Predicción de cobertura de vacunación contra H1N1 y gripe estacionaria

Inteligencia Artificial para las Ciencias e Ingeniería

Eliana Salas Villa, Marisol Correa Gutiérrez, Manuela Ospina Giraldo
Bioingeniería, Facultad de Ingeniería, Universidad de Antioquia, Medellín, Colombia
eliana.salas@udea.edu.co, marisol.correag@udea.edu.co, manuela.ospinag@udea.edu.co

Septiembre, 2023

1. Descripción del Problema Predictivo

En este proyecto, abordaremos el desafío de prever la probabilidad de que las personas sean vacunadas contra el virus H1N1 y la gripe estacional, lo cual es crucial en el ámbito de la salud pública para combatir enfermedades infecciosas y prevenir su propagación.

Este problema predictivo consiste en anticipar dos probabilidades para cada individuo en el conjunto de datos:

- La probabilidad de recibir la vacuna contra el virus H1N1 (h1n1_vaccine)
- La probabilidad de recibir la vacuna contra la gripe estacional (seasonal_vaccine).

Ambas variables objetivo son binarias, con valores posibles de 0 (No) y 1 (Sí), reflejando si las personas recibieron una, ambas o ninguna de las vacunas.

El objetivo principal es prever estas probabilidades basándonos en características personales, demográficas, económicas y de comportamiento. Al resolver este desafío, buscamos comprender cómo estas características se relacionan con las decisiones de vacunación y proporcionar orientación para futuras estrategias de salud pública con el fin de llegar a una “inmunidad de rebaño” o inmunidad colectiva de tal manera que se proporciona una protección indirecta a personas no inmunizadas o susceptibles a enfermedades estacionales o H1N1.

2. Descripción de la Base de Datos

La base de datos a utilizar es tomada desde la plataforma *Kaggle (https://www.kaggle.com/datasets/arashnic/flu-data)* y corresponde a un conjunto de registros tomados mediante la Encuesta Nacional de Gripe H1N1 de 2009 realizada telefónicamente en Estados Unidos con el objetivo de determinar si la persona encuestada recibió la vacuna contra el virus H1N1 o una vacuna contra un virus estacional. Dicho *DataSet* contiene 6437 respuestas a encuestas y 35 columnas enlistadas a continuación:

- *age_group:* Grupo de edad del encuestado
  - 18 – 34 years
  - 35 – 44 years
  - 45 – 54 years
  - 55 – 64 years
  - 65+ years
- *education:* Nivel de educación autodeclarado
  - < 12 years
  - 12 years
  - College Graduate
  - Some College
- *race:* Raza del encuestado
  - Black

    - Hispanic
    - Other or Multiple
    - White
- *sex:* Sexo del encuestado
    - Male
    - Female
- *income_poverty:* Ingresos anuales del hogar del encuestado con respecto a los umbrales de pobreza del censo de 2008
    - ≤ $75.000, Above Poverty
    - > $75.000
    - Below Poverty
- *marital_status:* Estado civil del encuestado.
    - Married
    - Not Married
- *rent_or_own:* Situación de vivienda del encuestado.
    - Rent
    - Own
- *employment_status:* Estado laboral del encuestado.
    - Employed
    - Not in Labor Force
    - Unemployed
- *h1n1_concern:* Nivel de preocupación por la gripe H1N1
    - 0 = Nada preocupado
    - 1 = Poco preocupado
    - 2 = Algo preocupado
    - 3 = Muy preocupado.
- *h1n1_knowledge*: Nivel de conocimiento sobre la gripe H1N1.
    - 0 = Sin conocimiento
    - 1 = Un poco de conocimiento
    - 2 = Mucho conocimiento.
- *behavioral_wash_hands:* Ha lavado frecuentemente las manos o usado desinfectante de manos.
    - 0 = No
    - 1 = Si
- *behavioral_large_gatherings:* Ha reducido el tiempo en reuniones grandes.
    - 0 = No
    - 1 = Si
- *behavioral_antiviral_meds:* Ha tomado medicamentos antivirales
    - 0 = No
    - 1 = Si
- *behavioral_avoidance:* Ha evitado el contacto cercano con otras personas con síntomas similares a la gripe.
    - 0 = No
    - 1 = Si
- *behavioral_face_mask*: Ha comprado una mascarilla facial.
    - 0 = No
    - 1 = Si
- *behavioral_outside_home*: Ha reducido el contacto con personas fuera de su hogar.
    - 0 = No
    - 1 = Si
- *behavioral_touch_face*: Ha evitado tocarse los ojos, la nariz o la boca.
    - 0 = No

    - 1 = Si
- *doctor_recc_h1n1*: El médico recomendó la vacuna contra la gripe H1N1.
    - 0 = No
    - 1 = Si
- *doctor_recc_seasonal*: El médico recomendó la vacuna contra la gripe estacional.
    - 0 = No
    - 1 = Si
- *chronic_med_condition*: Tiene alguna de las siguientes afecciones médicas crónicas: asma u otra afección pulmonar, diabetes, una afección cardíaca, una afección renal, anemia de células falciformes u otra anemia, una afección neurológica o neuromuscular, una afección hepática o un sistema inmunológico debilitado causado por una enfermedad crónica o por medicamentos tomados para una enfermedad crónica.
    - 0 = No
    - 1 = Si
- *child_under_6_months:* Tiene contacto cercano regular con un niño menor de seis meses.
    - 0 = No
    - 1 = Si
- *health_worker:* Es trabajador de la salud.
    - 0 = No
    - 1 = Si
- *health_insurance:* Tiene seguro de salud.
    - 0 = No
    - 1 = Si
- *opinion_h1n1_vacc_effective:* Opinión del encuestado sobre la efectividad de la vacuna contra la gripe H1N1.
    - 1 = Nada efectiva
    - 2 = Poco efectiva
    - 3 = No sabe
    - 4 = Algo efectiva
    - 5 = Muy efectiva
- *opinion_h1n1_risk*: Opinión del encuestado sobre el riesgo de enfermarse de gripe H1N1 sin vacuna.
    - 1 = Muy bajo
    - 2 = Algo bajo
    - 3 = No sabe
    - 4 = Algo alto
    - 5 = Muy alto
- *opinion_h1n1_sick_from_vacc*: Preocupación del encuestado por enfermarse al tomar la vacuna contra la gripe H1N1.
    - 1 = Nada preocupado
    - 2 = Poco preocupado
    - 3 = No sabe
    - 4 = Algo preocupado
    - 5 = Muy preocupado
- *opinion_seas_vacc_effective*: Opinión del encuestado sobre la efectividad de la vacuna contra la gripe estacional.
    - 1 = Nada efectiva
    - 2 = Poco efectiva
    - 3 = No sabe
    - 4 = Algo efectiva

    - 5 = Muy efectiva
- *opinion_seas_risk:* Opinión del encuestado sobre el riesgo de enfermarse de gripe estacional sin vacuna.
    - 1 = Muy bajo
    - 2 = Algo bajo
    - 3 = No sabe
    - 4 = Algo alto
    - 5 = Muy alto
- *opinion_seas_sick_from_vacc*: Preocupación del encuestado por enfermarse al tomar la vacuna contra la gripe estacional.
    - 1 = Nada preocupado
    - 2 = Poco preocupado
    - 3 = No sabe
    - 4 = Algo preocupado
    - 5 = Muy preocupado
- *hhs_geo_region*: Residencia del encuestado utilizando una clasificación geográfica de 10 regiones definida por el Departamento de Salud y Servicios Humanos de EE. UU. Los valores se representan como cadenas de caracteres cortas y aleatorias.
    - Región 1: Región de Nueva Inglaterra
    - Región 2: Región de Nueva York y Nueva Jersey
    - Región 3: Región de la Costa Atlántica Media
    - Región 4: Región del Sudeste
    - Región 5: Región del Medio Oeste
    - Región 6: Región del Sudoeste
    - Región 7: Región de las Llanuras Centrales
    - Región 8: Región de las Montañas Rocosas
    - Región 9: Región del Pacífico
    - Región 10: Región del Noroeste
- *census_msa*: Residencia del encuestado dentro de áreas metropolitanas estadísticas (MSA) según lo definido por el Censo de EE. UU.
    - MSA, Not Principle City
    - MSA, Principle City
    - Non-MSA
- *household_adults*: Número de otros adultos en el hogar, con un límite superior de 3.
- *household_children*: Número de niños en el hogar, con un límite superior de 3.
- *employment_industry*: Tipo de industria en la que el encuestado está empleado. Los valores se representan como cadenas de caracteres cortas y aleatorias.
- *employment_occupation*: Tipo de ocupación del encuestado. Los valores se representan como cadenas de caracteres cortas y aleatorias.

En la base de datos, más del 90% de las columnas se clasifican como categóricas, lo que significa que representan características cualitativas o etiquetas en lugar de valores numéricos. Por otro lado, el conjunto de datos presenta un 6.32% de valores faltantes distribuidos entre 30 columnas. No obstante, la gran mayoría de estos valores faltantes se encuentran concentrados en tres columnas específicas:

- Employment_occupation
- Employment_industry
- Health_insurance

Estas tres columnas en conjunto representan el 4.06% del total de valores faltantes. La base de datos puede ser consultada en el siguiente enlace: *H1N1 and Seasonal Flu Vaccines.*

3. Métricas de Desempeño

Las métricas de desempeño a utilizar en el proyecto, teniendo en cuenta que se trata de una problemática de clasificación binaria, son:

Accuracy: Predicciones correctas / Total de predicciones. Resulta útil para tener una idea general de la relación que hay entre la cantidad de verdaderos positivos o predicciones correctas con respecto al total de predicciones realizadas por el algoritmo dando cuenta del desempeño del modelo en términos de clasificación correcta. Al obtener un buen *accuracy*, se garantiza que la población que el modelo dice que tiene cobertura de vacuna, realmente sí tenga esa posibilidad de ser vacunado.

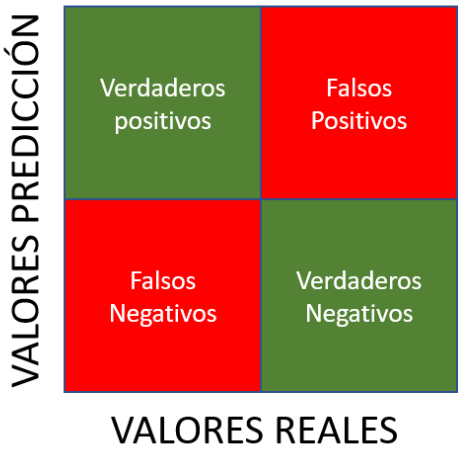
Precision (precisión): Verdaderos Positivos / (Verdaderos Positivos + Falsos Positivos). Evalúa cuántas de las predicciones positivas del modelo son verdaderamente correctas. Es relevante en situaciones donde los falsos positivos tienen un peso importante, en este caso particular cobra real importancia puesto que se debe garantizar que las comunidades que tienen cobertura de vacunación realmente sí la posean con el fin de realizar planes de distribución de vacunas con respecto a esa cobertura.

Recall (recuperación): Verdaderos Positivos / (Verdaderos Positivos + Falsos Negativos). Mide cuántos casos positivos reales el modelo ha identificado con precisión. Se usa en escenarios donde es crucial detectar la mayoría de los casos positivos.

F1-score: combina la precisión y la recuperación en una sola medida. Se utiliza para medir el equilibrio entre la precisión y la capacidad del modelo para recuperar ejemplos positivos correctamente.

$$F1 - score = 2 \cdot \frac{Precision + Recall}{Precision \cdot Recall}$$

Matriz de confusión: La matriz de confusión muestra la distribución de las predicciones del modelo en comparación con las clases reales. Es una herramienta útil para comprender las tasas de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.



Estas métricas de rendimiento ofrecen una visión completa de los puntos fuertes y débiles de un modelo de clasificación. Aunque la exactitud es un punto de partida habitual, la precisión, la recuperación, la puntuación F1 y la matriz de confusión ofrecen una visión más matizada. Estas métricas nos podrán decir si podemos confiar en el algoritmo para desarrollar un plan de vacunación de acuerdo a la decisión que se esperaría que tomara la población, basándose en sus decisiones previas.

4. Desempeño Deseable en Producción

Con el modelo de predicción propuesto se pueden identificar poblaciones que requieren de un mayor apoyo o de creación de nuevos programas por parte de los entes encargados de la distribución de vacunas con el fin de llevar dichos antídotos a toda la población de manera equitativa.

Según estudios realizados por epidemiólogos de la región, las gripes estacionales son generadas principalmente por tipos o subtipos de virus de influenza y se requiere alcanzar una cobertura de vacunación de alrededor del 50% al 70% para garantizar una inmunidad colectiva.

Para el caso del virus H1N1 (También conocida como Gripe Porcina), los epidemiólogos tomaron como base el porcentaje de cobertura de vacunación presentada en la pandemia del 2009 estimando que la inmunidad de rebaño puede ser conseguida si el 40% al 50% de la población es vacunada.