Ejercicio de evaluación 1

- **¿cuál es el problema de clasificacion supervisada que reflejan? ¿cuál es la variable clase ("problem class", "class to be predicted"), y ésta que distintos valores toma? ¿Está el problema de clasificación "desbalanceado": son la mayoría de casos de una clase y unos pocos del resto?**

- El data challenge que me ha llamado la atención es el "Humpback Whale Identification Challenge". El problema trata de establecer las condiciones adecuadas en cuales el aterrizaje automático de una nave espacial seria preferable a un aterrizaje manual. La variable clase es "Class" y toma las valores "no-auto" y "auto". Aunque las frecuencias de las clases no son iguales (hay 6/15 casos que corresponden a un aterrizaje manual y 9/15 casos cuales corresponden a un aterrizaje automático), opino que esta distribución de clases no corresponde a un problema de clasificación "desbalanceado", porque la distribución de las clases acerca una distribución perfectamente balanceada en cual cada clase tiene una probabilidad de p=0.5.

- **¿cuántas variables predictoras hay en total? Trata de entender, y luego de explicar en tu respuesta, unas 5-6 variables predictoras.**

- Hay en total 6 variables predictoras. Estas variables son la estabilidad de la nave, el grado de errores que pueden haber surgido hasta el momento del aterrizaje, la señal del error (positivo o negativo), la direction del viento, la fuerza del viento y la visibilidad. Asumo que estas variables reflejan las informaciones relevantes en el momento de decidir si la nave debe aterir en modo manual o automático. Las primeras tres variables predicadoras (stability, error, sign) pertenecen al estado físico y técnico de la nave. Interpreto la estabilidad como el grado de maniobrabilidad que tiene la nave en el moment de aterir y el error y su señal como el rango del error (de major a pequeño) y su valor asociado (positivo o negativo). La dirección y fuerza del viento, como la visibilidad, reflejan las condiciones meteorológicas en el momento de tomar la decision sobre el modo de aterizaje

- ¿cuántos casos contiene el "dataset"? ¿te parecen suficientes casos para realizar predicciones fiables con un modelo que se aprenda a partir de ellos?

- En el dataset publicado hay en total 15 casos. Este numero de casos es muy bajo, y creo que en este caso seria un reto aplicar algoritmos clásicos de aprendizaje estadístico como por ejemplo una red neuronal o un support vector machine. Pero, no creo que sea imposible y que de alguna manera se puede aplicar algoritmos de aprendizaje sobre-visado. En el enlace abajo, por ejemplo, se pueden ver algunos intentos de clasificación con estos datos:

- https://www.openml.org/d/172

- comenta cualquier otro punto o ángulo del problema de clasificación que haya llamado tu atención: ¿por qué han llamado tu atención, por qué te "gustan"? ¿por qué has escogido éstas y no otras?... abierto a tus comentarios.

- Me ha llamado la atención este “dataset” porque el tema de la navegación aerospacial me interesa mucho. Es un campo en cual el desarrollo de interfaces hombre-computador es muy importante, y en cual los sistemas inteligentes tienen ya una historia muy larga. Están establecidos como una herramienta muy importante para aumentar las capacidades humanas durante la exploración del espacio, que es un medio ambiente muy complejo e en cual la soberbia del ser humano todavía no es garantizada. Ademas son datos recogidos por los expertos de la NASA, que tienen una fama internacional y que son considerados como los mejores investigadores del mundo en este campo.

Ejercicio de evaluación 2

- ¿qué te ha motivado a escogerla, qué te ha llamado la atención?
  Este dataset me ha llamado la atención porque esta relacionado con el tema de las ballenas, que pertenecen al grupo de los mamíferos marinos, una especia en cual estoy muy interesado por dos motivos 1) Los mamíferos marinos pertenecen al las especias que tienen cerebros muy desarrollados y que tienen comportamientos complejos. 2) Las ballenas pertenecen a las especias que están amenazadas de desaparecer , y el “data challenge” quiere fomentar el desarrollo de. nuevas herramientas para poder monitorizar el estado de diferentes poblaciones de ballenas a través del globo.

- ¿en qué consiste el problema de análisis de datos? Esto es, ¿qué "pregunta" se les hace a los datos?
  El problema pertenece al campo del reconocimiento automático de imágenes. Consiste en reconocer/detectar automáticamente la especie a cual pertenece una ballena gracias a una imagen de su cola. Cada especia de ballena tiene una forma de cola muy característica , y se trata de entrenar un modelo predictivo que explota esas regularidades.

- ¿qué tipo de datos alberga? Describe brevemente las variables predictoras (de algún ejemplo), la variable clase a predecir y sus valores (si es que existe), la cantidad de ejemplos de que se dispone en las particiones de train/test, tamaño del dataset (MegaBytes, GigaBytes), etc…

- El dataset alberga imágenes de colas de ballena en formato JPEG. Las variables predicadoras son matrices 2D que reflejan los niveles de color para cada pixel. La variable clase es la identidad (Id) a la que pertenece la ballena en la imagen. Puede por ejemplo tomar el valor “humpback”, “ballena jorobada”. Existen 9850 casos training, y 15610 casos de test. Los datos training tienen un tamaño de 283,5MB y los datos test tienen un tamaño de 442,1 MB.

- esto es, "empápate" del dataset y describe los aspectos que serían críticos para un posterior análisis de datos

  Un aspecto critico es que solo existen pocos ejemplos para cada una de las mas de 3000 especies de ballenas. Por lo tanto, se podría computar las probabilidades de pertenecer a una especia para tener una medida de confianza sobre el resultado de la clasificación. Otro aspecto criticó seria probablemente de usar un algoritmo de tipo red neuronal, porque ya se ha demostrado que son eficaces en problemas de reconocimiento de imagines.

- ¿cuál de ellos te ha llamado la atención? ¿por qué?
- El dataset que mas he ma llamado la atención dentro de los datasets disponibles es el "NOAA Global Historical Climatology Network Daily" dataset. Me ha llamado la atención porque se trata de datos sobre el climate que han sido grabados por estaciones meteorológicas localizadas en de todo el globo (mas de 100,000 estaciones en 180 países). El clima y la meteorología son temas que me interesan mucho, porque el clima es un sistema complejo, y se pueden observar muchos fenomenal cíclicos. También estos datos provienen de una red de sensores y permiten de estudiar las dinámicas tiempo-espaciales del clima desde el día de hoy hasta 175 años en el pasado (el primer año registrado es el año 1763!). Por lo tanto, este dataset contiene mucha información sobre la evolución de nuestro clima en los últimos años y el desarrollo de un modele productivo sobre como podría cambiar en el futuro.
-
- ¿qué tipo de análisis propondrías hacer sobre ellos: esto es, qué "preguntas" te gustaría hacer a esos datos?
  Un análisis que me gustaría aplicar es un análisis predictivo. Es decir me gustaría poder predecir cual seria la evolución del clima en 5, 10, 25 años en el futuro. Como los datos vienen en forma de seria de tiempo (time series) se podría considerer la aplicación de métodos de la regresión lineal o no-lineal.

- esas "preguntas" que propones: ¿suponen aprender algún "modelo" desde los datos o simplemente se trataría de "visualizar" los datos de una forma atractiva?
  Se trataría de hacer las dos cosas: 1) aprender un modelo y 2) visualizar los resultados en un mapa del mundo. Como existen solo 180 nodos en la red de estaciones a través del globo, igual se podrían interpolar los datos entre los nodos, para obtener poder obtener una mejora. visualización de los resultados.