Taller 2: Aprendizaje Supervisado -Random forests.

Docente: Nicolás Abuhadba.

Objetivo

- Predecir si un pasajero *sobrevivió al hundimiento del Titanic* o no, usando los algoritmos de *aprendizaje supervisado (Random forests o AdaBoost*).
- Para cada uno en el conjunto de prueba (test), debe predecir un valor de 0 o 1 para la variable (*survival*).

Tipo de actividad	Taller grupal
Taller	*Random forests*
Evaluación	Formativa
Tiempo estimado de entrega	24/10/2021

Actividades para desarrollar:

1. Breve contexto

El hundimiento del Titanic es uno de los naufragios más conocidos de la historia.

El 15 de abril de 1912, durante su viaje inaugural, el RMS Titanic, ampliamente considerado "insumergible", se hundió después de chocar con un iceberg. Desafortunadamente, no había suficientes botes salvavidas para todos a bordo, lo que resultó en la muerte de 1502 de los 2.224 pasajeros y la tripulación.

Si bien hubo algún elemento de suerte involucrado en sobrevivir, parece que algunos grupos de personas tenían más probabilidades de sobrevivir que otros.

- En este taller, usted deberá crear un modelo predictivo que responda a la pregunta: "¿Qué tipo de personas tenían más probabilidades de sobrevivir?" utilizando datos de pasajeros (es decir, nombre, edad, sexo, clase socioeconómica, etc.). Muy importante, deberá usar exclusivamente los algoritmos de aprendizaje supervisado *(Random forests o AdaBoost*)

2. Descripción de datos

Visión general

Los datos se han dividido en dos grupos:

- conjunto de entrenamiento (train.csv)
- conjunto de prueba (test.csv)

El conjunto de entrenamiento debe usarse para crear sus *modelos de aprendizaje automático*. Para el conjunto de entrenamiento, proporcionamos el resultado (*también conocido como la " "ground truth"*) para cada pasajero.

Su modelo se basará en "características" como el *género* y la *clase de los pasajeros*. También puede utilizar la ingeniería de funciones para crear nuevas funciones.

El *conjunto de prueba* debe usarse para ver qué tan bien se desempeña su modelo con datos invisibles. Para el equipo de prueba, no proporcionamos la "ground truth" para cada pasajero. Es su trabajo predecir estos resultados. Para cada pasajero en el conjunto de prueba, use el modelo que entrenó para predecir *si sobrevivieron o no al hundimiento del Titanic*.

Diccionario de datos

Variable	Definición	Llave
Supervivencia (survival)	Supervivencia	0 = No, 1 = Sí
pclass	Clase de entrada	1 = 1. °, 2 = 2. °, 3 = 3. °
Sexo (Sex)	Sexo	
Edad (Age)	Edad en años	
sibsp	# de hermanos / cónyuges a bordo del Titanic	
Parch	# de padres / hijos a bordo del Titanic	
Billete (Ticket)	Numero de ticket	
Tarifa (Fare)	Tarifa de pasajero	
Cabina (Cabin)	Número de cabina	
Embarcado (Embarked)	Puerto de embarque	C = Cherburgo, Q = Queenstown, S = Southampton

Notas variables

pclass: un proxy para el estatus socioeconómico (SES)
- 1st = Upper
- 2nd = Middle
- 3rd = Lower

(age) edad: la edad es fraccionaria si es menor que 1. Si se estima la edad, ¿está en la forma de xx.5

sibsp: el conjunto de datos define las relaciones familiares de esta manera ...
- Sibling = hermano, hermana, hermanastro, hermanastra
- Spouse = esposo, esposa (se ignoraron las amantes y los novios)

parch : El conjunto de datos define las relaciones familiares de esta manera ...
- Parent = madre, padre
- Child = hija, hijo, hijastra, hijastro
- Algunos niños viajaban solo con una niñera, por lo tanto, parch = 0 para ellos

Entregable:

- ✓ Usando la plataforma *Colab o RStudio*, elabore un reporte con el desarrollo y creación de un *modelo predictivo* que determine *si un pasajero sobrevivió al hundimiento del Titanic o no*, usando los algoritmos de *aprendizaje supervisado* (*Random forests o AdaBoost*).

- ✓ Recuerde que es tan importante el *modelo creado,* como *la explicación de los resultados* paso a paso (no solo código) y una conclusión general. Además, no olvidar responder las preguntas "¿Qué tipo de personas tenían más probabilidades de sobrevivir?"

- ✓ El archivo (*Taller2.pynb o Taller2.R*) debe ser subido a la plataforma *Classroom (pestaña "Evaluación Talleres")*. El archivo debe indicar "Taller 2 - Grupo 0x".