Plan Overview

A Data Management Plan created using DMPTool

Title: Modelo Predictivo para Diagnóstico de Accidente Cerebrovascular

Creator: Nasly Posada

Affiliation: Universidad de Los Andes (uniandes.edu.co)

Principal Investigator: Paul Guzman, Juan Carlos Acosta, Sonia Olaya, Nasly Posada

Funder: Universidad de Los Andes (uniandes.edu.co)

Template: Digital Curation Centre

Project abstract:

El ataque cerebrovascular (ACV) es un trastorno vascular que afecta a millones de personas en todo el mundo. Según la Organización Mundial de la Salud, aproximadamente 15 millones de personas sufren un ACV cada año, de los cuales el 43% mueren y el 33% quedan con discapacidad permanente. El ACV es considerado como la segunda causa de muerte y la primera causa de discapacidad en todo el mundo.

Lo más alarmante es que para el 2022 la Organización Mundial de Accidentes Cerebrovasculares (WSO) reporto que el riesgo de desarrollar un ACV a lo largo de la vida ha aumentado en los últimos 17 años. Actualmente, se estima que 1 de cada 4 personas sufren un derrame cerebral en su vida.

Una identificación temprana en los riesgos y las intervenciones preventivas pueden reducir significativamente el riesgo sufrir un ACV, alrededor del 80% de los ACV se pueden prevenir si se identifican los factores de riesgo. A través de técnicas de machine learning, es posible analizar grandes conjuntos de datos para identificar factores de riesgo y generar predicciones más precisas sobre la probabilidad de que un individuo sufra un ACV. Este enfoque facilita una toma de decisiones más efectiva por parte de los profesionales de la salud y puede ser una herramienta clave para el diseño de estrategias de prevención más eficaces.

Start date: 10-08-2024

End date: 11-30-2024

Last modified: 10-20-2024

Modelo Predictivo para Diagnóstico de Accidente Cerebrovascular

Data Collection

What data will you collect or create?

Los datos de este proyecto provienen del conjunto de datos Stroke Prediction Dataset de Kaggle el cual contiene resultados de 11 características clínicas para predecir eventos de accidente cardiovascular.

El conjunto de datos contiene 5.110 observaciones con los siguientes 12 atributos:

- id: Identificador único
- gender: Género del paciente (Hombre, Mujer u Otro).
- age: Edad del paciente.
- hypertension: si el paciente tiene o no hipertensión.
- heart_disease: si el paciente tiene o no alguna enfermedad cardiaca.
- ever_married: si el paciente ha estado alguna vez casado.
- work_type: tipo de trabajo (trabajador del gobierno, privado, autónomos o nunca ha trabajado).
- residence_type: Tipo de residencia (rural o urbana).
- avg_glucose_level: nivel promedio de glucosa en la sangre.
- bmi: Índice de masa corporal
- smoking_status: si el paciente fuma o no. La categoría “Desconocido” significa que la información no está
- disponible para el paciente.
- stroke: si el paciente tuvo un accidente cerebrovascular o no.

How will the data be collected or created?

El conjunto de datos inicial se extrae directamente de la plataforma Kaggle y se compone de un único archivo .csv de 316.97 kB.

No se conoce la fuente de los datos originales dado que no se proporciona información en Kaggle.

Durante el proyecto, los datos serán organizados en una estructura de carpetas clara y jerárquica:

- /raw_data: para almacenar los datos crudos descargados desde Kaggle.
- /processed_data: para los datos procesados y transformados para su análisis.
- /models: para almacenar los diferentes modelos predictivos entrenados.

Los archivos seguirán convenciones de nomenclatura claras, como stroke_raw.csv y stroke_processed.csv, para asegurar que cualquier miembro del equipo pueda identificar el propósito de cada archivo fácilmente.

Se implementará un sistema de control de versiones utilizando GitHub y la herramienta DVC (Data Version Control) para asegurar que todos los cambios en los datos y modelos sean documentados adecuadamente. Cada versión del conjunto de datos será registrada con etiquetas descriptivas que reflejen los cambios realizados, garantizando la trazabilidad y la integridad de los datos durante todo el ciclo del proyecto.

Documentation and Metadata

What documentation and metadata will accompany the data?

El conjunto de datos original cuenta con descripción general de las variables medidas, se espera investigar el origen de los datos y lograr crear una descripción más detallada del origen de estos.

La documentación del proyecto incluirá un archivo README.md con los detalles sobre la construcción del conjunto de datos, una guía de variables, y proceso de limpieza de datos implementado. Todo estará organizado en un repositorio de GitHub para facilitar la lectura, interpretación y acceso a largo plazo.

Ethics and Legal Compliance

How will you manage any ethical issues?

El conjunto de datos a utilizar está completamente anonimizado por lo que no hay riesgos para los participantes.

En el desarrollo del proyecto, en la construcción del modelo de Machine Learning se evitará que introduzca sesgos o discriminaciones hacia algún grupo con características similares. La capacidad predictiva del modelo a construir podrá ser usada como herramienta de apoyo en entidades que busquen mitigar la problemática y no como una solución a la misma.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

El conjunto de datos tiene licencia autores originales y fuente confidencial por lo cual se pueden solo para propósitos educativos. Por lo tanto, es fundamental que todas las fuentes de datos utilizadas en este proyecto sean citadas correctamente y que se reconozca debidamente la autoría del conjunto de datos.

Storage and Backup

How will the data be stored and backed up during the research?

Durante el proyecto, el conjunto de datos será almacenado en GitHub, utilizando el sistema de control de versiones para mantener un registro detallado de los cambios en los datos y el código. El tamaño del conjunto de datos es manejable dentro de las capacidades de almacenamiento de GitHub, lo que garantiza suficiente espacio para almacenar los datos y versiones futuras del proyecto.

El respaldo de los datos se realizará automáticamente cada vez que se realice un cambio significativo, aprovechando las capacidades de GitHub para gestionar y versionar datos y código mediante la herramienta DVC. Esto permitirá un seguimiento detallado de las versiones de los archivos, lo que facilita la recuperación de cualquier estado anterior del proyecto.

Todos los miembros del equipo garantizarán que los respaldos se realicen automáticamente después de cada cambio significativo. En caso de incidentes, GitHub permitirá restaurar versiones anteriores de manera rápida y efectiva. Al centralizar todo en GitHub, se minimizan los riesgos de pérdida de datos y se asegura su integridad y trazabilidad a lo largo del proyecto.

How will you manage access and security?

Dado que el conjunto de datos proviene de una fuente pública (Kaggle) y no contener información personal sensible, los riesgos de seguridad son relativamente bajos. Sin embargo, se implementarán medidas de seguridad para proteger la integridad del proyecto, como evitar accesos no autorizados y garantizar la protección del control de versiones en GitHub.

Los colaboradores tendrán acceso seguro al repositorio a través de GitHub, usando cuentas autenticadas y restringidas a usuarios aprobados.

No se recolectarán nuevos datos en campo, ya que los datos utilizados son públicos y ya procesados. Por lo tanto, no es necesario implementar mecanismos adicionales para la transferencia segura de datos desde fuentes externas.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

La informacion original del conjunto de datos es de alto valor ya que puede ayudar a identificar tendencias anuales de salud en los Estados Unidos pero es de confiedencial. Por lo cual no debe ser información almacenada y ser usada solo para propositos educativos. La labor de recolección, almacenamiento y disponerla a disposición del dominio publico es realizada por el autor del conjunto de datos. Cualquier información nueva generada por este proyecto no será preservada ya que se considera que con los datos originales disponibles en Kaggle y el código almacenado en Github se podrán replicar los resultados sin incurrir en costos o tareas de almacenamiento

What is the long-term preservation plan for the dataset?

Durante el desarrollo del proyecto toda la información estará almacenada en los servidores de AWS, y el código con el que se realice el proyecto estará almacenado en GitHub, los datos empleados están disponibles en Kaggle y por tal motivo no se considera la necesidad de almacenarlos en algún sistema propio.

Data Sharing

How will you share the data?

Al ser datos de uso público no se considera la necesidad de crear canales especializados para compartir la información y solamente se referenciarán las fuentes de datos, en cuanto a los hallazgos y desarrollos del proyecto estos estarán almacenados en Github y serán de acceso público.

Are any restrictions on data sharing required?

Si, el conjunto de datos solo puede ser usado para propósitos educativos, así mismo determinamos que el desarrollo realizado debe ser para los mismos propósitos.

Responsibilities and Resources

Who will be responsible for data management?

Todo el equipo será responsable por el manejo de datos. La implementación y revisión periódica del DMP será colaborativa.

What resources will you require to deliver your plan?

Se utilizará AWS para el almacenamiento de datos y de código durante la duración del desarrollo del proyecto, adicionalmente se debe contar con un ambiente de desarrollo donde este configurado Python.

Planned Research Outputs

Model representation - "Modelo Predictivo Supervisado"

Un modelo de aprendizaje supervisado diseñado para predecir el riesgo de desarrollar diabetes tipo 2, acompañado de una evaluación de su desempeño (precisión, recall, AUC) y la documentación de los experimentos realizados, incluyendo ajustes de hiperparámetros y configuraciones del modelo.

Audiovisual - "Prototipo Funcional y Tablero Interactivo"

Un prototipo funcional que implementa el modelo predictivo y permite la interacción con los datos mediante un tablero interactivo. El dashboard mostrará visualizaciones de los resultados y predicciones, facilitando la toma de decisiones en base a las predicciones del modelo.

Data paper - "Documentación Completa"

Un manual de usuario para el tablero interactivo, un manual de instalación que detalla el proceso de despliegue del sistema, y un informe final que aborda el problema de negocio, la pregunta de investigación y el análisis exploratorio de los datos utilizados.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Modelo Predictivo Supervisado	Model representation	2024-11-19	Open	None specified		None specified	None specified	No	No
Prototipo Funcional y Tablero Interactivo	Audiovisual	2024-11-19	Open	None specified		None specified	None specified	No	No
Documentación Completa	Data paper	2024-11-19	Open	None specified		None specified	None specified	No	No