**Analítica Computacional para la toma de Decisiones**

**Proyecto 2 – Predicción de enfermedades cardíacas**

**Integrantes: Nicolas Ricaurte y Juan Camilo Bello**

**Introducción**

Este informe presenta los resultados del proyecto que busca predecir si una persona padece de enfermedad cardiaca mediante la implementación y evaluación de distintos modelos de aprendizaje automático. Para lograr este objetivo, se aplicaron diversas técnicas de preprocesamiento y transformación de variables a los datos recopilados. Los modelos implementados fueron evaluados utilizando varias métricas de rendimiento, como exactitud, precisión, recall, K2score y matriz de confusión. Se compararon los resultados para identificar los modelos que mejor se ajustaban a los datos y presentaban el mejor rendimiento. Este informe presenta los resultados de la evaluación y comparación de los modelos implementados, así como las conclusiones obtenidas con relación a la solución propuesta.

**Punto de partida**

Como se mencionó en la introducción, en la primera parte del proyecto se ha construido un modelo de redes bayesianas. El objetivo de este modelo es que, a partir de datos como la edad, el sexo, el reconocimiento de dolor en el pecho, los niveles de azúcar, colesterol y presión arterial, que son algunos de los exámenes más comúnmente realizados según el gremio médico, se pueda determinar el nivel de probabilidad de existencia o presencia de enfermedades cardiacas, sin importar su grado. Para ello, se ha tomado la variable de interés como binaria. La aplicación de este modelo podría permitir a los pacientes acceder con mayor prioridad a los servicios de asistencia médica y recibir exámenes más específicos relacionados con enfermedades del corazón. Se ha modificado la relación entre las variables para que tenga la siguiente forma:

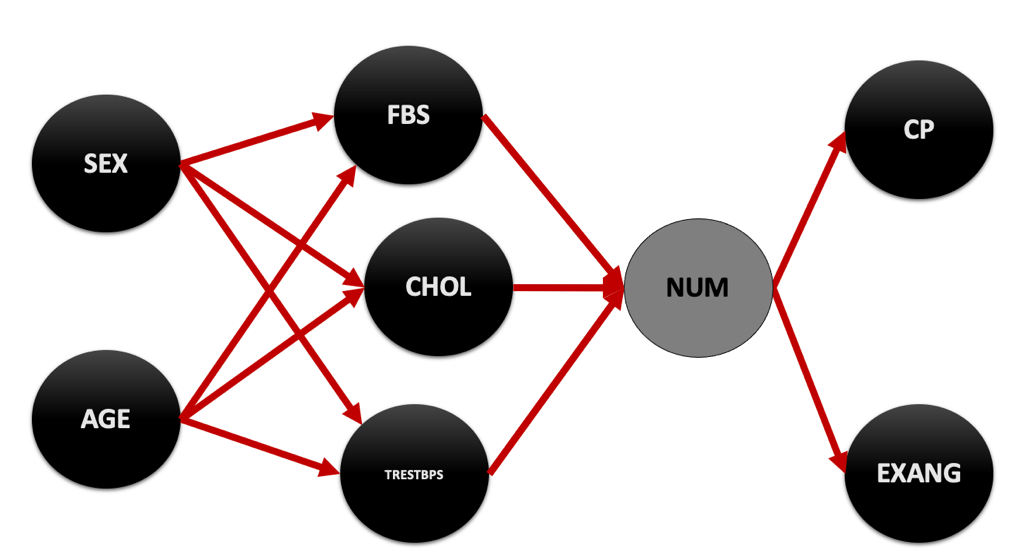


Ilustración 1

Además, se ha tenido en cuenta la importancia de categorizar de forma correcta las variables continúas convirtiendo las variables de edad, colesterol y presión arterial en 10 rangos.

**Resultados entrenamiento**

Se utilizó el conjunto de datos "Processed Cleveland Data Set" del UCI Machine Learning Repository para entrenar y evaluar el rendimiento del modelo bayesiano. Este conjunto de datos contiene información sobre pacientes con problemas cardíacos, y se dividió en dos partes: una para entrenamiento y otra para evaluación. El conjunto de prueba comprendió el 30% de los datos originales, mientras que el conjunto de entrenamiento incluyó el 70% restante.

Es importante destacar que tratando de abordar el problema de valores faltantes en un conjunto de datos dividido en un conjunto de entrenamiento y otro de prueba. Primero se identifican los valores faltantes en el conjunto de prueba en comparación con el conjunto de entrenamiento, y luego se agregan filas al conjunto de entrenamiento para cualquier valor faltante encontrado en el conjunto de prueba. Asegurándonos que el conjunto de entrenamiento tenga todos los valores que se presentan en el conjunto de prueba para que el modelo pueda ser entrenado y evaluado adecuadamente en todos los posibles escenarios.

Se construyó un modelo de red bayesiana con nodos y aristas para representar las relaciones de dependencia entre las variables aleatorias. Se estimaron las distribuciones de probabilidad de las variables utilizando el método de máxima verosimilitud (MLE) y BayesianEstimator, este último considera la incertidumbre en las estimaciones mediante el uso de distribuciones de probabilidad previas. La clase VariableElimination del paquete pgmpy.inference se utilizó para hacer inferencias en el modelo y obtener información sobre las relaciones de dependencia entre las variables, utilizando solo los datos de entrenamiento.

Para evaluar el rendimiento del modelo, se realizaron predicciones sobre los datos de prueba utilizando el modelo bayesiano previamente creado. Se utilizó una matriz de confusión, una herramienta comúnmente utilizada para evaluar el rendimiento de un modelo de clasificación, en este caso, sobre un conjunto de datos de prueba. La matriz de confusión compara las etiquetas de clase reales de un conjunto de datos con las etiquetas de clase predichas por el modelo y muestra la cantidad de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

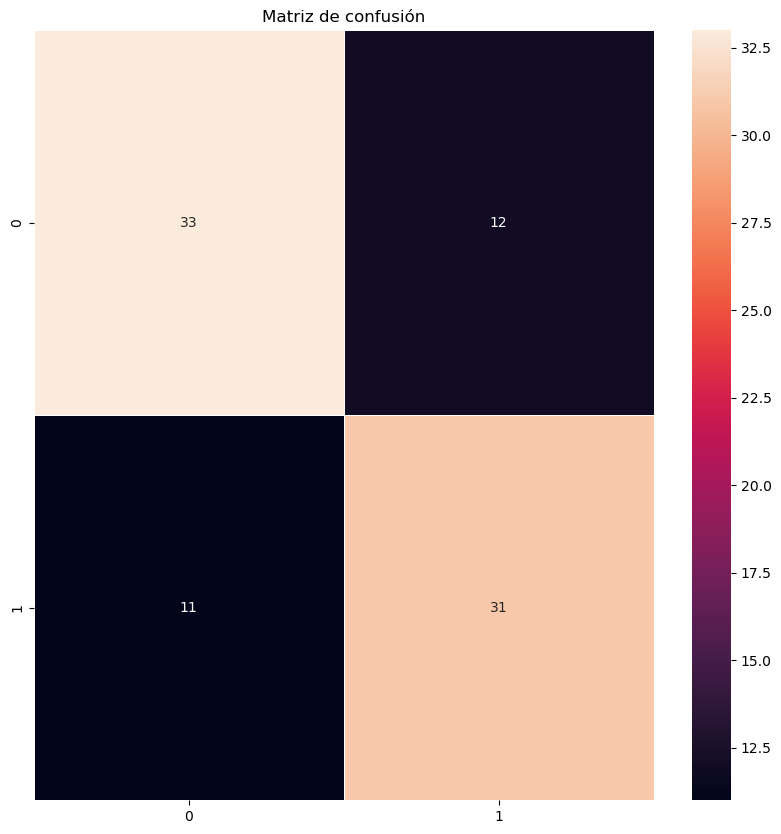


Ilustración 2

La exactitud del modelo, es decir, su capacidad para predecir correctamente los verdaderos positivos y verdaderos negativos se mide a través de la exactitud o accuracy, que en este caso es del 73.56%. Esto indica que el modelo clasifica correctamente alrededor del 73.56% de los casos.

La precisión del modelo, que mide la proporción de predicciones positivas que son verdaderamente positivas, es de 72.09%. Esto sugiere que cuando el modelo predice que un paciente tiene la enfermedad, hay un 72.09% de probabilidades de que la predicción sea correcta.

El recall, también conocido como sensibilidad, que mide la proporción de verdaderos positivos que se identifican correctamente, es de 73.8%. Esto indica que el modelo clasifica correctamente el 73.8% de los pacientes que realmente tienen la enfermedad.

La especificidad, que mide la proporción de verdaderos negativos que se identifican correctamente, es de 73.33%. Esto significa que el modelo clasifica correctamente el 73.33% de los pacientes que realmente no tienen la enfermedad.

Los resultados muestran que el modelo tiene una precisión decente y una buena tasa de recall, lo que indica que el modelo es capaz de identificar correctamente la mayoría de los casos positivos. Además, la especificidad es alta, lo que significa que el modelo es capaz de identificar correctamente la mayoría de los casos negativos. Sin embargo, la exactitud del modelo sugiere que el modelo podría estar clasificando incorrectamente algunos casos. Por lo tanto, es importante tener en cuenta que, aunque los resultados son buenos, se deben interpretar con precaución.

De forma adicional utiliza el K2Score para evaluar el ajuste del modelo a los datos. El K2Score es una medida de bondad de ajuste que se utiliza en el aprendizaje de redes bayesianas. Cuanto menor sea el valor del puntaje K2, mejor será el ajuste del modelo a los datos. En este caso, el puntaje K2 es -3069.783, lo que sugiere que el modelo se ajusta bien a los datos y es mejor que un modelo que se ajuste de manera similar pero que no tenga en cuenta las relaciones entre las variables aleatorias. Sin embargo, es importante tener en cuenta que el puntaje K2 solo proporciona una medida de la bondad de ajuste de un modelo y no es una medida absoluta de la calidad del modelo.

**Nuevo modelo**

Para generar un nuevo modelo utilizando métodos de aprendizaje de estructura, se consideraron varios enfoques, incluyendo los métodos basados en restricciones, basados en puntajes y los híbridos. Para este trabajo, se escogió el método de búsqueda de escalada de HillClimb, que es una técnica comúnmente utilizada en la construcción de redes bayesianas.

El método HillClimb comienza con un grafo vacío y se construye iterativamente, añadiendo o eliminando arcos hasta encontrar la mejor estructura que maximiza algún criterio de ajuste, como el puntaje K2. Es importante destacar que durante la implementación del método de HillClimb se puede especificar una lista negra de arcos que se deben evitar durante la construcción del grafo para excluir relaciones conocidas entre variables que se sabe que no existen en el dominio del problema.

Asimismo, se puede establecer un límite en el número máximo de padres para cada nodo para controlar la complejidad del modelo y evitar el sobreajuste. Se buscaron iterar sobre estos parámetros para encontrar el mejor modelo que se ajuste a los datos.

Luego de aplicar el método de HillClimb, se obtuvo un modelo que se ajusta adecuadamente a los datos y se espera que pueda ser utilizado para hacer predicciones y obtener información valiosa. Es importante mencionar que la selección del método de aprendizaje de estructura y la iteración de los parámetros se realizaron con el objetivo de obtener el mejor modelo posible para el conjunto de datos en cuestión, resultando en el modelo de la ilustración 3.

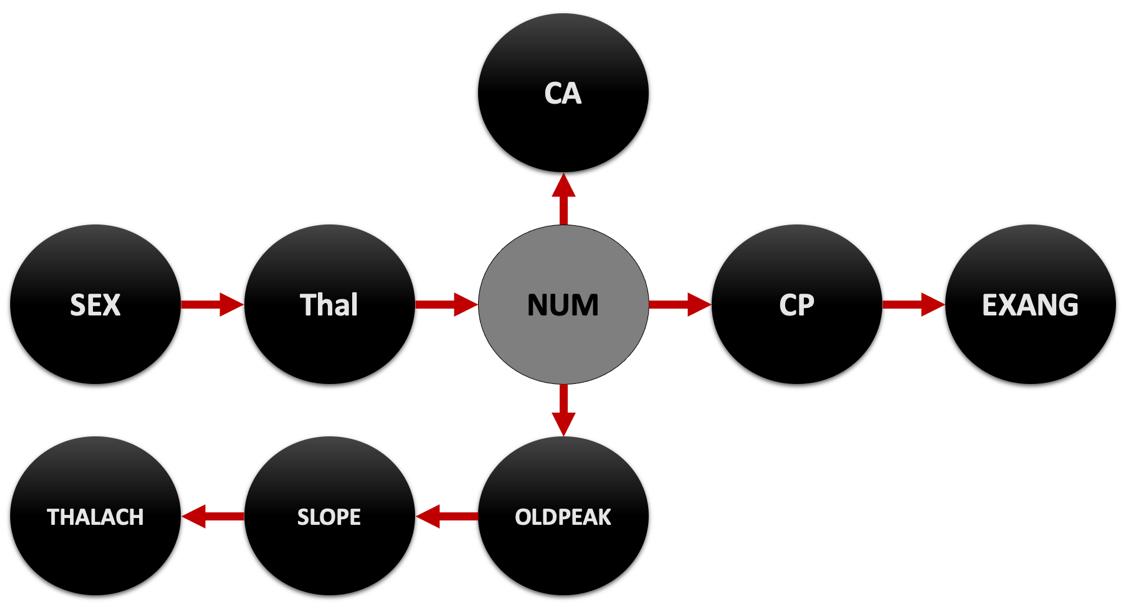


Ilustración 3

Existen diversas variables que influyen en la salud del corazón y que están interconectadas entre sí. Por ejemplo, la variable raíz es "sexo", que influye en la variable thal. La variable thal representa los resultados de la gammagrafía cardíaca con talio y se asume que está influenciada por el sexo del paciente. La variable cp, que representa el dolor torácico, influye en la variable exang, que indica si el paciente sufre de angina inducida por ejercicio. La variable oldpeak, que representa la depresión del segmento ST en un electrocardiograma, influye en la variable slope, que indica la pendiente del segmento ST en el mismo electrocardiograma. La variable slope influye en la variable thalach, que representa la frecuencia cardíaca máxima alcanzada durante una prueba de esfuerzo. La variable thal influye en la variable num, que representa la presencia de una enfermedad cardíaca (0 = no, 1 = sí). La variable num influye en las variables ca, cp y oldpeak, donde ca que representan el número de arterias coloreadas por fluoroscopia.

Analizando algunas relaciones creadas la influencia de sex en thal puede representar diferencias en la distribución de enfermedades cardíacas entre hombres y mujeres, mientras que la influencia de cp en exang puede indicar que ciertos tipos de dolor torácico están más relacionados con la angina inducida por ejercicio. La relación entre oldpeak y slope puede indicar que una mayor depresión en el segmento ST está relacionada con una pendiente más pronunciada en el segmento ST.

Por otro lado, fue necesario crear otro modelo nuevo para comparar 3 modelos en total, se decidió realizar una aproximación nueva a la creación del modelo teniendo en cuenta los valores deseados de cada variable, de acuerdo con la investigación realizada en la primera fase del proyecto buscamos asignando valores de 0 cuando la variable se encuentra entre los valores normales y 1 de lo contrario. Posterior a esto se realizó el mismo método de aprendizaje de estructura, con el fin de comparar el efecto de esta categorización de las variables. En la ilustración 4 se puede observar la red bayesiana resultante.

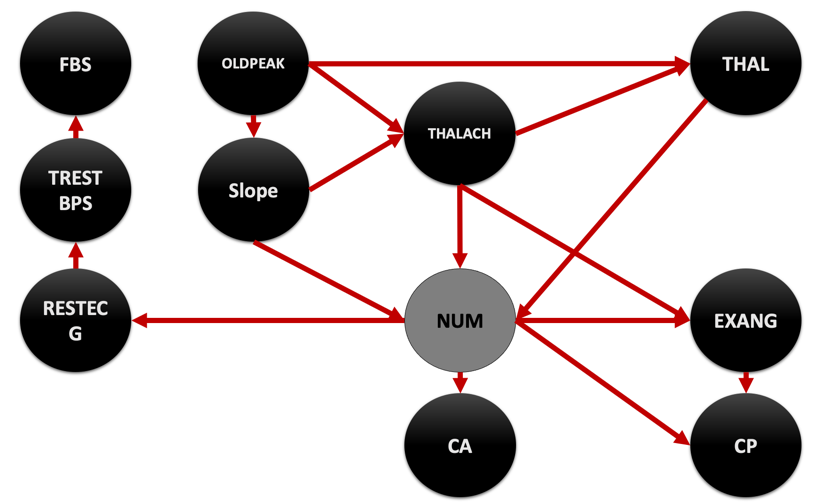


Ilustración 3

**Evaluación y comparación de modelos**

Al analizar los datos de prueba de los tres modelos, se compararon los índices de desempeño de la matriz de confusión, incluyendo la precisión, exactitud, sensibilidad y especificidad, así como el puntaje K2. La gráfica 1 proporciona una visualización clara de estas comparaciones.



Gráfica 1

Con base a los resultados obtenidos, se puede concluir que el Modelo 2 tuvo el mejor desempeño en términos de precisión, exactitud y especificidad, superando al Modelo 1 y Modelo 3. Estos hallazgos se corresponden también con el K2Score, el cual arrojó resultados de -3069.78 para el modelo 1, -4900.5 para el modelo 2 y -4215.46 para el modelo 3. Por lo tanto, se puede afirmar que el modelo 2 se ajusta mejor a los datos y muestra un mejor desempeño en general.

Al comparar el desempeño de los modelos creados mediante el método de HillClimb con el modelo basado en expertos y literatura, se observa una clara diferencia a favor de los primeros. Los modelos generados por HillClimb son más objetivos y consideran una mayor cantidad de variables, lo que les permite capturar de manera más precisa y completa la complejidad del problema en cuestión.

Por otro lado, los modelos basados en expertos y literatura pueden estar sesgados o limitados en su capacidad para abarcar toda la complejidad del problema. Esto se debe a que los expertos pueden tener limitaciones en su conocimiento o experiencia, lo que podría afectar la precisión y generalización del modelo.

Es importante destacar que, para la generación de los modelos 2 y 3, se necesitó utilizar conocimiento previo de la literatura para establecer qué arcos no debían formar parte de la red bayesiana. Esto se hizo para evitar relaciones sin sentido entre las variables y asegurar un mejor desempeño. También, teniendo en cuenta que el modelo planteado inicialmente está pensado para poder hacerse con exámenes básico, los cuales el usuario pueda hacer en su casa, el trade-off para el desempeño del modelo es prometedor, ya que, puede dar índices a los usuarios de una enfermedad médica con probabilidades aceptables.

En conclusión, la aplicación del método de HillClimb en la creación de modelos es una estrategia efectiva para obtener modelos más precisos y generalizables. Combinar esta técnica con el conocimiento de expertos y la literatura puede maximizar el desempeño del modelo en la resolución de problemas complejos. Como también, que siempre es importante analizar el usuario final de estas herramientas para evaluar si las variables de desempeño dan resultados prometedores o debería ajustarse el modelo.