**UNIVERSIDAD DE PALERMO**

**Comprendiendo el preprocesamiento de datos Caso Titanic**

Actividad Individual

Alumno:

Juan Cruz Rey

Materia:

Aprendizaje de Máquina  
Modalidad Online



**1) Explicar brevemente el dataset Titanic: features que lo componen, label, si es que tiene, tipos de datos de los features y de un ejemplo de tipo de aprendizaje que se podría utilizar para resolver un problema usando este dataset. Describir qué problema resolvería.**

Feature que componen al dataset:

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre de la feature** | **Tipo de dato** | **Descripción (opcional)** | **Ejemplos / Valores posibles (opcional)** |
| PassengerId | Entero (int) | ID único de cada pasajero | 1,2,3 |
| Pclass | Entero (int) | Clase del ticket que indica la clase social de la persona abordo | 1,2,3 |
| Name | Texto (object) | Nombre completo del pasajero que incluye título social | “Mr. Owen Harris” |
| Sex | Texto (object) | Género del pasajero | "male", "female" |
| Age | Numérico (float) | Edad del pasajero | 38.0, 26.0 |
| SibSp | Entero (int) | Nº de hermanos/esposo(a) que viajan con la persona abordo | 0,1,2,3,4… |
| Parch | Entero (int) | Nº de padres/hijos que viajan con la persona abordo | 0, 1, 2, 3... |
| Ticket | Texto (object) | Número del boleto | "A/5 21171", "PC 17599" |
| Fare | Numérico (float) | Precio pagado por el pasaje | 71.2833, 8.05 |
| Cabin | Numérico (float) | Indicaba la cabina de la persona a bordo | B42, C148 |
| Embarked | Texto(object) | Puerto desde donde embarcó | 'S', 'C', 'Q' |
| Survived | Entero (int) | Label que indica si sobrevivió o no | 1 = sobrevivió, 0 = falleció |

Podríamos utilizar ***aprendizaje supervisado*** con el objetivo de predecir el valor de la columna Survived, y así estimar la supervivencia de un pasajero que no pertenezca a este conjunto de datos.

**2) Enumerar y explicar brevemente los métodos utilizados para el preprocesamiento del Dataset. Observar:**

**a) ¿Qué features se seleccionaron? ¿Cuáles se descartaron? ¿Por qué?**

**b) ¿Qué features nuevos se crearon?**

**c) ¿Había features con valores faltantes? ¿Cuáles? ¿Cómo se trataron?**

**d) ¿Había features categóricos? ¿Cuáles? ¿Cómo se trataron?**

Los métodos utilizados para el pre-procesamiento son los siguientes:

|  |  |
| --- | --- |
| **método** | **Explicación** |
| pd.read\_csv() | Lee el dataset en formato CSV |
| \_.head() | Usada para visualizar las primeras filas del DataFrame |
| \_.tail() | Usada para visualizar las últimas filas del DataFrame |
| \_.shape() | Permite observar el tamaño del dataframe dándonos sus filas y columnas |
| \_.dtypes() | Permite conocer los tipos de cada feature |
| \_.describe() | Permite analizar estadísticamente los features numéricos |
| .isnull().sum() | Nos permite conocer que feature tienen valores faltantes |
| \_ .set\_index() | Permite elegir una feature como índice del dataframe |

**2) a) ¿Qué features se seleccionaron? ¿Cuáles se descartaron? ¿Por qué?**

Feature seleccionados:

|  |  |
| --- | --- |
| **Feature** | **¿Por qué se seleccionaron?** |
| Pclass | La clase social/económica estaba fuertemente asociada a mayores probabilidades de sobrevivir. A mayor clase social, mayor prioridad se tenía |
| Sex | Se priorizaba a mujeres y niños, por lo cual el género impactaba en las probabilidades de supervivencia |
| Age | Los niños tenían prioridad |
| SibSp | Al parecer el tamaño y la composición familiar influía en la evacuación |
| Parch | Tener hijos o padres cerca podría influir en decisiones de evacuación y supervivencia conjunta. |
| Fare | Si la tarifa era mayor, entonces era más probable que seas de clase mayor y por lo tanto, aumentaba las posibilidades. |
| Embarked | El punto de embarque afectó las posibilidades de supervivencia |
| Survived | Feature que queremos predecir |

Feature descartados:

|  |  |
| --- | --- |
| **Feature** | **¿Por qué se descartaron?** |
| Ticket | No aporta información relevante para predecir la supervivencia |
| Cabin | Al tener tantos valores faltantes su información no es fiable |
| Name | Se utiliza para extraer el título del nombre y luego se elimina ya que no es útil para la predicción |
| Sex | Se elimina tras codificación por one-hot encoding |
| Embarked | Se elimina tras codificación por one-hot encoding |
| Title | Se elimina por redundancia |
| SibSp | Se combina con Parch para crear FamilySize y se elimina por estar correlacionada a Ella |
| Parch | Se combina con SibSp para crear FamilySize y se elimina por estar correlacionada a Ella |

**2) b) ¿Qué features nuevos se crearon?**

|  |  |
| --- | --- |
| **Feature** | **Motivo de creación** |
| FamilySize | Se crea debido a que el tamaño de la familia a bordo influía en las posibilidades de supervivencia |
| Title | Se utiliza para conocer las jerarquías sociales, recordando que a mayor clase social, más posibilidades de supervivencia. |

**2) c) ¿Había features con valores faltantes? ¿Cuáles? ¿Cómo se trataron?**

|  |  |
| --- | --- |
| **Feature** |  |
| Age | Tenía datos faltantes que se rellenan con la edad promedio de los pasajeros que NO sobrevivieron. |
| Embarked | Se reemplazan sus 2 valores faltantes por el valor más frecuente: 'S' (Southampton). |
| Cabin | Se elimina por completo debido a su cantidad de valores nulos. |

**2) d) ¿Había features categóricos? ¿Cuáles? ¿Cómo se trataron?**

|  |  |
| --- | --- |
| **Feature** |  |
| Sex | Se convierte a numéricos mediante one-hot encoding |
| Embarked | Se convierte a numéricos mediante one-hot encoding |
| Title | Se convierte a numéricos mediante one-hot encoding |
| Pclass | Es un feature categórico, aunque sus valores ya estaban codificados ordinalmente por lo que no se trataron |