Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Inteligencia Artificial Avanzada para la Ciencia de Datos I

Ingeniería en Ciencias de Datos y Matemáticas

Análisis del contexto y la normatividad
Portafolio Análisis

Profesores
Daniel Otero Fadul
Hugo Terashima Marin

Andrea Galicia Jimenez A01177643

Monterrey, Nuevo León. 06 de septiembre de 2023

1. Introducción

Al usar bases de datos obtenidos de alguna pagina externa, por mas confiable que sea, siempre ese debe de tener en cuenta el no entrar en algún tipo de robo de información o en este caso copyright, para esto existen las normativas.

Una normativa es un conjunto de reglas establecidas por una autoridad para regular actividades en un área específica, como leyes, estándares de seguridad o requisitos de calidad. Se usan para garantizar orden, seguridad y cumplimiento de estándares en la sociedad y en diversas industrias. Cumplirlas es importante para evitar sanciones, y se actualizan regularmente para adaptarse a cambios en la sociedad. [1]

Para el trabajo realizado en el modulo de Aprendizaje Automático, decidí trabajar con una base de datos llamada "Salary Dataset"la cual contiene información sobre los salarios de los empleados en una empresa. Cada fila dentro del dataset representa a un empleado diferente y las columnas son variables que incluyen distinta información sobre este empleado teniendo variables como la edad, el género, el nivel educativo, el título de trabajo, los años de experiencia y por supuesto, el salario. En este caso, como se trabajo con una regresión lineal opte por eliminar las columnas que no iba a necesitar y quedarme solamente con las dos que considere mas importantes, los años de experiencia (siendo mi variable x) y el salario (siendo mi variable y) para con esto poder crear un modelo de regresión lineal sin librerías que me ayudara a predecir cuanto seria el salario de un empleado basándonos en los años de experiencia que lleva. [2]

Independientementemente de los cambios y/o modificaciones que se hayan hecho con el dataset, este cuenta con una licencia que da base a la normativa sobre como usar la información proporcionada en este dataset o bien que si esta correcto hacer y que no. La licencia que tiene esta base de datos se llama .Attribution 4.0 International"(CC BY 4.0) y hace referencia a una licencia de Creative Commons (CC) que permite a los creadores de contenido compartir su trabajo de manera publica siempre y cuando se les de el crédito adecuado en caso de utilizarlo. De igual manera, esta licencia permite que cualquier usuario que tenga acceso a ella (en este caso cualquiera ya que se encuentra publica en la plataforma de kaggle) pueda utilizar, compartir, adaptar, modificar y distribuir el contenido dentro de ella con ciertas condiciones especificas como lo son [3]:

1. Se le debe de dar el crédito adecuado al creador original de la base de datos mencionando su nombre y proporcionando un enlace a la licencia mencionada anteriormente (CC BY 4.0)
2. No puedes modificar o agregar restricciones adicionales a las que la licencia ya establece; por ejemplo, no se permite utilizar tecnología de gestión de derechos digitales para limitar cualquier acceso a esta información de manera que viole los términos de la licencia propia. [3]

Como mencionado anteriormente, este dataset fue utilizado meramente para propósitos educativos al implementarlo dentro de una asignatura donde el propósito principal era crear una predicción mediante regresión

lineal de manera manual en el lenguaje de python, específicamente en Google Colab. Al entregar el proyecto me asegure de no manipular de manera indebida la base de datos y mucho menos mencionar que yo soy la propietaria, en este caso se agrego un pequeño párrafo donde explicaba que la información fue obtenida de la plataforma de Kaggle; sin embargo, de manera no intencional, no agregue en si el nombre del propietario de esta base de datos el cual puede llegar a hacer ruido en la normativa establecida ya que aunque haya dado crédito a la plataforma de donde lo obtuve lo correcto seria agregar el nombre del creador.

Una vez que ya se analizo un poco mas la normativa aplicada en el dataset, nos enfocaremos un poco mas en analizar la herramienta en donde se aplico esta información la cual para este caso fue una regresión lineal y se deben de considerar algunas cosas importantes como por ejemplo:

- Asegurarse de mantener la privacidad de datos en todo momento y que estos cumplan con las reglas que el propietario establece sin importar lo que tu modelo pueda llegar a modificar en ellos, esto incluye proteger la información personal y obtener permisos cuando sea necesario.
- Como mencionado anteriormente, aunque la herramienta haya sido trabajo hecho por uno mismo siempre se debe de pedir permiso al usar datos protegidos por derechos de autor o en general respetar las reglas de propiedad intelectual.

Como en cualquier trabajo o escenario, es muy sencillo caer en un sesgo ético sin tener esa intención o inclusive sin darse cuenta por lo que se pueden tomar en consideración los siguientes aspectos para no incurrir en ningún sesgo ético:

- Evitar usar características que puedan causar discriminación injusta, como la raza o el género, en tu modelo.
- Limpiar los datos para eliminar problemas que puedan afectar la precisión del modelo, como valores extraños o datos faltantes.
- Comprobar si el modelo muestra resultados sesgados para diferentes grupos y toma medidas para corregirlo si es necesario.
- Asegurarse de que los resultados sean claros y explicables, para que se puedan identificar y solucionar problemas éticos si aparecen.

El mal uso de una herramienta de regresión lineal puede dar lugar a faltas éticas en escenarios como la discriminación injusta, la manipulación de datos para fines maliciosos, la violación de la privacidad, la desinformación, la falta de transparencia, la falta de responsabilidad y el incumplimiento normativo. Para evitar estas faltas éticas, es esencial utilizar la herramienta de manera responsable, garantizar la integridad de los datos y la equidad en las decisiones, ser conscientes de la información que se esta utilizando así como sus licencias, y establecer políticas y prácticas éticas sólidas en su uso. La ética en el análisis de datos es responsabilidad de los usuarios, desarrolladores y reguladores.

Bibliografía

[1] E. Trujillo, "Normativa." [Online]. Available: https://economipedia.com/definiciones/normativa.html

[2] K. RATTANAPORN, "Salary prediction dataset." [Online]. Available: https://www.kaggle.com/datasets/rkiattisak/salaly-prediction-for-beginer

[3] C. Commons, "Attribution 4.0 international — cc by 4.0. (2023)." [Online]. Available: https://creativecommons.org/licenses/by/4.0/