Alejandra Erazo – 201821038
Juliana Mendoza – 201822235
Proyecto 2

Sistema de Predicción de Enfermedades Cardíacas

Para el desarrollo de la herramienta de apoyo para la predicción de enfermedades cardiacas, se seleccionaron como usuarios finales los médicos que quieran incorporar un apoyo de analítica de datos en el proceso de evaluación de pacientes y la toma de decisiones asociada (solicitud de exámenes, chequeos y otros procedimientos). Se busca que esta herramienta sirva de apoyo para determinar el tratamiento de los pacientes que pueden sufrir de enfermedades cardiacas. Para esto, se seleccionaron 7 variables luego de realizar un análisis exploratorio y búsqueda en literatura. Estas variables corresponden a la edad del paciente (AGE), el valor de colesterol total en sangre en mg/dL (CHOL), el nivel de glucosa en sangre en ayunas (FBS), el tipo de talasemia que presenta el paciente (THAL), la angina inducida por el ejercicio (EXANG), el valor de la depresión del segmento ST en el electrocardiograma (OLDPEAK) y la presencia de una enfermedad cardíaca (HD), como variable de respuesta. Para esta última, se tenían valores de predicción de enfermedad cardiaca de 0, 1, 2, 3 y 4, sin embargo, se categorizaron en 3 grupos correspondientes a pacientes sin enfermedad cardiaca (0), pacientes con enfermedad cardiaca leve (1 y 2) y pacientes con enfermedad cardiaca severa (3 y 4).

Ahora, antes de realizar los modelos bayesianos, se generaron algunas visualizaciones que permiten obtener información relevante acerca de los datos. Estas se pueden observar en las *Figuras 1, 2,* y *3.* En primer lugar, la *Figura 1* permite evidencia el comportamiento de la variable que mide el colesterol total en sangre, en función de la edad de los pacientes. A partir de esta gráfica es posible afirmar que a medida que aumenta la edad, los valores de colesterol tienden a aumentar. Esto, como consecuencia de la disminución de actividad física en las personas mayores, lo que genera una acumulación de triglicéridos, una diminución en la concentración de colesterol HDL (bueno) y un aumento en la concentración de colesterol LDL (malo) [1].

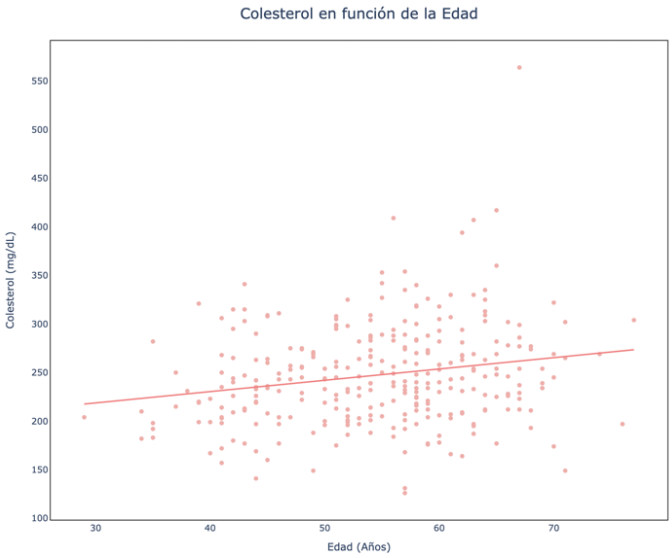

Figura 1. Gráfica de dispersión del colesterol en función de la edad.

En segundo lugar, la *Figura 2* evidencia los efectos de presentar un valor de glucosa en sangre mayor a 120 mg/dL (FBS =1) y angina inducida por el ejercicio (EXANG = 1), en el desarrollo de enfermedades cardíacas. Durante la búsqueda de literatura fue posible determinar que las personas que tienen valores altos de glucosa sérica en ayunas cuentan con una mayor predisposición a tener niveles elevados de colesterol y, por ende, una mayor probabilidad de sufrir enfermedades cardíacas [2]. Sin embargo, luego de analizar los datos de este estudio, se pudo evidenciar lo contario. En la *Figura 2* se puede observar la mayor cantidad de pacientes que presentan niveles de glucosa altos

son aquellos que no presentan ningún tipo de enfermedad cardíaca, mientras que son pocos los pacientes que cuentan con glucosa alta y enfermedad cardíaca severa. Por otra parte, en cuento a la angina inducida por el ejercicio, se puede que ver que esta se presenta en mayor medida en pacientes que padecen una enfermedad cardíaca leve o severa.

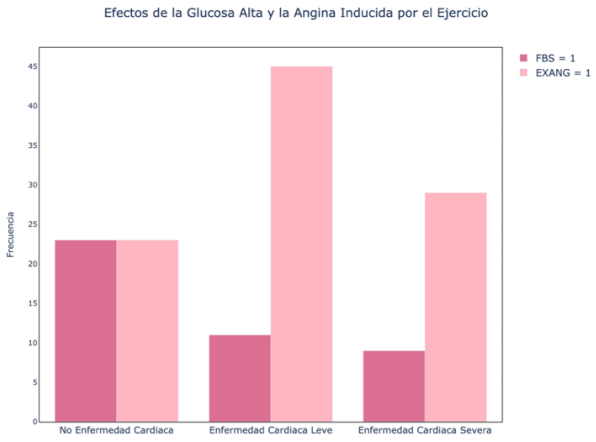

Figura 2. Gráfica de barras del efecto de la glucosa alta y la angina inducida por el ejercicio en las enfermedades cardíacas.

En tercer lugar, se realizó una gráfica que expone el nivel de gravedad de enfermedad cardíaca asociado a cada tipo de talasemia, como se muestra en la *Figura 3.* De esta forma, es posible ver que las personas que no presentan ninguna enfermedad cardíaca tienden a tener talasemia tipo normal, mientras que en las personas que presentan enfermedad cardíaca leve o severa predomina el tipo de talasemia con defecto reversible.

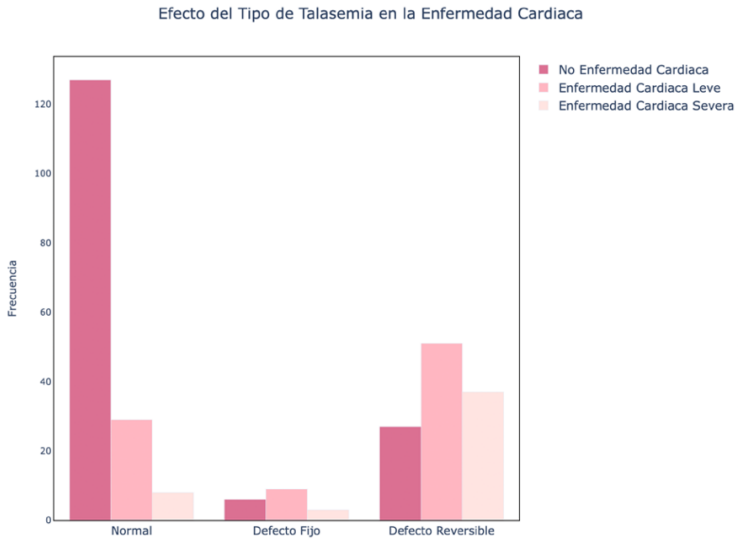

Figura 3. Gráfica de barras del efecto del tipo de talasemia en las enfermedades cardíacas.

Una vez analizados los datos, se continuó con la elaboración de dos modelos bayesianos, el modelo presentado en el Proyecto 1 y un modelo bayesiano que emplea métodos de aprendizaje de estructura. Para esto, en primer lugar, se llevó a cabo la importación de la base de datos dada y la eliminación de datos faltantes, con el propósito de evitar errores posteriores. Con esto, se obtuvo un valor final de 297 pacientes. Posteriormente, se realizó la discretización de las variables continuas utilizadas en el grafo, las cuales corresponden a colesterol (CHOL), edad (AGE) y depresión del segmento ST (OLDPEAK). Luego, se definieron los datos de entrenamiento, correspondientes a los primeros 250 pacientes y, de igual manera, los datos de validación, correspondientes a los últimos 47 pacientes de

la base de datos. Esto con el propósito de entrenar los modelos con datos diferentes a los que se utilizan para validar su funcionalidad.

Así bien, el primer modelo que se realizó fue aquel presentado en el proyecto 1. El grafo correspondiente a este modelo se puede observar en la *Figura 4.* Una vez definido el grafo, se realizó el modelo entrenado en Python. Se elaboró la red Bayesiana y se estimaron las distribuciones de probabilidad condicional (CPDs) de las variables asociadas. Lo anterior, se hizo utilizando el estimador de máxima verosimilitud.

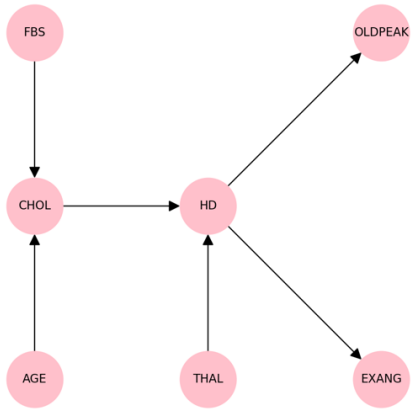

Figura 4. Grafo definido para la red bayesiana del Proyecto 1.

De esta forma, se obtuvo la CPD que se expone en la *Tabla 1,* correspondiente a la probabilidad de sufrir enfermedad cardíaca, teniendo como nodos padres el colesterol en sangre y el tipo de talasemia.

Tabla 1. P(HD|CHOL, THAL).

HD	CHOL	0			1			2		
	THAL	3	6	7	3	6	7	3	6	7
0		0.8	0.4	0.33	0.82	0.5	0.28	0.74	0	0.18
1		0.16	0.4	0.33	0.16	0.33	0.39	0.19	0.83	0.51
3		0.04	0.2	0.33	0.02	0.17	0.33	0.07	0.17	0.31

Posteriormente, se llevó a cabo la validación del modelo utilizando datos conocidos de 47 pacientes. Se implementó el modelo con estos datos y se realizó la clasificación correspondiente del nivel de enfermedad cardíaca, considerando la probabilidad mayor dada por el algoritmo. Con esto, se calculó la matriz de confusión que se presenta en la *Figura 5,* que permite observar el rendimiento del sistema de predicción.

A partir de esta matriz fue posible determinar el porcentaje de falsos y verdaderos positivos y negativos, como se muestra en la *Tabla 2.* Con esto, se calcularon los porcentajes de precisión y sensibilidad generales del modelo, obteniendo valores de 63% para ambas métricas.

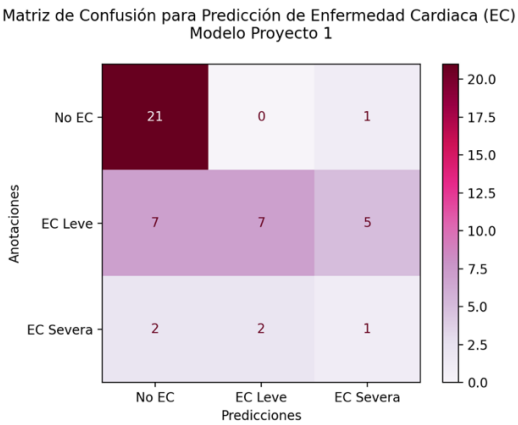

Figura 5. Matriz de Confusión del modelo realizado en el Proyecto 1.

Tabla 2. Porcentaje de falsos y verdaderos positivos y negativos del modelo del Proyecto 1.

	No EC	EC Leve	EC Severa
VP	45.7%	15.2%	2.2%
FP	2.2%	26.1%	8.7%
FN	19.6%	4.3%	13.0%
VN	32.6%	54.3%	76.1%
Precisión	**95.5%**	**36.8%**	**20.0%**
Sensibilidad	**70.0%**	**77.8%**	**14.3%**

Se puede notar que el sistema presenta un rendimiento alto a la hora de predecir que el paciente no presenta ningún tipo de enfermedad cardíaca, sin embargo, al momento de tener que predecir si la enfermedad cardíaca es leve o modera, el sistema incurre en bastantes errores. Por esto, se decidió evaluar el sistema de predicción con base en la detección de enfermedad o no, es decir, de forma binaria. Así, la *Figura 6* muestra la matriz de confusión obtenida en este caso y la *Tabla 3* expone los resultados del modelo.

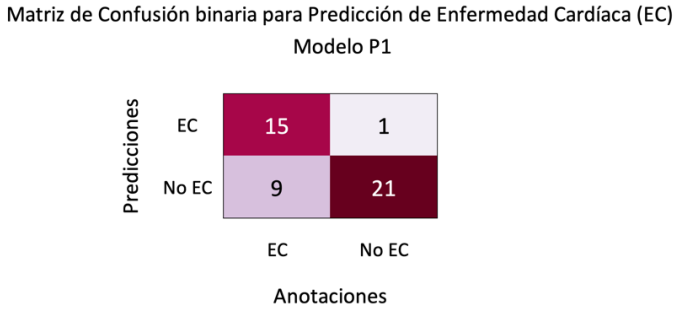

Figura 6. Matriz de Confusión binaria del modelo realizado en el Proyecto 1.

Tabla 3. Porcentaje de falsos y verdaderos positivos y negativos del modelo binario del Proyecto 1.

VP	15%
FP	1%
FN	9%
VN	21%
Precisión	**93.8%**
Sensibilidad	**62.5%**

Ahora bien, se realizó un segundo modelo bayesiano empleando métodos de aprendizaje de estructura. En este caso, se utilizó el método de búsqueda Hill Climbing y el puntaje K2. Es importante mencionar que para el parámetro *max_indegree* se empleó un valor de 8, el cual indica el número máximo de padres que puede tener cada variable en el modelo resultante. Por otro lado, se emplearon 10000 iteraciones, con el fin de obtener un modelo con mejor ajuste. Con esto en mente, la *Figura 7* permite observar el grafo generado por este método.

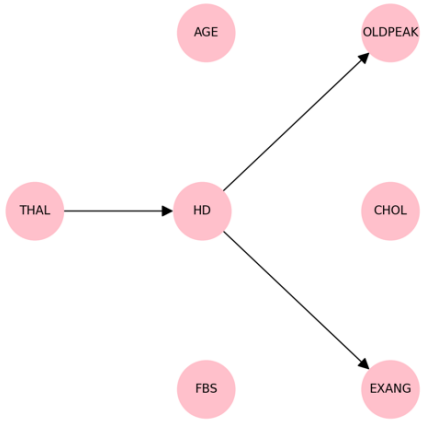

Figura 7. Grafo definido para la red bayesiana del modelo por Puntaje K2.

Para seleccionar el modelo final entrenado, se calcularon los puntajes K2 de cada uno de los modelos realizados, al variar los parámetros principales de la función. Esto, dado que un puntaje K2 alto indica que el modelo se ajusta de mejor manera a los datos dados. Así, el modelo seleccionado presenta un puntaje K2 de -1397.084. De esta forma, se obtuvo la CPD que se expone en la *Tabla 4,* correspondiente a la probabilidad de sufrir enfermedad cardíaca, teniendo como nodo padre únicamente el tipo de talasemia.

Tabla 4. P(HD| THAL).

HD	THAL	3	6	7
0		0.78	0.45	0.25
1		0.18	0.45	0.40
3		0.04	0.09	0.37

Una vez entrenado el modelo, se prosiguió a realizar la evaluación con los datos de validación establecidos previamente, de la misma manera como se realizó con el modelo del Proyecto 1. A partir de esto se obtuvo la matriz de confusión multiclase que se presenta en la *Figura 8*.

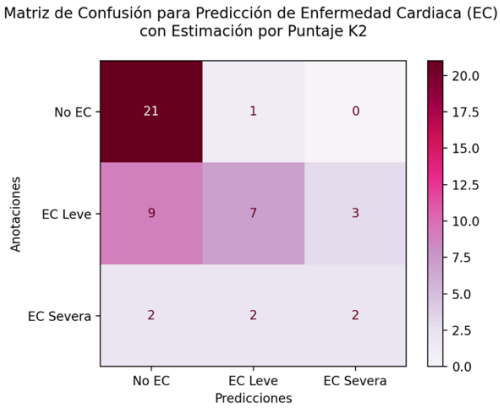

Figura 8. Matriz de Confusión del modelo por Puntaje K2.

A partir de esta matriz fue posible determinar el porcentaje de falsos y verdaderos positivos y negativos, como se muestra en la *Tabla 4.* Con esto, se calcularon los porcentajes de precisión y sensibilidad del modelo, obteniendo valores de 64% para las dos métricas. Como se puede observar, en ambos modelos la clase correspondiente a No Enfermedad Cardíaca presenta la mayor precisión, mayor al 90%, lo que indica que el sistema es muy bueno a la hora de predecir que el paciente no tiene enfermedad cardíaca. Sin embargo, los valores de precisión para las clases que indican que sí se tiene una enfermedad cardíaca son muy bajos. Por esta razón, se decidió evaluar nuevamente el sistema de predicción con base en la detección de enfermedad o no, es decir, de forma binaria. Así, la *Figura 9* muestra la matriz de confusión obtenida en este caso y la *Tabla 5* expone los resultados del modelo binario.

Tabla 4. Porcentaje de falsos y verdaderos positivos y negativos del modelo por Puntaje K2.

	No EC	EC Leve	EC Severa
VP	44.7%	14.9%	4.3%
FP	2.1%	25.5%	8.5%
FN	23.4%	6.4%	6.4%
VN	29.8%	53.2%	80.9%
Precisión	**95.5%**	**36.8%**	**33.3%**
Sensibilidad	**65.6%**	**70.0%**	**40.0%**

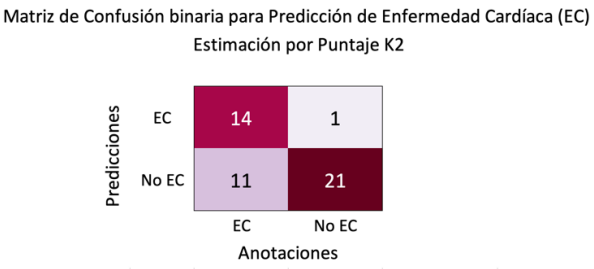

Figura 9. Matriz de Confusión binaria del modelo por Puntaje K2.

Tabla 5. Porcentaje de falsos y verdaderos positivos y negativos del modelo binario por Puntaje K2.

VP	14%
FP	1%
FN	11%
VN	21%
Precisión	**93.3%**
Sensibilidad	**56.0%**

Se obtuvieron valores de precisión y sensibilidad muy similares en ambos modelos, siendo aquellos del modelo del Proyecto 1 un poco más altos. En otras palabras, el modelo del Proyecto 1 es un poco más eficiente a la hora de predecir cuántos de los casos predichos correctamente resultaron verdaderamente positivos (precisión) y cuántos de los casos positivos reales se pudieron predecir correctamente con el modelo (sensibilidad). Esto puede ser consecuencia de la reducción en las variables que emplea el modelo por Puntaje K2, pues no incluye información que puede llegar a ser útil para predecir la enfermedad cardíaca.

Finalmente, se realizó una comparación de los modelos presentados previamente con el modelo generado por otros integrantes del curso. Este modelo de comparación fue elaborado mediante una red bayesiana que emplea métodos de aprendizaje de estructura, específicamente, el método de búsqueda Hill Climbing y el puntaje K2. El modelo entrenado generado por este grupo presenta la estructura que se muestra en la *Figura 10.* Es importante resaltar que este modelo fue generado con un mayor número de variables, con respecto a los modelos anteriores.

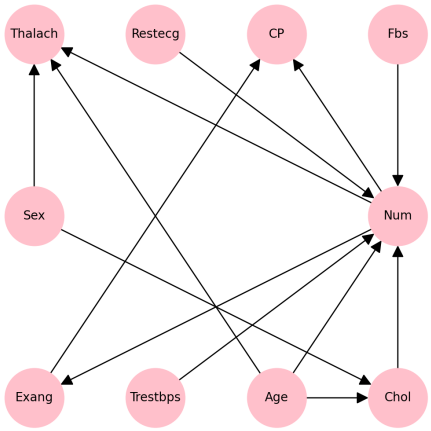

Figura 10. Grafo definido para la red bayesiana del modelo del otro grupo.

Una vez entrenado el modelo, se prosiguió a realizar la evaluación con los datos de validación establecidos previamente, de la misma manera como se realizó con los modelos anteriores. A partir de esto se obtuvo la matriz de confusión que se presenta en la *Figura 11*.

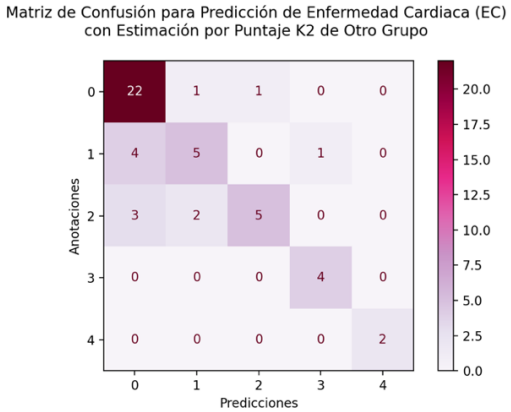

Figura 11. Matriz de Confusión del modelo de otro grupo

En este caso se puede observar que la variable de respuesta para la enfermedad cardíaca se encuentra discretizada en 5 clases, a diferencia de los modelos anteriores que presentaban 3 clases. Por esto, se decidió evaluar el sistema de predicción con base en la detección de enfermedad o no, es decir, de forma binaria. Así, la *Figura 12* muestra la matriz de confusión obtenida en este caso y la *Tabla 6* expone los resultados del modelo binario.

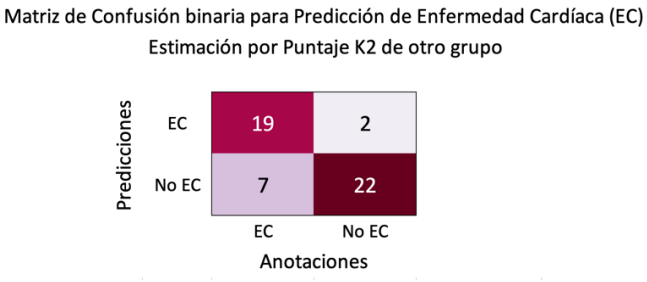

Figura 12. Matriz de Confusión binaria del modelo por Puntaje K2 de otro grupo.

Tabla 6. Porcentaje de falsos y verdaderos positivos y negativos del modelo binario por Puntaje K2 de otro grupo.

VP	19%
FP	2%
FN	7%
VN	22%
Precisión	**90.5%**
Sensibilidad	**73.1%**

La *Tabla 7* muestra un resumen de las métricas obtenidas para cada uno de los modelos estudiados. A partir de esta se puede concluir que le mejor modelo es el realizado por otro grupo, basado en puntaje K2. Esto, dado que el grupo definió desde un inicio el grafo del modelo y, posteriormente, realizaron en entrenamiento. Por el contrario, el modelo que obtuvimos con este método fue diseñado completamente por el algoritmo, lo que causó una reducción significativa en el número de variables empleadas y, por ende, en la eficacia del modelo. Ahora, el modelo realizado en el Proyecto 1 presenta una precisión más alta, con respecto al modelo de comparación, pero una sensibilidad más baja. De esta forma, al calcular el valor de F1, el cual representa una media armónica entre la precisión y la sensibilidad, se puede ver que el valor más alto lo obtuvo el modelo de comparación. Se puede concluir que este modelo es más completo y más efectivo y que los errores en los modelos del Proyecto 1 y por Puntaje K2 pueden ser consecuencia de una selección inadecuada de variables predictoras.

Tabla 7. Métricas de evaluación para cada modelo realizado.

Modelo	Precisión	Sensibilidad	F1 Score
P1	94%	63%	75%
K2	93%	65%	70%
Comparación	90%	73%	81%

Link del Repositorio:
https://github.com/jmendoza1705/P2_Erazo_Mendoza

Referencias

[1] H. Roshanmehr, M. Rostami, M. Seyedtabib, and N. Kamyar, "Determination of Fasting Blood Sugar and Cholesterol Levels in Gotvand City," *Int. J. Med. Lab.*, vol. 7, no. 4, pp. 267–279, 2020, doi: 10.18502/ijml.v7i4.4797.

[2] H. Roshanmehr, M. Rostami, M. Seyedtabib, and N. Kamyar, "Determination of Fasting Blood Sugar and Cholesterol Levels in Gotvand City," *Int. J. Med. Lab.*, vol. 7, no. 4, pp. 267–279, 2020, doi: 10.18502/ijml.v7i4.4797.