

Instituto Tecnológico y de Estudios Superiores de Monterrey



**Tecnológico
de Monterrey**

**Inteligencia artificial avanzada para la ciencia de datos I
(Gpo 101)**

Equipo 4

**Momento de Retroalimentación: Reto Limpieza del
Conjunto de Datos**

Para este proyecto, fue necesario realizar una extensa limpieza de datos para poder encontrar la solución a la problemática del pronóstico de supervivencia de los pasajeros del Titanic. Después de un análisis exhaustivo, el equipo decidió que los atributos más pertinentes para el análisis y uso posterior serían: Sex, Age, P Class, Sibsp, y Parch.

Sex: Indica el género del pasajero, distinguiendo entre hombres y mujeres. (Género)

Age: Representa la edad de los pasajeros. (En años)

P Class: Identifica la clase en la que viajaba el pasajero, ya sea primera, segunda, o tercera clase. (1 = 1 clase, 2 = 2 clase, 3 = 3 clase)

Subsp: Informa sobre la cantidad de hermanos o esposos con los que viajaba el pasajero. (# de hermanos o esposos)

Parche: Indica si el pasajero estaba acompañado de algún padre o hijo. (# de padres o hijos)

Estos atributos fueron seleccionados por su relevancia para determinar las relaciones familiares de los pasajeros y su posible impacto en la probabilidad de supervivencia, esto debido al contexto del proyecto.

Estamos quitando la columna de cabina, embarcación, fare y ticket ya que estos datos realmente no son pertinentes para el desarrollo de nuestro pronóstico dentro del proyecto

Para la transformación de los datos del proyecto, se utilizaron tres maneras para llenar los datos faltantes;

- Regresión Lineal y de Predicción, ayuda a obtener una estimación precisa de la edad media utilizando variables relacionadas como el género, la clase de pasajero y el tamaño de la familia. La regresión lineal permite modelar la relación entre la edad y otros factores, proporcionando valores estimados que rellenan los datos faltantes de manera informada.

- En base a parámetros definidos por los datos que analizamos, siendo estos más que nada obtenidos por los índices de supervivencia de distintos atributos tales como tamaño de familia y género, distribución de clases por grupo de edad, supervivencia por número de hermanos y esposos, etc.

- Vecinos media busca rellenas los datos faltantes comparando cada entrada con los registros similares en el dataset, donde se obtienen los demás datos cercanos

cercanos en el espacio de características para estimar el valor faltante. Al encontrar las instancias más similares (vecinos) y promediar sus valores, se obtiene un resultado que toma en cuenta esas variables.