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

Entrega Final

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):

Teniendo en cuenta que el cliente del modelo es el gobierno de los Estados Unidos, se define como métricas de negocio la identificación de las condiciones socioeconómicas, culturales, entre otras, que llevan a una persona a vacunarse o no. De esta manera, se pueden planear campañas de vacunación dirigidas a la población identificada.

Métricas del modelo:

- ✓ ROC-AUC como la principal métrica para comparar el rendimiento de los modelos.
- ✓ Recall: Identificar correctamente los porcentajes de probabilidad de vacunación para la mayor cantidad de casos.

Ideación

En los EE. UU. existe una tendencia de movimientos antivacuna, lo que ha generado diversas problemáticas de salud que podrían ser evitadas mediante la vacunación de los ciudadanos, por esto se hace necesario la generación de campañas centralizadas para las diversas poblaciones.

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. Este modelo se acompaña con un dashboard, donde las autoridades de salud podrán filtrar por niveles de educación, raza, edad y localización, para esto se parte de un mockup (ver anexo 1) en el que se define el diseño para cada componente del dashboard y con esto obtener el porcentaje de personas vacunas contra la influenza y el H1N1.

Requerimientos del producto:

- Modelo de Machine Learning de clasificación entrenado y validado.
- Dashboard

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 – Preparación de los datos

Para la limpieza de los datos empezaremos con eliminación de columnas que tengan un porcentaje de nulos mayor al 40% 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

[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/

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.

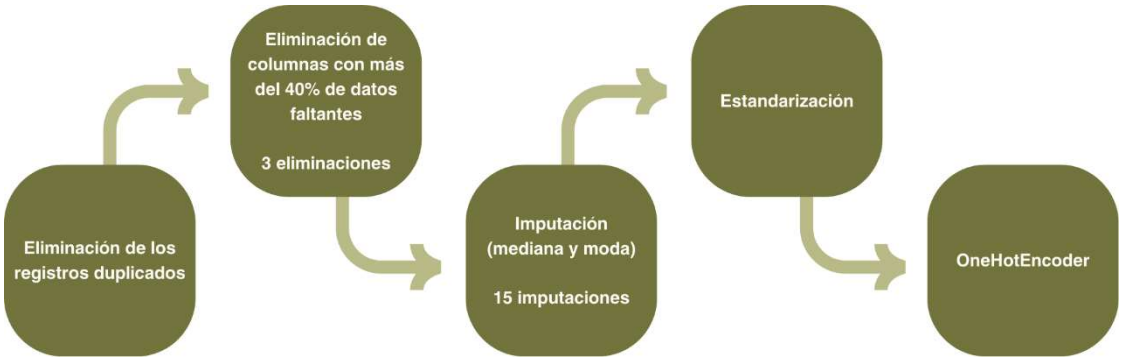


Figura 1. Flujo de datos

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

Implementaremos un dashboard con el fin de poder analizar los resultados de las predicciones y que así los analistas logren hacer un sesgo de las poblaciones con menor probabilidad de ser vacunados y logren tomar las medidas necesarias para realizar campañas de vacunación.

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).

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. 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 1. 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:

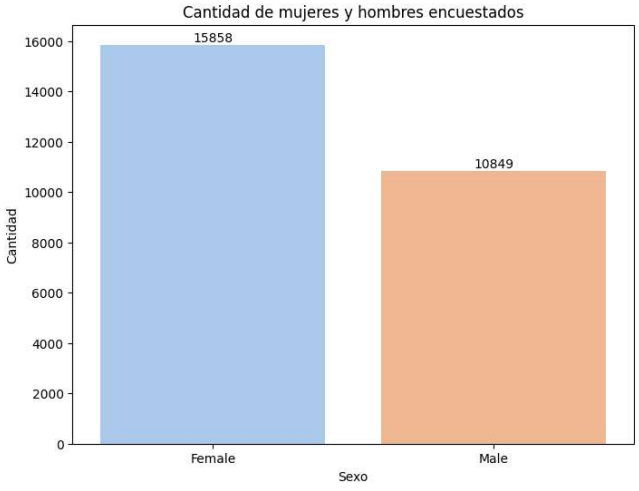


Figura 2. Distribución hombre y mujer

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

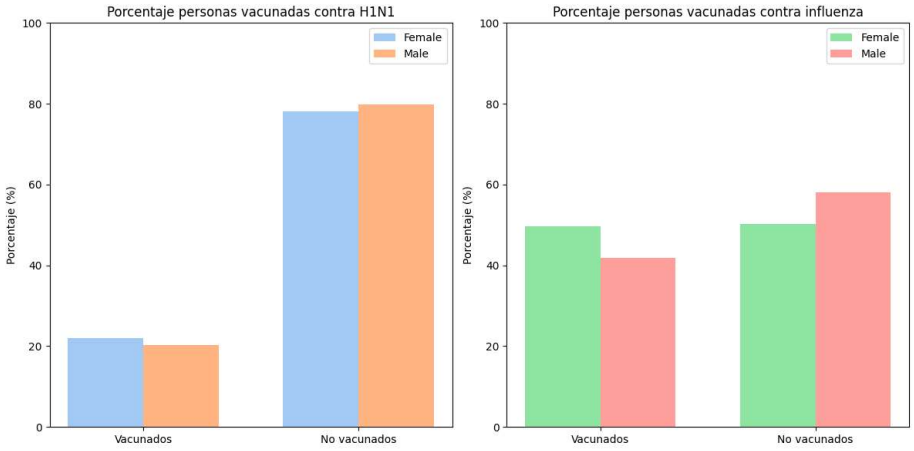


Figura 2. Vacunación por hombres y mujeres

Se puede apreciar que, para la vacuna de la influenza, hay un importante 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:

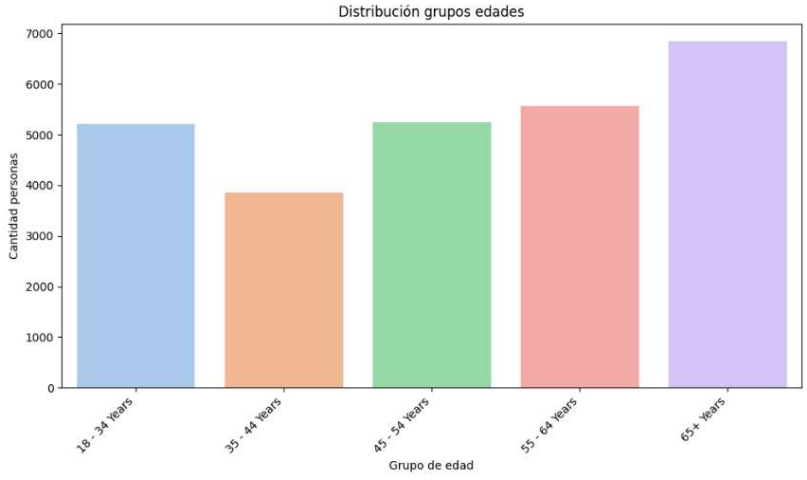


Figura 3. Grupo de edades

Hay una mayoría de personas mayores de 54 años, representando un 47% de las personas encuestadas. Observando ahora los datos de vacunación por grupos de edad, se encuentra que la edad no influye en la vacunación contra H1N1 (alrededor del 21% de encuestados sin importar el grupo de edad se vacuna):

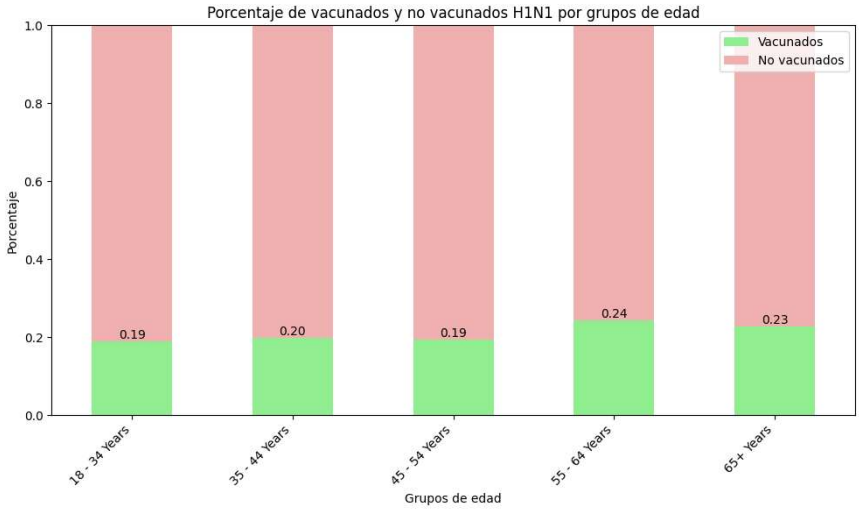


Figura 3. Vacunación H1N1 por grupos de edad

Con respecto a la vacunación contra la influenza, la tendencia cambia a favor de vacunación a mayor edad. La siguiente figura muestra un crecimiento en los encuestados en términos de vacunación a medida que el grupo de edad incrementa:

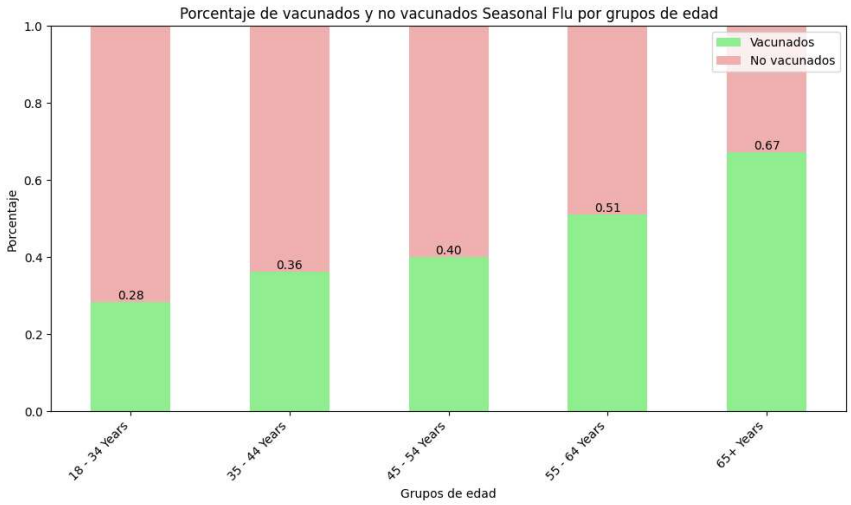


Figura 4. Vacunación influenza por grupos de edad

El set de datos proporciona el nivel educativo del encuestado, indicando el último nivel educativo alcanzado. Para la vacunación de H1N1 se tiene la siguiente información:

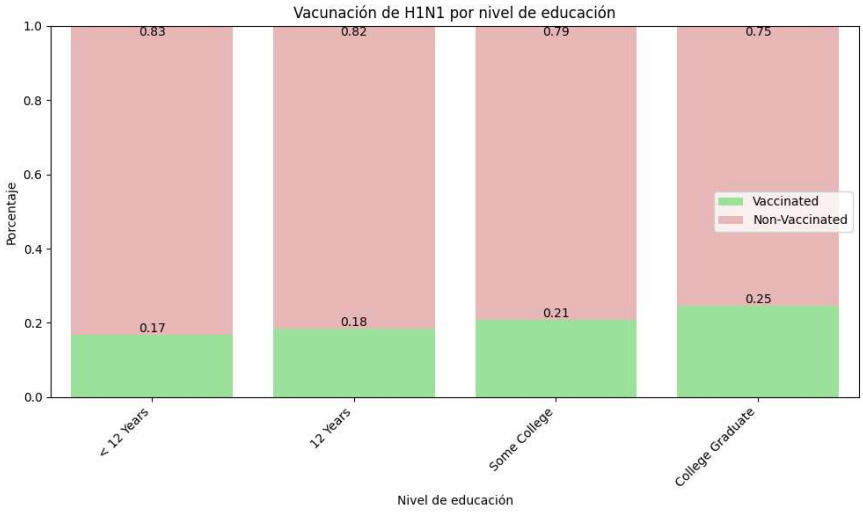


Figura 5. Vacunación H1N1 por nivel de educación

Se observa un pequeño crecimiento en la vacunación contra el H1N1a medida que el nivel educativo incrementa, aumentando 7% entre graduados de secundaria y graduados de universidad. Con respecto a la influenza se presenta un crecimiento similar:

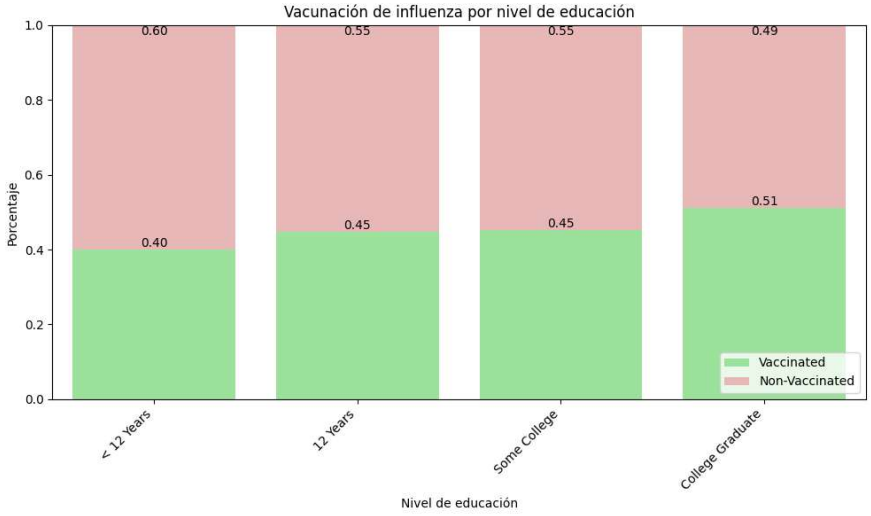


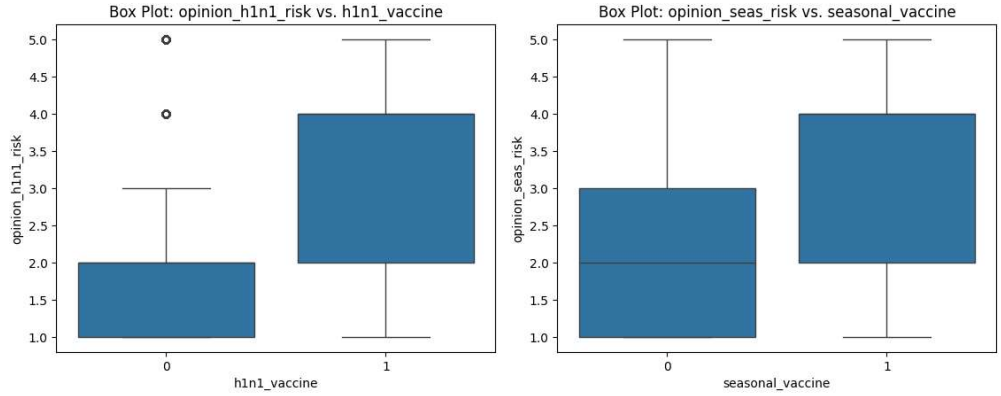
Figura 6. Vacunación influenza por nivel de educación

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). Estas variables se miden, como se menciona anteriormente, en una escala de 1 a 5:

- 1: Sin preocupación alguna
- 2: Sin gran preocupación
- 3: No sabe
- 4: Algo de preocupación
- 5: Muy preocupado

Relacionando estas opiniones, en este caso de preocupación del riesgo de contraer H1N1 o influenza, con la variable objetivo de vacunación (0 para no vacunados y 1 para vacunados) se obtiene el siguiente gráfico:



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. Esta relación puede visualizarse (en parte) con las siguientes gráficas:

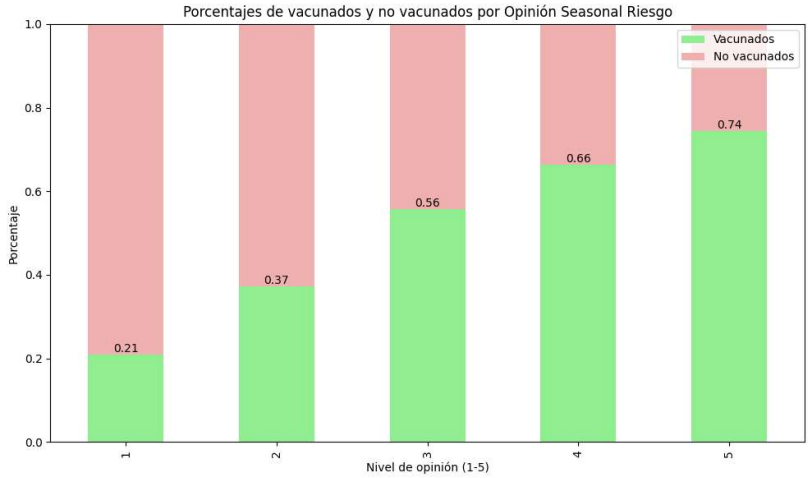


Figura 7. Vacunación influenza por opinión de riesgo

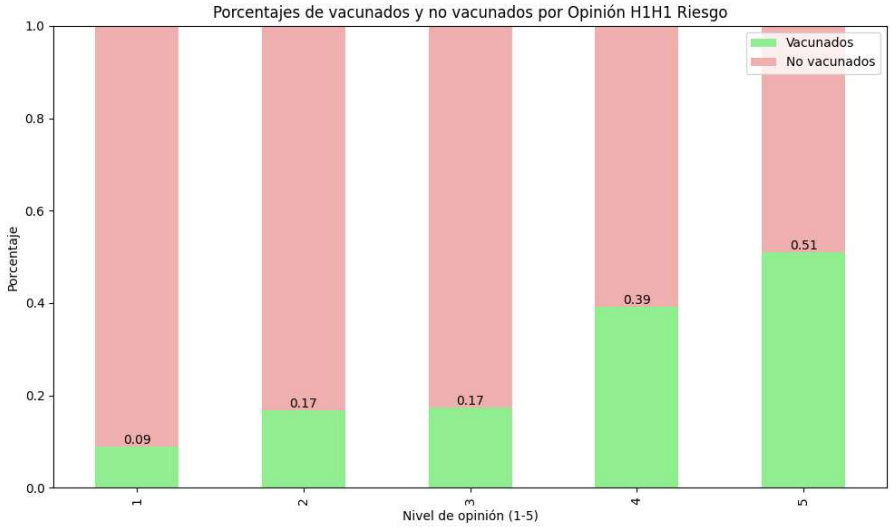
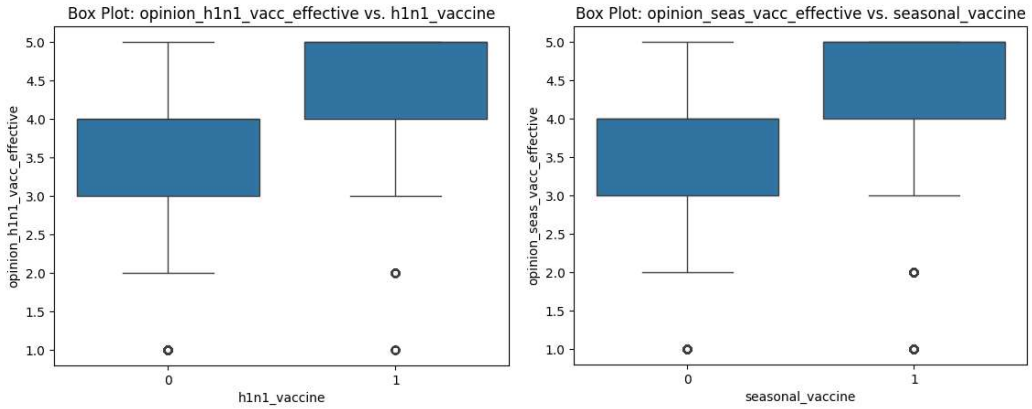


Figura 8. Vacunación H1N1 por opinión de riesgo

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. Esta relación puede visualizarse (en parte) con las siguientes gráficas:

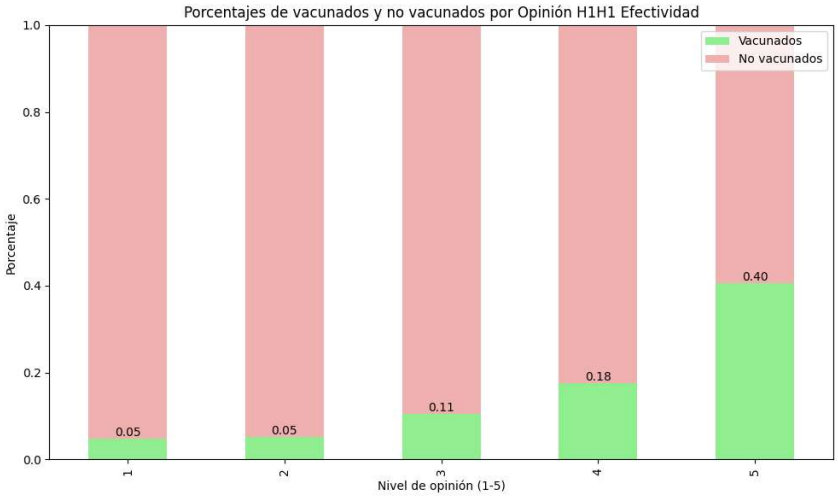


Figura 9. Vacunación de H1N1 por opinión de efectividad

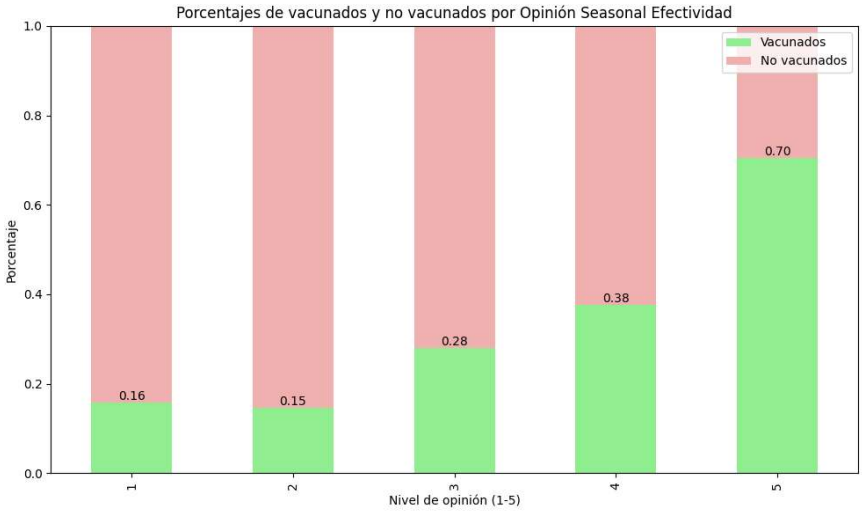


Figura 10. Vacunación influenza por opinión de efectividad

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.

Estrategia de validación y selección de modelo

Construcción de los modelos:

Para la implementación de los modelos primeramente se eliminaron las variables las cuales contaran con un porcentaje mayor al 50 % de datos faltantes, en este caso fueron tres (employment_industry, employment_occupation, health_insurance), teniendo estas variables eliminadas se aplicó una imputación de las variables categóricas por la moda y de las variables numéricas por la media.

Para la estandarización de las variables categóricas se realizaron pruebas con el LabelEconder y el OneHotEnconder, obteniendo una mejora del 4% en el ROC AUC, para las variables numéricas se realizó una estandarización por medio del StandardScaler.

Para las pruebas en los modelos se utilizaros tres modelos de regresión RandomForest, DecisisionTree y LogisticsRegresion , teniendo configurados los hiper parámetros en random_state de 42 y un máximo de iteraciones en 100.

Al finalizar como nuestro modelo es de dos salidas, cada una por las diferentes vacunas, se aplicó en el modelo MultiOutputClassifier para que nos entregue en un arreglo cada una de las predicciones, para los resultados se realizar las predicciones bajo el formato de probabilidad.

Elección del modelo:

Para la elección del modelo tuvimos en cuentan como métrica principal el ROC AUC y como métricas secundarias validamos el recall, Precision, f1 y Accuracy, obteniendo estos resultados.

Random Forest :

- Accuracy: - Test: 0.6553725196555598
- Precision: - Test: 0.7213887413535094
- Recall: - Test: 0.6316671320860094
- F1: - Test: 0.6670636517558142
- ROC AUC micro-promedio: 0.8515

Logistic Regression:

- Accuracy: - Test: 0.6712841632347435
- Precision: - Test: 0.7468101754546421
- Recall: - Test: 0.6375314158056409
- F1: - Test: 0.680516733497174
- ROC AUC micro-promedio: 0.8587

Desicion Tree :

- Accuracy: - Test: 0.5119805316360914
- Precision: - Test: 0.5630795719842744
- Recall: - Test: 0.5772130689751466
- F1: - Test: 0.5700427483761767
- ROC AUC micro-promedio: 0.6745

Con estas pruebas decidimos que la mejor opción por la combinación de las pruebas sería el Logistic Regression, con el cual realizamos las predicciones entregadas en el producto de datos.

Construcción del producto de datos

El producto de datos se materializa en un dashboard que cuenta con dos visualizaciones principales:

1. **Primera Visualización**: Esta sección (ver Anexo 2) muestra las probabilidades de vacunación para cada uno de los registros disponibles en la base de datos. Además, integra filtros interactivos que permiten segmentar la información según los niveles de educación, edad, ubicación geográfica y raza. Esto permite al cliente final personalizar su análisis y focalizarse en poblaciones específicas según los criterios seleccionados.
2. **Segunda Visualización**: En esta sección (ver Anexo 3), se proporciona una visión integral del estado de la vacunación, partiendo de la probabilidad de vacunación para cada una de las dos vacunas analizadas. A través de un gráfico tipo radar, se facilita un análisis detallado en cuatro dimensiones clave: grupo de edad, ubicación geográfica, raza y nivel de escolaridad. Este diseño permite al cliente identificar rápidamente qué dimensiones contribuyen de manera significativa a las probabilidades seleccionadas, promoviendo una comprensión clara y ágil de los datos.

El objetivo general del dashboard es permitir al cliente final tomar decisiones informadas para optimizar las estrategias de vacunación, asegurando que estas sean inclusivas y efectivas.

Para la elaboración de este dashboard se utiliza la herramienta Power BI, con una integración a una Power App, esto con el objetivo de que nuestro cliente final puede consultar la información de forma sencilla y con todas las medidas de seguridad e infraestructura que brinda Microsft.

De igual forma se hace entrega del manual, este contiene el paso a paso que debe seguir el cliente para poder ver las predicciones actualizadas en el Dashboard

Conclusiones

1. Se evidencia un bajo porcentaje de vacunación tanto para el virus H1N1 como para la influenza estacional en el grupo de edad de 18 a 35 años y un porcentaje más alto en los mayores de 65.
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. 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.
5. El algoritmo de Radom Forest y la Regresión Logística fueron los algoritmos con el mejor desempeño, ambos tuvieron un recall del 63%, lo que nos dice que el porcentaje de verdaderos positivos es significativo.
6. Mediante la métrica ROC AUC podemos concluir que el modelo de Logistics Regression, determina sigficativamente si una persona se va a vacunar o no, ya que, nos dio una media superior al 80%

Anexos

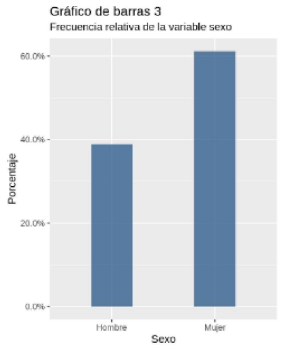
Columna	Tipo	Vacíos	Porcentaje vacíos
respondent_id	int64	0	0.00
h1n1_concern	float64	92	0.34
h1n1_knowledge	float64	116	0.43
behavioral_antiviral_meds	float64	71	0.27
behavioral_avoidance	float64	208	0.78
behavioral_face_mask	float64	19	0.07
behavioral_wash_hands	float64	42	0.16

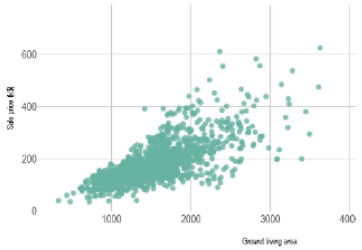
behavioral_large_gatherings	float64	87	0.33
behavioral_outside_home	float64	82	0.31
behavioral_touch_face	float64	128	0.48
doctor_recc_h1n1	float64	2160	8.09
doctor_recc_seasonal	float64	2160	8.09
chronic_med_condition	float64	971	3.64
child_under_6_months	float64	820	3.07
health_worker	float64	804	3.01
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

Count	H1N1	Seasonal

Education filter | Race filter | Location filter | Age filter





Vaccines Report

respondent_id
26707 53414

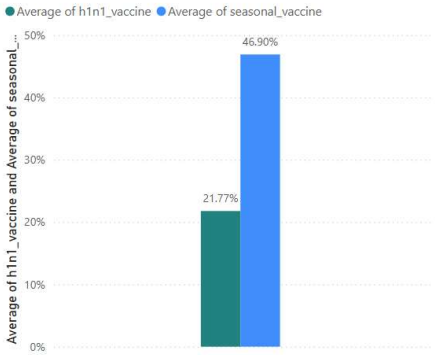
Features / Vaccines

Recuento ID	h1n1_vaccine	seasonal_vaccine
2432	0.00	0.00
2398	0.00	0.10
1255	0.00	0.20
1060	0.10	0.20
958	0.10	0.10
840	0.00	0.30
831	0.10	0.30
693	0.10	0.40
640	0.10	0.50
601	0.00	0.80
589	0.00	0.90
581	0.00	0.40
512	0.10	0.60
507	0.00	0.50
506	0.00	0.70
503	0.10	0.70
497	0.00	0.60
494	0.10	0.80
476	0.20	0.30
472	0.20	0.40
464	0.20	0.20
410	0.20	0.50
372	0.20	0.60
26708		

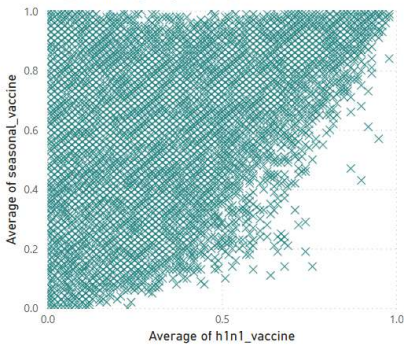
Filters

Education	Race	hhs_geo_region	Age_group
Select all	Select all	Select all	Select all
< 12 Years	Black	atmpeygn	18 - 34 Years
12 Years	Hispanic	bhuqouqj	35 - 44 Years
College Graduate	Other or Multiple	dqpwygqj	45 - 54 Years
Some College	White	fpwskwrf	55 - 64 Years
Without education		kbazzjca	65+ Years

Average of h1n1_vaccine and Average of seasonal_vaccine



Average of h1n1_vaccine and Average of seasonal_vaccine by respondent_id




Choose the effective percentage
h1n1_vaccine
0.25
0.50
seasonal_vaccine
0.25
0.50
H1n1_vaccine and Seasonal_vaccine by education
Axis H1n1_vaccine Seasonal_vaccine
College Graduate
Some C...
12 Years
< 12 Years
Without ...
H1n1_vaccine and Seasonal_vaccine by race
Axis H1n1_vaccine Seasonal_vaccine
White
Hispa...
Other or Multiple
Black
H1n1_vaccine and Seasonal_vaccine by age_group
Axis H1n1_vaccine Seasonal_vaccine
18 - 34 Years
35 - 4...
55 - 64 Years
45 - 54 Years
65+ Y...
H1n1_vaccine and Seasonal_vaccine by hhs_geo_region
Axis H1n1_vaccine Seasonal_vaccine
lzgpxyit
bhuqouqj
oxchjgsf
qufhixun
fpwskwrf
mlyzmhmf
kbazzjca
lrircsnp
atmpeygn
dqpwygqj