:45 – 6:30 (45 segundos)** |
| **Guion** | La aplicación de MLOps en esta fase garantiza dos pilares fundamentales: **trazabilidad y reproducibilidad**. Gracias a DVC, tenemos un registro inmutable de la versión de los datos que usamos. El uso de **Pipelines** de Scikit-learn asegura que las transformaciones se apliquen de forma idéntica en el futuro y que los resultados obtenidos sean consistentes. Toda la estrategia del proyecto, desde la propuesta de valor hasta los requerimientos de la solución de ML, fue documentada utilizando el *framework* **ML Canvas**. |
| **Visuales** | *Mostrar Diapositiva 8 (MLOps). Enfatizar los logos de DVC y Scikit-learn.* |

**9. Conclusiones (Juan Carlos Garcés)**

|  |  |
| --- | --- |
| **Tiempo Estimado** | **6:30 – 7:00 (30 segundos)** |
| **Guion** | En resumen, hemos aprendido que la **calidad de los datos** es el factor más crítico. Los retos principales fueron lidiar con los tipos de datos ambiguos y la alta presencia de atípicos que requirieron limpieza rigurosa. Como recomendación, continuaremos extendiendo la trazabilidad, aplicando DVC al modelado y a las métricas para la próxima fase, garantizando así un desarrollo de ML completamente reproducible y auditable. Gracias. |
| **Visuales** | *Mostrar Diapositiva 9 (Conclusiones).* |