**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**

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".

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. A finales del 2009 y principios del 2010, los Estados Unidos llevaron a cabo la Encuesta Nacional sobre la Influenza H1N1 2009, esta encuesta telefónica preguntó a los encuestados si habían recibido las vacunas contra la gripe H1N1 y la gripe estacional, junto con preguntas sobre ellos mismos.

Estas preguntas adicionales abarcaron sus antecedentes sociales, económicos y demográficos, opiniones sobre los riesgos de enfermedad y la efectividad de la vacuna, y comportamientos para mitigar la transmisión.

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

Objetivo: Diseñar un modelo que permita predecir la probabilidad de que una persona reciba las vacunas contra la influenza 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 sería un modelo de machine Learning para que las autoridades de salud puedan prever qué grupos poblacionales tienen mayor o menor probabilidad de vacunarse.

Los usuarios potenciales serían las autoridades de salud, para diseñar campañas de vacunación más dirigidas.

Requerimientos del producto:

* Modelo predictivo entrenado y validado.
* Visualización (dashboard o reporte).

**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), los Centros para el Control y la Prevención de Enfermedades (CDC), pueden usarse únicamente con fines de informes estadísticos de salud y análisis.

**Enfoque analítico**

Para la limpieza de los datos empezaremos con eliminación de columnas que tengan un porcentaje mayor de nulos 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 usaremosLabel Encoding que es muy útil para columnas con ordenamiento de valor, pero lo escogimos por su cualidad de asignar un numero 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 si una persona fue vacunada o no, este algoritmo no 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 de en árboles nos entregaran 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 nos centraremos en la medida **Recall** que se basa en medir la proporción de positivos que el algoritmo ha detectado correctamente, esto porque el costo de un falso positivo puede llegar a determinar que una persona no sea vacunada lo cual implicaría un costo alto en términos de salud, por lo tanto, esta medida nos ayudara a determinar la eficiencia del modelo para predecir las personas realmente con posibilidad de ser vacunadas y enfocarnos en el plan de mejora para las personas que no serán vacunadas.

**Recolección de datos**

Los datos están proporcionados por DrivenData y contienen características demográficas, de salud y económicas de las personas. 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.

**Entendimiento de los datos**

Análisis de calidad y limpieza de los datos.

Análisis univariado y multivariado.

**Conclusiones**