Facultad de Ciencias - Grado en Estadística

Trabajo de Fin de Grado

VNiVERSiDAD
D SALAMANCA

Redes bayesianas para comprender la complejidad del genoma en Leucemia Mieloide Aguda

Bayesian networks for understanding the complexity of the genome in Acute Myeloid Leukaemia

Autora:

Carla Fernández Martínez

Tutores:

Dr. José Manuel Sánchez Santos

Javier Martínez Elicegui

Julio, 2024

Facultad de Ciencias - Grado en Estadística

Trabajo de Fin de Grado

VNiVERSiDAD
D SALAMANCA

Redes bayesianas para comprender la complejidad del genoma en Leucemia Mieloide Aguda

Bayesian networks for understanding the complexity of the genome in Acute Myeloid Leukaemia

Autora:

Firmado por FERNANDEZ MARTINEZ CARLA - ***7928** el día 03/07/2024 con un certificado emitido por AC FNMT Usuarios

Carla Fernández Martínez

Tutores:

Dr. José Manuel Sánchez Santos

Javier Martínez Elicegui

Julio, 2024

VNiVERSiDAD D SALAMANCA
CAMPUS DE EXCELENCIA INTERNACIONAL

Facvltad D Ciencias
VNiVERSiDAD D SALAMANCA

Certificado del/los tutores TFG

D. José Manuel Sánchez Santos, profesor del Departamento de Estadística de la Universidad de Salamanca y D. Javier Martínez Elicegui, investigador del Centro de Investigación del Cáncer (CiC-IBMCC, USAL/CSIC),

HACEN CONSTAR:

Que el trabajo titulado "*Redes bayesianas para comprender la complejidad del genoma en Leucemia Mieloide Aguda*", que se presenta, ha sido realizado por Carla Fernández Martínez, con DNI 71479288ª, y constituye la memoria del trabajo realizado para la superación de la asignatura Trabajo de Fin de Grado en Estadística en esta Universidad.

Salamanca, 3 de Julio de 2023

SANCHEZ SANTOS, JOSE MANUEL (FIRMA)
2024.07.02 18:08:49 +02'00'

Fdo.: José Manuel Sánchez Santos

MARTINEZ ELICEGUI FRANCISCO JAVIER - 17437279J
Digitally signed by MARTINEZ ELICEGUI FRANCISCO JAVIER - 17437279J
Date: 2024.07.02 19:54:13 +02'00'

Fdo.: Javier Martínez Elicegui

Índice

1. Introducción

La Leucemia Mieloide Aguda (AML, por sus siglas en inglés) es un trastorno hematológico maligno que conlleva un desafío clínico por su complejidad genómica y la variabilidad en las respuestas al tratamiento. Está caracterizado por la proliferación no controlada de células mieloides inmaduras en la médula ósea y la sangre periférica. Las interacciones entre genes, proteínas y otros elementos moleculares pueden ser difíciles de discernir con métodos de análisis tradicionales. Sin embargo, la correcta comprensión de la complejidad genómica implícita en la AML es fundamental para mejorar las técnicas de diagnóstico, pronóstico y tratamiento de dicha enfermedad.

Gracias a los estudios de secuenciación masiva se han podido analizar gran cantidad de genes y sus mutaciones, y es necesario distinguir aquellas que forman parte del contexto de la enfermedad y que son clave para controlar los procesos celulares de la patología (Lagunas-Rangel, 2016).

El campo de la bioinformática ha avanzado exponencialmente en los últimos años, lo que ha dado lugar a herramientas poderosas para comprender la base genética de enfermedades complejas como la AML. Entre estas herramientas, las redes bayesianas (BN, por sus siglas en inglés: “bayesian networks”) destacan como una técnica de inteligencia artificial prominente para modelar, en este caso, relaciones probabilísticas entre variables moleculares y fenotípicas, comprendiendo mejor la dinámica genética de enfermedades. Al integrar información previa y evidencia experimental, las BN pueden inferir redes de interacción genética que subyacen a la AML, y proporcionan una herramienta gráfica orientativa en la exploración de las diferentes relaciones (Angelopoulos et al., 2022).

1.1. Leucemia Mieloide Aguda (AML)

La leucemia aguda se refiere a un tipo de cáncer de progresión rápida que se origina en el tejido responsable de la producción de células sanguíneas, como la médula ósea, desencadenando una sobreproducción de glóbulos blancos que entran en el torrente sanguíneo.

Concretamente, la leucemia mieloide aguda es un cáncer de sangre y de médula ósea. Este es el tipo más común de leucemia aguda en adultos y si no se trata, tiende a empeorar rápidamente. También se llama leucemia mielógena aguda o leucemia no linfocítica aguda (*Tratamiento de la leucemia mieloide aguda - NCI*, 2024). La AML representa del 15 al 20% de las leucemias agudas en niños y el 80% en adultos. Es la forma predominante de leucemia en el período neonatal y en adultos, aunque representa solo una pequeña proporción de casos en la infancia y la adolescencia (Lagunas-Rangel, 2016).

En condiciones normales, la médula ósea genera células madre sanguíneas, que son células inmaduras. Con el paso del tiempo, estas células se desarrollan y se convierten en células sanguíneas maduras. Una de estas células madre sanguíneas puede dar lugar a una célula madre mieloide o a una célula madre linfoide, la cual se convierte en glóbulo blanco. Mientras que, una célula madre mieloide se transforma en glóbulos rojos, granulocitos o plaquetas, que son tipos de células sanguíneas maduras. Sin embargo, en AML, las células madre mieloides normalmente se transforman en un tipo de glóbulo blanco inmaduro llamado mieloblasto. Los mieloblastos son glóbulos blancos anormales que no se convierten en glóbulos blancos sanos. Los glóbulos blancos, glóbulos rojos o plaquetas anormales, también tienen el nombre de células o blastocitos leucémicos. Cuando las células leucémicas proliferan en exceso en la médula ósea y la sangre, pueden inferir en la producción normal de

glóbulos blancos, glóbulos rojos y plaquetas saludables. Este suceso puede conducir a un desequilibrio en la composición sanguínea, lo que aumenta el riesgo de padecer infecciones, anemia y hemorragias con mayor facilidad.

La AML no se restringe únicamente a la médula ósea y la sangre periférica, ya que puede afectar diversos órganos debido a la infiltración de células leucémicas o a complicaciones metabólicas asociadas. Aunque la piel y los huesos son los sitios comúnmente más afectados, también pueden formarse sarcomas granulocíticos, que son tumores leucémicos que se desarrollan fuera de la médula ósea y la sangre, en cualquier órgano. La dificultad respiratoria en pacientes con AML generalmente se atribuye a infecciones, pero aquellos con un elevado recuento de células blastoides circulantes pueden experimentar disnea grave e hipoxemia en los capilares pulmonares (Devine & Larson, 1994).

La base de la leucemogénesis, que es el proceso mediante el cual se desarrolla la leucemia, reside en el daño genético no letal. En el caso de la AML, hay varios factores que contribuyen a su desarrollo, no obstante, los más importantes son la exposición a radiaciones ionizantes, altas concentraciones de benceno, agentes quimioterapéuticos e inhalación crónica de humo de cigarro (Lagunas-Rangel, 2016). Estos riesgos tienen la capacidad de producir daños en el ADN, principalmente por daño oxidativo.

1.1.1. Biomarcadores

Los biomarcadores son eventos que se producen en un sistema biológico y se utilizan para detectar enfermedades o como indicadores del estado de salud. En el contexto de la AML, son características biológicas, como proteínas, genes, mutaciones genéticas o niveles de expresión génica, que se pueden medir en muestras biológicas, como la sangre o la médula ósea.

Desde una perspectiva clínica, hay varios aspectos importantes a considerar con respecto al estudio de alteraciones genéticas en AML. Según la clasificación actual de la OMS, es necesario buscar defectos genéticos en pacientes con AML, porque cada defecto podría definir diferentes entidades clínicas y procesos patológicos. Ciertas anomalías cromosómicas y marcadores moleculares son importantes para la estratificación de riesgo (Prada-Arismendy et al., 2017).

El progreso en fisiopatología celular, inmunología y biología molecular, impulsado por las nuevas tecnologías, ha mejorado la percepción de los marcadores biológicos que distinguen entre células hematopoyéticas normales y leucémicas. Estas tecnologías permiten la detección de marcadores moleculares adicionales, como mutaciones puntuales, y la caracterización de perfiles epigenéticos y proteómicos, lo que facilita la clasificación concisa de un clon maligno, que puede ser, mieloide, linfoide B, linfoide T o bifenotípico habitualmente.

Detectar estos nuevos biomarcadores ayuda a entender mejor las causas moleculares de la enfermedad, lo que facilitará la toma de decisiones precisas en el diagnóstico, seguimiento y tratamiento. Esto no solo mejora los resultados para los pacientes, sino que también permite prever cómo responderá cada individuo al tratamiento. Por eso, las mutaciones genéticas se están incluyendo cada vez más en las guías de clasificación y estratificación del riesgo en AML (Tazi et al., 2022).

El análisis citogenético permite ampliar la comprensión de la leucemia al identificar genes clave involucrados en la transformación celular. En pacientes con AML y citogenética normal, las alteraciones moleculares con mayor relevancia pronóstica son las mutaciones de los genes FLT3-ITD, NPM1 y CEBPA. Mientras que la presencia de FLT3-ITD se asocia con

un pronóstico desfavorable, las mutaciones en NPM1 y CEBPA se relacionan con pronósticos favorables. Concretamente, el gen NPM1 funciona como supresor de tumores y el gen CEBPA codifica un factor de transcripción implicado en la diferenciación de los progenitores hematopoyéticos hacia la línea mieloide madura. Por otra parte, los cariotipos complejos se distinguen por la escasa presencia de reordenamientos en los genes NPM1, FLT3, CEBPA, RAS, o KIT. El gen AML1 es uno de los genes implicados con más frecuencia en la patogénesis de las leucemias y tiene un papel fundamental en la hematopoyesis (Quintero Sierra et al., 2021). Otros oncogenes y mutaciones importantes son: TP53, IDH1/2, DNMT3A, ASXL1, GATA, KIT y NRAS.

1.1.2. Clasificación de riesgo genético de ELN

Las siglas ELN-2017 se refieren a las recomendaciones de estratificación de riesgos de la AML establecidas por la Red Europea de Leucemia (European Leukemia Net) en el año 2017 (Döhner et al., 2017). Estas recomendaciones son utilizadas en la práctica clínica para clasificar a los pacientes con AML en diferentes grupos de riesgo con el fin de guiar el tratamiento y predecir el pronóstico. Estos grupos de riesgo permiten a los médicos tomar decisiones informadas sobre el tratamiento, como la intensidad de la quimioterapia, la posibilidad de un trasplante de médula ósea u otras terapias específicas. Además, esta clasificación ayuda a los pacientes y a sus familias a comprender mejor qué esperar en términos de resultados y pronóstico.

Las recomendaciones ELN-2017 para la estratificación de riesgos de AML han tenido una amplia influencia en la práctica clínica y se adoptaron en todo el mundo. Por lo tanto, es probable que los cambios introducidos por las pautas ELN-2022 (Döhner et al., 2022) también se incorporen en ensayos clínicos y prácticas rutinarias. Sin embargo, en 2023 sólo el 15% de los pacientes con AML son reclasificados por las nuevas recomendaciones, por lo que los resultados de los grupos de riesgo individuales, así como la precisión pronóstica general, siguen siendo en gran medida similares, ELN-2022 representa un cambio incremental respecto a la clasificación anterior. Aun así, para aquellos pacientes afectados por los cambios propuestos, sigue siendo de suma importancia evaluar si ese cambio incremental es un paso hacia una predicción de riesgo más precisa.

Al igual que ELN-2017, ELN-2022 es un sistema robusto de estratificación de riesgos aplicable tanto en pacientes jóvenes como mayores que reciben tratamiento intensivo. La asociación entre el sexo masculino y el riesgo genético adverso, que ya ha sido observada para ELN-2017, sigue siendo válida para el nuevo clasificador (Rausch et al., 2023). Aunque este suceso se debe en gran medida a diferencias de sexo en la frecuencia de mutaciones en los genes ASXL1, NPM1 y RUNX1. En particular, una proporción significativamente mayor de genética de riesgo adverso entre los pacientes masculinos se atribuye a una menor prevalencia de mutaciones en NPM1 y una mayor prevalencia de mutaciones en RUNX1 y ASXL1 en los hombres. Además, esta diferencia se ve aumentada por la introducción de mutaciones adicionales relacionadas con la mielodisplasia (MR, por sus siglas en inglés: “myelodysplasia-related mutations”) como definidoras de riesgo adverso, ya que la presencia de estas mutaciones también se asocia con el sexo masculino. La mielodisplasia es un trastorno de la médula ósea en el cual las células sanguíneas no se desarrollan ni funcionan correctamente, debido a que la médula ósea no produce suficientes células sanguíneas maduras y saludables. En algunos casos, la mielodisplasia puede progresar a leucemia, como la AML.

ELN-2022 reconoce mutaciones de MR en los genes BCOR, EZH2, SF3B1, SRSF2, STAG2, U2AF1 y ZRSR2 como marcadores independientes de riesgo adverso. La idea de que estas mutaciones reflejan mielodisplasia se refleja en la clasificación de la OMS y la

Clasificación Internacional de Consenso, que, con la excepción de RUNX1 en la clasificación de la OMS, también ven estas mutaciones como definitorias de AML con cambios genéticos relacionados con la mielodisplasia. Aunque estas mutaciones ocurren principalmente en el contexto de AML, su significado pronóstico no está del todo claro (Rausch et al., 2023).

La revisión de la clasificación de riesgo genético de ELN es útil para la aplicación de BN porque posibilita la selección de variables relevantes y focalizar el análisis en las características genéticas más significativas. Se reconocen tres grupos de riesgo citogenético: favorable, intermedio y desfavorable.

Tabla 1. Correlación citogenética y molecular para los grupos de riesgo pronósticos en la AML (ELN 2022)

Grupo de riesgo	Anomalía genética
Favorable	• t(8;21)(q22;q22.1); RUNX1-RUNX1T1 • inv(16)(p13.1q22) o t(16;16)(p13.1;q22)/ CBFB::MYH11 • NPM1 mutado, sin FLT3-ITD • CEBPA mutado en marco bZIP
Intermedio	• Mutación de NPM1 y FLT3-ITD • NPM1 de tipo salvaje con FLT3-ITD (sin lesiones genéticas de alto riesgo) • t(9;11)(p21.3;q23.3)/MLLT3::KMT2A • Anomalías citogenéticas y/o moleculares no clasificadas como favorables o adversas
Desfavorable	• t(6;9)(p23;q34.1)/ DEK::NUP214 • t(v;11q23.3)/ KMT2A reorganizado • t(9;22)(q34.1;q11.2); BCR::ABL1 • t(8;16)(p11.2;p13.3)/KAT6A::CREBBP • inv(3)(q21.3q26.2) o t(3;3)(q21.3;q26.2)/ GATA2, MECOM(EVI1) • -5 o del(5q); -7; -17/abn(17p) • Cariotipo complejo, cariotipo monosómico • ASXL1, BCOR, EZH2, RUNX1, SF3B1, SRSF2, STAG2, U2AF1, y/o ZRSR2 • TP53 mutado

1.2. Objetivos

El objetivo principal de este trabajo es el de conocer con cierto detalle la teoría general de redes y en particular la de las redes bayesianas, materia que no se estudia durante la realización del grado. Como primer objetivo específico, está el de tratar de descubrir por medio de redes bayesianas las relaciones estadísticas complejas entre las variables que representan las anomalías genéticas de la enfermedad AML. Esto se llevaría a cabo desarrollando un modelo de red bayesiana en R que integre datos genómicos de pacientes con AML, con el segundo objetivo específico de identificar patrones genéticos y clínicos que influyan en el curso y pronóstico de la AML para descubrir nuevas afirmaciones sobre la supervivencia y factores de riesgo en la enfermedad.

2. Materiales y métodos

2.1. Base de datos

Los datos utilizados para el estudio se han obtenido a través de la plataforma GitHub, que es un servicio de alojamiento de repositorios de código fuente y otros recursos relacionados con el desarrollo de software. Se utiliza para almacenar proyectos de programación colaborativa y proporciona herramientas para control de versiones, seguimiento de problemas, integración continua y más.

En este caso, los datos consisten en 2113 pacientes adultos con AML, inscritos en los ensayos del UK NCRI (Instituto Nacional de Investigación Clínica del Reino Unido). Estos ensayos reclutan exclusivamente hasta un 80% de los pacientes del Reino Unido aptos para recibir tratamiento intensivo o no intensivo, lo que los hace representativos de la población de pacientes en el mundo real, en contraposición a estudios limitados por criterios estrictos de entrada a ensayos clínicos (Tazi et al., 2022).

La mayoría de los participantes (83%, n=1755) recibió tratamiento intensivo, con una edad media de 56 años. Además, se ha utilizado también un conjunto de datos de 1540 pacientes con AML del AML-SG (Grupo de Estudio de Leucemia Mieloide Aguda), con una edad media de 50 años y con anotación molecular comparable al momento del diagnóstico.

Se obtuvo el consentimiento informado de todos los pacientes que participaron. La evaluación molecular de la cohorte del UK-NCRI incluyó la determinación de cariotipos, alteraciones en el número de copias y mutaciones oncogénicas en todo el cuerpo génico de 128 genes implicados en la patogénesis de neoplasias mieloides al momento del diagnóstico.

Para tener una cohorte de datos de tamaño mayor y poder mejorar la robustez de los análisis, he combinado ambos conjuntos de datos, el de UK-NCRI, que consta de 2113 pacientes y 207 variables, y el de AML-SG, con 1540 pacientes y 131 variables. Esta fusión se ha llevado a cabo mediante la identificación y el emparejamiento de las variables compartidas entre los dos conjuntos. Posteriormente, se procedió a la selección de las variables de interés, las cuales serán utilizadas en los análisis pertinentes.

Las variables de la base de datos final son las siguientes:

- **Os_status:** variable dicotómica que indica si el paciente ha fallecido (1) o si sigue vivo (0).
- **Os**: variable numérica que recoge la supervivencia del paciente en años.
- **Age**: variable numérica que indica la edad del paciente en años
- **Gender**: variable dicotómica que indica si el paciente es mujer (0) o hombre (1).
- Anomalías del cariotipo, son variables categóricas en las que 1= tenerlo, y 0= no tenerlo: **t_8_21, inv(16), t_9_11, inv(3), t_v_11**.
- Genes recodificados de la misma forma que las anomalías: **ASXL1, BCOR, CEBPA_bi, CEBPA_mono, EZH2, ITD, GATA2, NPM1, RUNX1, SF3B1, SRSF2, STAG2, TP53, ZRSR2**.

La selección de genes y anomalías del cariotipo se ha realizado de acuerdo con las recomendaciones de estratificación de riesgos de la AML establecidas por la Red Europea de Leucemia del año 2022, como se detalla en la Tabla 1 expuesta anteriormente. Esta selección incluye los genes relevantes para los diferentes grupos de riesgo de la enfermedad tratada. De este modo, se asegura que los análisis sean consistentes con los estándares internacionales y clínicamente relevantes, dando lugar a una interpretación precisa de los resultados en el contexto de la práctica clínica actual.

2.2. Redes bayesianas

Las redes bayesianas son herramientas para modelar fenómenos complejos mediante la representación de variables y sus interdependencias. Iniciado por Pearl, el modelo de redes bayesianas es una representación que integra los aspectos cualitativos y cuantitativos de las relaciones entre los atributos de un dominio de manera intuitiva (Sangüesa i Solé, s. f.).

En términos específicos, una red bayesiana es un modelo probabilístico que conecta un conjunto de variables aleatorias a través de un grafo dirigido, lo que facilita la resolución de problemas de decisión en situaciones de incertidumbre. Se trata de una representación gráfica de dependencias para el razonamiento probabilístico, donde los nodos representan variables aleatorias y los arcos indican relaciones de dependencia directa entre ellas (Lozano, 2011).

En el contexto de entornos biológicos o médicos, estas redes ofrecen una forma estructurada de explorar y comprender las relaciones entre diferentes variables. Aprovechando la inferencia bayesiana, permiten estimar la probabilidad de variables desconocidas utilizando información de variables conocidas.

En el ámbito de la investigación sobre la AML, se utilizarán las redes bayesianas para buscar patrones en las mutaciones que suceden juntas o que se excluyen entre sí para entender mejor cómo afectan a la enfermedad. Por ejemplo, a veces se encuentra que ciertas mutaciones no pueden ocurrir al mismo tiempo en el mismo tipo de cáncer, lo que da pistas sobre cómo las células cancerosas están funcionando. También se puede descubrir que algunas mutaciones son muy comunes en un tipo de cáncer, mientras que otras ocurren juntas con menos frecuencia.

2.2.1. Probabilidad condicionada y enfoque bayesiano

Las redes bayesianas, también conocidas como redes probabilísticas, se basan en la teoría de la probabilidad y fusionan el poder del teorema de Bayes con la representación visual de grafos dirigidos (Tirado Ríos et al., 2016).

La probabilidad condicional es relevante en las aplicaciones de la Estadística, ya que permite modificar nuestra creencia sobre sucesos aleatorios a medida que obtenemos nueva información. De esta forma, su adecuada comprensión es un requisito en el estudio de la inferencia estadística clásica y bayesiana (Estruch et al., 2019)

La definición de Laplace afirma que, si todos los sucesos elementales del espacio muestral *E* son equiprobables, la probabilidad de un suceso *A* es el cociente entre el número de resultados favorables de *A* y el número total de resultados posibles del experimento, que es el tamaño del espacio muestral *E* (Canals L., 2019):

$$P(A) = \frac{casos\ favorables}{casos\ posibles} \tag{1}$$

La probabilidad "a posteriori" o condicional *P(A|B)* de un suceso *A*, dado otro suceso *B*, es la probabilidad de que suceda *A*, sabiendo que *B* ha ocurrido. Formalmente, se define mediante la expresión (2).

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, siempre\ que\ P(B) > 0 \tag{2}$$

La probabilidad condicionada ayuda a introducir la noción de sucesos independientes. Dos sucesos *A y B* son independientes si, y solo si, *P(A|B) = P(A) y P(B|A) = P(B)*. Matemáticamente se puede deducir a partir de la regla del producto de probabilidades (3).

$$A\ y\ B\ son\ independientes\ si\ y\ solo\ si\ P(A \cap B) = P(A) \cdot P(B) \tag{3}$$

En los problemas enfocados a la probabilidad se suele suponer la hipótesis de independencia condicional. Esta afirma que los sucesos X e Y son independientes dado el suceso Z si y solo si,

$$P\big((X,Y)\big|Z\big) = P\big((X \cap Y)\big|Z\big) = P(X|Z)P(Y|Z) \tag{4}$$

De la misma forma, se puede definir la independencia condicionada entre X e Y, dado Z, si y solo si, se cumple cualquiera de las siguientes condiciones (Susi García, 2007):

$$\begin{aligned} &\text{i.} && P(X,Y|Z) = P(X|Z)P(Y|Z)\ \ con\ P(Z) > 0 \\ &\text{ii.} && P(X,Y,Z) = P(X|Z)P(Y|Z)P(Z)\ \ con\ P(Z) > 0 \\ &\text{iii.} && P(X,Y,Z) = \frac{P(X|Z)P(Y|Z)}{P(Z)}\ \ con\ P(Z) > 0 \end{aligned} \tag{5}$$

La probabilidad conjunta *P(A,B)* se denomina también probabilidad de intersección de sucesos *P(A∩B).* Ya que los sucesos se comportan como conjuntos, tenemos que *A∩B = B∩A*, se tiene:

$$P(A \cap B) = P(B \cap A) = P(B|A)P(A) \tag{6}$$

Se puede deducir el Teorema de Bayes:

$$P(A|B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{7}$$

Este, se puede generalizar a lo que se denomina partición del espacio muestral *E*, que es un conjunto de sucesos *{A_1, A_2, ..., A_n}*, tales que *$\forall_{i,j} \in \{1,2, ..., n\}$, $i \neq j$, $A_i \cap A_j = \emptyset$,* y $U_{i=1}^{n} A_i = E$. Es decir, n sucesos mutuamente excluyentes, cada uno con probabilidad distinta de cero y tales que su unión es el espacio muestral. Así que, dado *B*, un suceso cualquiera del que son conocidas las probabilidades condicionales *P(B|A_i)*, la probabilidad *P(A_i|B)* viene dada por la expresión:

$$P(\,A_i \mid B\,) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)} \tag{8}$$

En la que *P(A_i)* son las probabilidades a priori, *P(B|A_i)* es la probabilidad de *B* para la hipótesis *A_i*, y *P(A_i|B)* son las probabilidades a posteriori (Fernández Regalado, 2009). Esto quiere decir que el Teorema de Bayes permite establecer la relación entre la probabilidad a priori y a posteriori (Estruch et al., 2019).

El enfoque bayesiano es una vía para revisar creencias a medida que se obtienen nuevos datos. Está basado en el teorema de Bayes, explicado anteriormente, que establece

cómo actualizar una creencia inicial sobre cierto evento a partir de la evidencia observada. En el campo de las redes bayesianas, este enfoque bayesiano es utilizado para poder estimar las distribuciones de probabilidad condicional asociadas con cada nodo de la red. En otras palabras, las redes bayesianas utilizan el teorema de Bayes para inferir las relaciones probabilísticas entre variables, y este proporciona el fundamento matemático para la inferencia probabilística, permitiendo entender y modelar las relaciones entre variables en el genoma en AML.

2.2.2. Elementos de la teoría de grafos

La teoría de grafos es una rama de las matemáticas que estudia las relaciones entre objetos y ayuda a entender y analizar cómo están interconectados los elementos en diversos sistemas.

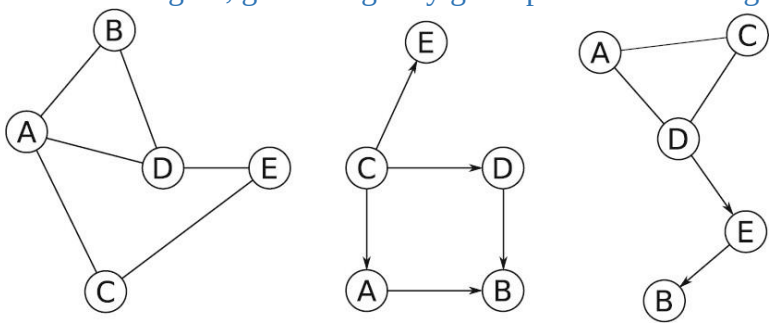
Una red bayesiana es un modelo probabilístico que conecta un conjunto de variables aleatorias mediante un grafo dirigido y acíclico. Para comprender mejor la definición de red bayesiana es necesaria la explicación de ciertos conceptos de teoría de grafos (Nagarajan et al., 2013).

Definición 1 (Grafo). *Un grafo es un par $G = (V, E)$, donde $V \neq \emptyset$ es el conjunto de vértices o nodos y $E \subseteq V \times V$ es el conjunto de arcos* (Mellado Cabrerizo, 2022).

Definición 2. *Si $(u,v) \in E$ y $(v,u) \in E$, se dice que (u,v) es un eje. Dos nodos $u,v \in V$ son adyacentes o vecinos si existe $(u,v) \in E$ o $(v,u) \in E$.*

Definición 3. *Un grafo $G = (V,E)$ se dice dirigido si todos los elementos de E tienen un sentido definido. En el caso en el que todos los arcos son ejes, se dice que G es un grafo no dirigido. Si contiene arcos dirigidos y ejes, se dice que es un grafo no dirigido.*

Figura 1. Grafo no dirigido, grafo dirigido y grafo parcialmente dirigido.



Nota: Grafos extraídos de Mellado Cabrerizo (2022)

A continuación, se tratarán únicamente los grafos dirigidos.

Definición 4 (Camino). *Se dice que $v_1v_2...v_n$ es un camino entre $v_1 \in V$ y $v_n \in V$ si $(v_i,v_{n+1}) \in E$ para todo $1 \leq i \leq n-1$.*

Definición 5 (Ciclo). *Un ciclo es un camino cerrado, es decir, aquel en el que $v_1 = v_n$.*

Definición 6 *(Grafo acíclico). Un grafo se dice acíclico si no contiene ciclos.*

Definición 7 (Relaciones entre nodos*). Dado un grafo $G= (V,E)$ y un nodo $u \in V$, se dice que $v \in V$ es antecesor de u si existe un camino dirigido de v a u. En tal caso, diremos que u es descendiente de v. Se dirá que v es padre de u si son adyacentes y v es antecesor de u.*

Análogamente, v es hijo de u si son adyacentes y v es antecesor de u. Análogamente, v es hijo de u si son adyacentes y v es descendiente de u.

Definición 8 (Conexiones elementales). *Dado un camino, independientemente de las direcciones de los arcos, se dice que un nodo es de arcos convergentes si solo tiene arcos incidentes, es decir, arcos que entran en un nodo. Similarmente, se dice que es un nodo de arcos divergentes si solo tiene arcos salientes, es decir, arcos que salen de un nodo. Por último, un nodo se dice de arcos en serie si posee un solo arco incidente y un solo arco saliente.*

Definición 9 (d-Separación). *Sea G= (V,E) un grafo y sean A, B, C ⊆ V subconjuntos de nodos disjuntos. Se dice que C d-separa A de B, denotando $A \perp_G B|C$, si a lo largo de cada secuencia de arcos entre un nodo de A y uno de B, existe un nodo v que satisface una de las siguientes condiciones:*

1.v tiene arcos convergentes y ni v, ni sus descendientes están en C.
2.v está en C, y no tiene arcos convergentes.

Definición 10 (v-estructura). *Se define una v-estructura como el conjunto de conexiones entre tres nodos en las que los nodos no adyacentes no son d-separados por el tercero.*

Definición 11. *Supongamos que tenemos una distribución de probabilidad conjunta P de un conjunto de variables aleatorias V y un grafo acíclico dirigido (GAD) G=(V,E). Decimos que (G,P) satisface la Condición de Markov si para cada variable X ∈ V, X es condicionalmente independiente del conjunto de sus no descendientes dado el conjunto de sus padres. Si (G,P) satisface la condición de Markov entonces (G,P) es llamada Red Bayesiana* (Hernández Leal, 2011).

Una BN *(G, P)*, por definición, es un GAD *G* y una distribución de probabilidad conjunta *P* que satisfacen la condición de Markov.

Teorema 1. *(G, P) satisface la condición de Markov (es una red bayesiana) si y solo si P es igual al producto de sus distribuciones condicionales de todos los nodos dados los padres en G, siempre que estas distribuciones condicionales existan. Si $V=\{X_1, \ldots, X_n\}$ para todos los valores posibles x_i de X_i, se tiene que:*

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i | PAx_i) \tag{9}$$

donde el conjunto de padres de X se denota como PAx y los no descendientes de X como NDx.

Esta igualdad es conocida como regla de la cadena. Además, si una distribución de probabilidad conjunta P cumple la condición de Markov en un DAG, entonces también se cumplirá la regla de la cadena (Domènech Pellejà, 2021).

Una manera de interpretar el Teorema 1 es que se puede representar una red bayesiana *(G, P)* utilizando un GAD con estructura *G* y las distribuciones condicionales de cada nodo como parámetros. Debido a esto, no es necesario mostrar todas las distribuciones conjuntas. Por lo tanto, se puede decir que una red bayesiana es una estructura que permite representar de manera sucinta una distribución de probabilidad conjunta. No obstante, el proceso debe ser de manera inversa. Inicialmente, se parte de un GAD, *G*, el cual representa las relaciones de dependencia entre las variables del problema, V. Luego, se procede a encontrar o estimar las distribuciones de probabilidad condicionales. Finalmente, se llega a la conclusión de que el

producto de estas distribuciones constituye una distribución conjunta, P, que cumple con la condición de Markov respecto a algún GAD. El teorema propuesto a continuación facilita este procedimiento (Domènech Pellejà, 2021):

Teorema 2. *Se da un DAG G donde cada nodo es una variable aleatoria discreta y se especifica una distribución de probabilidad condiciona de cada nodo dados los valores de sus padres en G. Entonces el producto de estas distribuciones condicionales dan una distribución de probabilidad conjunta PO de las variables (según la regla de la cadena), y la pareja (G,P) satisface la condición de Markov* (Neapolitan, 2003).

El Teorema 2 establece que, si se proporciona la distribución de probabilidad condicional de cada nodo de un DAG dado, con respecto a sus padres, y luego se genera una distribución de probabilidad conjunta P sobre las variables del DAG mediante la regla de la cadena, entonces el par (G, P) constituye una red bayesiana. Este método representa la forma práctica en la que se elaboran las redes bayesianas en la práctica.

Un concepto relevante vinculado a la independencia condicional entre variables de BN es el *manto de Markov* o *Markov blanket*. Este manto se define para un nodo específico dentro del grafo. El manto de Markov de un nodo es el conjunto de nodos que causan su *d-separación* con respecto al resto del grafo. En cualquier BN, el manto de Markov de un nodo A incluye a los padres de A, los hijos de A, y los padres de los hijos de A (Layo González, 2022).

El manto de Markov permite comparar redes bayesianas con modelos gráficos basados en grafos no dirigidos, también conocidos como redes de Markov. Se puede transformar un DAG en el grafo no dirigido correspondiente a la red de Markov siguiendo estos pasos:

1. Conectar los nodos no adyacentes en cada v-estructura con un arco no dirigido. Esto implica añadir un arco no dirigido entre cualquier nodo de la envolvente de Markov y el nodo central de dicha envolvente.
2. Ignorar la dirección de los otros arcos, reemplazando así los arcos dirigidos.

Esta transformación se denomina moralización, ya que une a los padres no adyacentes que comparten un hijo común. El grafo resultante se conoce como grafo moral (Romero Núñez, 2020).

2.2.3. Definición formal

Una red bayesiana es un grafo dirigido y acíclico donde los nodos representan variables aleatorias, las cuales pueden ser tanto continuas como discretas (Tirado Ríos et al., 2016). Los nodos se denotan con letras mayúsculas (X), mientras que sus posibles estados se designan con letras minúsculas (x_i). Los estados de una variable deben cumplir dos propiedades: ser mutuamente excluyentes, lo que significa que un nodo solo puede estar en un estado a la vez, y formar un conjunto exhaustivo, lo que implica que un nodo no puede tener ningún valor fuera de ese conjunto.

Formalmente, una red bayesiana se define como un par $B = (G, \theta)$, en el que $G = (V, A)$ representa un grafo dirigido acíclico con un conjunto de nodos o vértices $V = \{1, \dots, n\}$ y un conjunto de aristas o arcos $A \subseteq \{(a, b) \in V\ x\ V : a \neq b\}$. Esto implica que cada arista *(a, b)* pertenece a *V*, donde *a* representa el nodo inicial de la arista y *b* el nodo final. La parte $\theta = \{P(xi|x_{Pa(i)}), i = 1, \dots, n\}$ es un conjunto de parámetros que define la distribución de probabilidad condicional para cada variable de la red, donde *Pa(i)* denota el conjunto de nodos padre del nodo *i* (Layo González, 2022).

Una red bayesiana factoriza la distribución de probabilidad conjunta *P(x)* de un vector de variables aleatorias $X = (X_1, \ldots, X_n)$. Cada variable *Xi* es condicionalmente independiente de los nodos no descendientes de *i* dado *Pa(i)*. Esta propiedad es conocida como propiedad local de Markov. El conjunto de nodos descendientes de i consiste en todos los nodos alcanzables siguiendo un camino directo a través de los arcos dirigidos desde i.

2.2.4. Tipos de redes bayesianas

Las distribuciones condicionadas y la función de distribución conjunta propias de las redes bayesianas varían según la naturaleza de los datos (Mellado Cabrerizo, 2022), y por lo tanto, existen diferentes tipos de redes bayesianas. Si todas las variables son discretas, las distribuciones condicionadas son multinomiales, y el modelo correspondiente se denomina red bayesiana discreta o red bayesiana multinomial. Por otro lado, si las variables recogen valores continuos, la distribución conjunta es normal multivariante y las distribuciones condicionadas son variables aleatorias normales univariantes. Estas redes se llaman redes bayesianas gaussianas. Sin embargo, aún se está trabajando en definir un mecanismo de inferencia para otros tipos de variables continuas que no sean Gaussianas (Susi García, 2007). Además, existen las redes bayesianas mixtas, compuestas por variables discretas y Gaussianas. Pero en el caso de disponer de datos mixtos, también se puede optar por la discretización de los datos, es decir, transformar las variables continuas en discretas y, de esta forma, utilizar redes bayesianas discretas.

2.2.4.1. Redes Bayesianas Discretas

Las redes bayesianas discretas se distinguen por el carácter discreto de todas las variables en el modelo, lo que implica que cada variable solo puede tomar un conjunto finito de valores. Cuando las variables del problema, además de discretas, son binarias, siguiendo los procesos de Bernoulli, la red se identifica como una red bayesiana multinomial (Susi García, 2007). Este trabajo de fin de grado se basará en este tipo de redes.

Para las variables a incorporar que no son de tipo discreto, se realizará su discretización agrupando los valores en un conjunto de rangos o intervalos. A menudo, es conveniente representar un fenómeno continuo en la naturaleza mediante variables discretas. Para ello, las medidas continuas deben ser discretizadas, proyectando la escala de valores continua en un conjunto finito de intervalos. Los valores que caen en el mismo rango se consideran como un mismo estado. La discretización implica dividir el rango de las variables continuas en un número finito de intervalos exhaustivos y mutuamente excluyentes. Sin embargo, al discretizar se pierde información dependiendo del dominio y el número de intervalos. A pesar de ello, este método es comúnmente utilizado, ya que los modelos de redes bayesianas continuas están limitados a variables gaussianas y relaciones lineales (Molina Serrano et al., 2018).

Existen dos tipos de técnicas de discretización:

- No supervisada: no considera la clase, y se distingue entre intervalos iguales, intervalos con los mismos datos y basado en el histograma.
- Supervisada: basada en la clase, considerando los posibles “cortes” entre clases. Este tipo de discretización plantea un problema de complejidad computacional, y puede involucrar:
    - Probar clasificadores con diferentes datos.
    - Utilizar medidas de información, como reducir la entropía.

2.2.4.2. Redes Bayesianas Gaussianas

Cuando las variables aleatorias en cuestión siguen una distribución normal, la red bayesiana que las modela se denomina red bayesiana gaussiana. En estas redes, la distribución conjunta de las variables del problema $X = \{X_1, \dots, X_n\}$ es una distribución normal multivariante $N(\mu, \Sigma)$, donde μ es el vector de medias de dimensión n y Σ es la matriz de covarianzas definida positiva de dimensión $n\ x\ n$. La función de densidad se expresa como

$$f(x) = (2\pi)^{-\frac{n}{2}}|\Sigma|^{-1/2}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\} \tag{10}$$

Es importante destacar que la condición de normalidad para las distribuciones condicionadas de cada hijo según sus padres no implica una distribución conjunta normal multivariante, a menos que se exijan varianzas condicionales constantes y regresiones lineales (Arnold et al., 1999). Según la definición de red bayesiana, se debe verificar que la probabilidad conjunta sea jerárquica, factorizándose mediante:

$$f(x) = \prod_{i=1}^{n} f(x_i|pa(X_i)) \tag{11}$$

Por lo tanto, partiendo de la densidad conjunta del problema $N(\mu, \Sigma)$, la densidad condicionada $f(x_i|pa(X_i))\ \forall X_i$, es también normal y viene dada por

$$f(x_i|pa(X_i)) \sim N\left(\mu_i + \sum_{j=1}^{i-1}\beta_{ij}(x_j - \mu_j), v_i\right) \tag{12}$$

donde β_{ij} es el coeficiente de regresión de X_j en la regresión de X_i sobre sus padres, y v_i es la varianza condicionada de X_i dados sus padres.

$$v_i = \Sigma_i - \Sigma_{ipa(X_i)}\Sigma_{pa(X_i)}^{-1}\Sigma_{ipa(X_i)}^{T} \tag{13}$$

Es importante notar que el coeficiente de regresión es cero ($\beta_{ij} = 0$) si y solo si no hay una arista dirigida de X_j a X_i (Susi García, 2007).

2.2.4.3. Redes Bayesianas Mixtas

Las redes bayesianas mixtas también se conocen como redes bayesianas discretas-gaussianas, y se distinguen por la inclusión de variables tanto discretas como continuas en un modelo gráfico probabilístico dirigido.

Para poder definir el modelo, las variables discretas deben asumir un número finito de estados, mientras que las variables continuas deben seguir una distribución gaussiana. Es importante que, en el grafo, las variables discretas precedan a las variables continuas.

En estas redes, el conjunto de nodos $V = \{X_1, \dots, X_n\}$ se divide según si representan variables discretas (Δ) o continuas (Γ), de modo que $V = \Delta \cup \Gamma$. De esta manera, se designa el conjunto total de variables aleatorias como:

$$X = (x)_{\alpha \in V} = (i, \varsigma) = ((i_\delta)_{\delta \in \Delta}, (\varsigma_\gamma)_{\gamma \in \Gamma} \quad (14)$$

La distribución conjunta de las variables que componen una red bayesiana mixta es una distribución condicionada gaussiana, cuya densidad se expresa como:

$$f(x) = f(i, \varsigma) = exp\{g(i) + h(i)^T \varsigma - \varsigma^T K(i) \varsigma / 2\} \quad (15)$$

Aquí, *i* representa las variables discretas y ς las continuas. En esta fórmula, *g(i)* es un escalar, *h(i)* es un vector, *K(i)* es una matriz definida positiva, y $h(i)^T$ denota el vector *h(i)* traspuesto.

Para que la densidad conjunta pueda presentarse de manera jerárquica y sea factorizable, se deben considerar las densidades condicionadas de las variables dado que sus padres en el grafo acíclico dirigido han ocurrido (Susi García, 2007).

2.2.5. Modelado de las redes bayesianas

Una red bayesiana está compuesta por una estructura gráfica y tablas de probabilidad condicional entre los nodos, lo que permite representar el conocimiento que muestra la red de la siguiente forma:

- Conjunto de nodos: cada nodo, denotado como $\{X_i\}$, representa una variable del modelo, y cada una de ellas tiene un conjunto exhaustivo de estados $\{x_i\}$ mutuamente excluyentes.
- Enlaces o arcos: los enlaces, representados como (X_i, X_j), conectan nodos que tienen una relación causal, lo que asegura que todas las relaciones estén explícitamente representadas en el grafo.
- Tablas de probabilidad condicional: cada nodo X_i tiene asociada una tabla de probabilidad condicional que indica la probabilidad de sus estados para cada combinación de estados de sus padres. Si un nodo no tiene padres, se indican sus probabilidades a priori (Tirado Ríos et al., 2016).

La estructura de una red bayesiana se determina de la siguiente manera:

- Se atribuye un nodo a cada variable (X_i) y se identifica de qué nodos es una causa directa, formando el conjunto πXi llamado "padres de X_i".
- Se establecen flechas que parten de cada padre y llegan a sus hijos, indicando así las relaciones causales.
- A cada variable X_i se le asigna una matriz ($X_i \mid \pi X_i$) que estima la probabilidad condicional de un evento *Xi=xi*, dada una combinación de valores de los πXi (Molina Serrano et al., 2018).

Una vez la estructura de la red y las tablas de probabilidad condicional están definidas, se puede calcular la probabilidad de una variable dado el estado de cualquier combinación del resto de las variables de la red. Este cálculo se realiza mediante el cálculo de las probabilidades a posteriori de cada variable condicionada a la evidencia.

El proceso se simplifica mediante la aplicación de la propiedad de independencia condicional, que facilita el cálculo de la probabilidad conjunta a partir de las probabilidades condicionales de cada nodo en función de sus padres.

La construcción de redes bayesianas presenta una notable flexibilidad que permite la aplicación de enfoques manuales y/o automatizados. Esto implica la determinación de la estructura del modelo gráfico de la red, incluyendo la disposición de los nodos y los enlaces, así como la representación de la dependencia e independencia entre las variables (Ma et al., 2023).

2.2.6. Aprendizaje de redes bayesianas

El aprendizaje de redes bayesianas consiste en inducir un modelo, estructura y parámetros asociados, a partir de datos. Esto se realiza en dos etapas: aprendizaje estructural y aprendizaje de parámetros.

El aprendizaje de parámetros en redes bayesianas implica el uso de conjuntos de datos para determinar las probabilidades condicionales en cada nodo. Además, se han desarrollado enfoques automáticos para deducir tanto la estructura como los parámetros de las redes bayesianas a partir de datos. Es común combinar enfoques manuales y automáticos para el aprendizaje de parámetros, fusionando las probabilidades proporcionadas por expertos con aquellas aprendidas de los datos (Ma et al., 2023).

2.2.6.1. Aprendizaje estructural

A partir del conjunto de datos, se pretenden descubrir las dependencias existentes entre las variables para así poder construir el grafo representativo de la red. Es decir, obtener la estructura o topología de la red (Sucar, 2006).

Cuando el conjunto de variables no es extenso, es posible enumerar y evaluar exhaustivamente todos los posibles GAD y seleccionar aquel que obtenga el puntaje más alto. Sin embargo, el número de GAD que contienen n nodos se define mediante la siguiente recurrencia (Hernández Leal, 2011):

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} f(n-i) \qquad n > 2$$
$$f(0) = 1$$
$$f(1) = 0 \tag{16}$$

Concretamente, se ha comprobado que determinar la estructura óptima con variables discretas es un problema NP-difícil (Chickering et al., 2012), por lo que es habitual recurrir a técnicas heurísticas y aproximadas.

En términos generales, se pueden identificar tres enfoques para abordar esta problemática. El primero consiste en buscar independencias condicionales en los datos para derivar una estructura. El segundo enfoque emplea técnicas de programación dinámica y agrupamiento. Estos métodos suelen arrojar buenos resultados en redes pequeñas, con menos de 30 variables, aunque tienden a volverse ineficaces al aumentar el número de variables. Por último, existe un tercer grupo de métodos que realizan una búsqueda exhaustiva en el espacio de las redes bayesianas, haciendo uso de heurísticas y funciones de evaluación para obtener una estructura de red óptima.

Por otra parte, existen dos categorías para los algoritmos de aprendizaje de la estructura: los algoritmos basados en restricciones y los que están basados en puntaje.

- Algoritmos basados en restricciones.

Los algoritmos basados en restricciones determinan la estructura de la red evaluando las relaciones probabilísticas que se derivan de la propiedad de Markov en las redes bayesianas, mediante pruebas de independencia condicional. Posteriormente, construyen un gráfico que cumple con las declaraciones de d-separación asociadas.

Los algoritmos basados en restricciones se fundamentan en el algoritmo de causalidad inductiva (IC), el cual ofrece un marco teórico para el aprendizaje de modelos causales estructurales. Este proceso se puede desglosar en tres etapas:

1. Aprendizaje del esqueleto de la red. Inicialmente, se determina el esqueleto de la red. Debido a que una búsqueda exhaustiva es impracticable para conjuntos de datos grandes, los algoritmos emplean técnicas de optimización, como limitar la búsqueda a la manta de Markov de cada nodo.

2. Dirección de los arcos en v-estructuras. Se asignan direcciones a los arcos que forman una v-estructura, es decir, un conjunto de tres nodos con una conexión convergente $Xj \rightarrow Xi \leftarrow XkXj \rightarrow Xi \leftarrow Xk$.

3. Dirección de otros arcos para satisfacer la aciclicidad. Finalmente, se determinan las direcciones de los arcos restantes para asegurar que la red sea acíclica.

El paquete de R {*bnlearn*} ofrece seis algoritmos de aprendizaje basados en restricciones. En concreto, el algoritmo Grow-Shrink está enfocado en investigación genética y diagnóstico de enfermedades, y es el más sencillo dentro de los algoritmos de detección de Markov Blanket (Callejas Pinilla, 2020).

- Algoritmos basados en puntajes.

Los algoritmos basados en puntajes aplican diferentes métodos de búsqueda heurística de propósito general, tales como *hill-climbing* o *búsqueda tabú*. La función de puntaje generalmente es equivalente, lo que significa que las redes que representan la misma distribución de probabilidad reciben el mismo puntaje. {Bnlearn} implementa los algoritmos de aprendizaje basados en puntajes: Hill-climbing y búsqueda tabú.

Hill-climbing (hc) explora el espacio de los gráficos acíclicos dirigidos mediante la adición, eliminación e inversión de un solo arco, con reinicios aleatorios para evitar quedar atrapado en óptimos locales. La implementación optimizada utiliza almacenamiento de puntajes, descomposición de puntajes y equivalencia de puntajes para reducir el número de pruebas duplicadas. Este método se ha utilizado en problemas de optimización y algoritmos genéticos.

En cuanto a la búsqueda tabú, esta es una variante del hill-climbing que puede escapar de los óptimos locales seleccionando una red que disminuya mínimamente la función de puntaje. Este método se utiliza en problemas como el agente viajero, la secuenciación de producción y diversos problemas de diseño (Aboytes–Ojeda et al., 2013).

2.2.6.2. Aprendizaje de parámetros

Una vez se tiene la estructura de la red, el objetivo es calcular las distribuciones de probabilidad asociadas a la misma. Existen dos formas de llevarlo a cabo: la estimación de máxima verosimilitud y la estimación bayesiana (Andrés Mañas, 2017).

El aprendizaje de parámetros en una estructura de red fija es un desafío bien conocido en Estadística, especialmente en el contexto de las redes Bayesianas. En este enfoque, se

plantea el problema de la siguiente manera: se asume una distribución a priori sobre los parámetros de las funciones de densidad de probabilidad locales antes de utilizar los datos, y se busca una distribución previa conjugada, es decir, una que permita que el posterior sobre los parámetros pertenezca a la misma familia que la previa (Margaritis, s. f.). La conjugación de esta distribución a priori es deseable; una familia de distribución se llama prior conjugado a una distribución de datos cuando el posterior sobre los parámetros pertenece a la misma familia que el prior, aunque con diferentes hiperparámetros (los parámetros de una distribución sobre parámetros a veces se llaman hiperparámetros).

Los parámetros de una red bayesiana se encuentran en las distribuciones de probabilidad condicional (CPDs, por sus siglas en inglés) representadas por θ. Por lo tanto, el procedimiento de aprendizaje de parámetros para cada CPD depende de su tipo. Las dos CPDs más comunes utilizadas en la literatura son la categórica y la gaussiana lineal condicional (CLG, por sus siglas en inglés). Por esta razón, se incluye una descripción de los procedimientos de aprendizaje habituales para estas CPDs. Asumimos que se conoce la estructura de la red bayesiana *G*.

Una técnica estándar utilizada para estimar los parámetros es el criterio de estimación de máxima verosimilitud. La verosimilitud de un conjunto de datos *D*, dada una estructura de grafo *G*, y parámetros θ, se deriva de la siguiente ecuación (Atienza González, 2021):

$$P(x) = \prod_{i=1}^{n} P(x_i | x_{Pa(i)}) \tag{17}$$

Dando lugar a:

$$P(D|G,\theta) = \prod_{i=1}^{n} \prod_{j=1}^{N} P(x_i^j | x_{Pa(i)}^j) \tag{18}$$

Esta fórmula asume que las muestras en *D* son independientes e idénticamente distribuidas. El criterio de estimación de máxima verosimilitud selecciona el conjunto de parámetros θ que maximiza esta verosimilitud. Por conveniencia, la mayoría de las veces se utiliza el logaritmo de la verosimilitud, la log-verosimilitud, ya que la maximización de la verosimilitud y la log-verosimilitud devuelven el mismo conjunto de parámetros. La log-verosimilitud de unos datos *D* dada una estructura de grafo *G* y parámetros θ es:

$$L(G,\theta{:}\,D) = \sum_{i=1}^{n} \sum_{j=1}^{N} log P\left(x_i^j \middle| x_{Pa(i)}^j\right) = \sum_{i=1}^{n} L(X_i | X_{Pa(i)}, \theta_i{:}\,D) \tag{19}$$

donde $L(X_i|X_{Pa(i)}, \theta_i{:}\,D)$ es la log-verosimilitud local de la variable X_i, y θ_i es el conjunto de parámetros para la CDP $P(X_i|X_{Pa(i)})$. Entonces, el conjunto de parámetros de estimación de máxima verosimilitud se obtiene como:

$$\begin{aligned} \hat{\theta}^{MLE} = \arg\max_{\theta \in \Theta} L(G,\theta{:}\,D) = \{ & \arg\max_{\theta_1 \in \Theta_1} L(X_1 | X_{Pa(1),} \theta_1{:}\,D), \ldots, \\ & \arg\max_{\theta_n \in \Theta_n} L(X_n | X_{Pa(n),} \theta_n{:}\,D)\} \end{aligned} \tag{20}$$

donde Θ es el espacio de parámetros, es decir, el conjunto de parámetros permitidos, y Θ_i es el espacio de parámetros para la CDP $P(X_i|X_{Pa(i)})$. Esta optimización se realiza encontrando los parámetros de máxima verosimilitud de cada CPD de manera independiente

debido a que no hay parámetros compartidos entre las CPDs. Esta propiedad se denomina descomposición global.

Una alternativa a la estimación de máxima verosimilitud es la estimación bayesiana, que intenta tener en cuenta la incertidumbre en la selección de θ. Así, en lugar de realizar una estimación de punto único como dicho método, se consideran múltiples valores de parámetros en la estimación, calculando la probabilidad previa de cada estimación de θ. De esta manera, el conjunto de parámetros θ puede ser visto como una variable aleatoria que puede ser incluida en el modelo:

$$P(D, \theta|G) = P(D|G, \theta)P(\theta|G) \tag{21}$$

donde $P(\theta|G)$es una función de probabilidad previa para los parámetros θ dado el grafo *G*. La distribución previa define las creencias sobre la distribución de los parámetros θ antes de ver cualquier dato. Esta distribución previa sobre los parámetros se actualiza cuando hay datos disponibles para tener en cuenta la evidencia, generando la distribución posterior:

$$P(\theta \mid D, G) = \frac{P(D \mid G, \theta)P(\theta \mid G)}{P(D \mid G)} = \frac{P(D \mid G, \theta)P(\theta \mid G)}{\int_{\Theta} P(D \mid G, \theta)P(\theta \mid G)d\theta} \tag{22}$$

Una vez que se ha calculado la distribución posterior, la verosimilitud de un nuevo conjunto de datos D_J se puede calcular teniendo en cuenta la incertidumbre sobre los parámetros después de ver el conjunto de datos *D*:

$$P(D'|D, G) = \int_{\Theta} P(D'|G, \theta)P(\theta|D, G)d\theta \tag{23}$$

Esto se denomina comúnmente la distribución predictiva posterior. Estas dos últimas integrales pueden ser difíciles de calcular. Sin embargo, para algunas funciones de verosimilitud existen algunas distribuciones previas para las cuales la distribución posterior es de la misma familia y puede calcularse fácilmente. Estas distribuciones previas se llaman distribuciones previas conjugadas de la función de verosimilitud. Además, para muchas distribuciones previas conjugadas, la integral de la última ecuación también se puede calcular fácilmente con una fórmula cerrada. Por lo tanto, una técnica común es elegir la distribución previa sobre los parámetros θ utilizando distribuciones previas conjugadas para simplificar el proceso de aprendizaje.

2.2.6.3. Función de puntuación

Para poder evaluar la bondad de ajuste del modelo al conjunto de datos utilizado se emplea la función de puntuación, también conocida como score, por su nombre en inglés. Para poder explicar las funciones de puntuación más importantes, es necesario especificar ciertos aspectos de las redes bayesianas.

Se considera $D = \{c_1, \dots, c_M\}$ como un conjunto de datos de entrenamiento con M casos, y $V = \{X_1, \dots, X_n\}$ como un conjunto de variables aleatorias discretas. Para cada DAG Γ sobre V, se determina (Domènech Pellejà, 2021):

- Θ^{Γ}: vector de parámetros asociado al DAG.
- $\Theta^{\Gamma}_{\mathrm{MLE}}$: vector de parámetros estimado mediante máxima verosimilitud, es decir, el valor que tiene la mayor probabilidad de haber generado el conjunto de datos *D*.

- $L^D(\Theta^{\Gamma}_{\text{MLE}})$: el valor de la función de verosimilitud evaluada en $\Theta^{\Gamma}_{\text{MLE}}$ empleando el conjunto de datos *D*.
- d: la dimensión de Γ, que corresponde a los parámetros no redundantes en Θ^{Γ}, y representa la complejidad de la red bayesiana con DAG Γ.

El criterio de información bayesiano (BIC, por sus siglas en inglés: bayesian information criterion) es un tipo de función de puntuación que se emplea para seleccionar un modelo de entre un conjunto finito de modelos, prefiriendo aquel con el BIC más bajo. Al ajustar la estructura de una red bayesiana, es posible aumentar la verosimilitud añadiendo parámetros. Sin embargo, esto puede dar lugar a un ajuste excesivo y llevar a malas predicciones para nuevos casos fuera del conjunto de entrenamiento.

La puntuación BIC se calcula de la forma:

$$BIC(\Gamma, D) = \ln\left(L^D(\Theta^{\Gamma}_{\text{MLE}})\right) - \frac{d}{2}\ln(N) \tag{24}$$

En la primera parte de la fórmula se encuentra el logaritmo de la función de verosimilitud, que mide cómo de bien se ajusta el modelo (la red bayesiana) a los datos. La segunda parte de la fórmula representa la penalización por la complejidad del modelo.

Por otra parte, el criterio de información de Akaike (AIC, por sus siglas en inglés, Akaike information criterion) también se utiliza para seleccionar un modelo entre un conjunto finito de modelos, de forma que el elegido será el que tenga el AIC más bajo. Esta función balancea la bondad de ajuste y la complejidad del modelo. Aunque es similar al BIC, el AIC penaliza menos por la complejidad del modelo, lo que suele resultar en estructuras de red bayesiana más conectadas y complejas que las obtenidas con el score BIC.

El score AIC se define como:

$$AIC(\Gamma, D) = \ln\left(L^D\left(\Theta^{\Gamma}_{\text{MLE}}\right)\right) - d \tag{25}$$

Tanto el BIC como el AIC son fundamentales para evaluar y seleccionar modelos de redes bayesianas, cada uno con un enfoque para equilibrar el ajuste del modelo y la penalización por su complejidad (Domènech Pellejà, 2021).

2.2.7. Inferencia

La inferencia en una red bayesiana es el proceso computacional que utiliza el conocimiento probabilístico para calcular las probabilidades posteriores. Estas son las probabilidades de los estados del modelo después de presentar evidencia. Por ejemplo, al fijar los estados de ciertos nodos en la red, podemos calcular las probabilidades de los demás nodos. Esto nos permite realizar inferencia predictiva, diagnóstica o ambas a la vez.

En entornos donde se puede obtener evidencia incrementalmente, como en pruebas de laboratorio adicionales, podemos querer saber qué nueva información tendría el mayor impacto en la probabilidad de ciertas hipótesis. La red bayesiana puede calcular esto mediante el cálculo de la información mutua entre la hipótesis y la nueva observación. Esta información nos ayuda a identificar qué resultados de pruebas proporcionarían la mayor certeza al diagnóstico (Ma et al., 2023).

Existen diversos tipos de algoritmos para calcular las probabilidades posteriores, los cuales varían dependiendo del tipo de grafo y de si calculan la probabilidad de una variable a la vez o de todas simultáneamente. Los principales tipos de algoritmos de inferencia son los siguientes (Sucar et al., 2006):

- Algoritmo de eliminación. Utilizado para una variable en cualquier estructura.
- Algoritmo de propagación en Pearl. Aplicable a cualquier variable en estructuras simplemente conectadas.
- Para cualquier variable en cualquier estructura, se pueden emplear:
    - Agrupamiento (junction tree).
    - Simulación estocástica.
    - Condicionamiento.

Por otra parte, en redes con un gran número de nodos y dependencias, la propagación de probabilidades implica un alto coste computacional, lo que constituye un problema NP-complejo. Para representar una tabla de probabilidad condicional, que es utilizada para representar distribuciones condicionadas y describir la influencia entre variables, se necesitan 2^n valores si hay n variables booleanas o nodos. Sin embargo, una red bayesiana requiere de $n2^k$, siendo k el número de padres de la red y n el número de nodos. Por ejemplo, si se tienen n= 30 nodos, cada uno de ellos con k=5 padres, entonces la red requerirá 4960 números, mientras que la tabla de probabilidad conjunta 1,073,741,824 (Mappe Rojas, 2019).

Fue a partir de finales de los años 80, con el desarrollo de nuevos algoritmos de propagación, que se volvió posible utilizar redes capaces de modelar problemas del mundo real. Aunque no es necesario comprender los algoritmos de propagación para utilizar redes bayesianas, es importante entender los principios básicos para seleccionar los algoritmos y herramientas más adecuados. Los algoritmos de propagación pueden clasificarse en dos grandes grupos: algoritmos de propagación exactos, donde no hay error en las probabilidades calculadas, y algoritmos aproximados, donde las probabilidades de los nodos se estiman con cierto margen de error (Rodríguez & Dolado, s. f.). De esta forma, se explicará a continuación la inferencia exacta y la aproximada, para presentar un acercamiento a estos dos tipos de algoritmos.

2.2.7.1. Inferencia exacta

Se considera una red bayesiana con varios nodos y se supone que se ha observado el estado de una variable *Y*, que tomará un valor específico *y*. Existen múltiples causas posibles para este estado observado. Se va a calcular la probabilidad de cada uno de estos eventos usando el teorema de Bayes, que se expresa de la siguiente manera (Layo González, 2022):

$$P(A \mid y) = \frac{P(y \mid A) \cdot P(A)}{P(y)} \tag{26}$$

donde *A* es un nodo cuyo estado desconocemos e *y* es la evidencia observada en la red. Esta expresión es equivalente a la siguiente:

$$P(A, y) = P(A \mid y) \cdot P(y) \tag{27}$$

Utilizando la ecuación anterior, se pueden calcular las probabilidades condicionadas:

$$P(A = a|Y = y) = \frac{P(A = a, Y = y)}{P(Y = y)} = \sum_{B,C} P(B = b, A = a, C = c, Y = y) \tag{28}$$

Análogamente, para otra variable *B*:

$$P(B = b|Y = y) = \frac{P(B = b, Y = y)}{P(Y = y)} = \sum_{A,C} P(A = a, B = b, C = c, Y = y) \tag{29}$$

Donde *P(Y=y)* es una constante de normalización que asegura que la distribución de probabilidad *P(A|y)* suma 1.

Utilizando el teorema de Bayes:

$$P(\,A = a \mid Y = y\,) = \frac{P(Y = y \mid A = a)P(A = a)}{P(Y = y)} \tag{30}$$

Y mediante el Teorema de la probabilidad total:

$$\begin{aligned} P(Y = y) = {} & P(Y = y|A = a, B = b)\; P(A = a, B = b) + \\ & + P(Y = y|A = a^c, B = b)\; P(A = a^c, B = b) + \\ & + P(Y = y|A = a, B = b^c)\; P(A = a, B = b^c) + \\ & + P(Y = y|A = a^c, B = b^c)\; P(A = a^c, B = b^c) \end{aligned} \tag{31}$$

La inferencia que se ha llevado a cabo para calcular estas probabilidades es exacta. En la mayoría de los casos, dejando a un lado redes muy sencillas, calcular probabilidades a posteriori mediante la regla de Bayes es un problema intratable. Para este caso en concreto, el cálculo de esta constante *P*(*Y=y*) tiene que considerar cada uno de los escenarios a los que está condicionada la probabilidad de la variable en cuestión. Es decir, tiene en cuenta todos los posibles estados de las variables *A* y *B*.

A continuación, se introduce la eliminación de variables, un método que utiliza suposiciones de independencia condicional entre variables para acelerar el cálculo mediante inferencia exacta.

Si se aborda el problema del cálculo de lo que antes se llamaba constante de normalización y se simplifican las expresiones obtenidas mediante la regla de la cadena, se tiene:

$$P(Y = y) = \sum_{A} \sum_{B} \sum_{C} P(A = a, B = b, C = c, Y = y) \tag{32}$$

Esto se puede expresar como:

$$\begin{aligned} & P(Y = y) = \\ = & \sum_{A} \sum_{B} \sum_{C} P(A = a)\; P(B = b|A = a)\; P(C = c|A = a)\; P(Y = y|B = b, C = c) \end{aligned} \tag{33}$$

Es decir, cada nodo es dependiente de sus nodos padres, no de otros nodos ancestros. La estrategia clave aquí es llevar las sumas hacia atrás de la siguiente manera (Layo González, 2022):

$$\begin{aligned} & P(Y = y) = \\ = & \sum_{A} P(A = a) \sum_{B} P(B = b|A = a) \sum_{C} P(C = c|A = a)\; P(Y = y|B = b, C = c) \end{aligned} \tag{34}$$

Al realizar la suma más interna, se obtiene lo siguiente:

$$P(Y = y) = \sum_{C} P(C = c|A = a)\, P(Y = y|B = b, C = c) \tag{35}$$

Al realizar la suma restante se tiene:

$$P(Y = y) = \sum_{A} P(A = a)\, \tau_2(a, y) \tag{36}$$

Donde

$$\tau_2(a, y) = \sum_{B} P(B = b|A = a)\, \tau_1(a, y, b) \tag{37}$$

En general, en una cadena de nodos de longitud *n*, estos términos se calcularían un número exponencial de veces. Almacenando estos términos τ_i, se evita tener que calcularlos repetidamente. Así, se reduciría la complejidad de la inferencia exacta teniendo como base dos ideas:

- Debido a la estructura de una red bayesiana, algunas subexpresiones en la red solo dependen de un pequeño número de variables. Al calcular estos términos, se eliminan dichas variables del cálculo.
- Calculando estas expresiones una vez y almacenando los resultados, se evita generar dichos términos un número exponencial de veces.

2.2.7.2. Inferencia aproximada

Aunque la eliminación de variables es útil para resolver problemas de inferencia, la inferencia exacta se vuelve impracticable a medida que la red crece en tamaño. El problema de la inferencia en redes bayesianas es NP-complejo, y como se explicó anteriormente, esto significa que es muy difícil de resolver de manera eficiente para redes grandes. Por esta razón, se introduce la inferencia aproximada, que ofrece una forma de reducir significativamente el tiempo de cálculo a cambio de una precisión menor (Layo González, 2022). Existen diversos algoritmos que realizan inferencia aproximada:

- Muestreo directo:

  Una solución básica a este problema es el muestreo directo. Los algoritmos de muestreo generan una serie de muestras aleatorias basadas en la distribución de probabilidad de la red. Estas muestras se obtienen siguiendo un orden topológico de los nodos de la red: primero se consideran los nodos sin padres, luego los hijos de estos nodos, y así sucesivamente. Cada muestra representa un conjunto posible de valores para las variables de la red. Al generar estas muestras, se cuenta cuántas veces ocurre cada valor de cada variable. La probabilidad estimada de cada valor se calcula dividiendo el número de ocurrencias por el número total de muestras generadas.

  No obstante, este enfoque enfrenta problemas cuando se realizan consultas del tipo $P(A = a|B = b)$. Una forma de manejar esto es generar muestras de la misma manera, pero rechazando aquellas en las que $B \neq b$. Sin embargo, si $P(B = b)$ es pequeño, la mayoría de las muestras se rechazarán, lo que hace que el método sea ineficiente.

- Ponderación de verosimilitud:

  La ponderación de verosimilitud (también conocida por su nombre en inglés, Likelihood Weighting) es un método utilizado en la inferencia aproximada para generar muestras que sean más relevantes para la solución del problema. Este ajusta las probabilidades de las muestras según la evidencia disponible, mejorando así la precisión de la inferencia.

  Este método asegura que la evidencia observada tenga un impacto adecuado en la inferencia, haciendo que las muestras reflejen más fielmente la distribución objetivo. La ponderación de verosimilitud es especialmente útil cuando se dispone de evidencia que altera significativamente las probabilidades de las variables de la red.

- Muestreo por importancia:

  El muestreo por importancia (conocido también por su nombre en inglés, importance sampling) implica estimar el valor esperado de una función $f(x)$ de m variables en el dominio $\Omega \subset \mathrm{R^m}$ en relación con una distribución $P(X)$, conocida como la distribución objetivo, donde X es un conjunto de variables. El valor esperado de $f(x)$ puede estimarse tomando M muestras $x_1, \dots, x_M$ de P, y estimando de la siguiente manera:

$$E_P[f] \approx \frac{1}{M} \sum_{m=1}^{M} f(x_m) \tag{38}$$

  A veces es preferible utilizar otra distribución Q, llamada distribución de muestreo, debido a que puede ser costoso generar muestras a partir de P. Esto podría ser especialmente relevante si P es una distribución a posteriori para la red bayesiana.

  En general, la distribución de muestreo Q puede ser arbitraria, siempre y cuando $Q(x)>0$ para todo $P(x)>0$ y la generación de muestras a partir de Q sea computacionalmente simple. De esta manera, Q no pasa por alto ningún estado con probabilidad no nula en relación con P. Sin embargo, el rendimiento computacional de realizar inferencia aproximada a partir de Q dependerá de qué tan similar sea Q a P.

  Al generar muestras a partir de Q en lugar de P, no podemos estimar el valor de $f(x)$ tomando simplemente la media de las muestras generadas. Este estimador debe ajustarse para compensar el error introducido por la distribución de muestreo incorrecta, Q.

- Algoritmo de muestreo de Gibbs:

  Uno de los desafíos de los métodos anteriores es que, al insertar evidencia en un nodo o variable, el muestreo solo afecta a los nodos descendientes, mientras que los nodos no descendientes solo se consideran mediante pesos.

  El algoritmo de muestreo de Gibbs es un método de Monte Carlo de cadenas de Markov que aborda esta limitación para realizar inferencia aproximada en redes bayesianas. Este algoritmo genera una secuencia de muestras donde la primera se obtiene de la distribución de probabilidad a priori, y las muestras subsiguientes se generan de distribuciones que se acercan gradualmente a la distribución a posteriori.

  El objetivo es estimar la probabilidad de los nodos sin evidencia, dados ciertos nodos con evidencia fijada en una red bayesiana. El siguiente es el proceso general del algoritmo:

  1. Se genera inicialmente una muestra de las variables no instanciadas, es decir, sin evidencia, a partir de una distribución arbitraria.

2. Se asignan los valores correspondientes a la evidencia insertada a las variables instanciadas.
3. Para un gran número de iteraciones, para cada nodo *N* no instanciado:
   - Se calcula la probabilidad de los valores de *N* dados los valores actuales de los nodos en su vecindario de Markov.
   - Se asigna a los nodos *N* el valor de la muestra calculada.
   - Se almacena el valor de estos nodos.

Al finalizar este proceso, la proporción de iteraciones en las que se asigna a un nodo *N* un valor específico se aproxima a la probabilidad a posteriori de ese valor para *N*.

2.2.8. Clasificadores bayesianos

Construir una red bayesiana a partir de datos es útil para poder utilizarla como clasificador, es decir, como modelo que utiliza los valores de un conjunto de variables "input" para atribuir una clase a la variable predicha o "output" (Domènech Pellejà, 2021).

Los clasificadores bayesianos pueden cambiar según la estructura del clasificador, de la función score o del algoritmo de aprendizaje, entre otros. A continuación, se explican dos tipos de clasificadores bayesianos: el Naive Bayes y el Augmented Naive Bayes.

2.2.8.1. Clasificador Naive Bayes

El clasificador Naive Bayes destaca como uno de los más eficaces, debido a que tiene un buen rendimiento predictivo en comparación a otros clasificadores avanzados. Este, aprende la probabilidad condicional a cada atributo X_i, dada la clase C_k, a partir del conjunto de datos de entrenamiento *T*. La estructura del grafo acíclico dirigido de la red bayesiana se fija con anterioridad a la búsqueda de la distribución de probabilidad condicional, en la que se establece con obligatoriedad que las flechas partan desde la clase C_k a todos los atributos y no existan relaciones entre los atributos.

Este clasificador utiliza, como su nombre indica, la fórmula de Bayes para calcular la probabilidad de la variable predicha *C* en base a las probabilidades $P(X_i|C)$ y $P(C)$, y seleccionando como clasificación la clase con mayor probabilidad a posteriori.

2.2.8.2. Clasificador Augmented Naive Bayes

El clasificador augmented Naive Bayes está basado en la estructura del clasificador Naive Bayes, como su propio nombre indica. Sin embargo, en este caso, la variable predicha *C* también debe ser padre del conjunto de atributos, y se permiten arcos entre los diferentes atributos, aunque no son obligatorios. En otras palabras, la influencia entre los atributos se modela a partir de los datos. Por lo tanto, la estructura del DAG se fija con anterioridad con la obligación de que las flechas salgan de la variable predicha C hacia los atributos.

El aprendizaje en este caso tiene costos computacionales, en contraposición al clasificador Naive Bayes, debido a que la estructura sí ha de aprender (Domènech Pellejà, 2021).

2.2.9. Aplicaciones

Actualmente, las redes bayesianas tienen numerosas aplicaciones, especialmente cuando la cantidad de datos aumenta rápidamente, lo que requiere procesarlos e interpretarlos para extraer conocimiento preciso para una adecuada toma de decisiones (Rojas et al., 2012). Los clasificadores mencionados anteriormente también se pueden considerar una aplicación práctica de las redes bayesianas de mucha utilidad. Además, a menudo se necesita conectar la metodología de análisis de datos con el conocimiento de expertos en contenido. Las redes bayesianas ofrecen una oportunidad única para este diálogo, combinando análisis gráfico y métodos de estimación estadística. (Kenett, 2012).

El primer ejemplo de aplicación de redes bayesianas a mencionar es la usabilidad web. El diagnóstico de usabilidad web busca identificar deficiencias de diseño que afectan negativamente la experiencia de navegación. Para comprender la experiencia del usuario, se comparan la actividad del usuario y sus expectativas, estimadas mediante el procesamiento adecuado de datos de registros de servidor. Se utilizan intervalos de tiempo para medir la capacidad de respuesta de la página, el rendimiento de la página y el tiempo de lectura del contenido. Un análisis de redes bayesianas derivado de datos de registros web ayuda a predecir comportamientos y mejorar la puntuación de usabilidad.

Un segundo ejemplo, en el campo del análisis de riesgos operativos en tecnología de la información y comunicación (TIC), donde se fusionan diferentes conjuntos de datos, como datos de centros de llamadas, datos financieros y registros de monitorización de servicios de TI, las redes bayesianas permiten combinar datos cualitativos ordinales y datos nominales discretizados.

En el campo de las encuestas de satisfacción del cliente son herramientas clave en investigación social, gestión de clientes y marketing. Un ejemplo es una encuesta anual de satisfacción de clientes de un producto electrónico, evaluando aspectos del producto y servicios. Para dicho ejemplo, se puede emplear una red bayesiana derivada de los datos que permita realizar diagnósticos y diseñar indicadores tempranos de insatisfacción del cliente (Kenett, 2012).

En el campo de la psicología, el Instituto de Investigación en Salud Mental de Victoria utiliza esta herramienta para comprender, tratar y prevenir trastornos psicológicos como la esquizofrenia y los trastornos del estado de ánimo (Rojas et al., 2012).

Otros campos de utilidad también son la biomedicina, las investigaciones policiales, por ejemplo, para elaborar un sistema de predicción de la probabilidad de robo, y la agricultura entre otros.

2.2.10. Comparación con otras técnicas

Las técnicas empleadas para la estimación en el ámbito de la biología y la medicina suelen basarse principalmente en métodos estadísticos, especialmente en modelos de regresión. El análisis de regresión es una técnica utilizada para modelar la relación entre diferentes variables. Busca establecer cómo una o varias variables dependientes se comportan en relación con una o más variables independientes. Esta técnica permite obtener información sobre cómo varía una variable de interés, Y (la variable dependiente), cuando se produce un cambio en una de las variables independientes (Bouza, 2018). Sin embargo, este tipo de

modelos no son capaces de representar relaciones causales, lo que impide que sus predicciones sean precisas, ya que no incorporan todos los aspectos del dominio.

Las redes neuronales son modelos matemáticos que se inspiran en el comportamiento biológico de las neuronas y en la estructura del cerebro. También pueden ser vistas como sistemas inteligentes que llevan a cabo tareas de manera diferente a como lo hacen las computadoras actuales. Aunque estas últimas procesan la información con gran rapidez, hay tareas muy complejas, como el reconocimiento y la clasificación de patrones, que demandan demasiado tiempo y esfuerzo incluso en las computadoras más potentes de la actualidad. Sin embargo, el cerebro humano es más apto para resolverlas, muchas veces sin aparente esfuerzo (Tablada & Torres, 2009). Dichas redes, por su parte, han demostrado ser efectivas en conjuntos de datos con muchas instancias. Sin embargo, a diferencia de las redes bayesianas, las redes neuronales no manejan la incertidumbre. Además, actúan como cajas negras ya que no es posible entender cómo se han obtenido los resultados, ni interpretar los nodos intermedios. En contraste, en las redes bayesianas, todos los nodos y las tablas de probabilidad pueden interpretarse en relación con el dominio médico. Otra desventaja de las redes neuronales es que no pueden incorporar conocimiento de expertos médicos, ya que solo pueden ser entrenadas con bases de datos. Si se aumenta el número de registros, es necesario volver a entrenar la red. Las redes bayesianas, en cambio, pueden ser generadas por expertos médicos, por bases de datos o por una combinación de ambas, lo que permite la adquisición incremental de conocimiento (Rodríguez & Dolado, s. f.).

En cuanto a los sistemas basados en reglas, estos consisten en una serie de reglas del tipo: si “condición”, entonces “acción”. Estas reglas se utilizan para actuar según la información obtenida. La ventaja de estos sistemas es su simplicidad, aunque son más adecuados para entornos determinísticos, lo cual no siempre es aplicable en el campo médico. Para abordar esta limitación, se ha introducido la lógica difusa, permitiendo incorporar la incertidumbre de una manera simple e intuitiva. Además, las redes bayesianas operan de manera global, lo que significa que cualquier nodo puede recibir evidencia y las probabilidades se propagan globalmente, mientras que, en los sistemas basados en reglas, el orden de aplicación de las reglas está predefinido.

2.3. Bootstrapping

El bootstrapping es una técnica útil en el análisis de redes bayesianas para evaluar la estabilidad, precisión y confiabilidad de los modelos aprendidos. Al crear múltiples conjuntos de datos remuestreados (muestras Bootstrap) y aprender una red bayesiana de cada uno, ayuda a determinar qué arcos o conexiones entre nodos son más robustas y estables. Se introdujo en 1979 como una técnica estadística computacionalmente intensiva que permite a los investigadores hacer inferencias a partir de datos sin asumir distribuciones específicas. Existen dos tipos de distribuciones a considerar. La primera es la distribución subyacente de los datos, que generalmente se describe como una función de probabilidad (por ejemplo, normal, binomial o Poisson) que muestra todos los valores que las variables pueden tener y la probabilidad de que cada uno ocurra. La segunda es la distribución del estadístico (por ejemplo, la mediana) calculado a partir de los datos. Tanto los datos como el estadístico calculado pueden variar de formas que se pueden describir matemáticamente bajo la suposición de que se obtuvieron nuevos conjuntos de datos o "muestras" y, para cada conjunto de datos, se calculó un nuevo estadístico. Más precisamente, la distribución de muestreo del estadístico es la probabilidad de todos los posibles valores del estadístico estimado calculado

a partir de una muestra de tamaño n extraída de una población dada. El bootstrapping utiliza el remuestreo con reemplazo (también conocido como remuestreo de Monte Carlo) para estimar la distribución de muestreo del estadístico. Si se puede determinar la distribución de muestreo, esta puede usarse para estimar errores estándar e intervalos de confianza para ese estadístico en particular (Haukoos & Lewis, 2005).

Los pasos para estimar intervalos de confianza usando el bootstrap son los siguientes:

1. Primero, se utiliza el remuestreo con reemplazo para crear *m* conjuntos de datos remuestreados (también conocidos como muestras bootstrap) que contienen el mismo número de observaciones, *n*, que el conjunto de datos original. Para realizar el remuestreo con reemplazo, se selecciona aleatoriamente una observación o punto de datos del conjunto de datos original y se copia en el conjunto de datos remuestreados que se está creando. Aunque ese punto de datos se haya "utilizado", no se elimina del conjunto de datos original, o sea, se "reemplaza". Luego, se selecciona aleatoriamente otro punto de datos y se repite el proceso hasta crear un conjunto de datos remuestreados de tamaño *n*. Como resultado, la misma observación puede incluirse en el conjunto de datos remuestreados una, dos o más veces, o no aparecer en absoluto.
2. En segundo lugar, se calcula el estadístico descriptivo elegido para cada conjunto de datos remuestreados.
3. Después, se calcula un intervalo de confianza para el estadístico a partir de la colección de valores obtenidos para el estadístico. En este punto del análisis, existen varias opciones para calcular los intervalos de confianza, incluyendo el método de aproximación normal, el método percentil, el método corregido por sesgo (BC), el método corregido por sesgo y acelerado (BCa), y el método de confianza aproximada por bootstrap (ABC).

Cada muestra bootstrap debe tener el mismo tamaño de muestra que el conjunto de datos original. Si los tamaños de muestra del bootstrap difieren del tamaño de muestra del conjunto de datos original, la estimación calculada para el intervalo de confianza puede estar sesgada. Se ha descrito una corrección para este sesgo, aunque parece no haber ninguna ventaja práctica al realizar el análisis de esta manera. El método de aproximación normal calcula un error estándar aproximado usando la distribución de muestreo resultante de todos los remuestreos bootstrap. Luego se calcula el intervalo de confianza usando la distribución z (estadístico original ± 1.96 × error estándar, para un intervalo de confianza del 95%). El método percentil utiliza el histograma de frecuencia de los *m* estadísticos calculados a partir de las muestras bootstrap. Los percentiles 2.5 y 97.5 constituyen los límites del intervalo de confianza del 95%. El método BCa ajusta el sesgo en las distribuciones de muestreo bootstrap en relación con la distribución de muestreo real, y se considera una mejora sustancial sobre el método percentil. El intervalo de confianza BCa es un ajuste de los percentiles utilizados en el método percentil basado en el cálculo de dos coeficientes llamados "corrección de sesgo" y "aceleración". El coeficiente de corrección de sesgo ajusta la asimetría en la distribución de muestreo bootstrap. Si la distribución de muestreo bootstrap es perfectamente simétrica, la corrección de sesgo será cero. El coeficiente de aceleración ajusta las varianzas no constantes dentro de los conjuntos de datos remuestreados. El método ABC es una aproximación del método BCa que requiere menos conjuntos de datos remuestreados que el método BCa.

Como guía general, se deben usar 1,000 o más conjuntos de datos remuestreados al calcular un intervalo de confianza BCa. Al no tener que calcular la corrección de sesgo, se puede usar un valor más pequeño, en el rango de 250, cuando se utiliza el método percentil para estimar un intervalo de confianza. A medida que disminuye el número de conjuntos de

datos remuestreados, se introduce más variabilidad en la estimación del intervalo de confianza (es decir, la variabilidad está inversamente relacionada con el número de conjuntos de datos remuestreados) (Haukoos & Lewis, 2005).

2.4. Paquetes disponibles de R

2.4.1. Entorno R

R se define como un entorno de lenguaje de programación y computación, orientado al análisis estadístico, y está disponible para el público como software libre bajo la Licencia Pública General GNU que rige su código fuente. Esencialmente, R permite al usuario dar instrucciones al equipo para aplicar diversas técnicas estadísticas y generar gráficos (Angulo Montes, 2020).

La gestión y mantenimiento de los derechos de autor del software y la documentación están a cargo de los miembros del "Equipo Central de Desarrollo de R" a través de la Fundación R, una entidad sin ánimo de lucro. Entre los objetivos de esta fundación se encuentran el apoyo continuo al desarrollo de R, la exploración e implementación de nuevas metodologías y la enseñanza o capacitación en computación estadística, entre otros.

El entorno de R consiste en una distribución base que incluye funcionalidades principales como manipulación de datos, creación de gráficos, análisis de distribución de probabilidades, análisis estadístico de datos y utilidades para importar o exportar datos en diferentes formatos. A esta distribución base se le pueden añadir funcionalidades mediante paquetes creados por desarrolladores independientes. Estos paquetes deben ser enviados a la red de servidores conocida como "The Comprehensive R Archive Network" (CRAN), que verifica la calidad y los distribuye. Algunas herramientas que se pueden agregar mediante paquetes están centradas en el aprendizaje automático, la minería de texto, la visualización interactiva de datos y el procesamiento de imágenes, entre otros.

Es importante destacar que CRAN proporciona a los usuarios de R una guía llamada "task view", donde se pueden encontrar paquetes disponibles según la temática que se desee abordar. Además, estos paquetes están subagrupados según la tarea o actividad a desarrollar en relación con esa temática.

2.4.2. Paquete utilizado

En el task view de modelos gráficos del repositorio CRAN de R, existen más de 30 paquetes destinados a la representación, manipulación y aprendizaje de redes bayesianas y redes de Markov (Angulo Montes, 2020).

El paquete de R que se utilizará en este trabajo de fin de grado es {*bnlearn*}, este implementa algoritmos clave que cubren todas las etapas del modelado de redes bayesianas: el preprocesamiento de datos, el aprendizaje de estructuras combinando datos y conocimiento experto o anterior, y el aprendizaje de parámetros e inferencia. Este paquete tiene como objetivo ser un lugar de acceso a la realización de redes bayesianas en R, proporcionando las herramientas necesarias para aprender y trabajar con redes bayesianas discretas, redes bayesianas gaussianas y redes bayesianas gaussianas lineales condicionales en datos del mundo real. También se admiten datos incompletos con valores faltantes.

Además, la naturaleza modular de {*bnlearn*} facilita su uso para estudios de simulación. Los algoritmos de aprendizaje de estructuras implementados incluyen: algoritmos basados en restricciones, que utilizan pruebas de independencia condicional para aprender

restricciones de independencia condicional a partir de datos. A su vez, las restricciones se utilizan para aprender la estructura de la red bayesiana bajo la suposición de que la independencia condicional implica separación gráfica, es decir, dos variables que son independientes no pueden estar conectadas por un arco (Scutari et al., 2024).

Los algoritmos de aprendizaje de estructuras implementados incluyen:

- Algoritmos basados en puntajes, que son algoritmos de optimización de propósito general que clasifican las estructuras de red con respecto a un puntaje de bondad de ajuste, explicados anteriormente.
- Algoritmos híbridos que combinan aspectos tanto de los algoritmos basados en restricciones como en puntajes, ya que utilizan pruebas de independencia condicional (generalmente para reducir el espacio de búsqueda) y puntajes de red (para encontrar la red óptima en el espacio reducido) al mismo tiempo (Scutari et al., 2024)

El paquete {*bnlearn*} destaca sobre el resto de los paquetes disponibles en R por su capacidad para manejar datos reales, proporcionando herramientas que permiten a los usuarios combinar el aprendizaje automático a partir de datos con el conocimiento previo o experto. Esto se logra a través de funciones que facilitan la modificación de la estructura aprendida o la incorporación de información previa en el modelo.

{*bnlearn*} es una herramienta integral para la modelización de redes bayesianas. Además de facilitar el preprocesamiento de datos, aprendizaje de la estructura y parámetros, e inferencia, también soporta la construcción de clasificadores y la evaluación de modelos.

Para el aprendizaje de la estructura, dicho paquete implementa una variedad de algoritmos que permiten a los usuarios seleccionar el método más adecuado según sus necesidades específicas. Los algoritmos disponibles incluyen métodos basados en restricciones, métodos de búsqueda-puntaje y métodos híbridos, ofreciendo flexibilidad y adaptabilidad en el proceso de modelado.

Además, proporciona múltiples opciones para evaluar los modelos durante el proceso de búsqueda, utilizando puntajes como el Log-likelihood, AIC, BIC y otros, lo que permite una evaluación exhaustiva y precisa del rendimiento del modelo.

Para los procesos de inferencia, {*bnlearn*} cuenta con métodos como Logic Sampling y Likelihood Weighting, ambos enfocados en la inferencia aproximada, asegurando que los usuarios puedan realizar análisis robustos incluso con datos incompletos o valores faltantes.

En resumen, {*bnlearn*} es una herramienta versátil y robusta que soporta todas las etapas del modelado de redes bayesianas, desde el preprocesamiento y el aprendizaje hasta la inferencia y la evaluación, adaptándose a una amplia variedad de datos y necesidades analíticas (Angulo Montes, 2020).

También se utilizará el paquete {igraph}, que es una herramienta especialmente útil para el análisis y visualización de grafos y redes complejas. Tiene como principales objetivos proporcionar un conjunto de tipos de datos y funciones para la implementación sencilla de algoritmos de grafos y para el manejo rápido de grafos grandes, con millones de vértices y aristas.

Los grafos en {igraph} tienen una clase denominada “igraph”. Estos grafos aparecen en pantalla en un formato especial. En este paquete se presentan cuatro bits que denotan el tipo de grafo. El primero es “U” para grafos no dirigidos y “D” para grafos dirigidos. El segundo es “N” para grafos nombrados, es decir, si el grafo tiene el atributo de vértice “name” establecido. El cuarto es “B” para grafos bipartitos.

Para crear grafos, {*igraph*} ofrece muchas funciones, tanto para grafos deterministas como estocásticos. Además, es posible asignar atributos a los vértices o aristas de un grafo, o al grafo mismo.

En el código utilizado en este estudio, se emplea la función "*arc.strength*" para calcular las fortalezas de las aristas y normalizarlas, asignando luego estos valores como el ancho de las aristas en el grafo. Luego, el grafo se convierte a un objeto igraph con "*as.igraph*" y se personalizan varios atributos de los vértices y aristas. Por ejemplo, "*V(graph)$shape*" establece la forma de los vértices, "*V(graph)$color*" el color, "*V(graph)$label.color*" el color de las etiquetas, "*V(graph)$size*" el tamaño, "*V(graph)$label.cex*" el tamaño de las etiquetas y "*V(graph)$frame.color*" el color del marco. Finalmente, el grafo se visualiza con la función *plot()*, especificando los parámetros personalizados.

En resumen, este paquete proporciona una amplia gama de funciones que permiten crear, manipular, y visualizar grafos. En este estudio será necesario para crear redes fáciles de interpretar ya que ofrece múltiples opciones de visualización que permiten personalizar el color, tamaño, forma y grosores de los elementos del grafo.

3. Resultados

Antes de realizar los análisis pertinentes, es necesario realizar un preprocesamiento de los datos para asegurar su calidad y fiabilidad. Este proceso se ha llevado a cabo mediante diferentes etapas esenciales.

En primer lugar, se han combinado dos conjuntos de datos. Para llevar a cabo esto, fue necesario ajustar los nombres de las variables de ambos conjuntos de datos para crear una concordancia entre ellos y facilitar su unión. También se transformaron a factores las variables categóricas ya que son más adecuados para este tipo de variables. En cuanto a la dimensionalidad de los datos, seleccioné los genes y anomalías del cariotipo de acuerdo con las recomendaciones de estratificación de riesgos de la AML establecidas por la Red Europea de Leucemia del año 2022, para poder asegurarme de que las redes bayesianas a elaborar sean consistentes con la información previa sobre la AML.

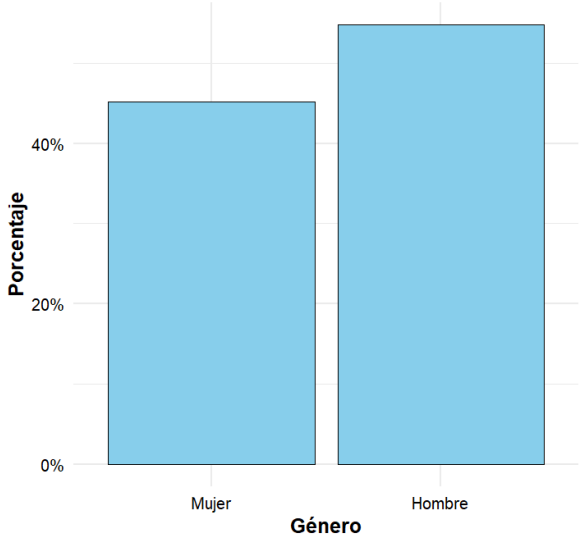
Después de combinar ambos conjuntos de datos, se han recodificado las variables continuas para poder emplear redes bayesianas discretas. La variable edad se ha recodificado en dos variables: pacientes mayores de 60 años y menores de esa edad. Estos grupos de pacientes tienen comportamientos clínicos muy diferentes.

Además, se han revisado los valores faltantes por cada variable, especialmente en las variables relevantes para la construcción de la curva de supervivencia de Kaplan-Meier debido a que la presencia de valores faltantes puede afectar significativamente a la validez de los análisis posteriores.

3.1. Estadística descriptiva

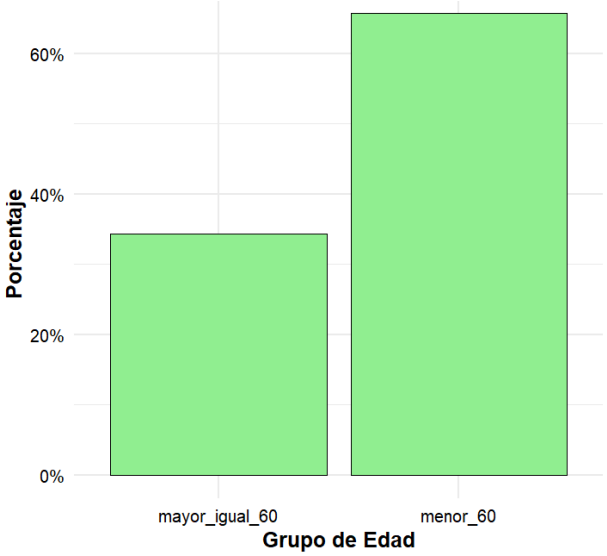
Para implementar los análisis descriptivos se utilizará todo el conjunto de datos, que está formado por 3653 pacientes y 23 variables, descritas anteriormente. De estos pacientes, el 45.22% son mujeres y el 54.78% son hombres, como se puede ver en la gráfica 1.

Gráfica 1. Diagrama de barras de la variable género



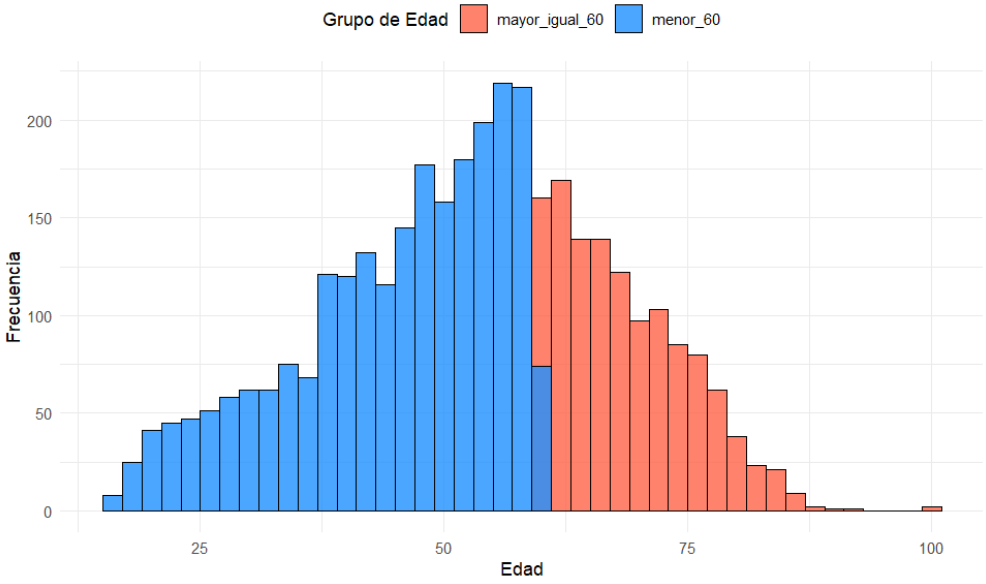
La variable edad, que recoge la edad de los pacientes en años, está distribuida en un intervalo entre los 15 hasta los 100 años. Esta variable, como se explicó anteriormente, está recodificada según si los pacientes son mayores o menores de 60 años.

Gráfica 2. Diagrama de barras de la variable edad recodificada



En la gráfica 2 se muestra que aproximadamente el 40% de las personas son mayores de 60 años o tienen esa edad, mientras que el 60% de los pacientes tienen menos de 60 años. Las barras destacan que la mayoría de la población en el estudio es menor de 60 años.

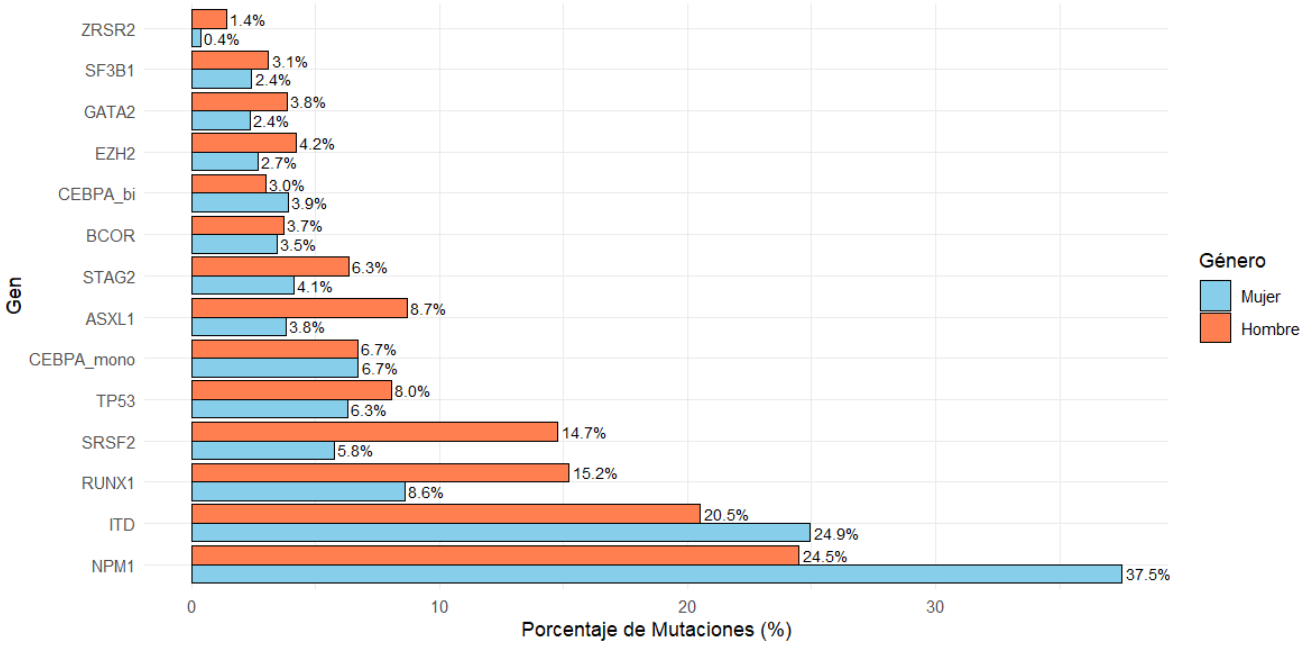
Gráfica 3. Histograma de la distribución de edad en pacientes



La gráfica 1 muestra una moda diferente en cada grupo de edad, con un pico en torno a los 55 años, en el caso de los pacientes menores de 60 años y otro pico en torno a los 65 años para los pacientes que son mayores de 60 años. En resumen, se observa que la mayoría de los pacientes se concentran en las edades cercanas a los 60 años.

A continuación, se plantea un gráfico de barras agrupadas utilizado para comparar el porcentaje de mutaciones entre hombres y mujeres para cada gen.

Gráfica 4. Porcentaje de mutaciones por gen y género



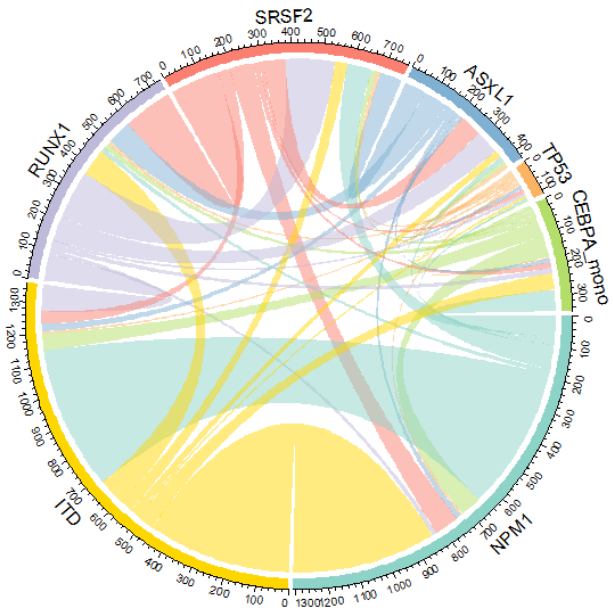
El gráfico muestra que el gen NPM1 tiene el mayor porcentaje de mutaciones, con una notable diferencia entre hombres (24.9%) y mujeres (37.5%). Este hallazgo concuerda con las normas European LeukemiaNet (ELN), que asocian una menor prevalencia de mutaciones en NPM1 con el género masculino. Los genes ITD y RUNX1 también presentan porcentajes altos de mutaciones, en el caso de ITD son similares para hombres y mujeres. Sin embargo, los resultados muestran que el gen RUNX1 presenta una mayor prevalencia de mutaciones en hombres (15.2%) en comparación con mujeres (8.6%). Esta tendencia es consistente con las normas de la ELN, que asocian las mutaciones de RUNX1 con el género masculino, y lo mismo ocurre con el gen ASXL1, que muestra una prevalencia de mutaciones más alta en hombres.

Este hallazgo no solo valida la importancia de considerar el género en los análisis genéticos de AML, sino que también subraya la precisión de las guías ELN en la identificación de patrones genéticos específicos en subgrupos de pacientes.

Por otra parte, los genes ZRSR2 y SF3B1 presentan porcentajes bajos de mutaciones y mínimas diferencias entre géneros, siguiendo una distribución más uniforme.

El diagrama de cuerdas o chord diagram, por su nombre en inglés, que se presenta a continuación, muestra las co-ocurrencias de mutaciones en los 7 genes con mayor porcentaje de mutaciones en el conjunto de pacientes. Los genes representados son: NPM1, ITD, RUNX1, SRSF2, ASXL1, TP53 y CEBPA_mono. Este diagrama es útil para visualizar las relaciones y los patrones de interacciones complejas entre los diferentes genes mutados. En él, cada sector del anillo externo del diagrama representa un gen, y las bandas que conectan los sectores indican la co-ocurrencia de mutaciones entre los genes en los mismos pacientes.

Gráfica 5. Diagrama de cuerdas de mutaciones genéticas



Observamos que los genes NPM1 e ITD muestran una fuerte conexión, indicando que sus mutaciones tienden a ocurrir juntas frecuentemente en los pacientes. Los genes RUNX1 y ASXL1 también presentan una significativa co-ocurrencia de mutaciones, lo que sugiere una interacción importante en la progresión de AML. Además, las mutaciones en SRSF2 y TP53 muestran una notable co-ocurrencia, destacando posibles subgrupos de alto riesgo. CEBPA_mono tiene múltiples conexiones con otros genes, reflejando la diversidad en las rutas de progresión de la enfermedad.

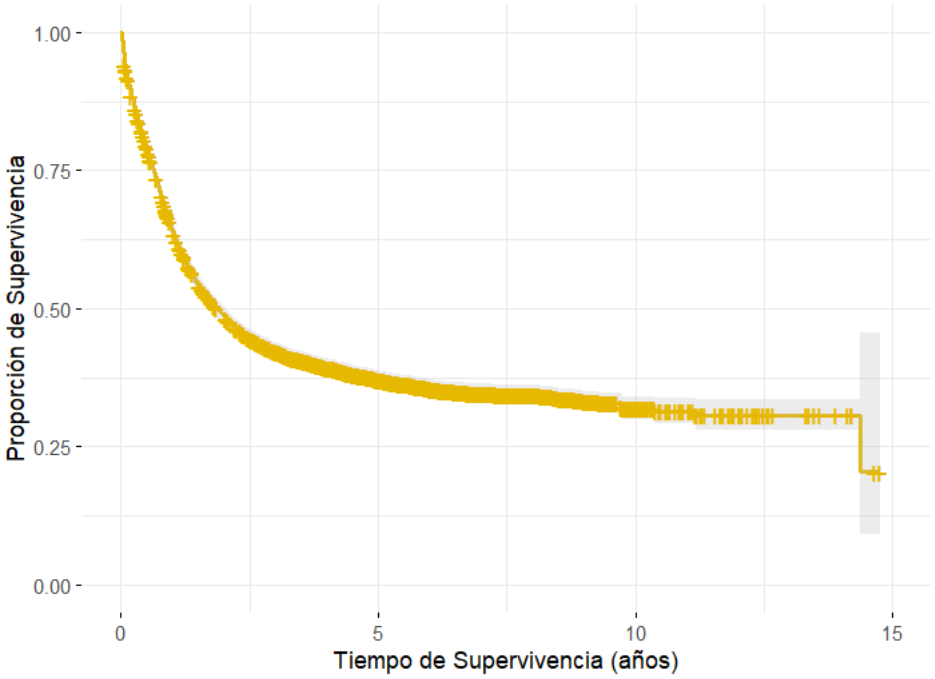
El análisis de las anomalías del cariotipo reveló que la anomalía inv(16) está presente en el 4.8% de los pacientes, siendo una de las más frecuentes, junto con t(8;21). En contraste, inv(3) se encontró solo en el 1.2% de los pacientes, siendo la anomalía menos común,

La supervivencia global de los pacientes se representa a través de la curva de Kaplan-Meier, este es un método no paramétrico que no asume que los datos sigan una distribución específica. En cambio, solo requiere que los sujetos censurados hubieran tenido el mismo comportamiento que aquellos seguidos hasta la ocurrencia del evento. En otras palabras, se asume que la censura es no informativa (Arias, 2022).

La curva de Kaplan-Meier se calcula mediante el riesgo instantáneo acumulado, que se obtiene multiplicando el riesgo instantáneo de todos los periodos anteriores a una fecha específica. El riesgo instantáneo se define como el número de pacientes que han fallecido en un periodo dividido por el número de pacientes disponibles al inicio de ese periodo. Estos intervalos de cálculo suelen ser día a día, lo que proporciona una visión detallada de cómo cambia la supervivencia a lo largo del tiempo.

Además, es importante destacar que la mediana de supervivencia, que es el tiempo en el que el 50% de los pacientes siguen vivos, no es lo mismo que la mediana de las fechas de muerte. La mediana de supervivencia tiene en cuenta los valores censurados y el riesgo acumulado, proporcionando una medida más precisa de la supervivencia esperada de la cohorte estudiada.

Gráfica 6. Curva de supervivencia



El gráfico 4 ilustra la proporción de pacientes con AML que sobreviven a lo largo del tiempo. Al inicio del seguimiento, la proporción de supervivencia es del 100%, ya que todos los pacientes están vivos en ese momento. En los primeros años, se observa un rápido descenso en la proporción de supervivencia, lo cual es típico en estudios de AML. Posteriormente, la curva tiende a estabilizarse, aunque sigue mostrando una disminución gradual, sugiriendo que, aunque la mortalidad continúa ocurriendo, lo hace a un ritmo más lento en los años siguientes.

Desde los 10 hasta los 15 años, la curva muestra una estabilización significativa, indicando que los pacientes que sobreviven hasta este punto tienen una mayor probabilidad de supervivencia a largo plazo. Sin embargo, el número de pacientes en esta fase es menor, como lo indica el ensanchamiento del intervalo de confianza hacia el final del seguimiento, reflejando la creciente incertidumbre sobre la proporción de supervivencia a medida que menos pacientes permanecen en seguimiento.

La curva de supervivencia sugiere que los tratamientos iniciales tienen un impacto significativo en los primeros años post-diagnóstico, donde se observa el mayor declive. Los tratamientos de consolidación y mantenimiento pueden ser clave para la estabilización de la supervivencia en los años siguientes. La estabilización de la curva en años posteriores resalta la importancia de las estrategias de manejo a largo plazo y la vigilancia continua en los pacientes supervivientes de AML. Estos hallazgos destacan la importancia del manejo intensivo inicial y las estrategias de mantenimiento a largo plazo para mejorar los resultados en estos pacientes, y pueden servir de base para futuras investigaciones y la optimización de los regímenes de tratamiento en AML.

3.2. Implementación de redes bayesianas

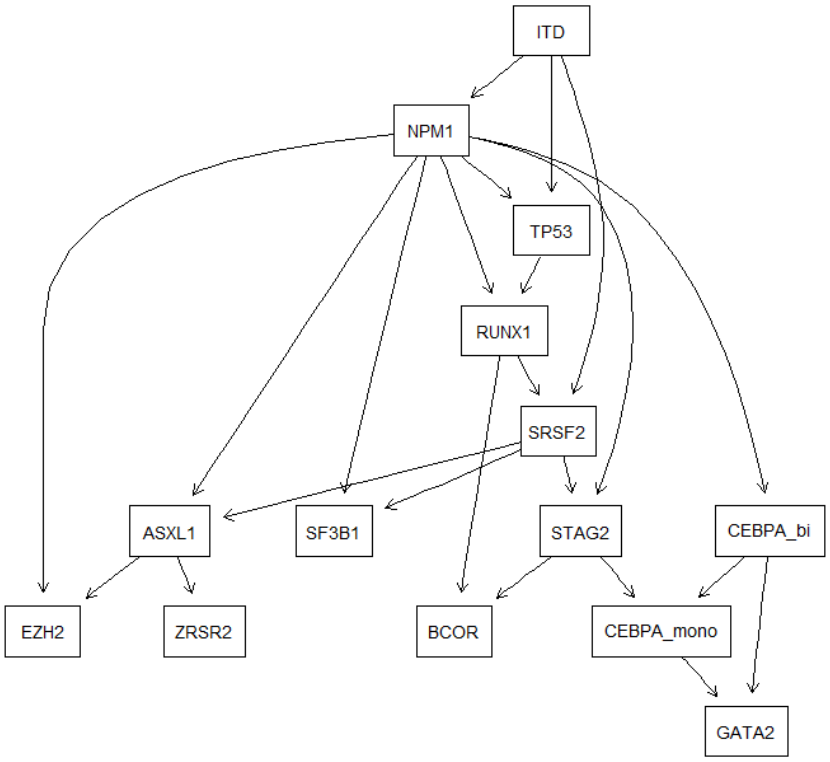
Como las variables son categóricas, y las que eran de carácter continuo se han recodificado, las redes bayesianas que se van a emplear son redes bayesianas discretas.

3.2.1. Red bayesiana asociada a los genes

La primera red construida (gráfica 5) está compuesta por los genes seleccionados y descritos anteriormente que están implicados en la AML. La estructura de la red ha sido aprendida utilizando el algoritmo de Hill-Climbing, optimizado para el criterio de información bayesiano (BIC) definido con anterioridad. Esto se ha realizado utilizando la función *hc* en R, que genera la estructura de la red basada en los datos proporcionados. Una vez obtenida la estructura, se ajusta la red a los datos mediante la función *bn.fit*, que toma dos parámetros: la estructura de la red y el conjunto de datos. Esta función ajusta los parámetros de la red bayesiana a los datos dados, permitiendo realizar inferencias posteriores y obtener una red más precisa.

La red resultante, como se puede observar, incluye 14 nodos, que representan a cada uno de los genes, y 23 arcos dirigidos. El tamaño promedio del manto de Markov es 3.86, es decir, en promedio, cada nodo tiene aproximadamente 3.86 nodos en su manto de Markov, que incluye tanto a sus padres, como a sus hijos y otros padres de sus hijos. El factor de ramificación promedio es 1.64, lo que significa que, en promedio, cada nodo influye directamente en 1.64 nodos. Este factor de ramificación nos dice cuántas conexiones directas se generan en promedio desde cada nodo hacia otros nodos, lo cual da una idea de cómo se propagan las dependencias a través de la red.

Gráfica 7. Red bayesiana de genes



La red generada revela varias relaciones clave entre los genes. El gen NPM1 aparece como un nodo central en la red, con múltiples conexiones entrantes y salientes. NPM1 está

condicionado por ITD y afecta a varios otros genes, incluyendo CEBPA_bi, TP53, RUNX1, SRSF2, SF3B1, STAG2, y ASXL1. Esto sugiere que NPM1 juega un papel crucial en la regulación de estos genes y podría ser un gen clave en la patogénesis de AML.

El modelo también sugiere vías de regulación específicas. Por ejemplo, CEBPA_bi influye en GATA2 a través de CEBPA_mono y STAG2. Este camino podría indicar una cascada regulatoria donde las mutaciones en CEBPA podrían tener efectos en cascada sobre la expresión y función de GATA2. Con un tamaño promedio del manto de Markov de 3.86 y un tamaño promedio del vecindario de 3.29, la red muestra que cada gen, en promedio, está directamente relacionado con aproximadamente tres a cuatro otros genes. Esto refleja la complejidad de las interacciones genéticas en AML y la importancia de considerar múltiples factores al estudiar la biología de esta enfermedad.

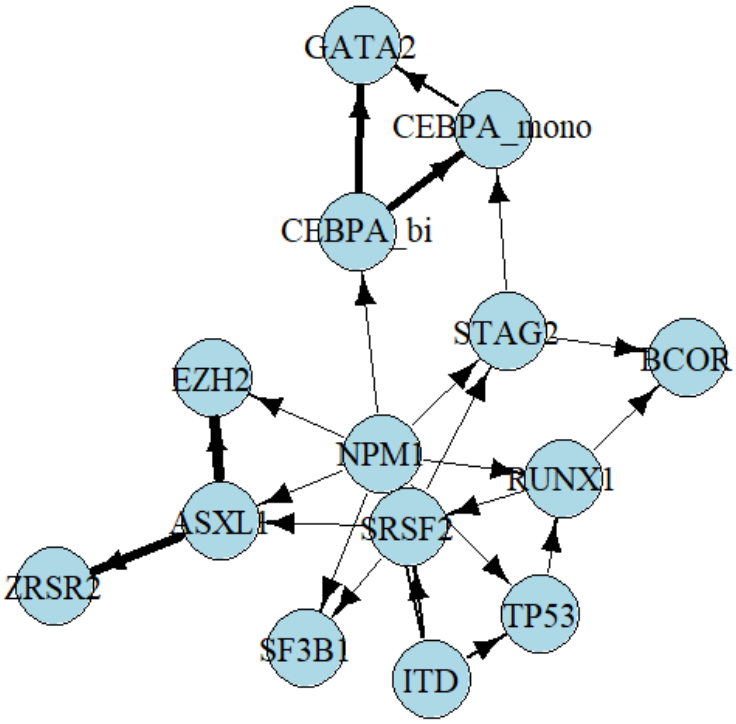
La red ajustada con grosores de flechas según la intensidad de las relaciones entre los genes representada en el gráfico 6 proporciona una visión más detallada de la fuerza de estas dependencias. Para realizar esta visualización se ha utilizado el paquete {igraph}. Primero calculé la fuerza de las aristas con la función *arc.strenght*, que toma como parámetros la estructura de la red, el conjunto de datos y el criterio de fuerza de las aristas, en este caso, *loglik* para el logaritmo de la verosimilitud. Luego, se convirtió la red aprendida a un objeto *igraph* y se ajustaron varias propiedades de los nodos y aristas para mejorar la interpretabilidad del gráfico.

Los grosores de las flechas indican la intensidad de la relación condicional entre los genes. Por ejemplo, los arcos más gruesos, como los que conectan a ASXL1 con ZRSR2, sugieren una relación fuerte entre estos genes. Esto puede indicar que la presencia de mutaciones en ASXL1 tiene una influencia significativa en la ocurrencia de mutaciones en ZRSR2.

De manera similar, el grosor del arco entre CEBPA_bi y GATA2 destaca una fuerte relación regulatoria, lo que refuerza la hipótesis de que las mutaciones en CEBPA pueden tener efectos en cascada importantes sobre la expresión y función de GATA2. Estas relaciones más fuertes pueden señalar posibles vías clave que merecen una investigación más profunda en futuros estudios.

Además, la variabilidad en los grosores de las flechas en la red sugiere que no todas las relaciones tienen la misma influencia.

Gráfica 8. Red bayesiana de genes con intensidad de relación entre ellos.



Para evaluar la robustez de la red bayesiana aprendida, se ha realizado una validación cruzada utilizando k-fold, que divide los datos en k subconjuntos y evalúa la red k veces, cada vez con un subconjunto diferente como conjunto de prueba. En este caso con 10 pliegues y 10 repeticiones utilizando la función *bn.cv*. Esta función toma varios parámetros: el conjunto de datos, la estructura de la red, el número de ejecuciones, el método de validación y el número de pliegues. La función de pérdida utilizada es la pérdida de log-verosimilitud (Log-Likelihood Loss) para datos discretos.

Los resultados de la validación cruzada muestran una pérdida promedio de 3.307346 con una desviación estándar de 0.0007469789. La baja desviación estándar indica una mínima variabilidad en la pérdida entre diferentes particiones de los datos, sugiriendo que el modelo es robusto y consistente. Esta validación confirma que la red bayesiana aprendida generaliza bien a diferentes subconjuntos de datos, proporcionando un modelo fiable para el análisis de interacciones genéticas en la AML.

Estos resultados refuerzan la confianza en la estructura de la red y las relaciones genéticas identificadas, destacando la importancia de NPM1 como un nodo central con múltiples conexiones entrantes y salientes. Las relaciones más fuertes, indicadas por los grosores de las flechas, sugieren vías reguladoras clave que merecen una mayor investigación.

Para poder validar algunas de las relaciones presentes en la red, se han realizado análisis estadísticos utilizando el test chi-cuadrado y el cálculo del odds ratio. El test chi-cuadrado se utiliza para determinar si existe una asociación significativa entre dos variables categóricas, comparando las frecuencias observadas con las esperadas. El odds ratio, por su parte, mide la fuerza de la asociación entre dos variables, indicando cuántas veces es más probable que ocurra un evento en presencia de otra variable.

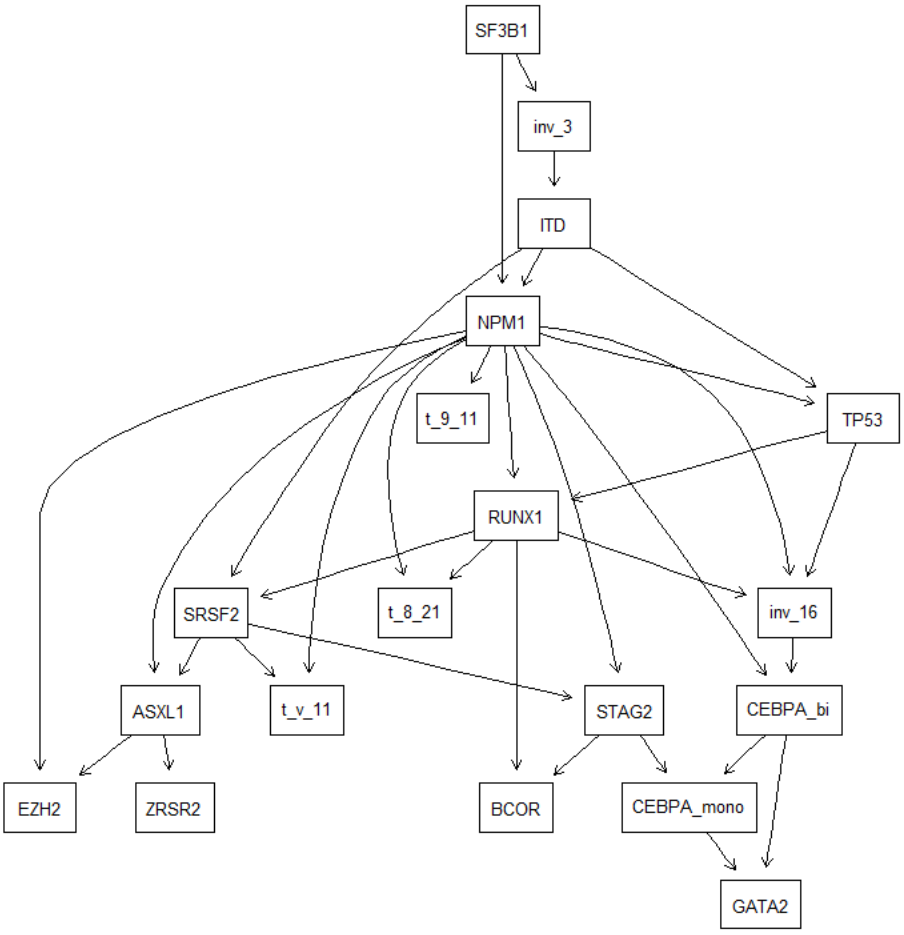
Los resultados del análisis han revelado asociaciones significativas entre varios genes en la AML. La fuerte asociación entre NPM1 e ITD, con un odds ratio de 4,44 y un p-valor

extremadamente bajo, indica que estos genes están altamente relacionados y pueden influir mutuamente en la progresión de la enfermedad. Asimismo, la relación entre ASXL1 y ZRSR2, con un odds ratio de 7,22 y p-valores muy bajos, sugiere una fuerte conexión entre estos genes.

3.2.2. Red bayesiana asociada a los genes y las anomalías

A continuación, la red bayesiana (gráfica 7) recogerá los diferentes genes además de las anomalías del cariotipo para poder ver cómo interaccionan entre ellos. Por lo tanto, en ella, cada nodo representa un gen o anomalía del cariotipo relevante en el contexto de la AML.

Gráfica 9. Red bayesiana de genes y anomalías



El modelo generado es una red bayesiana que consta de 19 nodos y 22 arcos dirigidos. La estructura de la red, al igual que en la anterior, fue aprendida utilizando el algoritmo de escalada y ha sido optimizada con el criterio BIC para datos discretos. Durante el proceso de aprendizaje, se realizaron 801 pruebas y la optimización del modelo de confirmó como verdadera.

En la red, SF3B1 se presenta como el nodo raíz, asociado con la inversión en el cromosoma 3 (inv_3), que a su vez tiene una influencia significativa sobre ITD. Al igual que en la red anterior, NPM1 destaca como un nodo central en la red. NPM1 está relacionado con

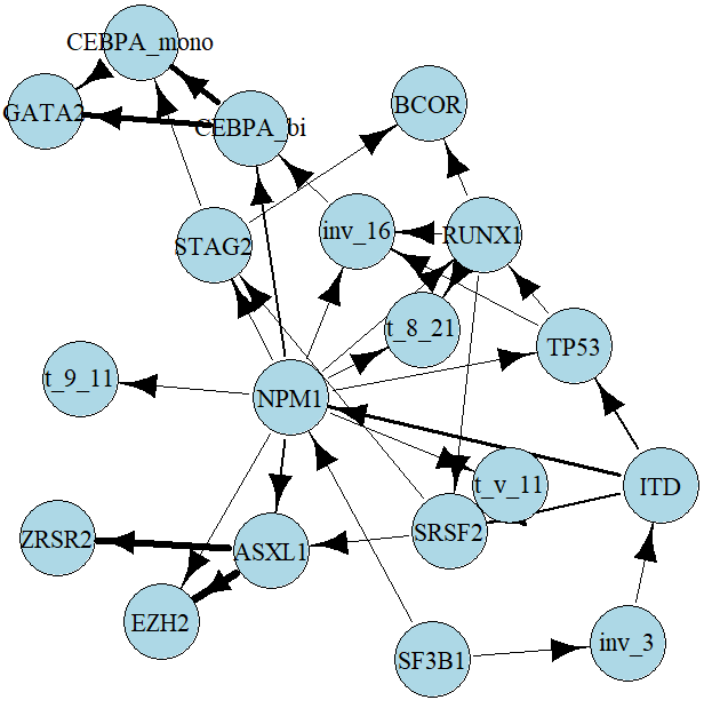
mutaciones comunes en AML y afecta a varios otros genes y anomalías como t_9_11, RUNX1, t_8_21, inv_16, ASXL1, CEBPA_bi, STAG2, t_v_11 y EZH2.

El gen TP53, que depende tanto de NPM1 como de ITD, tiene importantes conexiones con RUNX1 e inv_16, lo que subraya su papel en la progresión de la enfermedad. RUNX1, otro nodo central, afecta a SRSF2, t_8_21 y BCOR, demostrando su influencia en la red genética de la AML. SRSF2, dependiente de RUNX1 e ITD, está relacionado con ASXL1, STAG2 y t_v_11, indicando una red de interacciones complejas.

Esta red bayesiana también tiene implicaciones importantes cuando se relaciona con las normas de la European LeukemiaNet (ELN) 2022. Las normas ELN proporcionan guías para la clasificación y manejo de la AML, basándose en características genéticas y moleculares. La red puede alinearse con estas normas facilitando la comprensión del riesgo y pronóstico de los pacientes mediante la identificación de mutaciones específicas y sus interrelaciones. Además, puede guiar en la elección de terapias dirigidas, basándose en la identificación de combinaciones de mutaciones como ITD en FLT3 y mutaciones en NPM1 y TP53. Esto es esencial para personalizar el tratamiento y mejorar los resultados clínicos.

En la siguiente red (gráfica 8) se han añadido, al igual que antes, diferentes grosores en las flechas. Estos han sido calculados en función de la fuerza de los arcos calculada con el criterio de verosimilitud logarítmica. Estos valores se han normalizado y se aseguran para tener un mínimo grosor, lo que permite visualizar de manera efectiva la intensidad de las relaciones entre las variables.

Gráfica 10. Red bayesiana de genes y anomalías con diferentes grosores



En esta nueva red que se muestra en el gráfico 8, el nodo NPM1 sigue siendo el centro de la red, con múltiples conexiones a otros genes y anomalías cromosómicas. Las flechas gruesas que conectan NPM1 con ITD, ASXL1, y SRSF2 indican que las mutaciones en

NPM1 tienen una influencia particularmente fuerte sobre estas variables. Esto sugiere que los cambios en NPM1 pueden ser determinantes clave en la patogénesis y progresión de la LMA.

El nodo ASXL1 tiene arcos gruesos que lo conectan tanto con NPM1 como con SRSF2, lo que indica que las alteraciones en ASXL1 están fuertemente influenciadas por estos genes. Esto es consistente con la literatura sobre LMA, que identifica a ASXL1 como un marcador importante en la enfermedad.

SRSF2 está conectado de manera fuerte con NPM1 y ASXL1, como se indica por las flechas gruesas. Estas conexiones sugieren que SRSF2 es un nodo importante en la red de interacciones genéticas de la LMA, influenciado significativamente por mutaciones en otros genes críticos.

CEBPA_bi y CEBPA_mono están conectados con flechas gruesas, indicando una fuerte relación entre las mutaciones bialélicas y monoalélicas en CEBPA. Además, la conexión gruesa entre CEBPA_bi y GATA2 sugiere que estas interacciones también juegan un papel crucial en la función hematopoyética y la progresión de la AML, como se pudo ver también en la anterior red.

El nodo RUNX1 tiene una conexión importante con NPM1, resaltada por una flecha gruesa. Esta relación indica que RUNX1 es fuertemente influenciado por las mutaciones en NPM1, lo que es consistente con su papel conocido en la regulación de la diferenciación y proliferación celular.

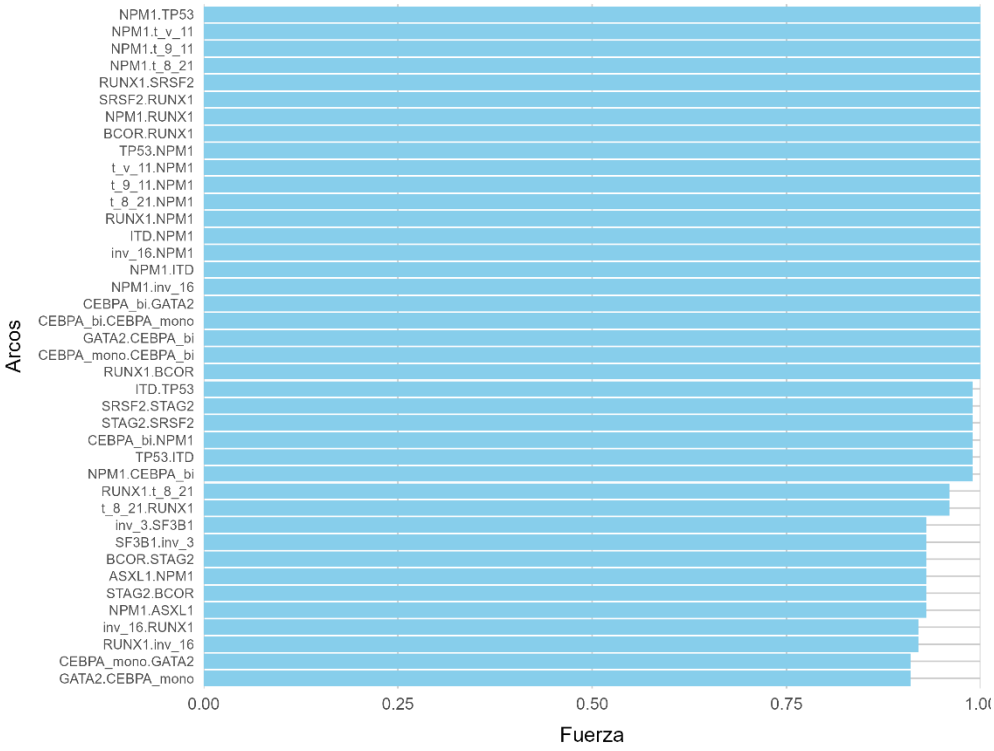
t_8_21 y inv_16 están conectados a RUNX1 y NPM1 con flechas gruesas, lo que indica que estas translocaciones cromosómicas son particularmente relevantes en la red genética de la LMA. Estas conexiones sugieren que t_8_21 e inv_16 son alteraciones críticas que deben ser monitorizadas de cerca en pacientes con estas características genéticas.

Después de analizar las relaciones dadas en la red, se ha realizado validación cruzada con 10 pliegues y 10 ejecuciones. La pérdida promedio ha sido de 3.867753, que indica que el modelo es bastante preciso en la predicción de las probabilidades condicionales. Además, la desviación estándar es extremadamente baja (0.002829), lo que sugiere que el modelo es consistente a través de diferentes particiones de los datos.

Se han utilizado técnicas de bootstrapping para evaluar la estabilidad de los arcos en la red. En este caso, se ha determinado que se generarán 100 iteraciones de bootstrapping, y que solo los arcos con una fuerza mayor o igual a 0.85 se incluirán en la red promedio (gráfica 9). Los arcos con una fuerza cercana a 1 deben considerarse los más confiables, mientras que los arcos con fuerzas bajas no son estables y deben ser revisados.

El gráfico 9 muestra los arcos más estables en la red bayesiana, determinados a través de un análisis de bootstrapping. El eje X representa la fuerza del arco, que varía de 0 a 1, indicando la proporción de veces que el arco aparece en las redes generadas a partir de muestras Bootstrap. El eje Y lista los arcos en forma de “nodo origen – nodo destino”, donde cada entrada corresponde a una relación condiciona entre dos genes o anomalías cromosómicas.

Gráfica 11. Arcos más fuertes en la red bayesiana (bootstrapping)



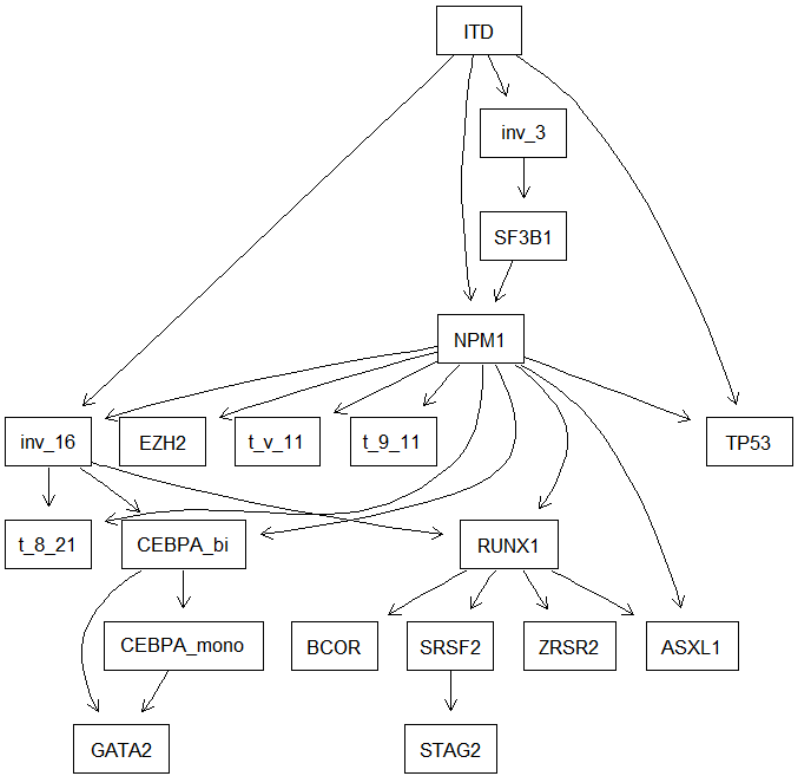
Los arcos más estables generalmente incluyen el gen NPM1, lo que destaca su importancia en la patogénesis de AML. Este, muestra múltiples conexiones fuertes con otros genes y anomalías cromosómicas como TP53, t_v_11, t_9_11 y t_8_21, subrayando su papel central en la red genética. RUNX1 también muestra conexiones fuertes con varios genes, sugiriendo su papel regulador significativo.

Por otra parte, las relaciones entre GATA2 y CEBPA_mono están representadas a través de arcos con fuerzas bajas, por lo que deben ser interpretados con cautela, ya que su aparición inconsistente sugiere una menor estabilidad. Esto indica una menor relevancia en la red general.

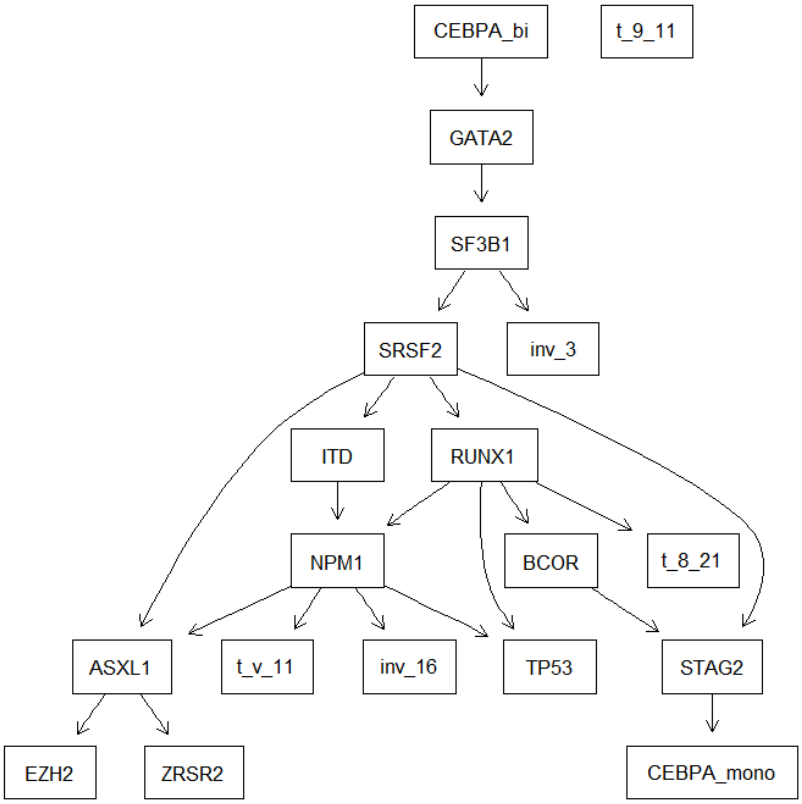
3.2.3. Red bayesiana por grupos de edad

Para poder ver si existen diferencias notables en la estructura y en las interacciones entre los genes y las anomalías del cariotipo para los pacientes de los diferentes grupos de edad, se van a analizar las redes bayesianas referidas a los pacientes menores y a los mayores de 60 años.

Gráfica 12. Red bayesiana para pacientes menores de 60 años



Gráfica 13. Red bayesiana para pacientes con 60 años o más



En este análisis comparativo se visualizan diferencias significativas en la estructura y las interacciones entre los genes y las anomalías cromosómicas. En la red correspondiente a los pacientes menores de 60 años, el gen NPM1 se posiciona como un nodo central, debido a sus numerosas conexiones entrantes y salientes. Esta centralidad sugiere que NPM1 desempeña un papel importante en la regulación genética de pacientes de este grupo de edad.

En esa misma red, el gen ITD se presenta como un nodo influyente con conexiones directas a otros genes, incluyendo NPM1. Esta interacción directa sugiere que las mutaciones en ITD podrían tener un impacto significativo en la actividad de NPM1 y, por lo tanto, en la regulación de otros genes relacionados con la AML. Además, NPM1 muestra interacciones complejas con otros genes importantes como RUNX1, EZH2 y CEBPA_bi, lo que indica una red densa de interacciones que podrían ser críticas para la progresión de la enfermedad. Las conexiones de NPM1 con varias anomalías cromosómicas, como inv_16 y t_9_11, también reflejan su papel central en la patogénesis de la AML en pacientes más jóvenes.

En esta red, los genes CEBPA_bi y CEBPA_mono están involucrados en rutas reguladoras significativas, destacando especialmente la influencia de CEBPA_mono sobre GATA2. Este tipo de interacción sugiere una cascada reguladora donde las mutaciones en CEBPA podrían afectar la expresión y función de GATA2, otro gen clave en la hematopoyesis. En general, la estructura de la red para los menores de 60 años es relativamente simple y directa, con menos capas de interacciones entre los genes, lo que podría reflejar una dinámica de regulación genética más lineal en este grupo de edad.

Por otro lado, la red bayesiana para los pacientes mayores de 60 años revela un panorama diferente. En esta red, SRSF2 emerge como un nodo central, sugiriendo que este gen tiene un papel más prominente en la patogénesis de la AML en pacientes mayores. La estructura de la red es notablemente más jerárquica y compleja, con múltiples capas de interacciones entre los genes. Por ejemplo, el gen CEBPA_bi influye en GATA2, que a su vez regula SF3B1, y este último tiene un impacto significativo en SRSF2. Esta jerarquía indica una regulación más compleja y posiblemente una mayor redundancia en las rutas genéticas en pacientes mayores.

RUNX1 es otro nodo influente en esta red, mostrando conexiones con genes críticos como NPM1, BCOR y TP53, además de varias anomalías cromosómicas como t_8_21 e inv_16. La red para los mayores de 60 años también presenta conexiones de anomalías cromosómicas más diversas. Por ejemplo, inv_3 tiene una conexión directa con SF3B1, y t_8_21 está directamente relacionado con RUNX1 y TP53. Estas conexiones más diversas y complejas pueden reflejar la acumulación de anomalías cromosómicas y mutaciones genéticas adicionales que se producen con el envejecimiento, lo que lleva a una red de regulación genética más complicada.

Además, aparecen nuevas conexiones en la red de los mayores de 60 años que no están presentes en la red de los menores de 60 años. Por ejemplo, la conexión directa de ASXL1 con EZH2 y ZRSR2 indica diferencias en los mecanismos genéticos de la enfermedad entre los dos grupos de edad. Estas nuevas interacciones pueden ser indicativas de rutas reguladoras adicionales o alternativas que se activan en pacientes mayores, sugiriendo una adaptación del sistema genético a las condiciones de envejecimiento y la progresión de la enfermedad.

En conclusión, NPM1 se destaca como el nodo central en la red de los menores de 60 años, mientras que SRSF2 y RUNX1 toman roles centrales en la red de los mayores de 60

años. La red para los mayores de 60 años es más jerárquica y compleja, lo que puede reflejar una mayor complejidad en la regulación genética y la progresión de la AML en este grupo de edad. Las diferencias observadas en las interacciones genéticas y las anomalías cromosómicas entre los dos grupos sugieren que los mecanismos subyacentes de la AML varían según la edad del paciente. Estas observaciones tienen implicaciones clínicas importantes, ya que sugieren que las estrategias de tratamiento para la AML podrían beneficiarse de una personalización basada en la edad del paciente. Por lo tanto, entender estas diferencias podría ayudar a desarrollar terapias más efectivas y específicas para cada grupo de edad, mejorando así los resultados clínicos para los pacientes con AML.

Después, se ha realizado la validación cruzada k-fold para los dos grupos de edad para evaluar la robustez y el rendimiento de las redes bayesianas. En la siguiente tabla se puede ver la pérdida promedio y la desviación estándar asociada a cada grupo de pacientes.

	Pérdida promedio	Desviación estándar
Menores de 60 años	3.607	0.00177
Mayores de 60 años	4.192	0.00344

Ambos modelos muestran una desviación estándar baja en la pérdida, lo que indica que son robustos y estables. Sin embargo, el modelo para los pacientes menores de 60 años tiene una pérdida promedio menor, sugiriendo una mayor precisión.

La red para los pacientes de 60 años o más es más jerárquica y compleja, lo que puede reflejar una mayor diversidad y acumulación de mutaciones genéticas y anomalías cromosómicas en este grupo de edad.

Además, las diferencias en las estructuras de las redes bayesianas y las pérdidas promedio entre los dos grupos de edad pueden tener implicaciones clínicas importantes. Los tratamientos para la AML podrían beneficiarse de una personalización basada en la edad del paciente.

4. Conclusiones

Durante la realización de este trabajo de fin de grado he tenido la oportunidad de profundizar en el estudio de las redes bayesianas, una herramienta poderosa y versátil para el análisis de datos complejos. Este tipo de redes permite modelizar relaciones de dependencia probabilística entre variables, proporcionando una forma intuitiva, y de carácter estadístico, de comprender la estructura y las interacciones dentro de los datos. Uno de sus aspectos más relevantes para el análisis del genoma en AML es su capacidad para manejar la incertidumbre inherente a los datos biológicos.

El uso de R como herramienta de análisis ha sido un aspecto crucial de este proyecto. A través de su aplicación, he aprendido a implementar redes bayesianas. Además, su extensa biblioteca de paquetes especializados en análisis estadístico y aprendizaje automático, han sido cruciales para alcanzar los objetivos concretos expuestos en este estudio. En resumen, la realización de este proyecto me ha permitido lograr los objetivos planteados inicialmente, adquiriendo nuevos conocimientos técnicos y analíticos.

Como conclusiones específicas del trabajo:

Las redes bayesianas construidas en este estudio han revelado varias relaciones clave entre los genes implicados en la AML. En particular, el gen NPM1 se destaca como un nodo central con múltiples conexiones fuertes con otros genes y anomalías cromosómicas. Esto sugiere que NPM1 juega un papel crucial en la regulación de estos genes y podría ser un gen clave en la patogénesis de la AML. Las conexiones fuertes de NPM1 con genes como ITD, ASXL1, y SRSF.

Además, el análisis de la red con grosores de flechas según la intensidad de las relaciones ha proporcionado una visión más detallada de la fuerza de estas dependencias. Las relaciones más fuertes, indicadas por los arcos gruesos, como las que están presentes entre ASXL1 y ZRSR2, o entre CEBPA_bi y GATA2, sugieren vías reguladoras clave que merecen una investigación más profunda. Esto puede indicar que la presencia de mutaciones en genes como ASXL1 y CEBPA tiene efectos en cascada importantes sobre otros genes, lo que podría influir en el pronóstico y tratamiento de la AML.

La robustez del modelo ha sido validada mediante una validación cruzada, mostrando una baja variabilidad en la pérdida de log-verosimilitud entre diferentes particiones de los datos, lo que sugiere que el modelo es consistente y fiable. Esto refuerza la confianza en la estructura de la red y las relaciones genéticas identificadas.

En cuanto al análisis comparativo con las redes bayesianas para pacientes menores de 60 años y para los mayores de esa edad, se observan diferencias significativas en la estructura y las interacciones entre los genes y las anomalías cromosómicas en pacientes con menos de 60 años y aquellos con más de 60 años. En los pacientes menores de 60 años, NPM1 emerge como un nodo central con múltiples conexiones, indicando su papel crucial en la regulación genética en este grupo de edad. La red para este grupo es relativamente simple y directa, reflejando una dinámica de regulación genética más lineal. Por el contrario, en los pacientes mayores de 60 años, SRSF2 y RUNX1 toman roles centrales en una red más jerárquica y compleja. Esto sugiere una mayor acumulación de anomalías cromosómicas y mutaciones con la edad, resultando en una regulación genética más intrincada.

Como futuras líneas de trabajo, para continuar avanzando en esta línea de investigación, se propone ampliar el análisis a otras enfermedades hematológicas y

oncológicas, lo que permitirá comparar las estructuras de redes bayesianas y las interacciones genéticas específicas de cada patología. Además, se sugiere incorporar variables adicionales como recaídas, trasplantes, medicamentos aplicados y tiempos de supervivencia para enriquecer el análisis y obtener una visión más completa de los factores que influyen en la progresión y tratamiento de la AML. También sería beneficioso incluir variables continuas de genes, como las frecuencias alélicas variantes (VAFs), para capturar de manera más precisa la heterogeneidad genética y su impacto en la red de interacciones.

5. Bibliografía

Aboytes–Ojeda, M., Laureano-Cruces, A. L., & Ramírez-Rodríguez, J. (2013). Algoritmo de búsqueda tabú para una variante del problema de coloración. *Revista de Matemática: Teoría y Aplicaciones*, *20*(2), 215-230. https://doi.org/10.15517/rmta.v20i2.11661

Andrés Mañas, M. Á. (2017). *Clustering en redes bayesianas. Implementación en R*. https://repositorio.ual.es/handle/10835/5909

Angelopoulos, N., Chatzipli, A., Nangalia, J., Maura, F., & Campbell, P. J. (2022). Bayesian networks elucidate complex genomic landscapes in cancer. *Communications Biology*, *5*(1), 1-11. https://doi.org/10.1038/s42003-022-03243-w

Angulo Montes, L. E. (2020). *Redes bayesianas en R: Análisis de los paquetes software disponibles* [Masters, E.T.S. de Ingenieros Informáticos (UPM)]. https://oa.upm.es/63644/

Arias, M. (2022, abril 20). *Estudios de supervivencia. Método de Kaplan-Meier*. *18*(2). https://evidenciasenpediatria.es/articulo.php?lang=es&id=7991&tab=

Arnold, B. C., Castillo, E., & Sarabia, J. M. (Eds.). (1999). Bayesian Analysis Using Conditionally Specified Models. En *Conditional Specification of Statistical Models* (pp. 293-336). Springer. https://doi.org/10.1007/978-0-387-22588-3_13

Atienza González, D. (2021). *Nonparametric models and bayesian networks: Applications to anomaly detection* (p. 1) [Http://purl.org/dc/dcmitype/Text, Universidad Politécnica de Madrid]. https://dialnet.unirioja.es/servlet/tesis?codigo=305042

Bouza, C. (2018). *MODELOS DE REGRESIÓN Y SUS APLICACIONES*.

Callejas Pinilla, P. F. (2020). *Identificación de parámetros operacionales críticos en el rendimiento de camiones mediante redes bayesianas: Los Bronces, Anglo American S.A.* https://repositorio.uchile.cl/handle/2250/175559

Canals L., M. (2019). Bases científicas del razonamiento clínico: Inferencia Bayesiana. *Revista médica de Chile*, *147*(2), 231-237. https://doi.org/10.4067/s0034-98872019000200231

Chickering, D. M., Meek, C., & Heckerman, D. (2012). *Large-Sample Learning of Bayesian Networks is NP-Hard* (arXiv:1212.2468). arXiv. https://doi.org/10.48550/arXiv.1212.2468

Devine, S. M., & Larson, R. A. (1994). Acute leukemia in adults: Recent developments in diagnosis and treatment. *CA: A Cancer Journal for Clinicians*, *44*(6), 326-352. https://doi.org/10.3322/canjclin.44.6.326

Döhner, H., Wei, A. H., Appelbaum, F. R., Craddock, C., DiNardo, C. D., Dombret, H., Ebert, B. L., Fenaux, P., Godley, L. A., Hasserjian, R. P., Larson, R. A., Levine, R. L., Miyazaki, Y., Niederwieser, D., Ossenkoppele, G., Röllig, C., Sierra, J., Stein, E. M., Tallman, M. S., … Löwenberg, B. (2022). Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood*, *140*(12), 1345-1377. https://doi.org/10.1182/blood.2022016867

Domènech Pellejà, B. (2021). *Homicidios y drogas: Un estudio basado en redes bayesianas*. https://ddd.uab.cat/record/237043

Estruch, V. D., Planes, F. J. B., Vidal, A., & Pastor, J. I. (2019). Redes bayesianas y diagnóstico médico. Una forma diferente de aprender probabilidades condicionadas. *Modelling in Science Education and Learning*, *12*(2), Article 2. https://doi.org/10.4995/msel.2019.10830

Fenton, N. E., & Neil, M. (1999). Software metrics: Successes, failures and new directions. Journal of Systems and Software, *47*(2), 149-157. https://doi.org/10.1016/S0164-1212(99)00035-7

Fernández Regalado, R. (2009). El teorema de Bayes y su utilización en la interpretación de las pruebas diagnósticas en el laboratorio clínico. *Revista Cubana de Investigaciones Biomédicas*, *28*(3), 158-165.

Haukoos, J. S., & Lewis, R. J. (2005). Advanced Statistics: Bootstrapping Confidence Intervals for Statistics with "Difficult" Distributions. *Academic Emergency Medicine*, *12*(4), 360-365. https://doi.org/10.1197/j.aem.2004.11.018

Hernández Leal, P. F. (2011). Algoritmo de Aprendizaje para Redes Bayesianas de Nodos Temporales. *2011*, 115.

Kenett, R. S. (2012). *Applications of Bayesian Networks* (SSRN Scholarly Paper 2172713). https://doi.org/10.2139/ssrn.2172713

Lagunas-Rangel, F. A. (2016). Leucemia mieloide aguda. Una perspectiva de los mecanismos moleculares del cáncer. *Gaceta Mexicana de Oncología*, *15*(3), 150-157. https://doi.org/10.1016/j.gamo.2016.05.007

Layo González, I. (2022, junio). *Análisis e implementación del operador de Bayes multivariante en redes Bayesianas, marginalización aproximada vía simulación* [Info:eu-repo/semantics/bachelorThesis]. E.T.S. de Ingenieros Informáticos (UPM). https://oa.upm.es/71469/

Lozano, M. R. (2011). *EL PAPEL DE LAS REDES BAYESIANAS EN LA TOMA DE DECISIONES*.

Ma, S. X., Dhanaliwala, A. H., Rudie, J. D., Rauschecker, A. M., Roberts-Wolfe, D., Haddawy, P., & Kahn, C. E. (2023). Bayesian Networks in Radiology. *Radiology: Artificial Intelligence*, *5*(6), e210187. https://doi.org/10.1148/ryai.210187

Mappe Rojas, K. A. (2019). *Evaluación del desempeño de tres algoritmos de inferencia bayesiana, implementados como sistema experto para la identificación de modos de falla en ejes*. https://repositorio.unal.edu.co/handle/unal/75649

Margaritis, D. (s. f.). *Learning Bayesian Network Model Structure from Data*.

Mellado Cabrerizo, E. (2022). *Estimación paramétrica y validación en redes bayesianas*. https://idus.us.es/handle/11441/134585

Módulo 7. Redes bayesianas—Redes bayesianas Ramon Sangüesa i Solé PID_ Introducción Las redes—Studocu. (s. f.). Recuperado 25 de abril de 2024, de https://www.studocu.com/ca-es/document/universitat-oberta-de-catalunya/mineria-de-datos/modulo-7-redes-bayesianas/6620026

Molina Serrano, B., González-Cancelas, N., Soler-Flores, F., Molina Serrano, B., González-Cancelas, N., & Soler-Flores, F. (2018). Gestión de la sostenibilidad portuaria basada en un modelo de redes bayesianas. Aplicación al sistema portuario español. *Ingeniare. Revista chilena de ingeniería*, *26*(4), 631-644. https://doi.org/10.4067/S0718-33052018000400631

Nagarajan, R., Scutari, M., & Lèbre, S. (2013). *Bayesian Networks in R: With Applications in Systems Biology*. Springer. https://doi.org/10.1007/978-1-4614-6446-4

Neapolitan, R. (2003). *Learning Bayesian Networks*. https://doi.org/10.1145/1327942.1327961

Prada-Arismendy, J., Arroyave, J. C., & Röthlisberger, S. (2017). Molecular biomarkers in acute myeloid leukemia. *Blood Reviews*, *31*(1), 63-76. https://doi.org/10.1016/j.blre.2016.08.005

Quintero Sierra, Y., Hernández Padrón, C., Concepción Fernández, Y., Quintero Sierra, Y., Hernández Padrón, C., & Concepción Fernández, Y. (2021). Leucemia mieloide aguda: Influencia pronóstico de algunos biomarcadores y la respuesta terapéutica en los pacientes menores de 60 años. *Revista Cubana de Hematología, Inmunología y Hemoterapia*, *37*(3). http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S0864-02892021000300006&lng=es&nrm=iso&tlng=es

Rausch, C., Rothenberg-Thurley, M., Dufour, A., Schneider, S., Gittinger, H., Sauerland, C., Görlich, D., Krug, U., Berdel, W. E., Woermann, B. J., Hiddemann, W., Braess, J.,

von Bergwelt-Baildon, M., Spiekermann, K., Herold, T., & Metzeler, K. H. (2023). Validation and refinement of the 2022 European LeukemiaNet genetic risk stratification of acute myeloid leukemia. *Leukemia*, *37*(6), 1234-1244. https://doi.org/10.1038/s41375-023-01884-2

Rodríguez, D., & Dolado, J. (s. f.). *Redes Bayesianas en la Ingeniería del Software*.

Rojas, J. C. S., Pérez, D. U., & Reyes, C. E. H. (2012, junio 21). *Definición de Redes Bayesianas y sus aplicaciones*. Revista Vinculando. https://vinculando.org/articulos/redes-bayesianas.html

Romero Núñez, M. (2020). *Introducción a las Redes Bayesianas*. https://idus.us.es/handle/11441/115167

Sangüesa i Solé, R. (s. f.). *Módulo 7. Redes bayesianas*. https://www.studocu.com/ca-es/document/universitat-oberta-de-catalunya/mineria-de-datos/modulo-7-redes-bayesianas/6620026

Scutari, M., Silander, T., & Ness, R. (2024). *bnlearn: Bayesian Network Structure Learning, Parameter Learning and Inference* (4.9.4) [Software]. https://cran.uvigo.es/web/packages/bnlearn/

Sucar, L. E. (2006). *Redes Bayesianas*. https://www.cs.cinvestav.mx/SeminarioComputo/2010/resumenEnriqueSucar.html

Susi García, R. (2007). *Análisis de sensibilidad en redes Bayesianas Gaussianas* [Http://purl.org/dc/dcmitype/Text, Universidad Complutense de Madrid]. https://dialnet.unirioja.es/servlet/tesis?codigo=17292

Tablada, C. J., & Torres, G. A. (2009). Redes Neuronales Artificiales. *Revista de Educación Matemática*, *24*(3), Article 3. https://doi.org/10.33044/revem.10280

Tazi, Y., Arango-Ossa, J. E., Zhou, Y., Bernard, E., Thomas, I., Gilkes, A., Freeman, S., Pradat, Y., Johnson, S. J., Hills, R., Dillon, R., Levine, M. F., Leongamornlert, D., Butler, A., Ganser, A., Bullinger, L., Döhner, K., Ottmann, O., Adams, R., … Papaemmanuil, E. (2022). Unified classification and risk-stratification in Acute Myeloid Leukemia. *Nature Communications*, *13*, 4622. https://doi.org/10.1038/s41467-022-32103-8

Tirado Ríos, N. R., Triana Litardo, F. E., & Saa Saltos, J. W. (2016). Optimización de Redes Bayesianas basado en técnicas de aprendizaje por inducción. *Revista Publicando*, *3*(9), 41-60.

Tratamiento de la leucemia mieloide aguda—NCI (nciglobal,ncienterprise). (2024, abril 5). [pdqCancerInfoSummary]. https://www.cancer.gov/espanol/tipos/leucemia/paciente/tratamiento-lma-adultos-pdq