Alumno: Héctor Hibran Tapia Fernández
Matrícula: A01661114

Reporte de Ética y Normatividad (TC3006C)

Portafolio de Análisis

Kaggle es una plataforma web de Ciencia de Datos que cuenta con más de 536 mil miembros activos en 194 países, y que recibe más de 150 mil publicaciones por mes. Kaggle brinda, herramientas, recursos, competiciones y más importantes para progresar al máximo en el campo de Ciencia de Datos. Kaggle, tiene una interfaz Jupyter Notebooks personalizable y sin configuración. Permite acceder de manera gratuita a GPUs y a una gran cantidad de datos y códigos publicados por la comunidad. Ahí, puedes encontrar códigos y datos que se necesiten para realizar proyectos de ciencia de datos.

Teniendo esto en mente es importante que esta plataforma cumpla con ciertas normativas que protejan a los usuarios y sobre todo a la información que estos brinden y/o suban a la plataforma, además por supuesto de no revelar información de terceros o información que sea delicada para el resto de los usuarios, para ello Kaggle se encarga de cumplir con los estándares de privacidad para proteger los datos de los usuarios, a continuación se muestran algunas normativas que está organización se apega:

- **Propiedad y Derechos de los Datos: El usuario** debe asegurarse de tener los derechos o permisos necesarios para subir los datos, ya sea si este mismo es el propietario o tiene autorización del dueño original, y elige una licencia adecuada que determine cómo pueden ser utilizados.

- **Cumplimiento con Normativas de Privacidad (GDPR, CCPA): Los datos** deben cumplir con leyes de privacidad, como el GDPR y CCPA, asegurando que la información personal esté anonimizada o que se cuente con el consentimiento adecuado para compartirla, especialmente si los datos incluyen información de personas.

- **Contenido Aceptable: No está permitido subir datos sensibles o ilegales**, por ejemplo, información confidencial o que promueva actividades ilícitas.

- **Transparencia y Descripción de los Datos: Proporcionar una descripción clara y detallada de la base de datos,** se debe explicar la fuente, estructura, propósito y cualquier limitación en el uso de los datos, para que los usuarios entiendan su contexto y cómo usarlos correctamente.

- **Responsabilidad Legal: Como propietario de los datos subidos**, el ususario es el responsable de que estos cumplan con las leyes aplicables y las normativas de Kaggle, lo que significa que Kaggle no será responsable por infracciones legales que resulten de su publicación.

Ahora teniendo en cuenta la solución del reto de “Titanic – Machine Learning For Disaster”, me di cuenta principalmente de dos cosas:

1. Desbalanceo de Clases

2. Data Awaresness.

Desbalanceo de Clases.

Después de realizar y aplicar al dataset modelos predictivos para determinar que persona dependiendo de su descripción física/social tienes probabilidad de sobrevivir, se debe tener en cuenta que este dataset es pequeño, y refleja la vida en 1914. Hace más de 100 años, la vida que vivimos ahora no era la misma, hablando principalmente de los derechos humanos. Retomando el tema de desbalanceo de clases, la primera parte que se notó simplemente por el análisis de datos fue que sí eras hombre y pertenecías a la tercera clase, tenías la probabilidad más alta de morir. En comparación por ejemplo con su contraparte socioeconómica de la primera clase que tenían una probabilidad más baja.

Al emplear los modelos predictivos, surgieron problemas en la correcta predicción de la supervivencia de ciertos grupos, particularmente debido al desbalance de clases en el dataset. Sobrevivieron aproximadamente 712 personas de las 2225 a bordo, lo que significa que la clase de los sobrevivientes estaba menos representada en los datos. Este desbalance provoca un sesgo en los modelos, ya que cuentan con menos información sobre los individuos que sobrevivieron, lo que afecta su capacidad para generalizar correctamente los patrones de supervivencia.

Es importante resaltar que este sesgo no solo está relacionado con la cantidad de datos, sino también con las circunstancias históricas, sociales y de género que influyeron en la probabilidad de supervivencia, lo que pone en evidencia las limitaciones de los modelos predictivos cuando se enfrentan a datasets que reflejan contextos históricos con fuertes disparidades sociales.

Data Awareness.

Para concluir, las personas a bordo no tenían idea de lo que harían con la información que los tripulantes o el personal del Titanic haría con sus datos, a este punto en la historia también se puede inferir que tal vez se obtuvo más información de los pasajeros una vez que sucedió la tragedia. También la forma de generar y almacenar los datos no era tan sencilla como lo es ahora, este problema tal vez sea uno de los problemas a los que nos enfrentamos hoy en día: aunque ahora tenemos tecnologías avanzadas para recolectar y analizar grandes cantidades de datos, también surgen dilemas importantes relacionados con la privacidad, el consentimiento y el manejo ético de la información. En la actualidad, los individuos están más conscientes de cómo se utilizan sus datos, pero la magnitud y velocidad a la que se recopila información plantea nuevos desafíos. Esto nos lleva a reflexionar sobre la responsabilidad que tenemos al usar y gestionar datos, tanto en el pasado como en el presente, y la importancia de equilibrar el avance tecnológico con consideraciones éticas.