Plan Overview

A Data Management Plan created using DMPTool

DMP ID: https://doi.org/10.48321/D1ABF92646

Title: Predicción de riesgo de insuficiencia cardíaca

Creator: Camilo Robayo - **ORCID:** 0009-0007-6073-0346

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

Principal Investigator: Jhon Tavares

Data Manager: Javier Abril

Project Administrator: Angel Benito

Funder: Heart Failure Clinical Records

Template: Digital Curation Centre

Project abstract:

El problema de la insuficiencia cardíaca representa un desafío significativo en la medicina moderna, ya que afecta a millones de personas en todo el mundo y es una de las principales causas de mortalidad. A pesar de los avances en el tratamiento y el manejo de esta enfermedad, la identificación temprana de pacientes con mayor riesgo de mortalidad sigue siendo un reto. Este proyecto busca desarrollar un modelo predictivo que permita estimar el riesgo (alto o bajo) de padecer insuficiencia cardíaca, utilizando un conjunto de variables clínicas, corporales y de estilo de vida.

El objetivo principal de este proyecto es crear un modelo basado en datos que pueda predecir si un paciente tiene mayor o menor probabilidad de sobrevivir dentro del periodo de seguimiento establecido. Al hacerlo, se busca ofrecer una herramienta de apoyo a los médicos, ayudándoles a identificar a los pacientes que presentan un mayor riesgo de mortalidad y que podrían beneficiarse de intervenciones o tratamientos personalizados y oportunos. Este enfoque podría mejorar significativamente el pronóstico de los pacientes y reducir la carga sobre los sistemas de salud al optimizar el uso de recursos.

Objetivo principal:

Diseñar una metodología basada en técnicas de aprendizaje automático que pueda estimar con precisión la probabilidad de supervivencia de los pacientes con insuficiencia cardíaca.

Objetivos secundarios:

Optimizar la selección de características relevantes para mejorar la precisión y eficiencia del modelo predictivo, empleando técnicas de ingeniería de características que permitan reducir la complejidad del modelo sin perder precisión en las predicciones.

Evaluar diferentes algoritmos de machine learning para determinar cuál proporciona el mejor rendimiento en términos de precisión, sensibilidad y especificidad, asegurando así que el modelo sea robusto y efectivo en un entorno clínico real.

Metodología:

Fase 1: Exploración de datos

Objetivo: Obtener claridad sobre la información disponible y la relevancia de las variables en relación con el problema de insuficiencia cardíaca. En esta fase se selecciona una muestra adecuada, se define el tamaño de la muestra, y se aplica un procedimiento de selección. Se analiza la calidad de la información mediante la preparación de los datos, eliminando duplicados, identificando valores faltantes, errores y valores atípicos. Se determina y aplica el preprocesamiento necesario para garantizar que la información sea adecuada para el modelado. El resultado de esta fase es un data set con información de calidad listo para ser utilizado en el desarrollo del modelo.

Fase 2: Análisis y selección de características

Objetivo: Identificar el conjunto de variables que mejor describen a los pacientes con insuficiencia cardíaca. Se analizan las variables actuales y se consideran otras características que podrían influir en el modelo. El análisis de las variables se lleva a cabo usando herramientas estadísticas de correlación y extracción de características que sean relevantes para el modelo predictivo.

Fase 3: Selección de estrategias de aprendizaje

Objetivo: Seleccionar las estrategias que encajen con el problema de predicción de supervivencia, basándose en el conjunto de datos disponible. Se analizan y seleccionan los modelos de Machine Learning más adecuados, considerando aspectos como el número de parámetros a ajustar, el tamaño de la base de datos, el costo computacional del entrenamiento y la validación de los modelos. Se evalúan algoritmos como la regresión logística, árboles de decisión y redes neuronales para determinar cuál ofrece el mejor rendimiento.

Fase 4: Validación del modelo

Objetivo: Validar el modelo mediante simulación con datos históricos. Se emplea una metodología de validación, como el bootstrapping, en la cual el conjunto de datos se divide en un 80% para el entrenamiento del modelo y un 20% para la validación. De esta forma, se asegura que el modelo tiene un buen rendimiento en su capacidad predictiva.

Fase 5: Implementación y despliegue

Objetivo: Valorar la posible implementacion del modelo seleccionado en un entorno de producción entendiendo la limitacion de disponibilidad continua de datos en el marco del ejercicio.

Start date: 10-10-2024

End date: 11-29-2024

Last modified: 10-20-2024

Predicción de riesgo de insuficiencia cardíaca

Data Collection

What data will you collect or create?

El conjunto de datos contiene los registros médicos de 299 pacientes con insuficiencia cardíaca, recolectados en dos hospitales de Faisalabad, Pakistán, durante 2015. El dataset incluye 13 características que abarcan información clínica, corporal y de estilo de vida, como la presencia de anemia, hipertensión, diabetes, los niveles de creatinina y sodio en sangre, y la fracción de eyección del corazón. El propósito principal del dataset es predecir, mediante machine learning, la probabilidad de supervivencia de los pacientes y determinar cuáles son las características más importantes que influyen en este resultado. La variable objetivo es un evento de muerte (si el paciente murió o sobrevivió antes del final del período de seguimiento).

How will the data be collected or created?

El data sheart_failure_clinical_records_dataset.csv es propocionado por Heart Failure Clinical Records, este conjunto de datos está licenciado bajo una licencia Creative Commons Attribution 4.0 International (CC BY 4.0). Esto permite compartir y adaptar los conjuntos de datos para cualquier propósito, siempre que se otorgue el crédito correspondiente.

Documentation and Metadata

What documentation and metadata will accompany the data?

Informacion adicional como tabla de descripcion de las variables e informacion detalladas de cada una de ellas es provista junto con el conjunto de datos.

Ethics and Legal Compliance

How will you manage any ethical issues?

El manejo de los problemas éticos en un proyecto que involucra el uso de informacion medica y de salud ha generado grandes debates y controversias desde hace varias decadas, por lo tanto; es esencial garantizar que todo el proceso se realice de manera informada, responsable y respetuosa. A continuación, un listado de consideraciones éticas y cómo se abordarían:

Privacidad de los datos: Dado que se están utilizando datos personales y clínicos sensibles, es fundamental proteger la privacidad de los pacientes. Para ello, los datos no serán compartidos con ninguna entidad externa al proyecto y se aplicarán estrictas políticas de acceso.

Anonimización de datos: Todos los datos serán anonimizados para proteger la identidad de los pacientes. Se garantizará que ningún dato pueda ser asociado directamente con un individuo específico.

Transparencia: Ser transparente en relación con los objetivos del proyecto y cómo se utilizarán los resultados es clave. Se comunicará claramente a los pacientes y al personal médico la finalidad del estudio y cómo se manejarán los resultados.

Seguridad de los datos: Los datos estarán protegidos mediante métodos de encriptación y otros mecanismos de seguridad para evitar accesos no autorizados o fugas de información.

Equidad y discriminación: Se evitará que el algoritmo de aprendizaje automático introduzca sesgos o discriminaciones en sus predicciones. Esto implica evaluar el modelo para detectar y corregir posibles sesgos que puedan afectar a grupos específicos de pacientes.

Uso responsable de los resultados: El modelo desarrollado será una herramienta de apoyo para los médicos, y sus resultados no deben ser utilizados como la única base para decisiones críticas sin una evaluación clínica adecuada.

Formación y capacitación: Se proporcionará capacitación ética al personal involucrado en el proyecto para garantizar que todos comprendan la importancia de la privacidad y el uso responsable de los datos y los resultados del modelo.

Ética en la comunicación: Los resultados del proyecto serán comunicados de forma ética y transparente, asegurando que tanto los médicos como los pacientes comprendan las limitaciones del modelo y cómo deben interpretarse los resultados.

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

El manejo de problemas de copyright y derechos de propiedad intelectual (DPI) se manejarían de la siguiente manera:

Derechos de los datos: Como ya se menciono anteriormente los datos están licenciados bajo una licencia Creative Commons Attribution 4.0 International (CC BY 4.0). Esto permite compartir y adaptar los conjuntos de datos para cualquier propósito, siempre que se otorgue el crédito correspondiente.

Licencias de software y herramientas: Se verificará que todas las herramientas de software utilizadas en el desarrollo del proyecto tengan las licencias y permisos requeridos. Esto evitará problemas legales relacionados con el uso indebido de software.

Derechos de autor en resultados: Los resultados del proyecto serán comunicados exclusivamente por el equipo de desarrollo. Además, se implementará una marca de identificación en todos los documentos y resultados generados para proteger la propiedad intelectual del proyecto.

Respetar los derechos de terceros: En caso de utilizar información que no provenga del proyecto o que no haya sido producida por el equipo, se citarán adecuadamente las fuentes y se dará el crédito correspondiente.

Conservación de registros: Se mantendrán registros detallados de los datos utilizados y de los resultados obtenidos para cumplir con los requisitos legales y de licencia. Esto ayudará a demostrar el cumplimiento de los derechos de propiedad intelectual.

Evaluación de riesgos: Se realizará una evaluación de riesgos para identificar posibles problemas de propiedad intelectual y tomar medidas preventivas para mitigar cualquier conflicto potencial.

Documentación adecuada: Se garantizará documentación de todos los procedimientos, licencias de software y datos utilizados en el proyecto, con el fin de mantener un control adecuado sobre los derechos de propiedad intelectual.

Generación de patentes: En caso de que se desarrollen procedimientos o herramientas innovadoras durante el proyecto, se evaluará la posibilidad de patentar dichas innovaciones para proteger el trabajo realizado y contribuir al avance del conocimiento científico.

Storage and Backup

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

El almacenamiento es suficiente porque el estudio es realizado en una muestra relativamente pequena 299 pacientes 13 variables. ¿Cómo se realizará la copia de seguridad de los datos? Descargaremos una imagen semanal al equipo local con copia en repositorio en la nube (github). ¿Quién será responsable de la copia de seguridad y la recuperación? Jesus Esteban Tabares es responsable de crear la imagen y además, crear la copia de seguridad. ¿Cómo se recuperarán los datos en caso de incidencia? Los datos serán recuperados localmente y compartidos a todo el equipo, adicionalmente se contara con respaldo de los datos en el repositorio creado para el proyecto.

How will you manage access and security?

Se ha creado un repositorio donde se hara seguimiento del proceso, a el tienen acceso solo los miembros del equipo de desarrollo, durante la fase de construccion.

Selection and Preservation

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

Todos los datos se pueden conservar pues estos son de uso abierto y en ellos no se cuenta con informacion personal de los pacientes para su posible identificacion. ¿Cómo decidirá qué otros datos conservar? Se conservaran solo aquellos datos escenciales. ¿Cuáles son los usos previsibles de los datos en la investigación? Caracterización de los pacientes de acuerdo a su grado de riesgo, e identificacion de factores de riesgo mediante la prevalencia de las variables en el modelo. ¿Durante cuánto tiempo se conservarán los datos? Los datos serán conservados solamente durante la implementación del proyecto

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

Los datos utilizados para el desarrollo del modelo pueden ser conservado en el largo plazo dada la naturaleza de los mismo, sin embargo; en una eventual fase de despliegue la naturaleza de todo lo anteriormente senalado en relacion con el manejo de los datos debera ser repensada.

Data Sharing

How will you share the data?

Los datos seran comunicados mediante la creacion de un paper que recoja los resultados de los modelos valorados y su desempeno, asi como graficas y tablas con las consideraciones asumidas en el desarrollo del proyecto. Como no se menciona informacion personal de los pacientes este podria ser publicado en diferentes medios para su aprovechamiento. Los datos son abiertos lo que no comprometen su uso mas alla del otorgamiento de los creditos por su uso.

Are any restrictions on data sharing required?

Para el desarrollo del proyecto no se evidencia ninguna restriccion inicial para la publicacion de la informacion, en la fase de despliegue sin embargo si se deben valorar las necesidades de estas pero eso esta mas alla del alcance de este trabajo. La unica restriccion asociada puede presentarse en el caso de generacion de conocimiento o procesos patentables pero no es la intencion de este proyecto.

Responsibilities and Resources

Who will be responsible for data management?

La responsabilidad sobre el manejo de los datos recae directa y exclusivamente sobre los integrantes del equipo de desarrollo, al no haber una injerencia externa, el compromiso con la gestion de toda la informacion es obligacion principal de todo el equipo.

What resources will you require to deliver your plan?

El proposito principal es el despliegue de la solucion construida, razon por la cual los recursos principales se asocian a este proceso, AWS y sus herramientas constituyen la base del proceso, pero es claro que antes de esto es neceasario desarrollar y valorar los modelos y su comportamiento para lo cual es fundamental el uso de pythom y sus librerias especializadas. En caso de ser necesario es posible utilizar alguna herramienta adicional para la visualizacion o presentacion de los resultados. Todo lo anterior hace referencia a recursos fisicos, adicionalmente, es clara la necesidad de las capacidades de cada uno de los miembros del equipo de desarollo y el acompanamiento del cuerpo docente para la optimizacion de los resultados.

Planned Research Outputs

Model representation - "Predicción de riesgo de insuficiencia cardíaca"

Modelo basado en datos que pueda predecir el riesgo (alto o bajo) de padecer insuficiencia

---

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?
Predicción de riesgo de insuficiencia cardíaca	Model representation	Unspecified	Open	None specified		Creative Commons Attribution 4.0 International	None specified	No	No