1. Definición del problema
2. Exploración inicial de las variables
3. Separación en train y test
4. Tratamiento de missing, outlier y correlaciones
5. Codificación de las variables categóricas
6. Escalado de los datos (si es necesario)
7. Selección de variables input del modelo (eliminación de colinealidad si es necesario)

Referencia: .html 01_exploracion_general

1. **Definición del problema a resolver**
   - ¿Cuál es el problema?
   - Acción que buscamos hacer para solucionar el problema
   - ¿Cuáles son las variables disponibles?
   - ¿En qué momento se va a implantar el modelo? ¿Qué variables hay disponibles en el momento de llamada al modelo?
   - ¿Cómo se va a validar el modelo?

2. **Exploración general / inicial**
   - Dimensiones de la tabla y variables
   - Analizar si las variables estarán disponibles en el momento de la llamada al modelo (sino se estaría introduciendo información a futuro en el modelo)
   - Exploración de la variable objetivo
   - Rápido análisis de valores nulos
   - Se explora el número de variables numéricas y categóricas y se decide qué proceso realizar para tratarlas
   - Transformaciones iniciales de algunas variables: formato de fechas, eliminar espacios de una variable string, etc
   - **Analizar la variable objetivo:** aislándola de las demás variables (quizás, a pesar de no tener en cuenta las variables tenemos que sacar un 80% de accuracy.

3. **Separación entre train y test**
   - Se hace la separación antes de realizar transformaciones de cálculos de la columna
   - Onehotencoding en vez de getdummies
   - Los valores missing y outlier se deben sustituir después de haber dividido en train y test
   - **Ejemplo:** Para imputar los valores missing por la media, los pasos a realizar son:
   - Obtener la media del conjunto de train
   - Reemplazar los valores missing por la media obtenida en la muestra de train.

4. **Xd**

5. **Tratamiento de missing**
    - Reemplazo por media
    - Reemplazo por valores significantes
    - Eliminacion de filas
    - **<u>Notebook</u>: 02_Tratamiento_correlaciones_missing_outlier.html**
      <u>Buscar</u>: Distribución del resto de variables (¡¡¡INCLUIR EN ANÁLISIS DESCRIPTIVO!!!)
    - Separación entre variables categóricas y variables continuas (en las variables categóricas no hay outliers)
    - Tratamiento de las variables continuas y análisis de estas
    - Tratamiento de las variables categóricas y análisis de estas (V-cramer, correlación de spearman…)

6. **Codificación de variables categóricas**
    - Label Encoding (para la variable objetivo (no lo usamos en esta práctica))
    - **OneHotEncoder**
    - **Target Encoding**
    - Librería category_encoders
    - OrdinalEncoder
    - Frequency/Count Encoder
    - **Mean / Target Encoding** (variables con muchas dimensiones)
    - **CatBoost Encoder** ()
    - HTML: **ejemplos_encoding_variables_categoricas.html**
    -

7.