Planilla de dudas sobre variables del dataset

Equipo ML

26 de octubre de 2025

Resumen del dataset

Nombre del dataset	RECUIMA (Registro Cubano de Infarto Agudo de Miocardio)
Versión	v0.2 (post-limpieza inicial)
Fuente / Sistema origen	Registro hospitalario de pacientes con infarto agudo de miocardio
Periodo de cobertura	2016–2025
Población y unidad de análisis	Pacientes ingresados con diagnóstico de IAM;unidad: episodio de internación
Número de registros / variables	3,112 registros / 185 variables (después de limpieza inicial)
Fecha de extracción	02/04/2025
Objetivo analítico	Predicción de mortalidad intrahospitalaria en pacientes con IAM mediante modelos de machine learning
Restricciones legales / privacidad	Se eliminaron variables de identificación personal (nombres, números de identidad, números de contacto) para cumplir con protección de datos
Notas generales de calidad (duplicados, faltantes, codificaciones especiales)	Se observa presencia significativa de valores faltantes en múltiples variables. Existen duplicaciones de variables clave que requieren aclaración. El dataset parece ser resultado de la fusión de múltiples registros o fuentes. Se eliminaron variables redundantes identificadas en la limpieza inicial: `anno`, `numero` (identificador), `unidad`.

Nota importante sobre duplicaciones: Se han identificado las siguientes variables que aparecen duplicadas en el dataset: `presion_arterial_sistolica`, `presion_arterial_diastolica`, `asa`, `betabloqueadores`, `ieca`, `estatinas`, `clopidogrel`, `furosemida`, `nitratos`, `anticoagulantes`, `otros_diureticos`, `fecha_egreso` y `fecha_ingreso`. Se desconoce el motivo exacto de estas duplicaciones, aunque se presume que puede deberse a: (1) registro en diferentes momentos temporales (ingreso vs. egreso), (2) fusión de múltiples fuentes de datos, o (3) diferencias entre prescripción y administración real. Se requiere aclaración urgente sobre la interpretación correcta de estas columnas duplicadas.

Guía rápida

- Tipo: numérico, categórico, booleano, fecha/hora, texto libre, identificador. - Códigos especiales: por ejemplo, -1, 9, 99, 999 = “desconocido/no aplica”. - Estados: Pendiente, Enviado, Resuelto, Rechazado, En progreso.

1. Tabla maestra de variables

Variable	Descripción	Tipo	Unidad / Rango o Dominio	Faltantes y códigos	Reglas / Validación	Dudas principales
numero	Identificador único del paciente	numérico	enteros positivos	ninguno	único por paciente	Eliminada varias veces: variable redundante
anno	Año de registro	numérico	2016–2025	ninguno	–	Eliminada varias veces: variable redundante
unidad	Código de unidad hospitalaria	numérico	códigos específicos	presentes	–	Eliminada varias: variable redundante
fecha ingreso	Fecha de ingreso hospitalario	fecha	formato dd/mm/yyyy	presentes	fecha válida	**DUPLICADA**: requiere aclaración
fecha egreso	Fecha de egreso hospitalario	fecha	formato dd/mm/yyyy	presentes	>= fecha_ingreso	**DUPLICADA**: asociada a reingresos
numero identidad	Número de documento de identidad	texto	–	–	–	Eliminada: protección de datos
numero contacto	Número telefónico de contacto	texto	–	–	–	Eliminada: protección de datos
nombre	Nombre del paciente	texto	–	–	–	Eliminada: protección de datos
primer apellido	Primer apellido del paciente	texto	–	–	–	Eliminada: protección de datos
segundo apellido	Segundo apellido del paciente	texto	–	–	–	Eliminada: protección de datos
edad	Edad del paciente en años	numérico	0–120 años	escasos	>0, <120	Ninguna
sexo	Sexo del paciente	categórico	masculino, femenino	presentes	dominio cerrado	Requiere codificación binaria
color piel	Etnia o color de piel registrado	categórico	blanca, mestiza, negra	presentes	dominio cerrado	Requiere codificación numérica
peso	Peso corporal del paciente	numérico	kg, 20–200	presentes	>0	Ninguna
talla	Estatura del paciente	numérico	cm, 100–220	presentes	>0	Ninguna
imc	Índice de masa corporal calculado	numérico	kg/m², 10–60	presentes	peso/(talla/100)²	Ninguna
provincia	Nombre de la provincia	categórico	nombres normalizados	presentes	–	Valores con caracteres especiales
municipio	Nombre del municipio	categórico	nombres normalizados	presentes	–	Valores con caracteres especiales
area_salud	Código de área de salud	categórico	códigos específicos	presentes	–	Valores con caracteres especiales
idprovincia	Identificador numérico de provincia	numérico	enteros positivos	ninguno	–	**DUPLICADA**: columna repetida
idmunicipio	Identificador numérico de municipio	numérico	enteros positivos	ninguno	–	Ninguna
idareasalud	Identificador numérico de área de salud	numérico	enteros positivos	presentes	–	Ninguna
atencion inicial	Tipo de atención inicial recibida	categórico	servicio, cuerpo, sala	presentes	–	Requiere documentación del significado
horario llegada	Horario de llegada al hospital	categórico	7am7pm, 7pm7am	presentes	–	Requiere codificación binaria
ecg_previo	Electrocardiograma previo realizado	booleano	si, no	presentes	–	Requiere codificación binaria
ecg	Código de hallazgo electrocardiográfico	numérico	enteros (5–35)	presentes	–	Requiere tabla de correspondencia
llamada emergencias	Llamada al servicio de emergencias	categórico	si, no	presentes	–	Requiere codificación binaria
tiempo respuesta	Tiempo de respuesta de emergencias	numérico	probablemente minutos	presentes (solo si llamada=si)	>=0	Requiere confirmación de unidad
tiempo llegada	Tiempo de llegada al hospital	numérico	probablemente minutos	presentes (solo si llamada=si)	>=0	Requiere confirmación de unidad
primera						

2. Registro de dudas y resoluciones

ID	Variable	Duda / Observación	Evidencia / Contexto	Impacto	Prioridad	Responsable (consulta a)	Estado	Fechas (sol./cierre)
001	Variables duplicadas (grupo)	Se identifican 12 variables que aparecen duplicadas en el dataset. Se requiere aclaración sobre el significado de cada columna duplicada y cuál utilizar para el análisis.	Variables afectadas: presion_arterial_sistolica, presion_arterial_diastolica, asa, betabloqueadores, ieca, estatinas, clopidogrel, furosemida, nitratos, anticoagulantes, otros_diureticos, fecha_ingreso. Posibles hipótesis: (a) medición en momentos diferentes (ingreso vs egreso), (b) fusión incorrecta de datasets, (c) prescripción vs administración real	Alto	Alta	Cardiólogo	Pendiente	2025-10-25 /
002	asa	¿La variable ASA representa dosis en mg, días de tratamiento, o simplemente presencia/ausencia?	Valores observados son numéricos diversos (rango 0–24). No se identifica patrón claro de dosificación estándar	Alto	Alta	Cardiólogo	Pendiente	2025-10-25 /
003	betabloqueadores	¿Qué escala o codificación se utiliza para betabloqueadores?	Valores numéricos diversos (rango 0–36). No se identifica si es dosis, días, o código categórico	Alto	Alta	Cardiólogo	Pendiente	2025-10-25 /
004	ecg	Solicitar tabla de correspondencia completa para códigos ECG	Valores observados: 5, 10, 12, 15, 20, 25, 30, 35. Sin documentación disponible sobre significado clínico	Alto	Alta	Cardiólogo	Pendiente	2025-10-25 /
005	scacest, scacest_ secundario	Requiere explicación clínica detallada de la diferencia entre SCACEST primario y secundario	Ambas variables booleanas (0/1), pero no se comprende la distinción clínica ni criterios diagnósticos	Alto	Alta	Cardiólogo	Pendiente	2025-10-25 /
006	indice_killip, indice_mkillip	¿Cuál es la diferencia entre clasificación Killip y Killip modificada? ¿Cuándo se utiliza cada una?	Ambas utilizan números romanos I–IV. En algunos registros difieren, en otros coinciden	Medio	Alta	Cardiólogo	Pendiente	2025-10-25 /
007	tiempo puerta_aguja	Confirmar definición clínica exacta y unidad de medida	Asumimos minutos desde llegada hospitalaria hasta inicio de trombolisis, pero requiere confirmación oficial	Alto	Alta	Cardiólogo	Pendiente	2025-10-25 /
008	tiempo isquemia	Confirmar definición clínica exacta y unidad de medida	Asumimos minutos desde inicio de síntomas hasta reperfusión, pero requiere confirmación oficial	Alto	Alta	Cardiólogo	Pendiente	2025-10-25 /
009	tiempo respuesta, tiempo_llegada	Confirmar unidades de medida y definiciones exactas de cada variable	Valores presentes solo cuando llamada_emergencias=si. Asumimos minutos pero sin documentación	Medio	Media	Cardiólogo	Pendiente	2025-10-25 /
010	escala_grace	Validar rango de valores observados contra rango teórico esperado (0–372)	Valores parecen consistentes pero requiere validación clínica del cálculo	Medio	Media	Cardiólogo	Pendiente	2025-10-25 /

ID	Variable	Duda / Observación	Evidencia / Contexto	Impacto	Prioridad	Responsable (consulta a)	Estado	Fechas (sol./cierre)
011	Derivaciones ECG (V1–V9, D1–D3, AVL, AVF, AVR, V3R, V4R)	¿Estas variables indican presencia de alteración en cada derivación? ¿Qué tipo de alteración (supradesnivel, infradesnivel, onda Q)?	Variables booleanas (0/1) con muchos faltantes. Requiere documentación de criterios diagnósticos utilizados	Alto	Alta	Cardiólogo	Pendiente	2025-10-25 /
012	avc, mpt, vam, mpp	Confirmar significado exacto de acrónimos	Asumimos: AVC=asistencia ventricular, MPT=marcapaso temporal, VAM=ventilación mecánica, MPP=marcapaso permanente	Medio	Media	Cardiólogo	Pendiente	2025-10-25 /
013	cabg	Confirmar acrónimo CABG (¿Coronary Artery Bypass Graft?)	Asumimos cirugía de revascularización coronaria pero requiere confirmación	Bajo	Baja	Cardiólogo	Pendiente	2025-10-25 /
014	reperfusion	Explicar diferencia entre categorías y criterios de clasificación	Valores: no, parcial, total, otro. Requiere definición de criterios clínicos utilizados	Alto	Alta	Cardiólogo	Pendiente	2025-10-25 /
015	coronariografia	Explicar diferencia entre valores observados	Valores: no, si, otro, centro. Especialmente aclarar significado de çentro"vs .ºtro"	Medio	Media	Cardiólogo	Pendiente	2025-10-25 /
016	Grupo variables de calidad	Grupo de variables con muchos faltantes que requieren contexto clínico	Variables: razones_documentadas, riesgo_beneficio, anti_agregacion_plaquetaria, proteccion_embolica, funcion_renal, volumen_contraste, prescripcion_optima. ¿Son indicadores de calidad o checklist?	Medio	Media	Cardiólogo	Pendiente	2025-10-25 /
017	Laboratorio (grupo)	Confirmar unidades de medida para todas las variables de laboratorio	Variables: colesterol, creatinina, filtrado_glomerular, trigliceridos, glicemia, leuco, hb, ck, ckmb. Especificar unidades (mg/dL, mmol/L, etc.)	Alto	Alta	Laboratorio	Pendiente	2025-10-25 /
018	Ecocardiografía (grupo)	Confirmar significado de acrónimos: ud, pat, insao, estao, insmit, estmit	Asumimos insuficiencias y estenosis valvulares. ud y pat sin identificar	Medio	Media	Cardiología	Pendiente	2025-10-25 /
019	arteria	Solicitar tabla completa de códigos de arterias coronarias	Códigos: cd, cx, ada. Asumimos: CD=coronaria derecha, CX=circunfleja, ADA=descendente anterior	Medio	Media	Cardiólogo	Pendiente	2025-10-25 /
020	abordaje	Solicitar tabla completa de tipos de abordaje/intervención coronaria	Valores: stent_farmaco, stent_metalico, ninguno. ¿Existen otros valores posibles?	Medio	Media	Cardiólogo	Pendiente	2025-10-25 /
021	resultado	Explicar categorías de seguimiento y su significado clínico	Valores: vivo_sin, vivo_con, noevaluado, alta, fallecido. Requiere explicación de "vivo_sin"vs "vivo_con"(¿con/sin complicaciones?)	Medio	Media	Cardiólogo	Pendiente	2025-10-25 /

ID	Variable	Duda / Observación	Evidencia / Contexto	Impacto	Prioridad	Responsable (consulta a)	Estado	Fechas (sol./cierre)
022	angina24h	¿Se refiere a angina en 24h previas al ingreso o durante primeras 24h de hospitalización?	Variable booleana con contexto temporal ambiguo que afecta interpretación clínica	Bajo	Baja	Cardiólogo	Pendiente	2025-10-25 /
023	provincia, municipio, area_salud	Normalizar caracteres especiales en nombres geográficos	Caracteres especiales mal codificados: Sancti Spíritus, Cabaiguán, Camagüey, Manatí, Güines	Bajo	Baja	Admin. datos	Pendiente	2025-10-25 /
024	insulina	¿Variable representa dosis, tipo de insulina, duración de tratamiento, o uso binario?	Valores numéricos diversos sin patrón identificable. Dificulta su uso en modelado	Medio	Media	Cardiólogo	Pendiente	2025-10-25 /
025	lugar trombolisis	¿Es necesaria esta variable para el objetivo analítico del proyecto?	Valores: ucie, sala, servicio. Considerada candidata a eliminación según relevancia clínica	Bajo	Baja	Equipo ML	Pendiente	2025-10-25 /

Acciones siguientes y próximos pasos

Consultas prioritarias

1. **Duplicación de variables (Alta prioridad):** Aclarar el significado y uso correcto de las 12 variables que aparecen duplicadas en el dataset. Esta es la duda más crítica que afecta la calidad del análisis.
2. **Escalas y codificaciones de medicamentos:** Definir las escalas utilizadas para asa, betabloqueadores, insulina y confirmar codificaciones de otros medicamentos.
3. **Variables electrocardiográficas:** Proporcionar tabla de correspondencia para códigos ECG y documentar criterios para derivaciones (V1–V9, D1–D3, AVL, AVF, AVR, V3R, V4R).
4. **Variables de tiempo:** Confirmar definiciones clínicas exactas y unidades de medida para tiempo_puerta_aguja, tiempo_isquemia, tiempo_respuesta, tiempo_llegada.
5. **Clasificaciones clínicas:** Explicar diferencias entre SCACEST primario/secundario, Killip/Killip modificado, y categorías de reperfusión.
6. **Acrónimos y términos técnicos:** Confirmar significado de acrónimos no documentados (avc, mpt, vam, mpp, cabg, acd, ada, acx, insao, estao, insmit, estmit, ud, pat).
7. **Tablas de códigos:** Solicitar tablas completas para arteria, abordaje, resultado, complicaciones.
8. Especificar unidades de medida para todas las variables de laboratorio (colesterol, creatinina, filtrado glomerular, triglicéridos, glicemia, leucocitos, hemoglobina, CK, CK-MB).
9. Confirmar rangos de referencia y límites de detección de los equipos utilizados.

Tareas del equipo de machine learning

1. Mantener actualizado el Registro de dudas y resolucionesçon estados y fechas.
2. Documentar todas las decisiones de preprocesamiento tomadas en ausencia de aclaraciones.
3. Priorizar variables según impacto en el objetivo analítico (predicción de mortalidad intrahospitalaria).
4. Preparar pipeline de preprocesamiento flexible que permita incorporar aclaraciones posteriores.
5. Normalizar caracteres especiales en nombres geográficos.

Nota sobre el proceso de limpieza inicial

Este dataset es resultado de una primera etapa de limpieza en la que se eliminaron:

- Variables de identificación personal (nombre, primer_apellido, segundo_apellido, numero_identidad, numero_contacto) para cumplir con protección de datos.
- Variables redundantes identificadas en análisis preliminar (anno, numero como identificador, unidad).

El dataset original parece ser resultado de la fusión de múltiples fuentes o registros hospitalarios, lo cual explicaría algunas de las duplicaciones observadas. Se requiere confirmación de esta hipótesis por parte del equipo responsable de la recolección de datos.