Modelo de predicción de las vacunas contra la influenza H1N1 y la influenza estacional

Andrés Fernando Delgado Pérez, David Esteban Fajardo Torres, Jairo Antonio Caro Vanegas, Lizeth Viviana Perdomo Castañeda

Primera Entrega

Definición de la problemática y entendimiento del negocio

En los últimos años, se han presentado a nivel mundial una serie de enfermedades respiratorias importantes. A partir de la primavera de 2009, una pandemia causada por el virus de la gripe H1N1, coloquialmente llamada "gripe porcina", se extendió por todo el mundo. Los investigadores estiman que, en el primer año, fue responsable de entre 151.000 y 575.000 muertes en todo el mundo.

En octubre de 2009 se puso a disposición del público una vacuna contra el virus de la gripe H1N1. Las vacunas proporcionan inmunización a las personas, y una inmunización suficiente en una comunidad puede reducir aún más la propagación de enfermedades a través de la "inmunidad colectiva".

A finales del 2009 y principios del 2010, en los Estados Unidos se llevó a cabo la Encuesta Nacional sobre la Influenza H1N1, esta encuesta telefónica preguntó a los encuestados si habían recibido las vacunas contra la gripe H1N1 y la gripe estacional, así como preguntas que abarcaron sus antecedentes sociales, económicos y demográficos, comportamientos para mitigar la transmisión, opiniones sobre los riesgos de enfermedad y la efectividad de la vacuna.

Una mejor comprensión de cómo estas características se asocian con los patrones de vacunación puede proporcionar una orientación clara para futuros esfuerzos de salud pública.

Objetivo: Diseñar un modelo de Machine Learning que permita predecir la probabilidad de que una persona reciba las vacunas contra el virus H1N1 y la gripe estacional.

Métricas de negocio (KPIs):

- ✓ Porcentaje mínimo para lograr inmunidad de rebaño correspondiente al 32% para el virus H1N1 y 29%para la gripe estacional.

Métricas del modelo:

- ✓ ROC-AUC como la principal métrica para comparar el rendimiento de los modelos.

- ✓ Exactitud (Accuracy): porcentaje de predicciones correctas.

Ideación

El producto de datos a diseñar es un modelo de Machine Learning para que las autoridades de salud pública puedan prever qué grupos poblacionales tienen mayor o menor probabilidad de vacunarse y que características son las que influyen directamente en la toma de decisión.

En los Estados Unidos el movimiento antivacunas es bastante fuerte, por lo tanto, los usuarios potenciales son las autoridades de salud pública, quienes utilizarán el modelo para diseñar campañas de vacunación más dirigidas y así lograr aumentar la cantidad de personas vacunadas para alcanzar la inmunidad colectiva.

Requerimientos del producto:

- Modelo predictivo entrenado y validado.

El formato de resultados del modelo se basa en tres columnas, *respondent_id, h1n1_vaccine y seasonal_vaccine.*

Las predicciones para las dos variables objetivo (*h1n1_vaccine y seasonal_vaccine)* deben ser probabilidades tipo float que oscilen entre 0,0 y 1,0. Es importante mencionar que, de acuerdo con la particularidad del problema, no es necesario que las probabilidades de cada fila sumen uno.

Responsabilidad legal

El conjunto de datos de origen viene con las siguientes restricciones de uso de datos:

La Ley del Servicio de Salud Pública (Sección 308(d)) establece que los datos recopilados por el Centro Nacional de Estadísticas de Salud (NCHS) y los Centros para el Control y la Prevención de Enfermedades (CDC), pueden usarse únicamente con fines de informes estadísticos de salud y análisis. [1]

Adicionalmente, la NCHS retiró de la data todos los datos relacionados con la identidad de los encuestados, por lo tanto, cualquier identificación o revelación intencionada de una persona o establecimiento viola las garantías de confidencialidad dadas a los proveedores de la información.[2]

Enfoque analítico

Para la limpieza de los datos empezaremos con eliminación de columnas que tengan un porcentaje de nulos mayor al 30% y las cuales determinemos que no son relevantes para el análisis con el fin de evitar posibles sesgos con la imputación. Para la imputación de los datos validaremos el tipo de datos con el que cuente la columna con el objetivo de decir el método de imputación.

Debido a los rangos de los valores en las columnas numéricas y la naturaleza del Random Forest, donde este divide los valores en función de las características, no vemos la normalización de estos valores como crucial, aunque teniendo en cuenta que contamos con algunas columnas que están entre valores de 1 al 10 y normalmente el resto es 0 o 1, podremos aplicar pruebas con estandarización Z-score para validar el rendimiento del modelo.

Como nuestras variables categóricas no cuentan con un orden realizaremos la codificación de estas por medio de One-Hot Encoding siempre y cuando no tengan muchas categorías y así no incrementar sustancialmente la cantidad de columnas en el data set. Para las otras columnas usaremos Label Encoding que, aunque es muy útil para columnas con ordenamiento de valor, lo escogimos por su cualidad de asignar un número a cada categoría en una sola columna.

Se utilizará un algoritmo de aprendizaje supervisado y de clasificación como los árboles de decisión, específicamente el Random Forest, para predecir la

[1] [2]DrivenData. (s/f). *Flu shot learning: Predict H1N1 and seasonal flu vaccines*. DrivenData. Recuperado el 27 de octubre de 2024, de https://www.drivendata.org/competitions/66/flu-shot-learning/page/213/

probabilidad de que una persona obtenga cada una de las vacunas, este algoritmo nos permitirá tener un mejor análisis de los factores que influyen realmente en una persona a la hora de vacunarse, esto debido a que al ser un algoritmo basado en árboles nos entrega métricas sobre la importancia de cada variable, con esto no solamente podremos mejorar el modelo si no tendremos variables objetivo para recomendar un plan de mejora.

Para validar el modelo una de las medidas que evaluaremos será el Recall que se basa en medir la proporción de positivos que el algoritmo ha detectado correctamente, esto porque un falso positivo, al determinar que una persona no sea vacunada, implicaría un costo alto en términos de salud, por lo tanto, esta medida nos ayudará a determinar la eficiencia del modelo para predecir las personas realmente con una alta probabilidad de ser vacunadas y enfocarnos en el plan de mejora para las personas con baja probabilidad de ser vacunadas.

Como medida principal utilizaremos la métrica ROC AUC que nos permitirá evaluar la eficacia del modelo, se selecciona esta métrica por la facilidad que tiene para permitirnos saber el rendimiento del modelo sin establecer un umbral especifico.

Recolección de datos

Los datos están proporcionados por DrivenData y contienen características demográficas, de salud y económicas de las personas encuestadas. Estos datos provienen de la Encuesta Nacional sobre la Gripe H1N1 2009 (NHFS).

La NHFS fue una encuesta telefónica de hogares asistida por listas y mediante marcación aleatoria de dígitos, diseñada para monitorear la cobertura de vacunación contra la influenza en la temporada 2009-2010.

La población objetivo de la NHFS fueron todas las personas de 6 meses o más que vivían en los Estados Unidos en el momento de la entrevista. Los datos del NHFS se utilizaron para producir estimaciones oportunas de las tasas de cobertura de vacunación tanto para la vacuna monovalente pH1N1 como para la trivalente contra la influenza estacional.

Cada fila del conjunto de datos representa a una persona que respondió a la Encuesta nacional sobre la gripe H1N1 2009.

El conjunto de datos cuenta con 36 columnas. La primera columna *respondent_id* es un identificador único y aleatorio, de las restantes 35 columnas, hay 4 cuantitativas y 31 cualitativas.

Entendimiento de los datos

Para el set de datos encontrado en el *challenge*, se encuentran 26.707 filas, o registros, y 35 columnas, o características. Cada registro corresponde a una persona encuestada y las características observadas miden diferentes aspectos relacionados con datos básicos del encuestado (como sexo, grupo de edad, datos de vivienda y trabajo), opiniones acerca de efectividad de la vacuna y de comportamiento (prácticas para evitar contagio). Las características se presentan a continuación:

Columna	Tipo	Vacíos	Porcentaje vacíos
respondent_id	int64	0	0.00
h1n1_concern	float64	92	0.34
h1n1_knowledge	float64	116	0.43
behavioral_antiviral_meds	float64	71	0.27
behavioral_avoidance	float64	208	0.78
behavioral_face_mask	float64	19	0.07
behavioral_wash_hands	float64	42	0.16
behavioral_large_gatherings	float64	87	0.33
behavioral_outside_home	float64	82	0.31
behavioral_touch_face	float64	128	0.48
doctor_recc_h1n1	float64	2160	8.09
doctor_recc_seasonal	float64	2160	8.09
chronic_med_condition	float64	971	3.64
child_under_6_months	float64	820	3.07
health_worker	float64	804	3.01
health_insurance	float64	12274	45.96
opinion_h1n1_vacc_effective	float64	391	1.46
opinion_h1n1_risk	float64	388	1.45
opinion_h1n1_sick_from_vacc	float64	395	1.48
opinion_seas_vacc_effective	float64	462	1.73
opinion_seas_risk	float64	514	1.92
opinion_seas_sick_from_vacc	float64	537	2.01
age_group	object	0	0.00
education	object	1407	5.27

race	object	0	0.00
sex	object	0	0.00
income_poverty	object	4423	16.56
marital_status	object	1408	5.27
rent_or_own	object	2042	7.65
employment_status	object	1463	5.48
hhs_geo_region	object	0	0.00
census_msa	object	0	0.00
household_adults	float64	249	0.93
household_children	float64	249	0.93
employment_industry	object	13330	49.91
employment_occupation	object	13470	50.44

Tabla 1. Columna set de datos

Con respecto a la completitud de los datos, se observan bajos porcentajes de datos vacíos en su mayoría, a excepción de los datos de empleo (*employment_industry* y *employment_occupation*) con un 50% de datos faltantes y con datos de seguro médico con un 45% de datos faltantes. Al ser imposible la reconstrucción, o estimación de estos valores, no serán considerados para el proceso más allá de exploración de datos.

Centrándose en los tipos de variables, se encuentran 4 de tipo cuantitativo y 31 de tipo cualitativo. Las primeras hacen referencia a datos como salario anual promedio, grupo de edad, raza y estatus de empleo. Las variables cualitativas representan las demás características, midiéndose de manera binaria y en una escala de 0 a 5 para preocupación, conocimiento y opinión.

Estas características miden si la persona encuestada ha recibido una vacuna de H1N1 o de influenza (o ambas). Esta información se encuentra en otra tabla la cual tiene las siguientes columnas:

Columna	Tipo	Vacíos	Porcentaje vacíos
respondent_id	int64	0	0.00
h1n1_vaccine	int64	0	0.00
seasonal_vaccine	int64	0	0.00

Tabla 2. Columnas objetivo set de datos

Para relacionar esta tabla con la tabla de características existe la llave primaria *respondent_id*, las otras dos columnas representan si el encuestado fue vacunado por H1N1 o influenza (o ambas). No se presentan datos faltantes en este caso.

Haciendo énfasis ahora en los datos existentes de personas encuestadas, se puede apreciar lo siguiente:

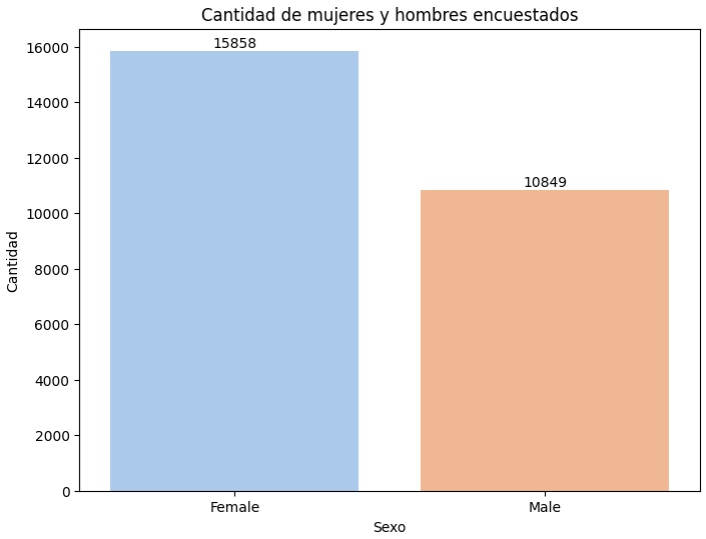


Figura 1. Distribución hombre y mujer

59% de encuestados son mujeres, con 41% hombres. De estos los vacunados son:

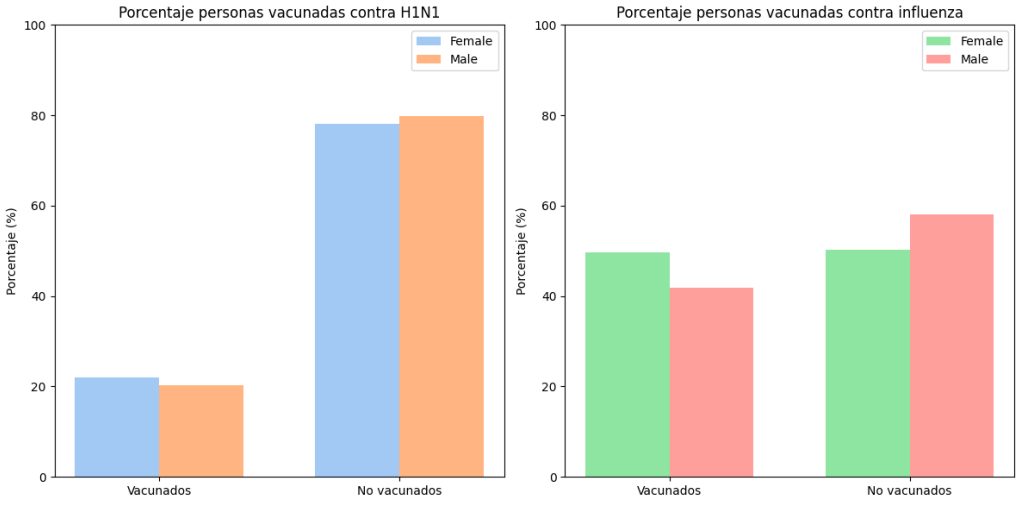


Figura 2. Vacunación por hombres y mujeres

Se puede apreciar que, para la vacuna de la influenza, hay un buen porcentaje de participación, tanto en hombres como mujeres, siendo las últimas las que más se vacunan (alrededor del 50%). En contraste, la vacunación contra H1N1 es más deficiente, alrededor de un 20% para hombres y mujeres. Observando la distribución de grupos de edades:

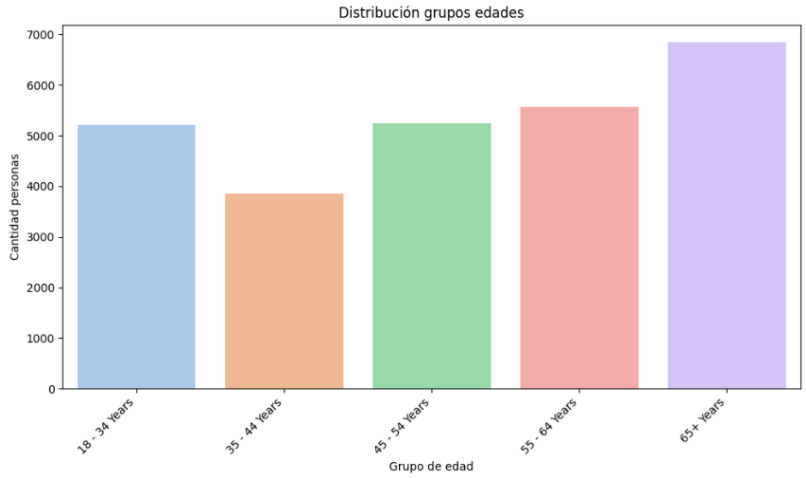
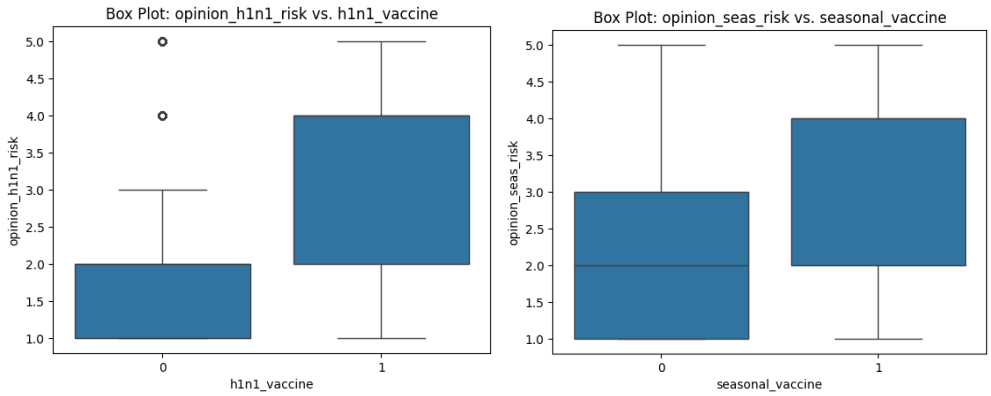


Figura 3. Grupo de edades

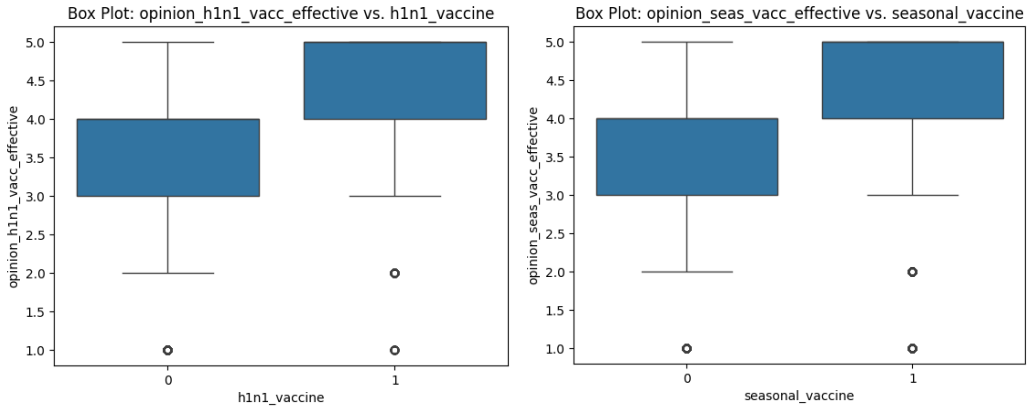
Hay una mayoría de personas mayores de 54 años, representando un 47% de las personas encuestadas.

Centrándose ahora con las características y cómo estas describen la vacunación, se identificaron las variables más significativas para el resultado objetivo de vacunas de H1N1 e Influenza, se han elegido la característica de opinión (riesgo, efectividad de la vacuna y preocupación a enfermarse por la aplicación de la vacuna).



En las gráficas anteriores se puede observar cómo la opinión del riesgo (sin vacunación) de H1N1 e Influenza determina, en parte, la toma de la vacuna, al tener un *box plot* concentrando los datos, en el caso de vacunación (valor 1), hacia el valor máximo (5, muy preocupado). Para profundizar en esto, encontrando la correlación de Pearson se observa que para la variable *opinion_h1n1_risk* y *h1n1_vaccine* se tiene un valor de r = 0.32, mostrando una correlación positiva. Este es el caso, de igual manera, para la variable de Influenza *opinion_seas_risk* y seasonal_*vaccine* con una correlación de r = 0.39.

Para la opinión de efectividad de las vacunas:



Se presenta la misma observación anterior, y encontrando la correlación se observa que para *opinion_h1n1_vacc_effective* y *h1n1_vaccine* el valor de *r* es de 0.26 y para *opinion_h1n1_seas_effective* y *seasonal_vaccine* r = 0.36.

Estas dos opiniones son de las variables que tienen relación (estadísticamente) más fuerte con respecto a la vacunación, con una relación positiva que indica que al tener una opinión positiva frente a la efectividad de las vacunas y una percepción de riesgo elevado sin vacuna hay mayor tasa de vacunación, independiente de sexo y edad.

Conclusiones

1. Se evidencia un bajo porcentaje de vacunación tanto para el virus H1N1 como para la influenza estacional, por lo que no se logró, en su momento, la inmunidad colectiva.

2. Se evidencia que las variables relacionadas con los hábitos de salud no tienen un impacto directo con la decisión de vacunación.

3. Se evidencia una fuerte correlación entre la opinión del riesgo que tiene contraer el virus H1N1 o la influenza estacional con la cantidad de personas vacunadas.

4. Las variables como el sexo, la raza o la ubicación geográfica no representan una relación significativa para determinar si una persona decide vacunarse. Del mismo modo, factores como el uso de tapabocas o el lavado de manos no están relacionados con la decisión de vacunarse.

5. Los encuestados que trabajan en el área de la salud muestran un comportamiento similar al del resto de los participantes, lo que sugiere que trabajar en el sector salud no implica una mayor probabilidad de vacunación.

6. Al usar el algoritmo de Random Forest en la problemática, se logrará una comprensión más profunda de los factores que influyen en la probabilidad de que una persona reciba la vacuna contra la gripe H1N1 y la gripe estacional.

7. La métrica AUC nos indicará si el modelo es capaz de distinguir de manera efectiva entre aquellos que probablemente se vacunarán y aquellos que no.