UNIVERSIDAD
DE ANTIOQUIA

TRABAJO ANALÍTICA DE DATOS APLICADA EN SALUD

SANTIAGO GÓMEZ BERRÍO
DIEGO ANDRÉS LUNA PATERNINA
MARIA CLARA SALAZAR DUQUE

JUAN CAMILO ESPAÑA LOPERA
ANALÍTICA DE DATOS

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
MEDELLÍN

2023

Análitica de Datos Aplicada en Salud

Descripción del problema:
Un reingreso hospitalario se produce cuando un paciente que ha recibido el alta hospitalaria vuelve a ingresar en un plazo determinado. Las tasas de readmisión hospitalaria para determinadas afecciones se consideran actualmente un indicador de la calidad hospitalaria, y también afectan negativamente al coste de la asistencia. Por este motivo, los Centros de Servicios de Medicare y Medicaid crearon el Programa de Reducción de Reingresos Hospitalarios, cuyo objetivo es mejorar la calidad de la atención a los pacientes y reducir el gasto sanitario mediante la aplicación de penalizaciones de pago a los hospitales que presenten tasas de readmisión superiores a las previstas para determinadas afecciones.

Un informe del Instituto de Liderazgo de Salud Global de Yale, señala que la atención médica real solo representa el 20% de lo que determina la salud de un paciente, el resto está definido por la genética en 20% y la mayoría por factores sociales, ambientales y de comportamiento que suman 60%. Así lo indica una nota de Healthcare Business & Technology.

Aunque la diabetes aún no está incluida en las medidas de penalización, el programa está añadiendo regularmente nuevas condiciones de enfermedad a la lista, que ahora asciende a 6 para el año fiscal 2018. En 2011, los hospitales estadounidenses gastaron más de 41.000 millones de dólares en pacientes diabéticos que fueron readmitidos en los 30 días siguientes al alta. Ser capaz de determinar los factores que conducen a una mayor readmisión en este tipo de pacientes y, en consecuencia, ser capaz de predecir qué pacientes serán readmitidos puede ayudar a los hospitales a ahorrar millones de dólares al tiempo que mejora la calidad de la atención. Así pues, con estos antecedentes en mente, utilizamos un conjunto de datos para responder a esta pregunta: ¿Cuáles son los factores que mejor predicen los reingresos hospitalarios en pacientes diabéticos?

El coste de los reingresos hospitalarios representa una gran parte del gasto en servicios de hospitalización. La diabetes no sólo es una de las diez principales causas de muerte en el mundo, sino también la enfermedad crónica más cara en Estados Unidos. Los pacientes diabéticos hospitalizados corren un mayor riesgo de reingreso que los no diabéticos. Por lo tanto, la reducción de las tasas de readmisión de los pacientes diabéticos tiene un gran potencial para reducir significativamente el coste médico. El objetivo de este estudio es predecir la probabilidad de que un paciente diabético sea readmitido. El conjunto de datos se obtuvo del Centro de Aprendizaje Automático y Sistemas Inteligentes de la Universidad de California, Irvine, y contiene más de 100.000 atributos y 50 características, como el número de procedimientos, el número de medicamentos, el tiempo de hospitalización, etc.

Diseño de la solución:
La propuesta de solución como se muestra en la *figura 1* será llevada a cabo por el equipo de analítica. Esta solución inicia desde el paso en que ya se ha detectado que el paciente tiene diabetes y que debe ser hospitalizado. Se requiere saber si viene para reingreso o si será una hospitalización por primera vez. Dicha información será guardada en una base de datos con la información del paciente para alimentar el modelo que clasificará los factores identificados con mayor relevancia al momento de incidir en un reingreso hospitalario. Posteriormente, se hará uso de redes neuronales con el fin de saber para cuántos posibles reingresos podrían haber en el hospital. Toda esta información se guardará y analizará para presentar un informe a los directivos del hospital.

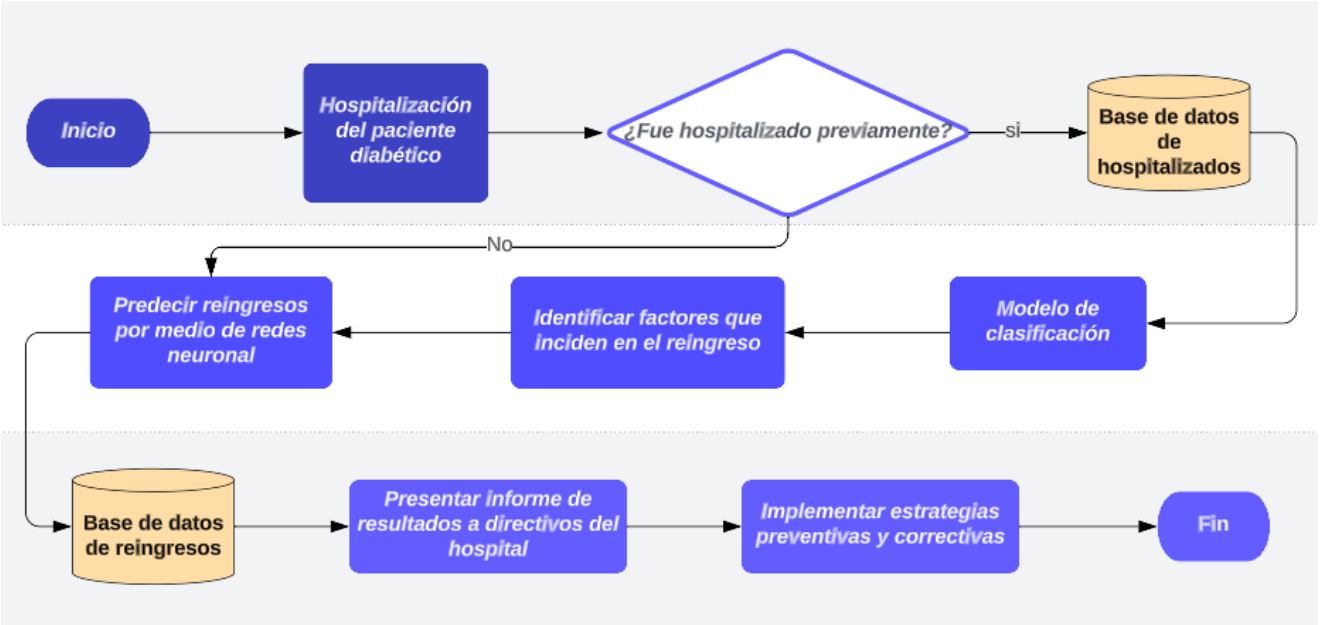

Figura 1. *Diagrama de solución propuesta*

Limpieza y transformación:

Se revisaron los tipos de variables, y se procedió a verificar que no hubiesen datos duplicados, después se identificó la presencia de información o valores nulos e inconsistentes en las variables weight, payer_code, medical_speciality, citoglipton, y examide, por lo que se optó por eliminarlas del data frame, ya que tenían casi 70 mil datos nulos, es decir, más del 50% de la base datos. En cuanto a la variable target llamada "Readmitted", se observó que tenía tres categorías, "<30", ">30", y "No", es decir, los pacientes son readmitidos en un periodo menor a 30 días, mayor a 30 días, o no ha sido readmitido, ya que se trata de una hospitalización nueva. Para tener un mejor enfoque en el estudio, se decidió dejar solo dos categorías, "<30 días" y "No", aclarando que esta última contiene también a los pacientes que son reingresados en un periodo mayor a los 30 días.

Análisis exploratorio:

Una vez finalizada la limpieza y transformación de datos, se procedió a analizarlos, encontrando la siguiente información relevante.

- La mayoría de los pacientes con diabetes son mayores de 30 años.
- La diabetes se considera una enfermedad genética, y la raza que más presenta esta enfermedad es la caucásica seguida de la afroamericana.
- Según la información presentada, el género femenino es el que más tiende a presentar diabetes.
- No se encontró relación alguna entre la edad del paciente con respecto a la estancia hospitalaria, dado a que presentan comportamientos aleatorios, pero la edad sí puede influir en la readmisión, ya que se presentan mayores reingresos en pacientes con más de 40 años de edad.
- En general la estancia de los pacientes dura entre 1 y 15 días salvo algunos casos atípicos donde dura un poco más
- Se encontró que aproximadamente 78.000 personas no toman medicamentos para la diabetes, y solo alrededor de 25.000 personas están medicadas para tratar esta enfermedad.

Selección del modelo y redes neuronales:
La utilización de redes neuronales en el análisis de datos de salud puede tener varias ventajas. Una de ellas es la capacidad de identificar patrones complejos y no lineales en los datos que podrían ser difíciles de detectar mediante métodos estadísticos convencionales. Las redes neuronales, son altamente tolerantes a errores, por lo que se comportan de forma excelente cuando existen imprecisiones en la información, como ocurre frecuentemente en medicina, lo que las convierte en una ayuda inestimable a la hora de tomar decisiones clínicas, minimizando de esta forma su incertidumbre (N. Sáenz & M. Álvaro, 2002).

Además, las redes neuronales pueden manejar datos de diferentes tipos y escalas sin necesidad de realizar una normalización rigurosa o una selección de características previa. Sin embargo, para responder la pregunta del caso de estudio, si es necesario recurrir a un modelo que clasifique los factores más relevantes al momento de predecir la readmisión de un paciente diabético.

Para ello, se construyó un modelo predictivo de Random Forest con una precisión del 89%, el cual, de acuerdo a la matriz de confusión no estaba prediciendo de la mejor manera cuantas personas estaban siendo readmitidas. En este caso, se tuvo en cuenta una técnica de balanceo como RandomUnderSampler y que de acuerdo a esto, se tiene una mejor predicción de las personas que el modelo considera readmitidas en el hospital. Aun así, se tiene un desempeño del 71% lo cual se considera aceptable. Posteriormente se hizo uso de las redes neuronales para perfeccionar la predicción de pacientes diabéticos que podrían llegar a ser readmitidos, dichas redes contienen la siguiente arquitectura e hiperparámetros:

Red Neuronal 1

Capa de entrada: esta capa tiene 128 neuronas y utiliza la función de activación ReLU. La capa recibe como entrada un vector de características de longitud igual al número de columnas del conjunto de datos de entrenamiento.

Capa oculta: esta capa tiene 64 neuronas y utiliza la función de activación ReLU. La capa recibe como entrada las salidas de la capa de entrada..

Capa de salida: esta capa tiene una única neurona y utiliza la función de activación sigmoide. La capa produce una salida binaria que representa la probabilidad de que un paciente sea readmitido en menos de 30 días.

Red Neuronal 2

Capa de entrada: esta capa tiene 64 neuronas y utiliza la función de activación ReLU. La capa recibe como entrada un vector de características de longitud igual al número de columnas del conjunto de datos de entrenamiento.

Capa de dropout: esta capa es opcional y ayuda a prevenir el sobreajuste al eliminar aleatoriamente algunas de las neuronas de la capa anterior durante el entrenamiento.

Capa oculta: esta capa tiene 32 neuronas y utiliza la función de activación ReLU. La capa recibe como entrada las salidas de la capa de entrada.

Capa de dropout: esta capa es opcional y ayuda a prevenir el sobreajuste al eliminar aleatoriamente algunas de las neuronas de la capa anterior durante el entrenamiento.

Capa de salida: esta capa tiene una única neurona y utiliza la función de activación sigmoide. La capa produce una salida binaria que representa la probabilidad de que un paciente sea readmitido en menos de 30 días.

El uso de dropout es opcional para ayudar a prevenir el sobreajuste del modelo.

Evaluación y análisis del modelo:
Se tuvieron en cuenta los siguientes gráficos para evaluar el desempeño de las redes neuronales:

Gráfico de pérdida: Esta gráfica muestra cómo disminuye la pérdida del modelo durante el entrenamiento. La pérdida es una medida de qué tan lejos están las predicciones del modelo de los valores reales. Se espera que la pérdida disminuya a medida que el modelo se ajusta a los datos de entrenamiento.

Gráfico de precisión: Esta gráfica muestra cómo mejora la precisión del modelo durante el entrenamiento. La precisión es una medida de qué tan bien el modelo puede predecir las etiquetas correctas para los datos de prueba. Se espera que la precisión aumente a medida que el modelo se ajusta a los datos de entrenamiento.

Matriz de confusión: Esta gráfica muestra cómo el modelo clasifica correctamente o incorrectamente las etiquetas de los datos de prueba. La matriz de confusión compara las etiquetas reales con las etiquetas predichas del modelo. Los valores de la diagonal principal de la matriz de confusión representan las predicciones correctas del modelo, mientras que los valores fuera de la diagonal principal representan las predicciones incorrectas.

Despliegue del modelo:
Es importante monitorear el rendimiento del modelo y actualizarlo según sea necesario para garantizar la precisión y la eficacia. En este caso el periodo de predicción en cuanto a readmisión de pacientes diabéticos se realizará cada tres días, dado que se considera lo más prudente debido a la recolección de información y manejo de recursos.

Una vez se tengan las predicciones de los pacientes diabéticos que serán readmitidos, se pasará un informe a los directivos del hospital, y ellos se encargarán de implementar estrategias y tomar decisiones, por ejemplo, si deben contratar especialistas en esta enfermedad, o brindar una mejor atención a los pacientes para evitar los costos y reprocesos de un reingreso hospitalario.

Referencia
N. Sáenz Bajoa, M. Álvaro Ballesteros. (2002, 30 junio). Redes neuronales: concepto, aplicaciones y utilidad en medicina. Atención Primaria.
https://www.elsevier.es/es-revista-atencion-primaria-27-articulo-redes-neuronales-concepto-aplicaciones-utilidad-13033737