Documentación del desarrollo de proyecto ISSSTELEON

edgar.hinojosas@uanl.edu.mx

September 2026

1 Introducción

El presente informe tiene como intención documentar requisitos, criterios, reglas de negocio y decisiones arquitectónicas que fuesen incorporándose sobre la marcha del desarrollo.

2 Variables derivadas

ID	Requisitos
VAR-01	El resultado de VPH necesita ser extraído por OCR de la columna `cit_vph_pdf_blob` de la tabla `citologia_cervical` para las columnas `cit_vph_resultado`, `cit_vph_genotipo_16`, `cit_vph_genotipo_18` de la misma tabla.
VAR-02	El resultado de biopsia cervical necesita ser extraído por OCR de la columna `bx_pdf_blob` de la tabla Y para la columna `bx_cin_grado` de la misma tabla.
VAR-03	Cada extracción de OCR debe contener un nivel de confianza que codifique en escala de 0 a 1 el estatus (fallido, o realizado), necesidad de inspección manual, método alternativo usado, método principal en las columnas `cit_vph_ocr_confianza`, `bx_ocr_confianza`.
VAR-04	Documentar modelo de extracción de OCR para citar artículo de desarrollo respectivo en entregable e identificar características disponibles.
VAR-07	Cada cadena de transformación en ETL o consulta cuente con una descripción de transformación de datos crudos, y validación de calidad de los datos (numeral 7).
VAR-08	Documentar la fuente de los datos, fecha de inicio y fin al momento de cargar los datos en crudo para poder describirlo en el reporte (numeral 5a).
VAR-09	Los métodos de extracción deben disponer de permiso de lectura y escritura de las tablas, esto último para prevenir corridas duplicadas innecesarias gracias a la persistencia del resultado.

Cuadro 1: Requisitos de extracción y trazabilidad de variables obtenidas por OCR de proyecto cacu

ID	Requisitos
VARM-01	El resultado de VPH necesita ser extraído por OCR de la columna `cit_vph_pdf_blob` de la tabla `citologia_cervical` para las columnas `cit_vph_resultado`, `cit_vph_genotipo_16`, `cit_vph_genotipo_18` de la misma tabla.
VARM-02	El resultado de biopsia cervical necesita ser extraído por OCR de la columna `bx_pdf_blob` de la tabla Y para la columna `bx_cin_grado` de la misma tabla.
VARM-03	El resultado de biopsia de mama necesita ser extraído por OCR de la columna `bx_grado_histologico` de la tabla `biopsia_mama` para la columna `bx_grado_histologicon` de la misma tabla.
VARM-04a	Cada extracción de OCR debe contener un nivel de confianza que codifique en escala de 0 a 1 el estatus (fallido, o realizado), necesidad de inspección manual, método alternativo usado, método principal.
VARM-05	Documentar modelo de extracción de OCR para citar artículo de desarrollo respectivo en entregable e identificar características disponibles.

Cuadro 2: Requisitos de extracción y trazabilidad de variables obtenidas de proyecto cama

Los detalles de las fuentes de cada dato también se encuentran en la documentación de los proyectos. Para el proyecto de TEA, aún no se definen las fuentes, las cuales están sujetas a documentos de muestra que se solicitarán.

3 Indicadores

Los indicadores se dividen en las categorias: clínicos y de oportunidad. La última se apega a la siguiente tabla

ID	Requisitos
IND-01	La fecha de nacimiento es requerida para que el sistema calcule la edad en cada evento de interés
IND-02	La únidad de análisis depende de a quién va dirigida el indicador. Para evaluación del desempeño del modelo la unidad de análisis será un episodio de tamizaje, diagnóstivo o tratamiento por paciente EUG.
IND-03	Para construir un episodio clínico se identifica código CIE10, especialidad médica, delimitación de ventana de tiempo fija apartir de evento anterior, y por último contenidos de registros escritos por los médicos.

Cuadro 3: Requisitos generales para la construcción de indicadores: unidad de análisis y definición de episodio clínico

4 Cohortes

Estos deben ser tablas materializadas seguidas de un proceso de exclusión registrado y versionado, así como lo estipulan los requisitos de TRIPODAI-6. Si nuevos pacientes son incluidos, o el periodo de análisis es aumentado la versión debe ser actualizada para replicabilidad de resultados de los modelos, por tanto es conveniente tener una versión fija y persistente separada de las tablas de origen. Los criterios ya implementados corresponden a la , formato, y plausibilidad de los datos. Para cada escenario clínico corresponde un modelo, para cada modelo, corresponde una tabla de cohorte definida.

ID	Requisitos
EL-01	La fuente de datos será apartir de la tabla de episodios clínicos según el escenario clínico de la cohorte.
EL-02	El criterio de elegibilidad clínica debe ser versionado.
EL-03	Cada tabla de cohorte deberá tener un hash que identifique su versión.
EL-04a	Los criterios de exclusión de la cohorte se aplican solo al construir la tabla de entrenamiento.
EL-04b	En producción, cada fila nueva se evalúa contra un subconjunto de criterios de aplicabilidad (edad, histerectomía previa, otro cáncer en el área). Los criterios que dependen de seguimiento futuro quedan fuera de esta evaluación.
EL-05	Cada criterio de exclusión debe tener un registro con un código de razón, se debe dejar registro junto a los artefactos del modelo.
EL-07	Cada construcción de la tabla de cohorte debe generar un conteo de exclusión por criterio, materializado junto con la tabla (numeral 20a).

Cuadro 4: Requisitos de construcción, versionado y exclusión de las tablas de cohorte por escenario clínico

No se usan las tablas de origen dado que estan sujetas a actualizaciones imprevistas, modificaciones por rutinas externas como de estracción, lo acuál atenta con la replicabilidad de los resultados de las rutinas. Supongase que se tiene una tabla de contenidos A, se ajusta el modelo y se tienen resultados satisfactorios, si mañana se cambia dicha tabla a contenidos B, el modelo no podrá volver replicar dichos resultados.

Ejemplo de Cohorte del Kaiser Permanente Northern California [Hedderson et al., 2022]

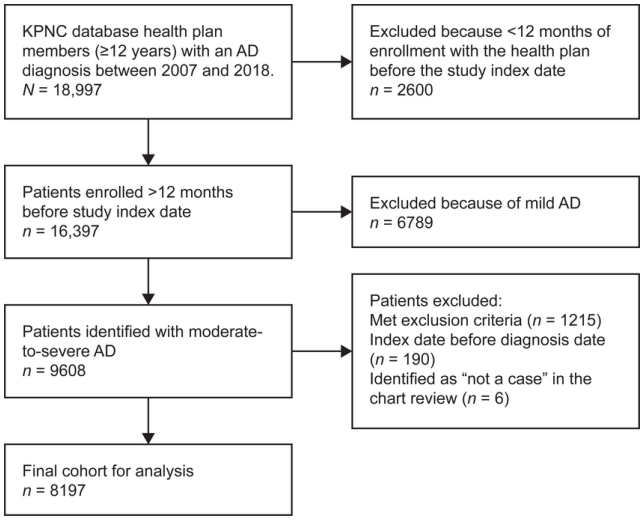


Figura 1: Ejemplo de cómo realizan cohortes para análisis. Si bien se cuenta con la fuente original de datos, se termina usando la versión final para el análisis.

5 Desarrollo de modelos

El desarrollo de modelos se llevará acabo en un repositorio dedicado por proyecto. Todo modelo se tendrá que entrenar con su respectiva tabla de cohorte. Distintos trabajos orbitales al desarrollo del modelo son necesarios de realizar para la satisfacción de los alcances del proyecto en apego a los numerales 10-14. Necesitamos tener control de ejecución y acceso a artefactos derivados de rutinas desarrolladas por nosotros: muestreo, tratamiento de datos faltantes, métodos análiticos, balance de clases, y estimación de nivel de confianza de las sálidas.

ID	Requisitos
ML-01a	Control de ejecución de rutinas en desarrollo y acceso a sus artefactos derivados
ML-01b	Caracterísitca de exportado de artefactos.
ML-02	Las versiones de desarrollo así como sus rutinas orbitales serán etiquetadas y distinguibles de la versión para producción.
ML-03	Cada conjunto de artefactos y métricas derivado del entrenamiento deberá ser enlazable al hash de la versión de la tabla de datos que se uso para derivarlos 18f).
ML-04	Los procesos aquí corriendo solamente serán con permiso de lectura.
ML-05	Dependiendo de la rutina a ejecutar cada trabajo puede requerirse aislar su entorno de ejecución para ligerar las dependencias de las demas rutinas.

Cuadro 5: Requisitos de control de ejecución, artefactos y trazabilidad durante el entrenamiento de modelos

La importancia de conocer la versión de la tabla fuente que entreno un modelo es por replicabilidad de los resultados y artefactos.

6 Despliegue de modelos

ID	Requisitos
MLD-01	Cualquier despliegue a producción de un modelo debe pasar por un conjunto de pruebas técnicas documentadas en función de sus métricas de desempeño y nivel de confianza casuístico de las estimaciones.
MLD-01b	Cualquier despliegue a producción de un modelo debe ofrecer validación clínica documentada.
MLD-03	Los pesos y el código para realizar una estimación serán reportados y visualizados para trazabilidad e interpretabilidad clínica (OR, HR, probabilidad acumulada).
MLD-4	La estimación predicha por el modelo debe contar con versión del mismo, de modo tal que toda la cadena por la que paso sea trazable y auditable.
MLD-5	Dicha estimación debe persistir en una tabla materializada propia junto con la información indicada en el punto anterior.

Cuadro 6: Requisitos de validación técnica y clínica previos al paso de un modelo a producción

7 Sincronización y actualización de las fuentes de información

La actualización de cada una de las variables por el equipo médico a lo largo de un episodio de tamizaje es documentado. El actual acceso a la información por cada uno de los departamentos es a través de hojas de excel compartidas y rellenadas periódicamente en un día programado por cada departamento asignado.

Responsable	Día
Enfermería	Viernes
Patología/laboratorio	Lunes
Trabajo Social	Martes – Viernes
Clínica de la Mujer	Martes – Viernes
Oncología	Martes – Viernes

Cuadro 7: Días de actualización de información por departamento responsable

Departamento	Variables capturadas
Patología/laboratorio	Resultado de citología; resultado de VPH
Trabajo Social	¿Se localizó a paciente?; motivo de localización; observaciones
Clínica de la Mujer y Colposcopia	Fecha de cita; diagnóstico colposcópico; toma de biopsia; resultado de biopsia; resultado crítico; fecha del resultado
Oncología	Plan terapéutico; estatus; prioridad en la cita; especialidad relacionada

Cuadro 8: Variables capturadas por departamento durante el episodio de tamizaje según excel proporcionado.

7.1 Fuentes normativas y metodológicas

Cuadro 9: Numerales de la lista TRIPOD+AI (checklist expandido, versión 11-enero-2024) citados en los requisitos técnicos de este documento [Cohen and Bossuyt, 2024]

Numeral	Texto (traducción)	Ref. usada en
Sección: Datos (Methods – Data)		
5a	Describir las fuentes de datos por separado para los conjuntos de desarrollo y de evaluación (p. ej., ensayo aleatorizado, cohorte, atención rutinaria o datos de registro), la justificación de su uso, y la representatividad de los datos.	§??
5b	Especificar las fechas de recolección de los datos de los participantes, incluyendo el inicio y el fin del reclutamiento y, si aplica, el fin del seguimiento.	EL-0X
Sección: Preparación de datos (Methods – Data preparation)		
7	Describir cualquier preprocesamiento y verificación de calidad de los datos, incluyendo si este fue similar entre los grupos sociodemográficos relevantes.	VAR-06
Sección: Ciencia abierta (Open science)		
8a	Indicar la fuente de financiamiento y el papel de los financiadores en el estudio.	—
8b	Declarar cualquier conflicto de interés y las declaraciones financieras de todos los autores.	—
8c	Indicar dónde puede consultarse el protocolo del estudio, o declarar que no se preparó uno.	—
8d	Proporcionar la información de registro del estudio (nombre del registro y número), o declarar que no fue registrado.	—
8e	Proporcionar los detalles de disponibilidad de los datos del estudio.	—
18f	Proporcionar los detalles de disponibilidad del *código analítico*[1].	ML-02, ML-03
Sección: Resultados (Results – Participants)		
20a	Describir el flujo de participantes a través del estudio, incluyendo el número de participantes con y sin el desenlace y, si aplica, un resumen del tiempo de seguimiento. Un diagrama puede ser útil.	EL-05, EL-07
20b	Reportar las características generales y, cuando aplique, por cada fuente de datos o entorno, incluyendo fechas clave, predictores clave (incluyendo datos demográficos), tratamientos recibidos, tamaño de muestra, número de eventos del desenlace y cantidad de datos faltantes. Una tabla puede ser útil. Reportar cualquier diferencia entre grupos demográficos clave.	—
20c	Para la evaluación del modelo, mostrar una comparación con los datos de desarrollo de la distribución de los predictores importantes (datos demográficos, predictores y desenlace).	—

[1] Se refiere al código de análisis: por ejemplo, limpieza de datos, ingeniería de características, construcción y evaluación del modelo (nota §del checklist original).

Referencias

[Cohen and Bossuyt, 2024] Cohen, J. F. and Bossuyt, P. M. (2024). Tripod+ ai: an updated reporting guideline for clinical prediction models.

[Hedderson et al., 2022] Hedderson, M. M., Asgari, M. M., Xu, F., Quesenberry, C. P., Sridhar, S., Geier, J., and Lemeshow, A. R. (2022). Rates of cardiovascular events among patients with moderate-to-severe atopic dermatitis in an integrated health care system: a retrospective cohort study. *PLoS One*, 17(11):e0277469.