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

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

Noviembre 25, 2023

Introducción

La salud pública es un componente esencial en la vida de cualquier sociedad, y la prevención de enfermedades a través de la vacunación desempeña un papel fundamental en este aspecto.

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. Se emplearon cuatro modelos de inteligencia artificial con el fin de determinar cuál tiene un mejor desempeño para la predicción: una Red Neuronal, una Máquina de Soporte Vectorial (SVM), K-Means + Red neuronal y K-Means + SVM. A continuación se describirá el proceso mediante el cual fueron hallados los hiperparámetros que mejor desempeño obtuvieron al resolver el problema predictivo, las curvas de entrenamiento de cada modelo y las conclusiones del presente trabajo

2. Base de Datos

E 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

- *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 multi-etiqueta (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.

3. Exploración descriptiva del dataset

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.

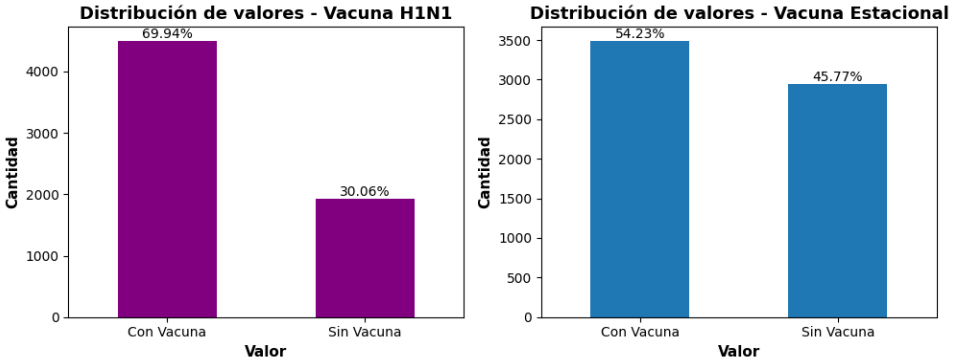

Figura 1. Distribución de resultados esperados para cada vacuna.

Iteraciones de desarrollo.

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.

División de datos en entrenamiento y prueba utilizando sobremuestreo

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. Luego se aplicó SMOTE al conjunto de entrenamiento y al conjunto de prueba para abordar el desequilibrio de clases generando instancias sintéticas de la clase minoritaria. 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.

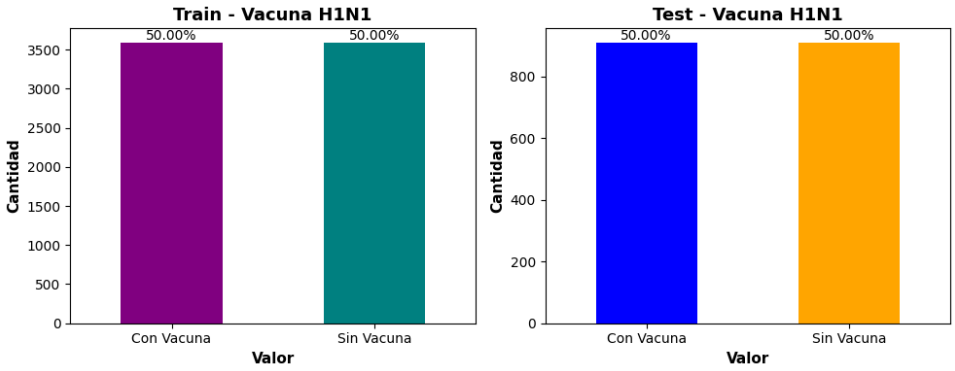

Figura 2. Proporción de sujetos vacunados y no vacunados contra H1N1 para sets de entrenamiento y prueba.

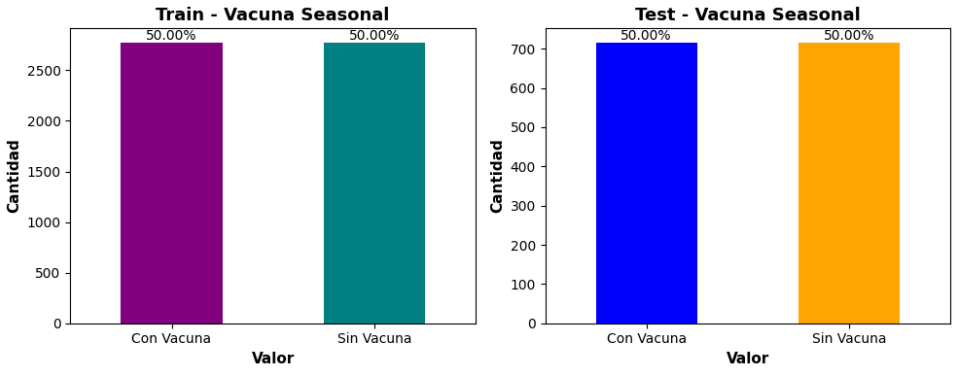

Figura 3. Proporción de sujetos vacunados y no vacunados contra Gripe estacional para sets de entrenamiento y prueba.

Validación cruzada

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

Modelos supervisados

Red Neuronal

La Red Neuronal es un modelo inspirado en la estructura del cerebro humano. Está compuesto por nodos llamados neuronas, organizados en capas, y cada conexión entre nodos tiene un peso que se ajusta durante el entrenamiento. Este modelo es capaz de

aprender patrones complejos y no lineales en los datos, lo que lo hace adecuado para problemas de clasificación como el propuesto en este estudio.

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). Fueron probados 7 modelos con combinaciones de hiperparámtros diferentes y los que proporcionaron la mejor precisión se presentan en la tabla 1 :

Tabla 1. Hiperparámetros de la red neuronal que mejor predijo H1N1 y Seasonal flu

Capas ocultas	Neuronas	Función de activación	Optimizador	Función de pérdida	Número de épocas	Tamaño del lote	Accuracy set prueba
1	64	Sigmoide	Adam	Binary crossentropy	15	64	79.09%

En las figuras 4 y 5 se muestran las matrices de confusión para cada set de datos y métricas de desempeño adicionales.

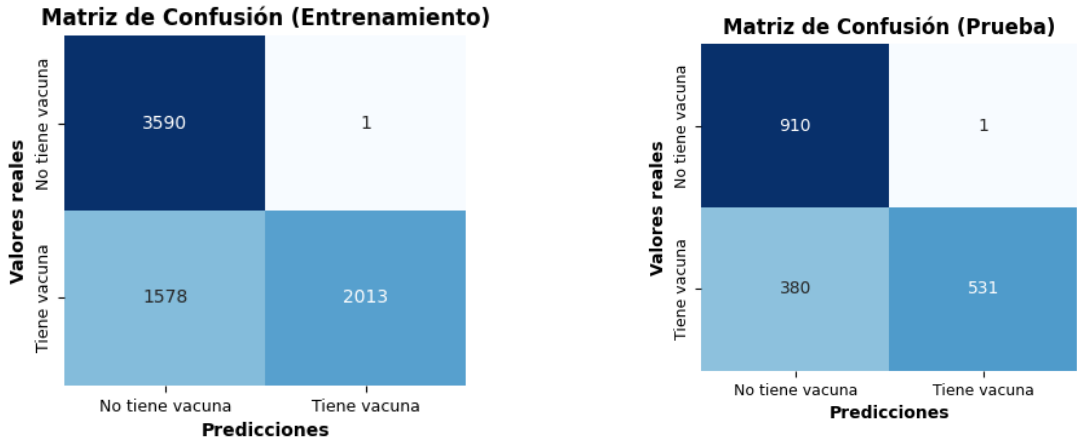

Figura 4. Matrices de confusión

	precision	recall	f1-score	support
0	0.74	0.88	0.80	716
1	0.85	0.69	0.76	716
accuracy			0.78	1432
macro avg	0.79	0.78	0.78	1432
weighted avg	0.79	0.78	0.78	1432

Figura 5. Métricas adicionales

Support Vector Machine

SVM busca encontrar el hiperplano óptimo que maximiza la separación entre clases. Este modelo es efectivo incluso en espacios de alta dimensionalidad y puede manejar datos no lineales mediante el uso de funciones de kernel.

Tabla 2. Hiperparámetros del modelo de SVM mejor predijo H1N1 y Seasonal flu

Kernel	Penalización	Accuracy set de prueba H1N1	Accuracy set de prueba Seasonal
Lineal	10	85%	74%

En las figuras 6 y 7 se muestran las matrices de confusión para cada set de datos y métricas de desempeño adicionales.

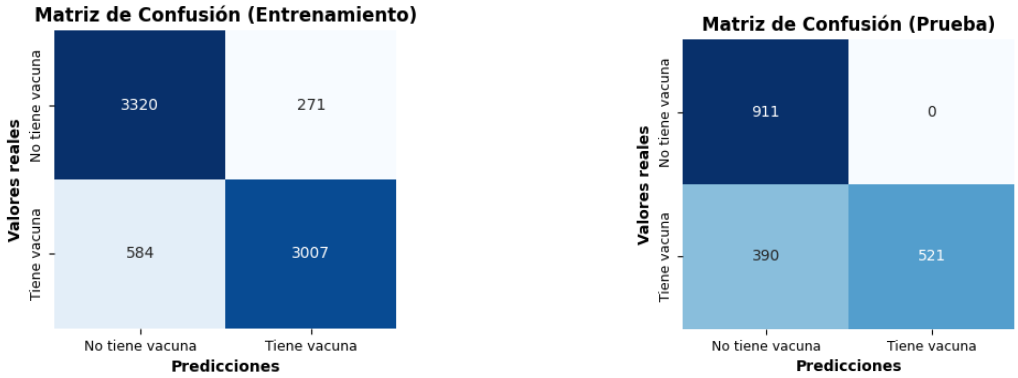

Figura 6. Matrices de confusión

	precision	recall	f1-score	support
0	0.70	1.00	0.82	911
1	1.00	0.57	0.73	911
accuracy			0.79	1822
macro avg	0.85	0.79	0.78	1822
weighted avg	0.85	0.79	0.78	1822

Figura 7. Métricas adicionales

Como este es el modelo que mejor predijo ambos grupos, se obtuvo adicionalmente una curva de aprendizaje.

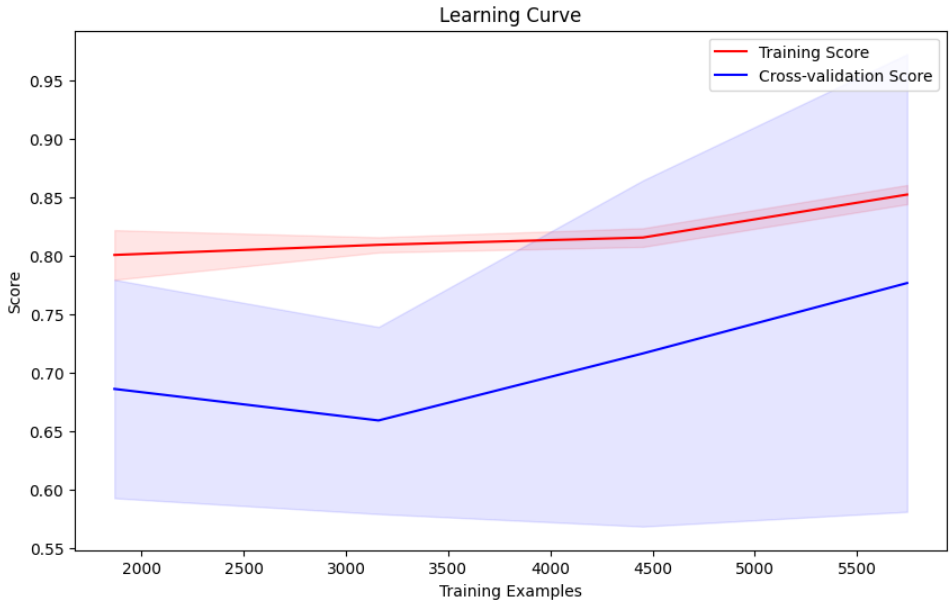

Figura 8. Curva aprendizaje SVM

Modelos compuestos

K-Means + Redes Neuronales

K-Means es un algoritmo de agrupamiento que divide un conjunto de datos en k grupos basándose en similitudes. Este método es útil para identificar patrones y segmentar la población en grupos homogéneos. Aunque K-Means es más comúnmente utilizado para tareas de agrupamiento, se explorará su capacidad para clasificar en este contexto.

Se aplicó una normalización de datos y luego una red neuronal que ayudarono a identificar estructuras en los datos que no eran evidentes en su forma original. Posteriormente, KMeans se uutilizó para agrupar las instancias de entrenamiento en clústeres y analizar la frecuencia y frecuencia relativa de los valores de la variable objetivo dentro de cada clúster, proporcionando una visión adicional sobre la distribución de las etiquetas en los grupos identificados por KMeans.

Los hiperparámetros que proporcionaron el mejor desempeño del modelo fueron:

Red neuronal:

Capas ocultas	Neuronas	Función de activación	Optimizador	Función de pérdida	Número de épocas	Tamaño del lote
1	64	Sigmoide	Adam	Binary crossentropy	15	64

K-means:

Número de clústeres	Número de inicializaciones	Número máximo de iteraciones
2	1000	100000

El accuracy proporcionado por este algoritmo compuesto fue de 82% para H1N1 y de 62% para Seasonal fu.

En las figuras 9 y 10 se muestra la matriz de confusión y las métricas adiconales.

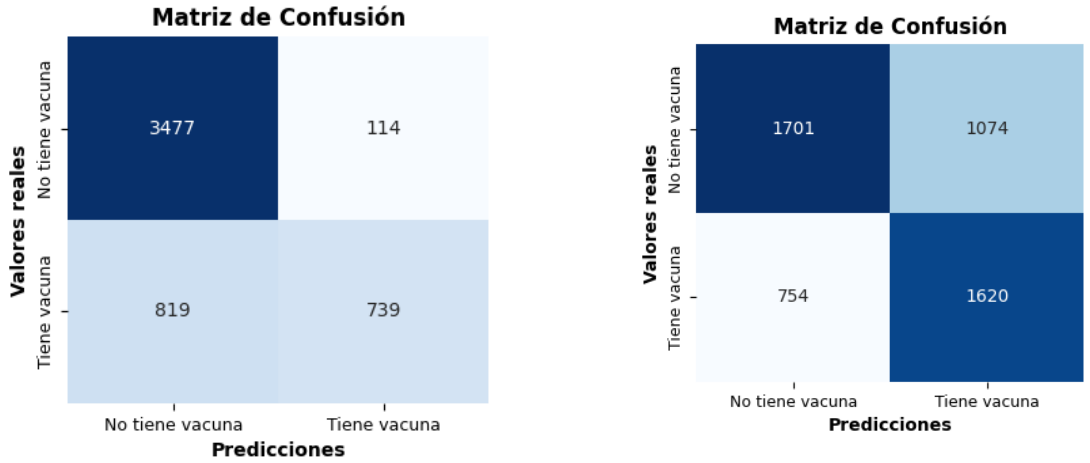

Figura 9. Matrices de confusión H1N1 y Seasonal

	precision	recall	f1-score	support
0	0.81	0.97	0.88	3591
1	0.87	0.47	0.61	1558
accuracy			0.82	5149
macro avg	0.84	0.72	0.75	5149
weighted avg	0.83	0.82	0.80	5149

	precision	recall	f1-score	support
0	0.69	0.61	0.65	2775
1	0.60	0.68	0.64	2374
accuracy			0.64	5149
macro avg	0.65	0.65	0.64	5149
weighted avg	0.65	0.64	0.65	5149

Figura 10. Métricas adicionales H1N1 y Seasonal

K-means + SVM

El objetivo de esta combinación de algoritmos era entender cómo se agrupaban las instancias en clústeres y cómo se distribuían las etiquetas de clase en esos clústeres, buscando patrones en la estructura subyacente de los datos que pudieran ser capturados tanto por SVM como por KMeans.

Se entrenó un modelo SVM con kernel lineal mediante validación cruzada para evaluar la precisión promedio del modelo. Después, aplicó KMeans al conjunto de entrenamiento para agrupar las instancias en dos clústeres. Examinó la distribución de las etiquetas de clase en cada clúster, calculó la frecuencia y frecuencia relativa de cada valor de la variable objetivo dentro de los clústeres y, finalmente, realizó una predicción basada en las etiquetas del clúster utilizando el modelo SVM entrenado.

El accuracy proporcionado por este algoritmo compuesto fue de 64% para H1N1 y para Seasonal fu.

En las figuras 9 y 10 se muestra la matriz de confusión y las métricas adiconales.

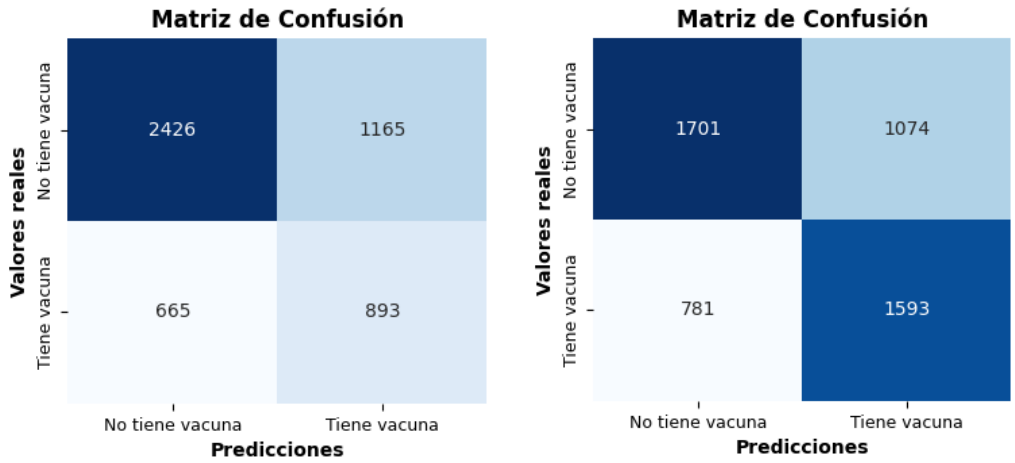

Figura 11. Matrices de confusión H1N1 y Seasonal

	precision	recall	f1-score	support
0	0.78	0.68	0.73	3591
1	0.43	0.57	0.49	1558
accuracy			0.64	5149
macro avg	0.61	0.62	0.61	5149
weighted avg	0.68	0.64	0.66	5149

	precision	recall	f1-score	support
0	0.69	0.61	0.65	2775
1	0.60	0.67	0.63	2374
accuracy			0.64	5149
macro avg	0.64	0.64	0.64	5149
weighted avg	0.64	0.64	0.64	5149

Figura 12. Métricas adicionales H1N1 y Seasonal

Retos y consideraciones de despliegue

En los algoritmos presentados como aquellos con mejor desempeño de cada modelo, no se dio overfitting ya que, aunque el desempeño en el set de entrenamiento fue siempre superior, a la hora de predecir en el set de prueba, su desempeño no decayó por completo. Se observó que en los algoritmos no supervisados la predicción para no vacunados y vacunados en cada caso fue similar, sin embargo, los algoritmos comppuestos presentaron una tendencia a predecir mejor los casos negativos (no tienen vacuna) que los positivos para H1N1.

Para mejorar el desempeño obtenido en cada modelo, lo que haría sería:

<u>Modelo de red neuronal:</u>

1. Introducir técnicas de regularización como capas de abandono o regularización L1/L2 para evitar el sobreajuste.
2. Experimentar con diferentes funciones de activación, arquitecturas de red y tasas de aprendizaje para mejorar el rendimiento.

<u>Modelo SVM:</u>

1. Optimizar los hiperparámetros de la SVM utilizando técnicas como la búsqueda en cuadrícula (grid search) o la búsqueda aleatoria (random search).
2. Explorar diferentes funciones de núcleo y parámetros de regularización para mejorar el rendimiento del modelo.

Modelo de agrupación K-means:

1. Experimentar con diferentes números de clusters y evaluar las puntuaciones de silueta para encontrar el número óptimo de clusters.

Consideramos que el desempeño mínimo para desplegar en producción nuestro modelo es que no se presente overfitting durante el entrenamiento y que tenga un accuracy en el set de entrenamiento superior a 82%.

Los modelos pueden desplegarse utilizando herramientas de contenedorización como Docker y desplegarse en plataformas en la nube o en servidores locales. Los procesos de supervisión del rendimiento pueden incluir la detección de desviaciones, la activación del reentrenamiento del modelo y la evaluación periódica del rendimiento del modelo en función de umbrales predefinidos.

Conclusiones

El modelo que presentó el mejorn desempeño fue el de máquina de soporte vectorial (SVM) con una precisión de predicción de 85% para H1N1 y de 74% para seasonal por lo que se propone como el modelo elegido para escoger como punto de base para puesta en producción únicamente en el caso de H1N1 que es en el cuál superó nuestras condiciones para esto descritas anteriormente.

Como el objetivo de la recolección del dataset era traspolarlo a la situación presentada durante la pandemia con el covid-19, la idea a futuro es que se realice nuevamente la encuesta en el mismo grupo focal recogiendo las mismas características presentadas en este dataset y evaluar el desempeño de nuestro modelo elegido para realizar comparaciones de desempeño y que pueda ser empleado a futuro como referente en cuánto a predicción de la probabilidad de vacunación para estudios en el campo de la salud pública.