Introducción al machine learning

Actividad 1

Situación

El RMS Titanic fue, en su momento, el mayor barco de pasajeros del mundo. Se hundió en su viaje inaugural de Southampton a Nueva York, en el año 1912. En el evento, fallecieron 1514 de las 2223 personas que iban a bordo, entre tripulación y pasajeros.

Hoy, en el año 2022, se ha decidido hacer un estudio de *machine learning* en el cual se requiere de nuestras labores como técnicos en Data Science, para identificar diversos patrones que permitan verificar si, basándonos en el entrenamiento de nuestro modelo de datos, la máquina es capaz de predecir si una persona logra sobrevivir o no.

Para ello, se utilizará una base muestral de 1220 datos, que forman parte del conjunto de datos, que figura en el archivo adjunto “Titanic_M1.xlsx”.

Las variables del conjunto de datos son:

Variables	Descripción
passengerId	- int, valor de identificación único de cada pasajero
name	- string, que hace referencia al nombre del pasajero
sex	- factor, con niveles (masculino y femenino)
age	- numeric, valor que se refiere a la edad de una persona determinada. La edad de los niños menores de 12 meses es dad en fracción de un año (1/mes)
class	- factor, especifica la clase para cada pasajero (tipo de servicio a bordo)
embarked	- factor, hace referencia al lugar de embarcamiento (puerto de embarque de las personas)
ticketno	- numeric, especifica el número de ticket (na para la tripulación)
fare	- numeric, valor con el precio del ticket (na para la tripulación, musicos, empleados y otros)
sibsp	- factor ordenado, especifica el número de hermanos/familiares
cabin	- factor, tipo de cabina que ocupa cada pasajero
parch	- factor ordenado, especifica el número de padres e hijos a bordo
survived	- factor 2 de dos niveles, que especifica (sí o no) la persona ha sobrevivido al hundimiento

Actividad Práctica Integradora

Consignas

Para realizar los futuros modelos de datos, el primer desafío será preprocesar la base de datos, según las siguientes instrucciones:

a) Describir brevemente la cantidad de valores faltantes para cada una de las variables de la base de datos.

b) En la columna Pclass, completar los valores NA, asignando a las personas a la clase 2.

c) En la clase “Sex”, imputar los datos de manera de poder completar la base solo con los valores “male” y “female”.

d) En los casos de las personas de las cuales no se tenga la edad, rellenar los datos con el promedio de edad de aquellos que sí tienen el dato.

e) En la columna embarked, asignar el valor “S” a aquellos datos que se encuentran con valor vacío.

Formato de entrega:

¿Cómo se debe presentar el trabajo?

La entrega debe ser mediante un archivo “ipynb” (formato de extensión de *notebook* para Python).

Cada una de las preguntas debe contener los códigos de cómo se lograron resolver las consignas solicitadas.

Entrega

¡Llegaste al final de la actividad de este módulo! Recuerda guardar tus respuestas y luego subirlas clicando en el botón "Enviar tarea".

Puedes consultar tus dudas con tus compañeros en el foro de la materia o con tu tutor.