Introducción al machine learning

Actividad 2

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 utilizarán varias bases para llevar a cabo nuestro análisis: las primeras denominadas “train.csv” y “test.csv”, que pertenecen a un set de datos de entrenamiento; y otra para testear nuestra información.

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

El desafío actual consiste en generar un modelo de regresión logística que permita calcular el grado de *accuracy* con el cual se pueda determinar, en el grupo de "prueba", quiénes son los sobrevivientes en la tragedia del Titanic.

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

b) Completar aquellas variables que se encuentran faltantes para las bases de *train* (“Age”,”fare”, “Cabin”) y test (“Age”, “Cabin”, “embarked”).

c) Ajustar el primer modelo de regresión logística.

d) Entrenar y determinar el nivel de *accuracy* del primer modelo.

Formatos 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.