INFORME CASO DE ESTUDIO SALUD

ELABORADO POR:

ANDRÉS FELIPE PENNA HERNÁNDEZ

JHONATAN VALENCIA OCAMPO

STEVEN ZAPATA ZULETA

PROFESOR:

JUAN CAMILO ESPAÑA LOPERA

ANALÍTICA



UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

1. **Diseño de solución propuesto.**

Problema de negocio

En la EPS SALUD es importante reducir los costos operativos con el fin de mejorar su eficiencia en el manejo de recursos, para cumplir el objetivo se ha realizado un análisis de los rubros que se deben cubrir, identificando qué oportunidades existen de reducción en diversos focos. Entre las estrategias evaluadas la que presenta un mayor interés de acuerdo con su impacto socio-sanitario y económico es el enfoque hacia la promoción y prevención de algunas enfermedades crónicas que requieren altos recursos financieros, estas enfermedades se caracterizan por ser de largo plazo y con baja posibilidad de tener cura.

Entre dichas enfermedades crónicas, la EPS enfatiza en la diabetes mellitus tipo II, tal interés se atribuye a la cantidad de factores o enfermedades que surgen y se desarrollan como consecuencia de la evolución de la enfermedad en cuestión al no ser tratada adecuadamente. Algunas de las causas asociadas que van desde la mala alimentación a malos hábitos de vida hacen factible la intervención a partir de la prevención.

Según datos de la cuenta de altos costos, en Colombia el 3,11% de la población tiene diabetes, con lo cuál las EPS del régimen contributivo invierten más de 600.000 millones de pesos en atención a personas con diabetes, logrando mantener controladas al 53% de estas.

Sin embargo, el reto asociado a esta enfermedad es identificar a las personas que aún no han sido diagnosticadas, se estima que son el 30-40% de los casos, esto se debe a que en la fase inicial no se presentan sintomatología.

Basándonos en la utilización de herramientas de inteligencia artificial, enfocados en la técnica de Machine Learning que se presenta como un elemento prometedor para mejorar la precisión del diagnóstico, reducir costos médicos asociados con la atención de pacientes diabéticos y capaz de analizar grandes cantidades de datos clínicos, es posible generar patrones que identifiquen tempranamente la diabetes y sus posibles causas, por ello este modelo también pueden proporcionar información valiosa para desarrollar recomendaciones basadas en hábitos saludables adaptadas a las necesidades que requieren los pacientes propensos a dicha enfermedad.

Debido al alto impacto de la enfermedad a la integridad de las personas y en los sistemas de salud que acarrean altos costos para la EPS es importante prevenir los factores que sean modificables asociados con la diabetes tipo II.

Problema analítico

Se establecen los siguientes apartados de analítica, estos son:

1. Definir las variables de información médica relevante que implique la generación de elementos característicos de personas con diabetes mellitus tipo II y con estos datos diseñar un sistema de prevención basados en el cuidado de los hábitos de los pacientes.
2. Realizar un modelo que prediga la propensión de un usuario a desarrollar la enfermedad de diabetes mellitus tipo II con el fin de tomar acciones preventivas, de acuerdo al tipo de edad.

**b. Análisis exploratorio y selección de variables**

Para la exploración de datos, se requiere establecer como primera instancia la procedencia de los mismos, esta base fue extraída desde el repositorio Kaggle con el nombre de *Diabetes detailed EDA with conclusion*. Una vez extraídos los datos se efectúa una revisión del contenido y para cada una de las variables se establecen posibles relaciones con la variable elegida como objetivo esto se realiza con el fin de determinar patrones que impliquen la generación de tendencias basados en personas que ya poseen un diagnóstico positivo en diabetes mellitus tipo II.

Para analizar las variables contenidas en la base de datos, se reconoce que hay en total 18 columnas y 70.692 datos, todos estos son de tipo float, las variables halladas son: edad, género, nivel de colesterol, chequeo en el nivel de colesterol, índice de masa corporal, frecuencia al fumar, historial de infartos, actividad física, consumo fruta, vegetales, alcohol, percepción de su salud en general, salud mental, salud física, si se posee dificultad para caminar, haber sufrido derrame cerebral, hipertensión y por último diabetes.

Se eliminan aquellas variables que se consideran innecesarias para el caso de estudio, el chequeo de colesterol es el caso de variable innecesaria. Además se valida si hay datos nulos.

A partir de estas variables se hallaron relaciones y elementos importantes para realizar análisis iniciales, algunas de ellas fueron:

* De acuerdo con la relación aislada de otras variables, se hace correspondencia de la posible causa de la diabetes es el auto-criterio sobre la salud de cada paciente escalados de 1 a 5 (1 = excelente, 5 = mala), los resultados hallados no dan un buen indicador de que la autopercepción indique un posible diagnóstico de diabetes, dónde la mayoría de datos arrojan como resultado un auto-criterio de buena salud y son diagnosticados con diabetes con un 33% del total.
* La relación entre la edad y diabetes acumula en los rangos de edades 9 - 10 - 11 correspondientes a las edades entre 60-64, 65-69, 70-74 respectivamente, esta variable puede ser un indicador importante que se asocia a la diabetes, focalizando estos rangos de edad como personas que propenden a sufrir de dicha enfermedad.
* Los géneros masculino y femenino fueron evaluados dando como resultado que las mujeres poseen mayor probabilidad 52.1% de sufrir dicha enfermedad respecto a los hombres.
* Bmi o índice de masa corporal clasifica a las personas con un índice entre 25-34 como las más propensas a ser diabéticas, además es posible afirmar que entre estos rangos de BMI es posible diagnosticar otra patología como la obesidad, por lo tanto es posible relacionar la obesidad con el diagnóstico o propensión a experimentar la enfermedad objetivo.
* En la base de datos se reconocen otras variables que pueden ser catalogadas como objetivos, no obstante la variable de importancia y de enfoque de estudio es la diabetes, por lo tanto es imprescindible conectar estas patologías (Hipertensión = ‘HighBP’, derrame cerebral='Stroke', Colesterol alto = 'HighChol') con la diabetes. De acuerdo con esta interacción de variables patológicas, la que más correlación posee es ‘HighBP’ con un 75.3 % de posibilidad de sufrir ambas enfermedades seguida de ‘HighChol' con un 77% respecto al total, la variable que menor impacto tuvo entre las tres enfermedades tratadas es derrame cerebral, no se tiene relación entre tener derrame cerebral y sufrir de diabetes, de acuerdo con el análisis exploratorio realizado. Por otro lado se observa que la frecuencia en las categorías que tenían como opción de posibles variables respuestas vemos que el derrame cerebral posee un desbalance muy claro, mientras que los datos de hipertensión y diabetes están balanceados e incluso para el caso de la variable diabetes existe la misma cantidad de datos tanto negativos y positivos.

**c. Definición de arquitectura y de los hiper parámetros de la red neuronal**

**Red neuronal profunda**

Este modelo de red neuronal se construye utilizando la biblioteca de aprendizaje profundo de Keras, y sigue la arquitectura de red neuronal feedforward.

El modelo comienza con una capa densa de 256 nodos con una función de activación "relu". La entrada de esta capa es de 9 dimensiones (input\_dim = 9), que significa que los datos de entrada tienen 9 variables explicativas. Luego, se agregan seis capas densas adicionales, cada una con una función de activación "relu". Estas capas tienen respectivamente 128, 64, 64, 32, 16 y 8 nodos. La capa final tiene un solo nodo con una función de activación "sigmoid", ya que el modelo se utiliza para clasificación binaria.

Después de definir la estructura de la red, se compila el modelo especificando la función de pérdida como "binary\_crossentropy" y el optimizador como "adam". Además, se agregan varias métricas de evaluación, incluyendo accuracy, AUC, recall y precision. Finalmente, se entrena el modelo utilizando los datos de entrenamiento (X\_train, Y\_train) durante 10 épocas (epochs) con un tamaño de lote (batch\_size) de 100. También se especifica un conjunto de datos de validación (X\_test, Y\_test) para monitorear el desempeño del modelo durante el entrenamiento.

**Red neuronal convolucional**

Se hace uso de redes neuronales convolucionales que extraen de forma automática características que luego son interpretadas como una combinación de las diferentes variables de entrada las cuales son usadas para asignar las predicciones asociadas a si una persona tiene propensión a la diabetes o no. La arquitectura de esta CNN consta de varias capas, para esta red se usan tres capas convolucionales con filtros de tamaño 3 y con diferente cantidad de neuronas, la primera consta de 32 neuronas y la entrada es de 9 dimensiones (variables explicativas), luego se tiene otras dos capas convoluciones de 64 y 128 neuronas respectivamente, para cada capa se utiliza la función “relu”.

En este caso se omiten las capas de polling, se implementa una capa totalmente conectada que se utiliza para realizar la tarea final y la capa final tiene un solo nodo con una función de activación "sigmoid", ya que la red neuronal se utiliza para clasificación binaria.

Para la compilación se define como función de pérdida “binary\_crossentropy” y el optimizador “adam”. La métrica de interés es “recall” ya que la prioridad para la red es su capacidad de acertar las personas con propensión a la diabetes, incurrir en falsos positivos no representa consecuencias graves para el problema de negocio planteado.

En el entrenamiento del modelo se usa 10 epochs y un batch\_size de 100, adicionalmente, se utiliza el refinamiento de hiperparametros poniendo a consideración tres optimizadores (“adam”, “sgd”, “rmsprop”).

**d. Evaluación y análisis del modelo**

**Red neuronal profunda**

Para evaluar y mejorar el desempeño y la efectividad del modelo, se considera la métrica Recall como base, ya que el objetivo principal del modelo es identificar correctamente los casos de diabetes.

Se analizan los resultados de la primera red neuronal propuesta y se observa que tanto la función de pérdida como las diferentes métricas establecidas presentan un bajo margen de mejora a lo largo de cada proceso de optimización. A pesar de esto, el modelo presenta un buen ajuste en las pruebas de entrenamiento y evaluación, lo que permite una cantidad significativa de aciertos en la predicción, ya que estos valores están por encima del 70%. Además, se nota un equilibrio en los resultados de ambas pruebas que señala que no hay problemas de overfitting.

Se observa, según la matriz de confusión, que el modelo es capaz de predecir correctamente, de 23,329 datos, 9.041 personas con diabetes y 8,363 normales. A pesar del buen desempeño, se considera que el modelo todavía tiene un margen importante de mejora, por lo que se determina que el modelo tiene underfitting. Por lo tanto, se llevan a cabo varias estrategias para combatir este problema y aumentar la capacidad del modelo para asimilar la complejidad de los datos.

Se han utilizado diversos métodos para mejorar el rendimiento del modelo, pero ninguno ha logrado producir resultados significativamente mejores que el modelo original. Se intentó aumentar el número de capas y neuronas para mejorar la capacidad del modelo para procesar la complejidad de los datos, pero los resultados no fueron efectivos. También se intentó aumentar el número de epoch de 10 a 30, pero el modelo se estancó después de varias iteraciones y no mejoró. Se cambió la función de activación de "relu" a una función tangente para resolver el problema de underfitting, pero el resultado fue similar a los métodos anteriores. Por último, se intentó aumentar el número de neuronas proporcionalmente al tamaño de la base de datos, comenzando desde 2048 neuronas, pero tampoco se obtuvo un resultado satisfactorio.

A partir de los resultados de las diversas estrategias implementadas, se concluye que este tipo de red neuronal no cuenta con la capacidad para seguir mejorando el desempeño del modelo. Por lo tanto, se considera que para mejorar el desempeño es necesario implementar otro modelo con mayor capacidad.

**Red neuronal convolucional**

Se evalúa el comportamiento de la red neuronal para los datos de entrenamiento y los datos de prueba teniendo en cuenta la métrica Recall, los resultados arrojan un comportamiento similar en el transcurso de las épocas evaluadas. Se obtienen valores para entrenamiento entre 78% - 79%, mientras que en los de prueba se encuentra entre 74% - 85%.

En búsqueda de mejorar la tasa de identificación de positivos se realiza un refinamiento de hiper parámetros donde se pone a consideración tres optimizadores, obteniendo que el optimizador “sgd” tiene el mejor desempeño con un Recall de 87,8%.

De esta manera, al evaluar los resultados por medio de la matriz de confusión, el modelo es capaz de predecir correctamente, de 23,329 datos, 9,976 personas con diabetes y 7,402 normales. Se observa una mejora del modelo en los casos positivos de diabete respecto al red neuronal profunda, por lo que se recomienda el uso de este modelo, a pesar de bajar los falsos verdaderos, de acuerdo al problema de negocio la prioridad es mejorar la capacidad del modelo con respecto a los verdaderos positivos (personas con propensión a la diabetes).

**e. Despliegue del modelo**

El modelo predice la diabetes en personas propensas a adquirirlo de acuerdo con algunos hábitos de vida que el paciente posea, esto se hace por medio de categorías acorde con la edad, las personas que estén dentro de un rango no crítico (edades de mediano y bajo impacto) requieren un tipo de atención diferentes a las de alto impacto, donde es preciso hacer la distinción entre los niveles de impacto, por lo tanto esta gestión se realiza almacenando dos bases de datos una que se usa para enviar al correo una serie de elementos informativos con aspectos a mejorar respecto a sus hábitos de vida, por otra parte, si la persona posee edad de alto impacto, la otra base se usa para asignar citas de seguimiento y control para la patología de diabetes. El sistema se retroalimenta con los resultados de los pacientes que se les asignó citas, en caso de ser puesto bajo supervisión médica por causa de diabetes, este sale del sistema de recomendaciones y seguimiento, y pasa a ser tratado por especialistas de esta índole. La actualización o cargue de datos se realiza de manera semanal retroalimentando el sistema.

**f. Conclusiones**

* La cantidad de datos generados en el área de salud hace factible el uso de la analítica de datos, sin embargo es necesario una correcta estructuración de los datos para lograr los resultados esperados, además de que los diseños de solución que se pueden plantear requieren el acompañamiento de profesionales de la salud para hacer robusto y confiable el alcance de los proyectos.
* El modelo de redes neuronales es una herramienta que permite procesar grandes cantidades de datos y asimilar su complejidad, logrando un rendimiento similar al de modelos robustos de clasificación como el gradient boosting. Al final, comparando los resultados obtenidos en cada modelo, se encuentra que el modelo de redes convolucionales es el que presenta el mejor desempeño en cuanto al recall. Por consiguiente, se toma este modelo como base para implementar la solución analítica planteada.

**Bibliografía:**

* **MinSalud. (2022)** *En el Día Mundial de la Diabetes: MinSalud promueve prácticas de vida saludable.* Recuperado de:

<https://minsalud.gov.co/Paginas/En-el-Dia-Mundial-de-la-Diabetes-MinSalud-promueve-praticas-de-vida-saludable.aspx#:~:text=En%20Colombia%20los%20reportes%20de,Antioquia%20y%20Valle%20del%20Cauca>

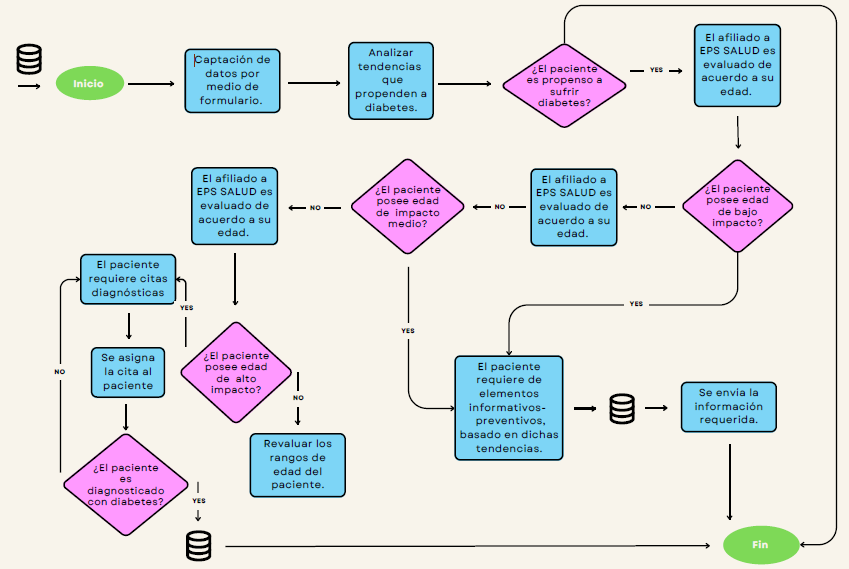
* **Cabrera-Pivaral,C., González-Pérez, G., Vega-López, M., Arias-Merino, E,.** (2004) *Impacto de la educación participativa en el índice de masa corporal y glicemia en individuos obesos con diabetes tipo 2.* Recuperado de:

<https://www.scielo.br/j/csp/a/zJCvhL9PRGnMwm5kkHXCJyq/?lang=es&format=html>

* **NOVA-RAMUDYA, W.**, (2022) *Diabetes Health Indicators.* Recuperado de:

<https://www.kaggle.com/code/bayunova/diabetes-health-indicators/input>

**Anexo:**

****