# Datathon II: Machine Learning

En esta segunda versión del Datathon de BaseTIS, el objetivo será conseguir la mejor puntuación posible en una competición de la plataforma Kaggle (link). En esta plataforma se juntan *Data Scientists* de todo el mundo para intentar resolver retos que proponen diferentes empresas para mejorar sus algoritmos y decisiones. En cada una de las competiciones se encuentran:

1. Train dataset: usado para entrenar nuestro modelo de predicción.
2. Test dataset: al que le lanzaremos el modelo y cuyas respuestas subiremos a la plataforma.
3. *Discussion*: apartado donde los participantes pueden comentar sus perspectivas a los problemas y subir, si quieren, su código.
4. *Kernels*: apartado donde se recogen todos los scripts/notebooks compartidos por los usuarios. Intentar dejarlo sólo como último recurso ya que hay gente que ha subido sus soluciones completas.
5. *Leaderboard*: para ver nuestra puntuación y posición en el reto.

La competición que vamos a seguir es la del: *Titanic: Machine Learning from disaster*, uno de los puntos de partida más reconocidos para aprender *Python* o *R* y las bases del *Machine Learning*. Con tal de hacer la actividad lo más llevadera y productiva posible, se ha preparado material que intentará guiaros desde un primer modelo hasta presentar preguntas sobre como mejorar ese primer modelo en sus puntos más críticos.

Los pasos a seguir serán:

1. Descargar los datasets y ENTENDERLOS. En este caso tenemos, entre otros, el ID del pasajero, la clase en que viajaba y la variable objetivo: si sobrevivió o no (que no tendremos en el *test dataset*). Este punto es, sin duda, el paso más importante. En este caso puede parecer trivial, ya que apenas contamos con variables o *features* y la mayoría podrían tener un peso importante en nuestro modelo. Aún así, fácilmente se puede jugar con 50 variables distintas, luego es vital estudiar su comportamiento.

Para aquellos que no tengan todavía *Python* instalado en su ordenador, podéis ir al punto 1.2 de la guía de instalación de las dashboards (link). Una vez esté instalado el software Miniconda y tengamos *Python* bien configurado, sólo nos faltará instalar las notebooks de Jupyter con un sencillo comando: "conda install jupyter" en una consola.

2. Ya que no todo el mundo está familiarizado con el lenguaje con el que se ha llevado a cabo el ejercicio, se han preparado unas **notebook** de Jupyter que pretenden ser un mini tutorial tanto a nivel de análisis y modelado como de código.

- Primera_aproximación.ipynb: Se prepara todo el camino a seguir, desde cargar los datos hasta generar el archivo final con las predicciones que colgaremos en *Kaggle*. Su finalidad es transmitir las primeras ideas detrás

del *Machine Learning* en su nivel más básico. Se estudian por encima los tipos de los datos, se presenta una primera limpieza de datos en un aspecto importante: los datos nulos y finalmente se aplica un modelo sencillo de predicción.

- Analysis_notebook.ipynb: Se presentan más a fondo un estudio de los datos, dejando abiertas ideas para mejorar los pasos realizados en el notebook anterior y centrándonos en librerías para la visualización de los datos, que es un punto muy importante a la hora de entender bien los datos pero que se ha obviado en el notebook anterior.
- FollowUpGuide.ipynb: Estudio y aplicación de distintos modelos de ML.
- Sample_solution.ipynb: Ejemplo de solución con un poco más de profundidad.