Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos (Gpo 101)

Análisis del contexto y la normatividad

Adrián Emmanuel Faz Mercado - A01570770

Septiembre 11, 2023

El set de datos que se analizó en esta ocasión es el **Diabetes Health Indicators Dataset**, el cual es un set de datos del Behavioral Risk Factor Surveillance System (BRFSS) de los Centros para el Control y la Prevención de Enfermedades (CDC). Esta encuesta se centra en recopilar información sobre comportamientos de riesgo relacionados con la salud, condiciones de salud crónicas y el uso de servicios preventivos. Las preguntas y variables del set de datos están diseñadas para obtener "insights" sobre indicadores de salud y factores de riesgo relacionados con enfermedades como la diabetes. Esta información puede utilizarse para generar un modelo que intentar predecir si una persona padece diabetes o no.

Este set de datos se obtuvo de la plataforma "Kaggle", la cual almacena todo tipo de conjuntos de datos para aprendizaje automático y ciencia de datos. En Kaggle, se indica que el set de datos cuenta con la licencia **"CC0: Public Domain"**. Esta licencia permite que cualquier persona pueda copiar, modificar, adaptar, distribuir y usar la información, incluso para fines comerciales y sin necesidad de pedir permiso. De cierta manera, el autor renuncia a los derechos de autor de la obra/información, para así contribuir a un común de obras/información para el público. (Creative Commons, 2009) Esto significa, que podemos utilizar este set de datos para el propósito que queramos, pero es importante utilizarlo de manera ética y responsable.

Puedo asegurarme de que no estoy violando la norma al entender que estos datos tienen una licencia de dominio público y que se pueden utilizar sin problema, pero por ello es importante primero revisar la licencia. Igualmente, para asegurarme de esto, revisé la documentación oficial de la licencia de "CC0", en donde se establecen los puntos específicos del funcionamiento de está licencia y de como el autor de cierta forma "dona" la obra para que se use tan libremente como sea posible en cualquier forma. Esta información se encuentra en el documento "CC0 1.0 Universal Public Domain Dedication". (Creative Commons, 2009)

Ahora, en cuestión de la naturaleza de los datos utilizados, estamos hablando de información que está relacionada con la salud. Esta encuesta recopila datos sobre comportamientos de riesgo para la salud, condiciones crónicas de salud y el uso de servicios preventivos. Por ello, las características reflejan aspectos médicos y de salud de la población. El hecho de que sean datos relacionados con la salud, nos lleva a tomar en cuenta ciertas normativas adicionales que son indispensables en este ámbito.

Dado que el conjunto de datos fue recopilado en Estados Unidos y fue una encuesta realizada a estadounidenses, la normativa asociada a ellos es la Ley de Portabilidad y Responsabilidad del Seguro de Salud (HIPAA). Esta ley tiene como objetivo proteger la información médica y de salud de los pacientes, garantizando que esta información sea mantenida en privacidad y que se limite su divulgación. Esta ley establece también, que para poder compartir o utilizar información de salud protegida de individuos sin tener el consentimiento explícito del individuo, es necesario que los datos sean desidentificados, lo que significa que se debe de eliminar cualquier identificador que permita vincular la información de salud con un individuo en específico. Dentro del documento oficial, se establece que 'Health information that does not identify an individual and with respect to which there is no reasonable basis to

believe that the information can be used to identify an individual is not individually identifiable health information.' (HIPAA, 1996). Esto significa que cualquier información de salud que no permita la identificación directa o indirecta de un individuo se considera desidentificada y, por lo tanto, su uso está permitido sin consentimiento explícito.

Al eliminar estos identificadores, entonces los datos se consideran despersonalizados y pueden ser compartidos sin el riesgo de violar esta ley. Para esta situación, podemos asegurarnos de que no se está violando la ley porque los datos que se encuentran en el set de datos no tienen ningún tipo de identificador o alguna manera que pueda hacer que se identifique la persona en específico que respondió la encuesta. Con esto en mente, podemos estar seguros de que se está cumpliendo con la ley al utilizar estos datos en nuestro proyecto.

En caso de que el modelo se utilizara en México, entonces la normativa que está asociada con este tema es la Ley Federal de Protección de Datos Personales en Posesión de los Particulares (LFPDPPP). En el Artículo 1 de está ley, se establece que su propósito es 'la protección de los datos personales en posesión de los particulares con la finalidad de regular su tratamiento legítimo, controlado e informado, a efecto de garantizar la privacidad y el derecho a la autodeterminación informativa de las personas.' (LFPDPPP, 2010). De acuerdo con este marco legal, es esencial garantizar que cualquier dato utilizado, especialmente en un contexto de salud, no pueda ser vinculado directamente a un individuo sin su consentimiento explícito. Por ello, al implementar el modelo en México, se deberían tomar precauciones adicionales para asegurar que los datos sean completamente anónimos y así se puedan alinear con las disposiciones de la ley. Cada país tiene una normativa distinta, pero lo esencial es conocerlas e informarse para asegurarnos de que se cumplan en su totalidad.

En este caso, lo que se busca es realizar un modelo que permita predecir si una persona tiene diabetes en base a las diferentes características que presenta su registro, incluyendo datos como edad, si fuma, si tiene colesterol alto, BMI, entre otras preguntas. Para ello, se pueden usar diferentes algoritmos para crear el modelo de clasificación, pero lo que se busca es que este modelo sea entrenado de la manera más balanceada posible y que no se incurra en ningún sesgo o discriminación de ningún grupo de personas, buscando que el modelo funcione adecuadamente para cualquier persona. Para reducir la probabilidad de que exista un sesgo en los datos, una de las técnicas que se utilizan sería revisar si nuestras clases se encuentran balanceadas, es decir, si tenemos la misma cantidad de registros de personas que sí tienen diabetes y de personas que no, pues si esto no está balanceado correctamente o si hay muchos más registros de una de las dos clases, podría afectar el desempeño de nuestro modelo y dar resultados incorrectos. Para esto, se pueden utilizar técnicas de oversampling, undersampling o de generación sintética, para crear más registros de la clase minoritaria, basados en los datos ya existentes, el punto es evitar que exista un sesgo hacia la clase mayoritaria. Igualmente, es fundamental realizar pruebas del modelo con datos que no sean los mismos de entrenamiento, y que exista una división de los datos desde el inicio, dejando algunos para pruebas, y que estos tengan también diversidad en sus características.

Ahora, si se crea un modelo para predecir si una persona tiene diabetes o no en base a los datos ingresados, este también podría ser usado de manera no ética, con la posibilidad de causar un daño a las personas o para que empresas se aprovechen de ello también. Por ejemplo, una manera en la que se podría utilizar este modelo de manera no ética sería por medio de una compañía de seguros, en donde ellos estén utilizando el modelo para aumentar las tasas de seguro o para negar la cobertura a personas que podrían ser identificadas como "de alto riesgo" para diabetes, basandose únicamente en el modelo y sin realizar estudios que verdaderamente lo comprueben.

También, el modelo se podría usar para discriminar ciertos grupos demográficos, como los de ciertas edades, géneros, pesos, o con ciertas características, insinuar que son "más propensos" a tener diabetes, y por medio de esto, negarles entrada a algún lugar, contratarlos en algún trabajo o asumir directamente que tienen diabetes solamente por el resultado del modelo, tomando decisiones definitivas cuando en realidad no es completamente preciso.

De igual forma, los doctores o las personas podrían utilizarlo como una decisión definitiva para recetar medicamentos o para tener una ganancia aprovechandose de algún sesgo que pueda existir en el set de datos. Por ejemplo, se podría manipular el modelo para que favorezca los resultados a que la mayoría de las veces obtenga de resultado que la persona tiene diabetes y en base a ello recete medicamentos que no son necesarios para obtener ganancias económicas a costa de la salud de los pacientes.

También, quizás podría usarse el modelo en un hospital, para definir la urgencia para atender a cierta persona en específico, y en caso de que salga que no tiene diabetes, se le niegue que lo atiendan con urgencia, cuando en realidad sí lo podría necesitar y su vida podría estar en riesgo.

A grandes rasgos, el uso ético y responsable de los datos es realmente importante en cualquier ámbito, pero especialmente cuando se tratan datos relacionados con salud, los cuales pueden tener implicaciones directas en el bienestar de las personas. Las leyes como la HIPAA y la LFPDPPPP en México, marcan pautas en la protección de los datos de los individuos, pero la responsabilidad final recae en las personas que utilizan estos modelos y datos. Este tipo de proyectos tienen un potencial increíble para beneficiar a la sociedad, pero también puede ser un arma de doble filo si cae en las manos equivocadas. Es esencial que como profesionistas, siempre velemos por el bienestar de la sociedad, tratando este tipo de información y nuestro conocimiento con ética y responsabilidad.

Referencias bibliográficas

1. Creative Commons. (2009) CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. Recuperado el 28 de Agosto del 2023, de: https://creativecommons.org/publicdomain/zero/1.0/legalcode.es

2. U.S. Department of Health and Human Services (1996) Health Insurance Portability and Accountability Act. Recuperado el 10 de Septiembre del 2023, de: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html

3. Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (2010). Guía para cumplir con los principios y deberes de la Ley Federal de Protección de Datos Personales en Posesión de los Particulares. Recuperado el 10 de Septiembre del 2023, de: https://home.inai.org.mx/wp-content/documentos/DocumentosSectorPrivado/Guia_obligaciones_lfpdppp_junio2016.pdf

4. Cámara de Diputados del Congreso de la Unión. (2010) Ley Federal de Protección de Datos Personales en Posesión de Particulares. Recuperado el 10 de Septiembre del 2023, de: https://www.diputados.gob.mx/LeyesBiblio/pdf/LFPDPPP.pdf