Reto Técnico: Clasificación de Artículos Biomédicos con Machine Learning

MISIÓN

En este desafío de Tech Sphere 2025, desarrollarás un sistema de inteligencia artificial que revolucionará la forma de clasificar literatura médica. Tu misión es construir un modelo de machine learning capaz de clasificar artículos médicos en uno o varios grupos temáticos, utilizando únicamente el título y el abstract como insumo.

El dataset proporcionado (challenge_data.csv) contiene 3565 registros provenientes de distintas fuentes incluyendo datos reales de dominio clínico: NCBI, BC5CDR y synthetic data.

Objetivo

Dado un artículo médico, tu modelo debe clasificar correctamente si pertenece a uno o varios de los siguientes grupos:

- Cardiovascular
- Neurological
- Hepatorenal
- Oncological

Tu solución deberá permitir cargar un archivo `.csv` con la misma estructura del dataset inicial. Aunque el archivo incluirá la columna `group`, esta no podrá ser utilizada como entrada para el modelo durante la clasificación, pero sí deberá emplearse posteriormente para evaluar su desempeño. El sistema tendrá que procesar el archivo, clasificar cada artículo en uno o varios grupos, generar un reporte con las métricas obtenidas y producir un archivo `.csv` de salida que contenga las predicciones generadas por el modelo.

Estructura del Dataset

- title → Título del artículo.
- abstract → Resumen del artículo.
- group → Grupo(s) al que pertenece el artículo.

Este proyecto simula un escenario real donde la tecnología puede acelerar la investigación médica y mejorar el acceso al conocimiento científico.

A quién va dirigido

Este reto está diseñado especialmente para **perfiles junior y personas que están dando sus primeros pasos en el mundo de los datos**. Es ideal para:

- Recién graduados en carreras STEM
- Profesionales en transición hacia ciencia de datos
- Desarrolladores que quieren incursionar en machine learning
- Analistas que buscan expandir sus habilidades en IA
- Estudiantes avanzados con conocimientos básicos en programación y estadística

Stack y habilidades requeridas

Conocimientos técnicos deseables:

- **Desarrollo de software**: Experiencia básica en programación (Python preferiblemente)
- **Ingeniería de datos**: Familiaridad con manipulación y limpieza de datasets
- **Ciencia de datos**: Comprensión básica de análisis exploratorio y visualización
- **Machine Learning**: Conceptos fundamentales de aprendizaje supervisado y clasificación

Tecnologías que podrás usar:

- Python (pandas, scikit-learn, numpy)
- Herramientas de NLP (NLTK, spaCy, transformers)
- Librerías de visualización (matplotlib, seaborn)
- Jupyter Notebooks
- Git para control de versiones
- AI Agents (Agentbricks, AutoGen, OpenAI, Claude)

No necesitas ser experto en todas estas tecnologías - el reto está diseñado para que puedas aprender sobre la marcha.

Condiciones de participación

- **Equipos de 3 personas**: La colaboración es clave. Busca compañeros con habilidades complementarias
- **Duración**: [Especificar duración del reto]
- **Modalidad**: Virtual - Sustentacion Presencial del Reto durante el TechSphere 2025
- **Entregables**: Modelo funcional, código documentado y presentación de resultados

¡Únete y contribuye al futuro de la tecnología en salud mientras desarrollas habilidades clave en inteligencia artificial!

Criterios y Rúbricas de Evaluación

🎯 Puntaje Total: 100 Puntos

---

1. 🛠️ Desempeño Técnico (30 Puntos)

1.1 Configuración del Proyecto (10 Puntos)

- Código conforme al estándar PEP 8
- Docstrings e instrucciones en línea descriptivas
- Estructura modular y reutilizable (funciones/clases)
- Manejo adecuado de errores y casos extremos
- Historial de Git limpio con mensajes descriptivos
- Evitar código duplicado
- Uso eficiente de algoritmos y estructuras de datos

1.2 Metodología (10 Puntos)

- Separación correcta del conjunto de datos (entrenamiento/validación/prueba o validación cruzada)
- Justificación teórica del enfoque
- Consideración de sesgos
- Resultados reproducibles con seeds fijos

1.3 Calidad de Implementación (10 Puntos)

- El código se ejecuta sin errores
- Buena eficiencia de ejecución y uso de memoria
- Independiente del entorno/plataforma

---

2. 📊 Profundidad Analítica (35 Puntos)

2.1 Análisis Exploratorio de Datos (EDA) (10 Puntos)

- Perfilado completo y diagnóstico de los datos
- Análisis de frecuencia de términos y n-gramas
- Detección de valores atípicos o textos irrelevantes
- Análisis de distribuciones de longitud y vocabulario
- Análisis de correlaciones con interpretación
- Evaluación de calidad y coherencia del dataset
- Visualizaciones claras e informativas

2.2 Ingeniería de Características (20 Puntos)

- Creación de variables innovadoras y relevantes
- Preprocesamiento de texto de alta calidad
- Aplicación de conocimiento del dominio
- Técnicas de selección de características
- Reducción de dimensionalidad
- Escalado o normalización de variables

---

3. Desarrollo del Modelo (20 Puntos)

3.1 Diseño y Evaluación del Modelo (15 Puntos)

- Evaluación usando F1-score y Accuracy
- Comparación con líneas baseline
- Justificación de elecciones de modelos
- Evaluación de trade-offs de rendimiento
- Interpretación de métricas en contexto

3.2 Optimización de Hiperparámetros (5 Puntos)

- Proceso de ajuste sistemático (Uso de Grid/Random/Bayesian search)
- Estrategia de validación adecuada
- Prevención de sobreajuste
- Uso eficiente de recursos computacionales

---

4. Valor de Negocio o Investigación (10 Puntos)

- Recomendaciones accionables
- Aplicaciones prácticas claras
- Sugerencias estratégicas basadas en resultados
- Discusión sobre el retorno de inversión
- Viabilidad de implementación
- Análisis de riesgos o limitaciones

---

5. 🗣️ Comunicación y Presentación (10 Puntos)

- Informe bien estructurado y claro
- Flujo lógico de ideas
- Inclusión de resumen ejecutivo
- Equilibrio entre tecnicismo y claridad
- Conclusiones respaldadas por evidencia
- Excelente gramática y estilo