Resumen

El análisis de la calidad de los datos abarca muchas dimensiones, desde aquellas tan obvias como la completitud y la coherencia, hasta otras menos evidentes como la correctitud o la capacidad de representar a la población objetivo. En general, es posible clasificar estas dimensiones como las producidas por un efecto externo y las que son inherentes a los propios datos. Este trabajo se centrará en la evaluación de aquellas inherentes a los datos en repositorios de datos sanitarios, como son la variabilidad temporal y multi-fuente. Los procesos suelen evolucionar con el tiempo, y esto tiene un impacto directo en la distribución de los datos. Análogamente, la subjetividad humana puede influir en la forma en la que un mismo proceso, se ejecuta en diferentes fuentes de datos, influyendo en su cuantificación o recogida.

La inteligencia artificial se ha convertido en uno de los paradigmas tecnológicos más extendidos en casi todos los campos científicos e industriales. Los avances, no sólo en los modelos sino también en el hardware, han llevado a su uso en casi todas las áreas de la ciencia. Es cierto que, los problemas resueltos mediante esta tecnología, suelen tener el inconveniente de no ser interpretables, o al menos, no tanto como otras técnicas de matemáticas o de estadística clásica. Esta falta de interpretabilidad, motivó la aparición del concepto de "inteligencia artificial explicable", que estudia métodos para cuantificar y visualizar el proceso de entrenamiento de modelos basados en aprendizaje automático.

Por otra parte, los sistemas reales pueden representarse a menudo mediante grandes redes (grafos), y una de las características más relevantes de esas redes, es la estructura de comunidades. Dado que la sociología, la biología o las situaciones clínicas, usualmente pueden modelarse mediante grafos, los algoritmos de detección de comunidades se están extendiendo cada vez más en el ámbito biomédico.

En la presente tesis doctoral, se han hecho contribuciones en los tres campos anteriormente mencionados. Por una parte, se han utilizado métodos de evaluación de variabilidad temporal y multi-fuente, basados en geometría de la información, para detectar la variabilidad en la distribución de los datos que pueda dificultar la reutilización de los mismos y, por tanto, las conclusiones que se puedan extraer. Esta metodología demostró ser útil tras ser aplicada a los registros electrónicos sanitarios de un hospital a lo largo de 7 años, donde se detectaron varias anomalías.

Además, se demostró el impacto positivo que este análisis podría añadir a cualquier estudio. Para ello, en primer lugar, se utilizaron técnicas de aprendizaje automático para extraer las características más relevantes, a la hora de clasificar la intensidad del dolor de cabeza en pacientes con migraña. Una

de las propiedades de los algoritmos de aprendizaje automático es su capacidad de adaptación a los datos de entrenamiento, en bases de datos en los que el número de observaciones es pequeño, el estimador puede estar sesgado por la muestra de entrenamiento. La variabilidad observada, tras la utilización de la metodología y considerando como fuentes, los registros de los pacientes con diferente intensidad del dolor, sirvió como evidencia de la veracidad de las características extraídas. En segundo lugar, se aplicó para medir la variabilidad entre los histogramas de los niveles de gris de mamografías digitales. Se demostró que esta variabilidad estaba producida por el dispositivo de adquisición, y tras la definición de un preproceso de imagen, se mejoró el rendimiento de un modelo de aprendizaje profundo, capaz de estimar un marcador de imagen del riesgo de desarrollar cáncer de mama.

Dada una base de datos que recogía las respuestas de una encuesta formada por escalas psicométricas, o lo que es lo mismo cuestionarios que sirven para medir un factor psicológico, tales como depresión, resiliencia, etc., se definieron nuevas arquitecturas de aprendizaje profundo utilizando la estructura de los datos. En primer lugar, se diseño una arquitectura, utilizando la estructura conceptual de las citadas escalas psicométricas. Dicha arquitectura, que trataba de modelar el grado de felicidad de los participantes, tras ser entrenada, mejoró la precisión en comparación con otros modelos basados en estadística clásica. Una segunda aproximación, en la que la arquitectura se diseñó de manera automática empleando detección de comunidades en grafos, no solo fue una contribución de por sí por la automatización del proceso, sino que, además, obtuvo resultados comparables a su predecesora.