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 **3**
- 1.1. Descripción del negocio . . . 3
- 1.2. Descripción del problema . . . 3
- 1.3. Objetivos de la minería . . . 4
  - 1.3.1. Objetivo general . . . 4
- 1.4. Diseño de solución . . . 4

2. Entendimiento de los datos **4**
- 2.1. Diccionario de datos . . . 4
- 2.2. Reglas de calidad . . . 5

3. Preparación de datos **5**
- 3.1. Selección de variables . . . 5
- 3.2. Descripción estadística . . . 5
- 3.3. Limpieza de atípicos . . . 5
- 3.4. Limpieza de nulos . . . 5
- 3.5. Creación de nuevas variables . . . 5
- 3.6. Análisis de relaciones . . . 6
- 3.7. Reducción de dimensiones . . . 6
- 3.8. Balanceo . . . 6

4. Modelamiento y evaluación **6**
- 4.1. Configuración y selección de métodos . . . 6
- 4.2. Ajuste de hiperparámetros con 70 % de datos + CV . . . 6
  - 4.2.1. Justificación de la métrica . . . 6
  - 4.2.2. Ajuste de modelos clásicos . . . 6
  - 4.2.3. Ajuste de modelos de ensamble . . . 6
- 4.3. Medida de calidad del modelo . . . 7
  - 4.3.1. Selección del mejor modelo . . . 7

5. Despliegue **7**
5.1. Predicción de datos futuros 7
5.2. Construcción del conjunto de entrada 7
5.3. Preparación del conjunto nuevo 7
5.4. Predicción e interpretación 7
5.5. Aplicación web con Streamlit 8

6. Conclusiones **8**

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 principal contiene variables demográficas, geográficas y síntomas asociados a dengue. La variable objetivo del problema es `pac_hos_` (1=Sí hospitalización, 2=No).

Variable	Descripción	Tipo	Rol
edad_	Edad del paciente	Numérica	Entrada
comuna	Comuna de residencia	Categórica	Entrada
fiebre	Presencia de fiebre	Categórica	Entrada
cefalea	Presencia de cefalea	Categórica	Entrada
dolor_abdo	Dolor abdominal	Categórica	Entrada
vomito	Vómito	Categórica	Entrada
hipotensio	Hipotensión	Categórica	Entrada
pac_hos_	Paciente hospitalizado (1=Sí, 2=No)	Categórica	Salida

Cuadro 2: Diccionario resumido de variables relevantes

2.2. Reglas de calidad

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

Variable numérica	Valor mínimo	Valor máximo
edad_	0	120

Cuadro 3: Reglas de calidad para variables numéricas

Variable categórica	Valores posibles
pac_hos_	1=Sí, 2=No
fiebre	1=Sí, 2=No, SD
cefalea	1=Sí, 2=No, SD
hipotensio	1=Sí, 2=No, SD

Cuadro 4: Reglas de calidad para variables categóricas

3. Preparación de datos

3.1. Selección de variables

Se eliminaron variables con baja contribución y/o redundancia, manteniendo la variable objetivo al final del flujo para facilitar el procesamiento.

3.2. Descripción estadística

Se realizaron inspecciones de tipos, conteos, distribución de clases y porcentaje de nulos por variable.

3.3. Limpieza de atípicos

No se aplicó una eliminación masiva de atípicos por tratarse mayoritariamente de variables categóricas codificadas y binarias. En variables numéricas se verificó consistencia de rangos.

3.4. Limpieza de nulos

Se eliminaron registros con alta proporción de nulos y se imputaron nulos categóricos usando KNN sobre codificación temporal.

3.5. Creación de nuevas variables

No se crearon variables sintéticas adicionales; se aplicó codificación one-hot para variables categóricas.

3.6. Análisis de relaciones

Se revisó comportamiento de la variable objetivo antes y después del balanceo, y relación de síntomas/atributos con la clase.

3.7. Reducción de dimensiones

En la versión actual no se aplicó PCA. La decisión se sustentó en mantener interpretabilidad y en dimensionalidad manejable tras selección de variables.

3.8. Balanceo

Se aplicó SMOTE para balancear las clases de hospitalización y reducir sesgo por desbalance de clase.

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. El ajuste de hiperparámetros se ejecutó exclusivamente en el 70 % de entrenamiento mediante `GridSearchCV` y validación cruzada estratificada.

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.

4.2.3. Ajuste de modelos de ensamble

Se ajustaron hiperparámetros de VotingClassifier, Random Forest y XGBoost.

4.3. Medida de calidad del modelo

La evaluación final se hizo sobre el 30 % de prueba no usado en tuning.

Modelo	Tipo	F1-macro (CV train)	F1-macro (test)
Regresión Logística	Clásico	—	—
Árbol de Decisión	Clásico	—	—
KNN	Clásico	—	—
SVM	Clásico	—	—
Red Neuronal MLP	Clásico	—	—
VotingClassifier	Ensamble (Votación)	—	—
Random Forest	Ensamble (Bagging)	—	—
XGBoost	Ensamble (Boosting)	—	—

Cuadro 5: Resultados comparativos de modelos

Nota: completar con valores finales exportados del notebook (bloque formal 70/30).

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.

5. Despliegue

5.1. Predicción de datos futuros

Se generó un conjunto de entrada nuevo y se aplicó el modelo final para predecir hospitalización.

5.2. Construcción del conjunto de entrada

El conjunto de entrada se estructuró con el mismo esquema de variables del modelo entrenado.

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

5.5. Aplicación web con Streamlit

Se implementó una app en `streamlit_app/app.py` para cargar CSV y obtener predicciones.

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 el mejor enfoque según `F1-macro` en test.

Como trabajo futuro se recomienda:

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