Predicción de Hospitalización por Dengue

Metodología CRISP-DM

José Ronaldo Duran Toloza

Rafael Eduardo Garcia Montes

Curso: Aprendizaje de Máquinas

Docente: John Fernando Vargas Buitrago

28 de febrero de 2026

Índice

1. Entendimiento del negocio

1.1. Descripción del negocio

El presente proyecto se desarrolla en el sector de salud pública, utilizando datos abiertos provenientes de la plataforma MEData de la Alcaldía de Medellín (Colombia). El estudio se enfoca en el análisis de pacientes diagnosticados con dengue, enfermedad viral transmitida por el mosquito *Aedes aegypti*, que constituye un problema recurrente de salud pública en regiones tropicales.

Desde la perspectiva del sistema de salud, la hospitalización de pacientes con dengue representa un alto costo en términos de recursos hospitalarios, talento humano y ocupación de camas. Adicionalmente, la identificación tardía de pacientes con riesgo de complicaciones puede derivar en desenlaces clínicos graves e incluso en mortalidad.

En este contexto, contar con un modelo predictivo que permita anticipar la probabilidad de hospitalización de un paciente con síntomas de dengue puede contribuir a:

- Optimizar la asignación de recursos hospitalarios.
- Implementar intervenciones médicas tempranas.
- Reducir costos asociados a hospitalizaciones evitables.
- Disminuir complicaciones clínicas.
- Apoyar la toma de decisiones médicas basada en datos.

Este proyecto se desarrolla con fines académicos en el marco del programa de maestría, aplicando técnicas de minería de datos bajo la metodología CRISP-DM.

1.2. Descripción del problema

El problema consiste en predecir si un paciente diagnosticado con dengue requerirá hospitalización, a partir de variables demográficas, administrativas y sintomatológicas registradas en el sistema de vigilancia epidemiológica.

La variable objetivo del modelo es:

- **pac_hos_**: Variable binaria que indica si el paciente fue hospitalizado (1) o no (0).

El problema se formula como una tarea de clasificación supervisada binaria, dado que el modelo aprenderá a partir de datos históricos etiquetados.

Actualmente, la decisión de hospitalización se toma con base en criterios clínicos establecidos por protocolos médicos. No obstante, un modelo predictivo puede servir como herramienta de apoyo para:

- Identificar pacientes con alto riesgo de hospitalización.
- Priorizar seguimiento clínico.
- Mejorar protocolos de atención temprana.

Un desafío adicional del problema es el desbalance de clases, dado que la proporción de pacientes hospitalizados es considerablemente menor que la de pacientes no hospitalizados.

1.3. Objetivos de la minería

1.3.1. Objetivo general

Desarrollar y evaluar un modelo de clasificación supervisada capaz de predecir la hospitalización de pacientes con diagnóstico de dengue, utilizando variables demográficas y clínicas registradas en datos abiertos de la ciudad de Medellín.

- Analizar la calidad y estructura del conjunto de datos.
- Realizar procesos de limpieza y transformación de datos.
- Construir modelos de clasificación supervisada para la variable objetivo `pac_hos_`.
- Optimizar hiperparámetros con validación cruzada para maximizar desempeño.
- Comparar modelos clásicos y ensambles para seleccionar el mejor.

1.4. Diseño de solución

Se implementó un flujo CRISP-DM: comprensión de datos, preparación, modelamiento con hiperparametrización, evaluación en *hold-out* y despliegue básico con Streamlit.

Tipo de análisis	Tipo de aprendizaje	Métodos	Evaluación
Predictivo	Supervisado	Clásicos + Ensambles	F1-macro, recall, precision, accuracy

Cuadro 1: Diseño general del análisis

2. Entendimiento de los datos

2.1. Diccionario de datos

El dataset de trabajo `dengue_mod.csv` contiene **53.813 registros** y **22 variables**. Incluye variables demográficas, geográficas y síntomas asociados al dengue.

La variable objetivo del problema es `pac_hos_`. En el origen está codificada como `1=Sí hospitalización, 2=No`; para modelado se transformó a binaria `1=Sí, 0=No`.

Variable	Descripción	Tipo	Rol
ID	Número consecutivo	Numérica	Entrada
semana	semanas del año de 1 a 53	Categórica	Entrada
edad_	Edad	Numérica	Entrada
uni_med_	Unidad de medida: 0= No aplica, 1=Años, 2=Meses, 3=Días, 4=Horas, 5=Minutos SD=Sin informacion	Categórica	Entrada
sexo_	M=Masculino, F=Femenino, SD=Sin informacion	Categórica	Entrada

Variable	Descripción	Tipo	Rol
nombre_barrio	Texto asociado a la tabla de barrios definidos por la entidad territorial, Vacios se diligencian con "Sin iNformacion", Sin ubicación en zona urbana.	Categórica	Entrada
comuna	Texto asociado a la tabla de comunas definidos por la entidad territorial, Vacios se diligencian con "Sin iNformacion", Sin ubicación en zona urbana.	Categórica	Entrada
tipo_ss_	Tipo de Régimen de seguridad social: C= Contributivo, S=Subsidiado, P=Excepción, E=Especial, N= No asegurado, I= Indeterminado/Pendiente, SD=Sin informacion.	Categórica	Entrada
cod_ase_	Codigo de la aseguradora	Categórica	Entrada
fec_con_	Fecha de Consulta	Categórica	Entrada
ini_sin_	Fecha de inicio de síntomas	Categórica	Entrada
tip_cas_	Tipo de caso: 1=Sospechoso, 2= Probable , 3=Confirmado por laboratorio , 4=Confirmado por clinica , 5= Confirmado por nexo epidemiológico.	Categórica	Entrada
pac_hos_	Paciente hospitalizado 1= Si, 2=No.	Categórica	Salida
cod_dpto_r	Código departamento residencia	Categórica	Entrada
cod_mpio_r	Código municipio residencia	Categórica	Entrada
cod_dpto_o	Código departamento ocurrencia	Categórica	Entrada
cod_mpio_o	Código municipio ocurrencia	Categórica	Entrada
desplazami	1= Si, 2=No. SD= Sin Informacion	Categórica	Entrada
cod_mun_d	Código municipio desplazamiento	Categórica	Entrada
clas_dengue	0= No aplica, 1=Dengue sin signos de alarma, 2=Dengue con signos de alarma 3=Dengue grave, SD=Sin Informacion	Categórica	Entrada
fiebre	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
cefalea	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
dolrretroo	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
malgias	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
artralgia	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
erupcionr	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
dolor_abdo	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
vomito	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
somnolenci	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
hipotensio	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
hepatomeg	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
hem_mucosa	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
hipotermia	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
aum_hemato	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
caida_plaq	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
acum_liquievento	1= Si, 2=No.SD=Sin Informacion	Categórica	Entrada
evento	Descripción del evento notificado	Categórica	Entrada

Variable	Descripción	Tipo	Rol
year_	Año de notificación	Categórica	Entrada

Cuadro 2: Diccionario completo de variables (fuente: diccionario_datos.json)

2.2. Reglas de calidad

Se definieron reglas de validación para variables numéricas y categóricas con base en el diccionario oficial y en los valores observados en los datos reales.

Variable numérica	Valor mínimo observado	Valor máximo observado
edad_	0	174

Cuadro 3: Reglas de calidad para variables numéricas (valores observados en dengue_mod.csv)

Variable categórica	Valores posibles (según diccionario de datos)
semana	semanas del año de 1 a 53
uni_med_	Unidad de medida: 0= No aplica, 1=Años, 2=Meses, 3=Días, 4=Horas, 5=Minutos Sin información=SD
sexo_	Masculino=M, Femenino=F, Sin información=SD
nombre_barrio	Texto asociado a la tabla de barrios definidos por la entidad territorial, Vacios se diligencian con "Sin iNformacion", Sin ubicación en zona urbana.
comuna	Texto asociado a la tabla de comunas definidos por la entidad territorial, Vacios se diligencian con "Sin iNformacion", Sin ubicación en zona urbana.
tipo_ss_	Tipo de Régimen de seguridad social: C= Contributivo, S=Subsidiado, P=Excepción, E=Especial, N= No asegurado, I= Indeterminado/Pendiente, Sin información=SD.
cod_ase_	Codigo de la aseguradora
fec_con_	Fecha de Consulta
ini_sin_	Fecha de inicio de síntomas
tip_cas_	Tipo de caso: 1=Sospechoso, 2= Probable , 3=Confirmado por laboratorio , 4=Confirmado por clinica , 5= Confirmado por nexo epidemiológico.
pac_hos_	Paciente hospitalizado Sí=1, No=2.
cod_dpto_r	Código departamento residencia
cod_mpio_r	Código municipio residencia
cod_dpto_o	Código departamento ocurrencia
cod_mpio_o	Código municipio ocurrencia
desplazami	Sí=1, No=2. SD= Sin Informacion
cod_mun_d	Código municipio desplazamiento

Variable categórica	Valores posibles (según diccionario de datos)
clas_dengue	0= No aplica, 1=Dengue sin signos de alarma, 2=Dengue con signos de alarma 3=Dengue grave, Sin información=SD
fiebre	Sí=1, No=2.Sin información=SD
cefalea	Sí=1, No=2.Sin información=SD
dolrretroo	Sí=1, No=2.Sin información=SD
malgias	Sí=1, No=2.Sin información=SD
artralgia	Sí=1, No=2.Sin información=SD
erupcionr	Sí=1, No=2.Sin información=SD
dolor_abdo	Sí=1, No=2.Sin información=SD
vomito	Sí=1, No=2.Sin información=SD
somnolenci	Sí=1, No=2.Sin información=SD
hipotensio	Sí=1, No=2.Sin información=SD
hepatomeg	Sí=1, No=2.Sin información=SD
hem_mucosa	Sí=1, No=2.Sin información=SD
hipotermia	Sí=1, No=2.Sin información=SD
aum_hemato	Sí=1, No=2.Sin información=SD
caida_plaq	Sí=1, No=2.Sin información=SD
acum_liquievento	Sí=1, No=2.Sin información=SD
evento	Descripción del evento notificado
year_	Año de notificación

Cuadro 4: Reglas de calidad para variables categóricas

3. Preparación de datos

3.1. Selección de variables

Sobre el dataset inicial se eliminaron seis variables con baja contribución en el análisis exploratorio: `artralgia`, `dolrretroo`, `erupcionr`, `hipotermia`, `malgias` y `sexo_`. Además, `pac_hos_` se mantuvo explícitamente como variable objetivo.

3.2. Descripción estadística

Se inspeccionaron tipos de datos, conteos de nulos por variable y distribución de clases de `pac_hos_` antes y después del balanceo.

3.3. Limpieza de atípicos

No se realizó poda agresiva de atípicos por tratarse en su mayoría de variables categóricas/binarizadas. Para `edad_` se verificó consistencia de rango observado en los datos.

3.4. Limpieza de nulos

Se eliminaron filas con tres o más nulos (regla `isnull().sum(axis=1) <3`), quedando **17.958 registros**. Luego se imputaron nulos en variables categóricas con `KNNImputer(n_neighbors=5)`

sobre una codificación temporal numérica.

3.5. Creación de nuevas variables

No se crearon variables sintéticas. Se aplicó codificación one-hot para transformar variables categóricas a formato apto para modelado.

3.6. Análisis de relaciones

Se analizó la distribución de la variable objetivo antes del balanceo (clase 2.0: 11.577, clase 1.0: 6.331) y después de SMOTE (11.577 por clase).

3.7. Reducción de dimensiones

No se aplicó PCA. Se mantuvo enfoque de interpretabilidad y se controló dimensionalidad mediante eliminación de variables redundantes posteriores al one-hot.

3.8. Balanceo

Antes del balanceo se reservaron **50 registros completos** para evaluación de predicción futura (fase de despliegue), dejando **17.908 registros** para entrenamiento/evaluación formal. Sobre este conjunto se aplicó `SMOTE` con `k_neighbors=5`, pasando de `(17908, 56)` a `(23154, 56)` y logrando balance exacto entre clases.

4. Modelamiento y evaluación

4.1. Configuración y selección de métodos

Se trabajó con dos familias:

- **Modelos clásicos (5):** Regresión Logística, Árbol de Decisión, KNN, SVM y MLP.
- **Ensembles (3):** VotingClassifier (votación), Random Forest (bagging) y XGBoost (boosting).

4.2. Ajuste de hiperparámetros con 70 % de datos + CV

Se realizó partición estratificada 70/30 sobre el conjunto balanceado final: `X_train=(16207,46)` y `X_test=(6947,46)`. El ajuste de hiperparámetros se ejecutó exclusivamente sobre entrenamiento mediante `GridSearchCV` con `StratifiedKFold(n_splits=3)`.

4.2.1. Justificación de la métrica

Se utilizó **F1-macro** como métrica principal para balancear desempeño entre clases y no favorecer una clase particular.

4.2.2. Ajuste de modelos clásicos

Se ajustaron hiperparámetros de los 5 modelos clásicos definidos con búsqueda exhaustiva por grilla.

4.2.3. Ajuste de modelos de ensamble

Se ajustaron hiperparámetros de VotingClassifier, Random Forest y XGBoost con el mismo esquema de validación cruzada estratificada.

4.3. Medida de calidad del modelo

La evaluación final se hizo sobre el 30 % de prueba no usado en tuning. Para el mejor modelo (XGBoost) la matriz de confusión en test fue:

$$\begin{bmatrix} 3025 & 449 \\ 793 & 2680 \end{bmatrix}$$

Modelo	Tipo	F1-macro (CV train)	F1-macro (test)
Regresión Logística	Clásico	0.7383	0.7488
Árbol de Decisión	Clásico	0.7670	0.7761
KNN	Clásico	0.7295	0.7394
SVM	Clásico	0.7850	0.7963
Red Neuronal MLP	Clásico	0.7737	0.7884
VotingClassifier	Ensamble (Votación)	0.7702	0.7855
Random Forest	Ensamble (Bagging)	0.7958	0.8077
XGBoost	Ensamble (Boosting)	0.8143	0.8208

Cuadro 5: Resultados comparativos de modelos (bloque formal 70/30)

Mejor modelo: XGBoost, con `test_f1_macro=0.8208` y `test_accuracy=0.8212`.

4.3.1. Selección del mejor modelo

El mejor modelo se seleccionó por el mayor valor de `test_f1_macro`. Además, se reportaron recall, precision y accuracy en test y se exportaron los artefactos `best_model_formal_7030.joblib` y `model_input_columns.joblib`.

5. Despliegue

5.1. Predicción de datos futuros

Se evaluó el modelo final sobre **50 registros reales reservados** y no usados en entrenamiento ni en el tuning.

5.2. Construcción del conjunto de entrada

El conjunto de entrada se construyó desde `df_future_raw`, conservando su etiqueta real para contrastar predicción vs realidad.

5.3. Preparación del conjunto nuevo

Se reindexaron columnas según `model_input_columns.joblib` para garantizar compatibilidad con el pipeline final.

5.4. Predicción e interpretación

Las salidas del modelo se interpretaron como:

- 1: Hospitalización Sí
- 0: Hospitalización No

En estos 50 registros se obtuvo:

- Aciertos: 39
- Fallos: 11
- Accuracy: 0.78
- F1-macro: 0.7329
- Recall clase positiva: 0.5625
- Precision clase positiva: 0.6923

5.5. Aplicación web con Streamlit

Se implementó una app en `streamlit_app/app.py` con formulario manual de variables clínicas/demográficas, que genera la predicción y la interpretación de hospitalización esperada.

Pendiente para entrega final: incluir captura de pantalla funcionando y URL pública de Streamlit Cloud.

6. Conclusiones

El proyecto permitió construir un sistema de clasificación para estimar hospitalización por dengue con base en síntomas y variables complementarias.

La comparación sistemática entre modelos clásicos y ensambles permitió seleccionar **XGBoost** como mejor enfoque, con `F1-macro=0.8208` y `accuracy=0.8212` en el conjunto formal de prueba (30 %).

En la evaluación de despliegue sobre 50 registros reales reservados, el modelo logró 39 aciertos (`accuracy=0.78`), lo que confirma utilidad práctica como herramienta de apoyo a la decisión clínica.

Como trabajo futuro se recomienda:

- ampliar validación temporal por periodos epidemiológicos,
- incluir calibración de probabilidades para soporte clínico,
- monitorear deriva de datos en operación.