Instituto Tecnológico y de Estudios Superiores de Monterrey

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

Momento de Retroalimentación: Análisis del contexto y la normatividad. (Portafolio Análisis - Individual)

Adrián Emmanuel Faz Mercado - A01570770

Agosto 29, 2023

El set de datos que analicé y con el que voy a realizar esta actividad es **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.
Podrían 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, 2023) 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".

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 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.

Referencias bibliográficas

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