Aldo Daniel Villaseñor Fierro A01637907
29/08/2023

Para el entrenamiento de una regresión logística utilicé una base de datos que contiene información sobre tumores encontrados en pacientes, con el objetivo de hacer la correcta clasificación de tumores malignos y benignos en base a la información que se tiene. Esta base de datos es de dominio publico y fue encontrada en la plataforma Kagel.

Las normas en torno a las bases de datos se refieren a las creencias y prácticas compartidas de una comunidad en relación con el intercambio y la reutilización de datos.

Cuando se trata de conjuntos de datos de dominio público, la dedicación de dominio público CC0 es de especial interés. Esta dedicación puede ser particularmente importante para el intercambio de datos y bases de datos, ya que puede ayudar a garantizar que los datos y bases de datos altamente factuales no estén restringidos por derechos de autor u otros derechos. Sin embargo, cabe señalar que las leyes de derechos de autor de muchas jurisdicciones cubren las compilaciones de hechos seleccionadas u organizadas de forma creativa y el diseño y la estructura creativos de las bases de datos, y algunas jurisdicciones han promulgado leyes sui generis adicionales que restringen los usos de las bases de datos sin tener en cuenta la legislación de derechos de autor aplicable[1].

En general, es importante garantizar que los conjuntos de datos de dominio público se utilicen de forma ética y responsable. Aunque puede que no existan restricciones legales al uso de datos de dominio público, sigue siendo importante seguir las mejores prácticas y normas en torno al intercambio y la reutilización de datos. Por ejemplo, es importante proporcionar la atribución adecuada cuando se utilizan datos de dominio público, y asegurarse de que los datos se utilizan de una manera respetuosa con la privacidad u otras consideraciones éticas[2].

El funcionamiento del modelo entrenado para la clasificación de tumores malignos y benignos solo toma en cuenta variables que demuestran relevancia para este objetivo. Además como en cualquier proyecto de machine learning se busca que el modelo pueda generalizar lo más posible la información contenida en la base de datos, es por esto que se utilizan métricas que no sesguen negativamente las predicciones del modelo.

En especifico para el proyecto al tratarse de un tema de salud y datos sensibles de pacientes reales, cometí el error de no incluir un archivo read.me con los metadatos de la base y la información sobre la fuente de los datos, los creadores de la base de datos y consideraciones que se deberían tener sobre los datos por ser información sensible para los pacientes de los cuales fueron recabados. En este sentido podría romperse alguna normativa de la industria que podría verse percibida como mala praxis.

Este proyecto al buscar predecir un diagnostico médico debe tratarse con mucho cuidado, recordando que un modelo de machine learning no es infalible y siempre tiene un porcentaje de error. Podría ser utilizado de manera negligente al tratar de abaratar costos en un hospital donde se busca diagnosticar a una gran cantidad de pacientes sin necesidad de pagar a un profesional para hacerlo. En áreas de la salud los modelos de machine learning siempre deben de ser supervisadas por expertos que puedan corroborar que las predicciones del modelo son correctas hasta cierto grado y se deben crear estrategias para permitir que los usuarios de

dichos modelos puedan darse cuenta de una manera efectiva cuando una predicción hecha por un modelo tenga un posible error.

Referencias:

[2] *CC0 use for data - Creative Commons*. (s. f.). https://wiki.creativecommons.org/wiki/CC0_use_for_data

[1] *Are data that can be obtained from the text of a report considered to be in the public domain?* (s. f.). Academia Stack Exchange. https://academia.stackexchange.com/questions/188248/are-data-that-can-be-obtained-from-the-text-of-a-report-considered-to-be-in-the