Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Campus Monterrey

TC3006C.102: Inteligencia artificial avanzada para la ciencia de datos I

# Momento de Retroalimentación: Reto Análisis del contexto y la normatividad. (Portafolio Análisis)

Jesús Daniel Martínez García - A00833591

Lunes 9 de Septiembre de 2024

## Análisis del Uso de Datos y Cumplimiento Normativo

Normativas Asociadas al Tipo de Datos Utilizados

El conjunto de datos (utilizado para evidencias de machine learning) proporcionado por Erdem Taha en Kaggle está clasificado como datos de acceso público y anonimizados. Es esencial garantizar que se cumplan las leyes como el Reglamento General de Protección de Datos (GDPR) de la Unión Europea y la Ley de Portabilidad y Responsabilidad de Seguros de Salud (HIPAA) de los Estados Unidos para que se utilicen correctamente. Los datos personales no contienen información personalmente identificable (PII), por lo que ambas regulaciones requieren que sean anonimizados o desidentificados.

1. GDPR (Artículos 5 y 6): El artículo 5 establece que los datos personales deben ser tratados de manera legal, justa y transparente, garantizando que sean procesados de acuerdo con los derechos del individuo. El artículo 6 define las bases legales para el tratamiento de los datos. En este caso, los datos anonimizados cumplen con estos principios, ya que no contienen información que pueda identificar directamente a una persona【1】.
2. HIPAA (Sección 164.502): Esta sección regula la desidentificación de información protegida de salud. Los datos anonimizados en el conjunto utilizado cumplen con este requisito, lo que asegura que su uso no infringe esta normativa【2】.

Método de Uso de los Datos y Garantías Normativas

Para utilizar este conjunto de datos, se descargó a través de la plataforma Kaggle, que impone términos de servicio que garantizan el cumplimiento de las leyes locales de privacidad. Kaggle exige que cualquier uso de datos cumpla con normas para evitar violaciones de privacidad, como la anonimización.

El cumplimiento de las normas depende del proceso de anonimización, y Kaggle explica que los datos anonimizados no contienen información que permita identificar a una persona. Por lo tanto, se garantiza que el uso de los datos en este caso no viola ni el GDPR ni la HIPAA.

Para evitar violaciones de la ley, el análisis de estos datos debe seguir las siguientes pautas:

- No intentar revertir la anonimización de los datos.
- No realizar acciones que busquen identificar a los individuos de los que provienen los datos.

Estos principios están alineados con los lineamientos éticos de investigación, como los del Cancer Research Data Commons (CRDC), que también subraya la importancia de la anonimización para la investigación médica【7】.

Cumplimiento Normativo de la Herramienta (Modelo de Machine Learning)

El modelo de machine learning desarrollado con este conjunto de datos también cumple con los principios éticos y normativos de la industria. Específicamente:

1. Principios Éticos de la IA (OMS): La Organización Mundial de la Salud establece que los modelos de IA en salud deben garantizar la transparencia y la equidad. En este proyecto, se asegura que el modelo no presenta sesgos basados en raza, género o estatus socioeconómico, lo cual se valida mediante el análisis de datos y la implementación de controles para evitar la discriminación【4】.
2. Normas de Seguridad de la Información (ISO/IEC 27001): Se implementan controles estrictos de seguridad para proteger los datos utilizados en el entrenamiento del modelo. Esta norma asegura que los datos sean almacenados y procesados de manera segura, evitando brechas de seguridad【5】.

Riesgos Éticos y Escenarios de Uso Indebido

Es esencial identificar los posibles escenarios en los que el modelo podría ser mal utilizado, tanto por negligencia como por malicia:

1. Negligencia: Un escenario de negligencia podría ocurrir si el modelo es desplegado sin una validación adecuada. Si el modelo genera predicciones inexactas, podría llevar a decisiones médicas incorrectas. Por ejemplo, un mal diagnóstico de cáncer podría resultar en un tratamiento inapropiado, poniendo en riesgo la salud del paciente. Para mitigar este riesgo, se debe implementar una validación cruzada rigurosa (realizada) y auditorías periódicas.
2. Malicia: En un escenario de malicia, los resultados del modelo podrían ser manipulados deliberadamente para beneficiar o perjudicar a ciertos individuos o grupos. Esto podría generar daños emocionales, financieros o de salud. Para prevenir este tipo de situaciones, se han implementado controles de acceso estrictos y medidas de transparencia, asegurando que solo personal autorizado tenga acceso a los resultados del modelo.

Reporte Final y Cumplimiento

Tanto el breve análisis de los datos como la evaluación del cumplimiento normativo del modelo se incluirán en el repositorio del proyecto. Esto garantiza la transparencia y el cumplimiento de los estándares de la industria para el uso de datos en la investigación del cáncer y en el desarrollo de herramientas tecnológicas basadas en estos datos.

El reporte detalla cómo la solución respeta las normativas como la GDPR, HIPAA, y otros marcos regulatorios, además de las medidas preventivas para evitar violaciones éticas, ya sea por negligencia o malicia.

El modelo de regresión logística creado cumple con todas las normas y principios éticos antes mencionados. Debido a que los datos utilizados son completamente anonimizados, no hay riesgo de exponer información personal, por lo que se puede cumplir con las regulaciones legales como el GDPR y la HIPAA. Al usar estos datos de acuerdo con los términos y condiciones establecidos por Kaggle, se garantiza que el uso de la información cumpla con las normas vigentes.

Además, el modelo cumple con las mejores prácticas de la industria, como los Principios Éticos de la IA de la OMS, que garantizan que el proceso de análisis y toma de decisiones sea transparente y sin discriminación.

En resumen, el modelo cumple con los estándares éticos y de la industria y es técnicamente sólido, lo que lo convierte en una solución confiable y responsable para su uso en el análisis de datos médicos anonimizados. Este enfoque completo combina las ventajas de su uso en la investigación del cáncer y otras aplicaciones médicas con los riesgos éticos y legales.

## Referencias


1. GDPR (2018). Artículos 5 y 6 del Reglamento General de Protección de Datos. Disponible en: https://gdpr-info.eu/art-5-gdpr/
2. HIPAA (2021). 45 CFR § 164.502 - Reglas de Privacidad. Disponible en: https://www.hhs.gov/hipaa/for-professionals/privacy/index.html
3. Kaggle (2023). Términos de Servicio y Normativas de Uso. Disponible en: https://www.kaggle.com/terms
4. OMS (2021). Principios Éticos para el Diseño y Uso de la Inteligencia Artificial. Disponible en: https://www.who.int/publications-detail/9789240029207
5. ISO/IEC 27001. Sistemas de Gestión de Seguridad de la Información. Disponible en: https://www.iso.org/isoiec-27001-information-security.html
6. Kaggle (2023). Terms and Cancer Dataset Information. Recuperado de https://www.kaggle.com/datasets/erdemtaha/cancer-data
7. Cancer Research Data Commons (CRDC). NCI Data Science Portal. Recuperado de https://datascience.cancer.gov/research