**Asunto: Inicio del Proyecto "Análisis de Supervivencia del Titanic"**

**Para:** Nuestro Científico de Datos

**De:** La Dirección de Proyectos

¡Bienvenido al equipo!

Como sabes, nuestra empresa está produciendo el documental definitivo sobre el hundimiento del Titanic. Queremos ir más allá de las anécdotas y usar la ciencia de datos para descubrir los factores que realmente determinaron la supervivencia de los pasajeros.

Nuestro objetivo principal es responder a la pregunta: **"¿Fue la supervivencia en el Titanic una cuestión de suerte, o hubo factores claros y predecibles (clase social, género, edad, etc.) que determinaron quién vivió y quién murió?"**

Tus análisis serán la columna vertebral de nuestro guion y del material visual para una exhibición interactiva. Confiamos en tu expertise para convertir estos datos crudos en una historia clara y veraz.

Aquí tienes tus primeras tareas, divididas en fases que nos permitirán avanzar de manera ordenada.

**Fase 1: Auditoría y Limpieza de Datos (Relacionado con 1. manejo\_datos\_nulos\_duplicados.ipynb)**

**Objetivo:** Antes de cualquier análisis, debemos asegurarnos de que nuestros datos son confiables. No podemos basar una producción de millones de dólares en datos erróneos.

**Tus Tareas:**

1. **Análisis Inicial**: Carga el archivo train.csv. Quiero un primer informe que incluya:
   * Las dimensiones del dataset (filas y columnas).
   * Un resumen de los tipos de datos de cada columna (.info()).
   * Un conteo exacto de los valores nulos para cada columna.
2. **Estrategia para la Columna Age (Edad)**: Esta columna es vital para nuestra narrativa. Tiene valores nulos.
   * **Acción Requerida**: Rellena los valores faltantes. Quiero que me justifiques en una celda de Markdown por qué eliges rellenarlos con la **media** o la **mediana**. (Pista: Revisa si hay outliers o cómo se distribuyen los datos de edad).
3. **Estrategia para la Columna Cabin (Camarote)**: Vemos que tiene muchísimos valores nulos.
   * **Acción Requerida**: Eliminar la columna no es ideal, ya que tener o no un camarote asignado podría ser un indicador de clase social. Propongo que transformes esta columna en una nueva llamada Tiene\_Camarote, donde el valor sea 1 si había un dato de camarote y 0 si era nulo.
4. **Estrategia para la Columna Embarked (Puerto de Embarque)**: Hay dos valores nulos aquí.
   * **Acción Requerida**: Rellena estos dos valores faltantes. Como es una columna categórica, la media o mediana no sirven. Investiga y aplica la estrategia más lógica (Pista: la **moda** suele ser una buena opción).

**Fase 2: Preparación de Datos para el Análisis (Relacionado con 2. transformacion\_datos\_normalizacion\_codificacion.ipynb)**

**Objetivo:** Ahora que los datos están limpios, debemos "traducirlos" a un formato numérico que los futuros modelos de IA puedan procesar.

**Tus Tareas:**

1. **Codificación de Variables Categóricas**: Las columnas Sex, Embarked y la nueva Tiene\_Camarote son texto o categorías.
   * **Acción Requerida**: Conviértelas en números. Te recomiendo usar **One-Hot Encoding**. En tu notebook, explica brevemente por qué esta técnica es preferible a simplemente asignar números como 0, 1, 2.
2. **Escalado de Variables Numéricas**: Las columnas como Age y Fare (Tarifa) se miden en escalas muy diferentes.
   * **Acción Requerida**: Aplica un escalado para estandarizarlas. Usa el StandardScaler que vimos en los ejemplos.
3. **Ingeniería de Características (Un pequeño reto)**: Nuestros historiadores creen que viajar solo o con familia era un factor clave.
   * **Acción Requerida**: Crea una nueva columna llamada Tamaño\_Familia. Su valor debe ser la suma de SibSp (hermanos/cónyuges) + Parch (padres/hijos) + 1 (para contarse a sí mismo).

**Fase 3: Creación de un Pipeline de Preprocesamiento (Relacionado con 3. pipelines\_preprocesamiento.ipynb)**

**Objetivo:** Tu trabajo hasta ahora es excelente, pero es manual. Necesitamos un proceso automatizado y profesional que podamos reutilizar.

**Tu Tarea:**

1. **Construir el Pipeline de Preprocesamiento**: Usando Pipeline y ColumnTransformer de Scikit-learn, crea un único objeto preprocessor que ejecute todas las acciones de las Fases 1 y 2 de forma automática (rellenar nulos de Age y Embarked, codificar las variables categóricas, y escalar las numéricas). Este es el paso más avanzado, ¡demuéstranos tu habilidad!

**Entregables Solicitados para esta Etapa**

Para esta primera entrega, necesito dos cosas, como acordamos:

1. **Un Notebook de Jupyter (.ipynb)**: Debe contener todo tu código, paso a paso, para las tres fases. Utiliza celdas de Markdown para explicar tus decisiones, especialmente donde te lo he pedido.
2. **Una Presentación Corta (3 Diapositivas)**:
   * **Diapositiva 1: El Estado de los Datos.** Un resumen de los problemas que encontraste (cuántos nulos, qué tipos de datos, etc.).
   * **Diapositiva 2: Estrategia de Limpieza.** Explica brevemente las decisiones clave que tomaste para limpiar los datos (ej. "Se imputó la edad con la mediana para evitar el efecto de outliers...").
   * **Diapositiva 3: Vista Previa de los Datos Preparados.** Muestra una pequeña tabla con el "antes" y el "después" de una fila de datos, para que los productores del documental puedan ver la transformación.

Estamos ansiosos por ver tus primeros resultados. Este análisis es crucial para el éxito del proyecto.

Atentamente,

**La Dirección** **Proyecto "Análisis de Supervivencia del Titanic"**