UNIVERSIDAD DE ANTIOQUIA
1803

Predicción de cobertura de vacunación contra H1N1 y gripe estacionaria – Entrega 2

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

Octubre 22, 2023

Introducción

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.

Base de datos

El objetivo es predecir la probabilidad de que los individuos reciban las vacunas contra la gripe H1N1 y la gripe estacional. En concreto, predecirá dos probabilidades: una para h1n1_vaccine y otra para seasonal_vaccine, así como cualquier análisis exploratorio de datos (AED) sofisticado. Cada fila del conjunto de datos representa las respuestas de una persona a la Encuesta nacional sobre la gripe H1N1 realizada en 2009 en Estados Unidos y hay dos variables objetivo:

h1n1_vaccine: Si el encuestado recibió la vacuna contra la gripe H1N1.

seasonal_vaccine: Si el encuestado recibió la vacuna contra la gripe estacional.

Ambas son variables binarias: 0 = No; 1 = Sí. Algunos encuestados no recibieron ninguna de las dos vacunas, otros recibieron sólo una y algunos recibieron ambas. Se trata de un problema multietiqueta (y no multiclase).

El conjunto de datos tomados consta de 36 columnas. La primera columna respondent_id es un identificador único y aleatorio. Las 35 columnas siguientes corresponden a características como edad del encuestado, nivel de preocupación por el H1N1 (de 0 a 3), si el doctor le ha recomendado o no que se ponga la vacuna (0 o 1), ocupación, región de residencia, entre otras.

La base de datos se encuentra desbalanceada, lo que puede llevar a un sesgo hacia las clases mayoritarias, dificultades en la generalización, y desafíos en la evaluación del desempeño. Para abordar este problema se plantea a futuro balancear los datos de manera manual, escogiendo aleatoriamente los datos usados para entrenamiento, el resto de los datos serán usados para la etapa de validación del modelo.

Exploración de datos

Para empezar, se realizó el gráfico de barras de la figura 1, en el que se puede observar cuál es la proporción inicial entre personas vacunadas y no vacunadas para cada una de las vacunas (H1N1 y gripe estacional) respectivamente en el dataset crudo, esto para tener una idea global del mismo previo al preprocesado.

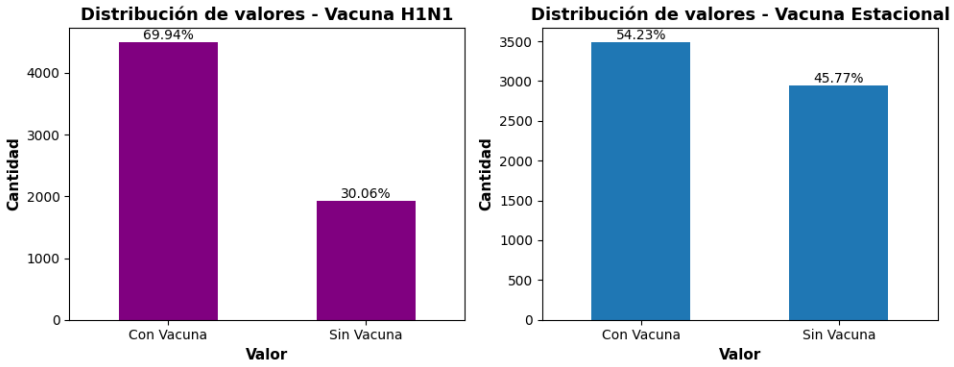

Figura 1. Distribución de resultados esperados para cada vacuna.

Preprocesado de datos

Para poder lograr que el dataframe sea apto para el entrenamiento de un algoritmo de machine learning se siguieron ciertos pasos:

1. Se eliminaron las últimas dos columnas que correspondían a las etiquetas de los datos para cada vacuna.
2. Se eliminó la primera columna que correspondía al ID de cada participante
3. Se eliminaron las filas que contenían caracteres nulos.
4. Se convirtieron todas las columnas del dataframe en variables categóricas.
5. Se realizó la codificación one-hot de las variables categóricas para que todas las entradas al algoritmo sean numéricas.

Entrenamiento y prueba del modelo

Se tomó el 80% de los datos para entrenamiento y el 20% para prueba. Se estableció una semilla fija de 42 para asegurarse de que la división de datos sea reproducible. En las figuras 2 y 3 se observa la proporción de datos para el set de entrenamiento y prueba respectivamente para H1N1 y Gripe estacional.

Vacuna H1N1

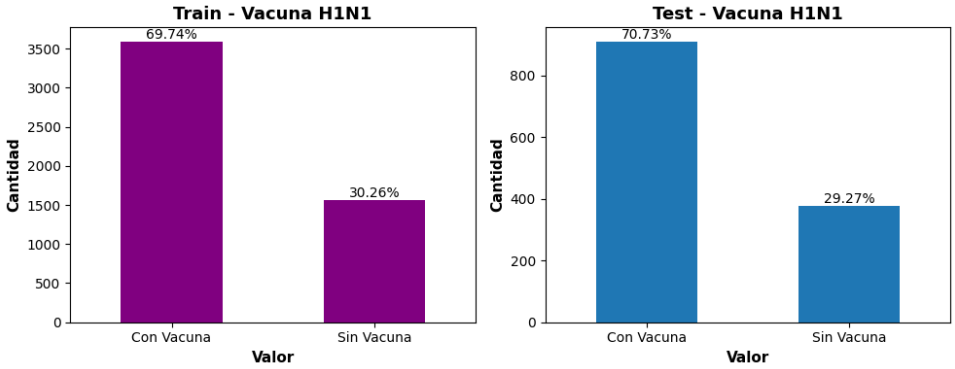

Figura 2. Proporción de sujetos vacunados y no vacunados contra H1N1 para sets de entrenamiento y prueba.

Gripe estacional

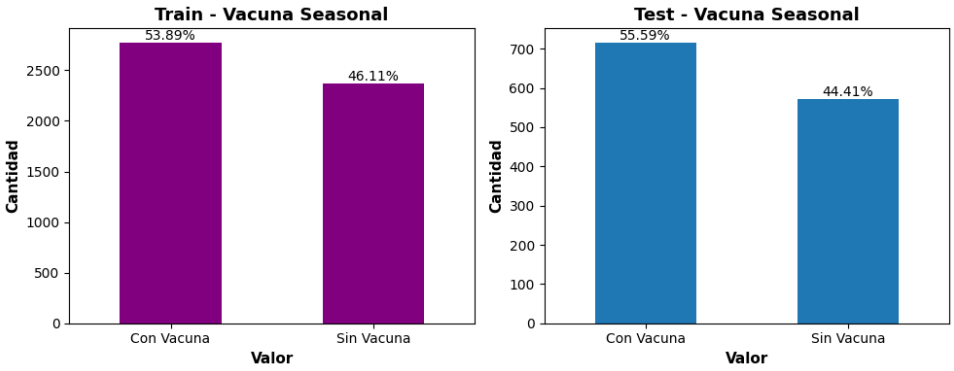

Figura 3. Proporción de sujetos vacunados y no vacunados contra Gripe estacional para sets de entrenamiento y prueba.

Algoritmo de ML

Se implementarán 3 algoritmos de aprendizaje supervisado (Redes ne para luego comparar su desempeño en la clasificación. Inicialmente se decidió aplicar una red neuronal secuencial y a continuación se describirá su desempeño.

Aplicación de Redes Neuronales

Previo a la creación de cada modelo, se creó una instancia de escalador estándar para que las características del set de entrenamiento tuviesen una media de 0 y desviación estándar de 1, ayudando a estabilizar y normalizar las características. Luego, se realizaron distintas variaciones de los hiperparámetros de la red neuronal (capas ocultas y número de neuronas) y empleando para todas, la función de activación sigmoide. Finalmente, para todos los casos se compiló el modelo escogido especificando el optimizador Adam, la función de pérdida 'Binary_crossentropy' para clasificación binaria y las métricas a seguir (en este caso, la precisión). En la tabla a continuación, se observa el valor de precisión (accuracy) para el set de entrenamiento y prueba para cada uno de los modelos escogidos, así como los hiperparámetros que corresponden a cada uno.

Variación de Estructura – Vacuna H1N1

Modelo	Capas Ocultas	Neuronas	Accuracy Entrenamiento	Accuracy Prueba
Modelo 1	1	32	83.54%	84%
Modelo 2	1	64	83.62%	84%
Modelo 3	1	128	83.39%	83%
Modelo 4	2	32, 32	82.69%	83%
Modelo 5	2	64, 32	83.23%	83%
Modelo 6	2	64, 64	83.23%	83%
Modelo 7	3	64, 64, 128	83.15%	83%

Variación de Estructura – Vacuna Gripe estacional

Modelo	Capas Ocultas	Neuronas	Accuracy Entrenamiento	Accuracy Prueba
Modelo 1	1	32	83.15%	85%
Modelo 2	1	64	83.31%	85%
Modelo 3	1	128	83.39%	85%
Modelo 4	2	32, 32	83.31%	85%
Modelo 5	2	64, 32	83.46%	85%
Modelo 6	2	64, 64	83.15%	86%
Modelo 7	3	64, 64, 128	80.36%	82%

Validación Cruzada (CrossValidation)

Se realizó validación cruzada k-fold con el objetivo de generalizar el modelo en datos no vistos. En esta técnica, el conjunto de datos se divide en k pliegues (folds), y el modelo se entrena y evalúa k veces. En cada iteración, un pliegue se utiliza como conjunto de prueba, mientras que los otros se utilizan como conjunto de entrenamiento. Las métricas de rendimiento se promedian a lo largo de las k iteraciones para obtener una estimación más precisa del rendimiento del modelo.

El código implementa la validación cruzada estratificada utilizando las bibliotecas scikit-learn y Keras en Python. Primero, normaliza los datos de entrenamiento con StandardScaler de scikit-learn. Luego, crea los modelos escogidos de red neuronal con Keras, que consta de una capa oculta con activación sigmoidal y una capa de salida para problemas de clasificación binaria. Después de entrenar el modelo con los datos de entrenamiento y validarlos con los datos de prueba, se implementa un bucle para la validación cruzada estratificada.

Para esto, se utiliza StratifiedKFold para dividir los datos en un número específico de pliegues (num_folds). El modelo se entrena y evalúa en cada pliegue, y los puntajes de precisión se almacenan en una lista. Luego, se calcula el promedio de los puntajes de precisión para cada número de pliegues y se almacena en un diccionario llamado results. Esto permite comparar el rendimiento del modelo en diferentes configuraciones de validación cruzada estratificada.

La elección de scikit-learn en lugar de Keras para la validación cruzada se debe a que scikit-learn ofrece herramientas más robustas para el manejo de datos y técnicas de validación cruzada, lo que facilita la implementación y evaluación del modelo en diferentes pliegues de datos.

A continuación, se presentan los resultados de las matrices de confusión para el modelo original escogido y para 3 y 8 iteraciones tanto para la gripa estacional como para el H1N1.

Vacuna H1N1

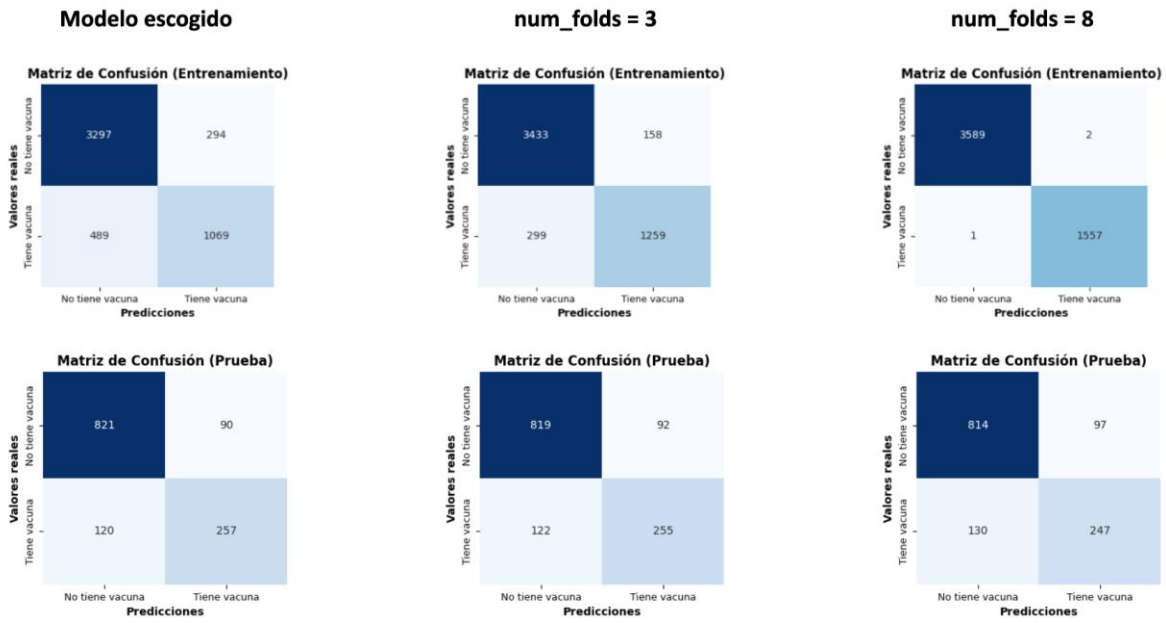

Figura 4. Matrices de confusión de los resultados de la validación cruzada para vacunas H1N1

Vacuna Gripe estacional

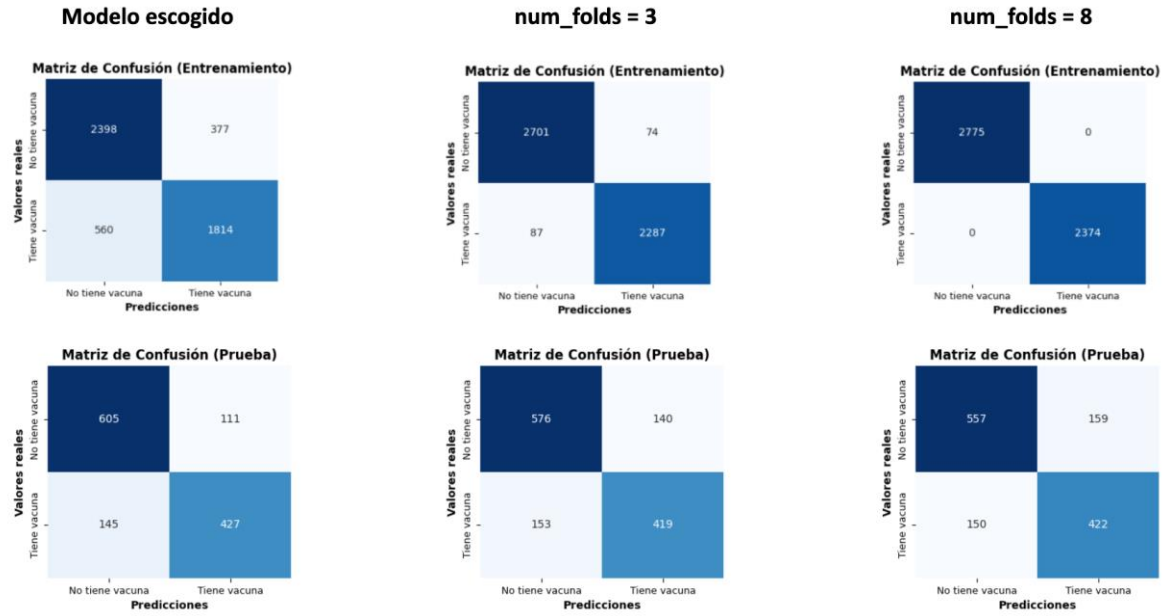

Figura 5. Matrices de confusión de los resultados de la validación cruzada para vacunas de la gripe estacional.

En el proceso de validación cruzada con 3 iteraciones, el modelo no muestra mejoras significativas. De hecho, se observa una ligera disminución en su rendimiento, indicando una estabilidad en los resultados. Sin embargo, al aumentar el número de iteraciones a 8, se presenta un fenómeno preocupante: el modelo comienza a sobreajustarse. Esto significa que el modelo se ajusta demasiado bien a los datos de entrenamiento específicos que se le han proporcionado, pero su capacidad para generalizar y hacer predicciones precisas en datos nuevos (los datos de prueba) se ve comprometida.

La sobreoptimización del modelo para los datos de entrenamiento puede llevar a una pérdida significativa de precisión cuando se enfrenta a datos no vistos anteriormente, lo cual es precisamente lo que se intenta evitar en cualquier tarea de modelado predictivo.

En este contexto, realizar una validación cruzada con un mayor número de iteraciones no está contribuyendo positivamente a mejorar el modelo. De hecho, esta técnica está revelando una debilidad fundamental en el modelo actual, mostrando su falta de capacidad para generalizar adecuadamente a partir de los datos de entrenamiento. En lugar de mejorar la calidad de las predicciones, el modelo está siendo afectado negativamente por el sobreajuste, lo que destaca la importancia de encontrar un equilibrio adecuado entre la complejidad del modelo y la cantidad de datos disponibles para entrenarlo.

Trabajo Futuro

En el futuro, se planea poner a prueba los algoritmos de K-Means y Support Vector Machine (SVM) para determinar si estos mejoran las métricas de desempeño. Además de esto, como se mencionó anteriormente, se tiene la intención de implementar estrategias de preprocesamiento para abordar el desbalanceo de los datos, con el objetivo de evaluar si estas medidas resultan en una mejora del rendimiento global de todos los modelos.

Otra medida contemplada es la modificación de los hiperparámetros del proceso de validación cruzada, con el propósito de reducir el sobreentrenamiento. Es crucial evitar que el proceso de validación cruzada ajuste en exceso los datos, ya que esto puede llevar a evaluaciones poco confiables de los modelos. Por lo tanto, se trabajará en ajustar cuidadosamente estos parámetros para garantizar una evaluación precisa y confiable del desempeño de los modelos.