Práctica 2: Limpieza y validación de los datos

Ismael Rosende Rey

26/12/2020

Índice

1. Descripción del dataset

Se escoge el dataset sugerido en el enunciado de la competición de Kaggle basada en el hundimiento del titanic. Este dataset se divide en un csv para entrenar el modelo que contiene la información completa de los pasajeros, indicando si han sobrevivido o no, y otro csv para test que omite la información de supervivencia.

Esta competición consiste en preparar un modelo que predice si un pasajero ha sobrevivido o ha muerto en base a la información disponible. En los datos que nos proporcionan, disponemos de los siguientes campos acerca de los pasajeros:

- **Survival (0=No, 1=Yes)**: Indica si ha sobrevivido.
- **Pclass (1=1st, 2=2nd, 3=3rd)**: Clase en la que viajaba.
- **Sex**: Si se trataba de un hombre o mujer.
- **Age**: Edad en años.
- **Sibsp**: Número de hermanos / esposas abordo del barco.
- **Parch**: Número de padres / niños abordo del barco.
- **Ticket**: Número del ticket.
- **Fare**: Tarifa que ha pagado por el ticket.
- **Cabin**: Número de cabina.
- **Embarked (C=Cherbourg, Q=Queenstown, S=Southampton)**: Puerto de embarcación.

```
train_data <- read.csv("./data/train.csv", stringsAsFactors = TRUE)
test_data <- read.csv("./data/test.csv", stringsAsFactors = TRUE)
```

2. Integración y selección de datos

3. Limpieza de datos

4. Análisis de los datos

5. Resultados

6. Conclusiones

Contribuciones	Firma
Investigación previa	isrosrey
Redacción de las respuestas	isrosrey
Desarrollo código	isrosrey