Instituto Tecnológico y de Estudios Superiores de Monterrey

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

Equipo 4

Momento de Retroalimentación: Reto Limpieza del Conjunto de Datos

Integrantes:

Eliezer Cavazos Rochin A00835194

Facundo Colasurdo Caldironi A01198015

Saul Francisco Vázquez A01198261

José Carlos Sänchez Gómez A01174050

Campus Monterrey **Fecha: 19/8/2024**

Limpieza 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: PassengerId, Sex, Age, PClass, Sibsp, y Parch.

PassengerId: Indica el Id del pasajero en la base de datos

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

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

PClass: 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)

Parch: 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 Cabin, Embarked, Fare y Ticket ya que estos datos realmente no son necesarios para el desarrollo de nuestro análisis de predicción de supervivencia.

Transformacion de Datos

Para la transformación de los datos del proyecto, se utilizaron tres maneras para llenar los datos faltantes del campo de **Edad** para ver cual genera la mejor prediccion.

-**Regresión Lineal**, ayuda a obtener una estimación precisa de la edad media utilizando variables relacionadas como el género, la clase social del 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 analizamo**s, 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:** busca rellenar los datos faltantes comparando cada entrada con los registros similares en el dataset, donde se obtiene los demás datos 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.