TRABAJO FIN DE GRADO

Introducción a las Redes Bayesianas

Realizado por: Marta Romero Núñez

Supervisado por: Eduardo Conde Sánchez

FACULTAD DE MATEMÁTICAS
DPTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Índice general

Abstract

One of the more relevant purposes of the Statistical Modeling is that of describing probabilistic relations among a set of random variables and show them in a meaningful format. Bayesian Networks (BNs) combine a modular representation of the joint statistical distribution of the random vector under study with a powerful graphical tool allowing the identification of statistical dependencies by direct observation of the network structure. Despite the inherent difficulties of managing large sets of interrelated variables with a huge set of parameters, BNs have been gaining increasing relevance in the set of the Statistical applications in fields as diverse as medical diagnosis, insurance management tools, decision making or engineering. The computational drawbacks encountered by early BN applications are also being overcame due to the numerical capabilities of modern computers. These factors, together with the flexibility of BNs to be integrated in general Decision-Making Support Systems, make us to think that these techniques will continue spreading into the field of Statistical and Operational Research applications.

In this study we present the introductory elements of BNs. First, we analyse the structure of the network and the statistical relations induced by its topology. We limit ourselves to the case of discrete random vectors. Although, they represent only a part, this set of random vectors cover a large amount of practical situations in the fields mentioned above. In the first chapter we also discuss the non-uniqueness of the graphical representation of a given joint distribution. In general, we have an equivalence relation in the set of possible networks and only a graphical representant of the corresponding equivalence class is needed in order to model the problem.

In actual applications, the BN can be given by experts or learned from data. In this last case, we distinguish between the topological learning of the structure and the quantitative inference of the set of parameters modeling the conditional probabilities. Different techniques can be used to learn the topological structure of the network. None of them is fully satisfactory. However, once the structure of the network has been fixed, the statistical estimation of the parameters is given by more consolidated techniques, usually *maximum likelihood estimation*.

The second chapter is devoted to explain how can be used a BN in order to implement *Probabilistic Reasoning and Evidences*. Basically, once that one or more random variables have been *instantiated*, that is, we have certain realizations for the corresponding variables, the network is used to compute the posterior probabilities given such an evidence. After the updating process one can infer the most probable explanation of that evidence. Maximum a posteriori queries are concerned with finding the configuration of the variables in a given set that has the highest posterior probability. This is the basis of the probabilistic reasoning we can develop through a BN.

In the third chapter, we show how to apply these concepts into an illus-

trative *academic* application in the context of the nowadays COVID-19 pandemic situation in Spain. We have based our application in the report of the Red Nacional de Vigilancia Epidemiológica, or RENAVE, of the middle of last April. We known the existing difficulties to have access to a reliable and extensive data-basis about the actual status of the disease, hence we used the data collected in that report to develop our experiment. By using the numerical tables of that report we identified a set of variables concerning different symptoms, treatment, patient features and final results of the disease. This quantitative information was also used to make a *subjective* estimation of the conditional probabilities in the same way as it could be done by an expert. Of course, the limit of our application is just to exemplify concepts and procedures previously discussed in this memory. The resulting BN was used to simulate a data set of realizations of the random vector. This data was taken as the input of the learning procedures in order to show the validity of these methods to rebuild the BN taken as target. We also develop tasks of probabilistic explanation of the evidence in our BN.

In the last part of this work, we have added some appendices containing definitions of graph elements, the R scripts written for the above numerical application and the technical report used in the study.

Capítulo 1

Introducción a las Redes Bayesianas

En este capítulo, nuestro propósito es encontrar una representación adecuada mediante un grafo o red de una estructura probabilística de un vector aleatorio. Queremos encontrar dicha representación para poder usar las propiedades topológicas que induce un grafo en el contexto del comportamiento aleatorio que existe entre las variables. Esta representación no es única, nosotros elegiremos una y trabajaremos con ella.

1.1. Definiciones y propiedades esenciales

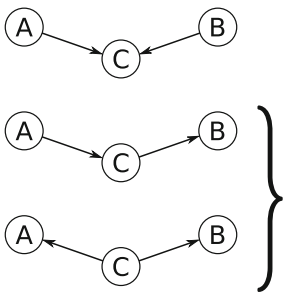
Las redes Bayesianas son una clase de grafos que permiten una representación concisa de las dependencias probabilisticas entre un conjunto dado de variables aleatorias X=$\{X_1,X_2,...,X_p\}$ como un grafo acíclico dirigido (DAG) G=(V,A). Cada nodo $v_i \in V$ corresponde a una variable aleatoria X_i.

1.1.1. Proponiendo una representación en forma de red

Partimos de una función de probabilidad conjunta factorizada de la siguiente forma

$$P_X(X) = \prod_{i=1}^{p} P_{X_i}(X_i|\Pi_{X_i}) \tag{1.1}$$

donde Π_{X_i} es un subconjunto de variables que condicionan el comportamiento probabilístico de la variable X_i (véase figura 1.1) .



$$\mathbf{A} \not\perp\!\!\!\perp_G \mathbf{B} \,|\, \mathbf{C}$$
$$\mathrm{P}(\mathbf{A},\mathbf{B},\mathbf{C}) = \mathrm{P}(\mathbf{C} \,|\, \mathbf{A},\mathbf{B})\,\mathrm{P}(\mathbf{A})\,\mathrm{P}(\mathbf{B})$$

$$\mathbf{A} \perp\!\!\!\perp_G \mathbf{B} \,|\, \mathbf{C} \Rightarrow \mathbf{A} \perp\!\!\!\perp_P \mathbf{B} \,|\, \mathbf{C}$$
$$\begin{aligned}\mathrm{P}(\mathbf{A},\mathbf{B},\mathbf{C}) &= \\ &= \mathrm{P}(\mathbf{B} \,|\, \mathbf{C})\,\mathrm{P}(\mathbf{C} \,|\, \mathbf{A})\,\mathrm{P}(\mathbf{A}) \\ &= \mathrm{P}(\mathbf{A} \,|\, \mathbf{C})\,\mathrm{P}(\mathbf{B} \,|\, \mathbf{C})\,\mathrm{P}(\mathbf{C})\end{aligned}$$

Figura 1.1: Conexión convergente, conexión en serie y conexión divergente.

En el figura 1.1 tenemos dos familias de funciones de probabilidad, donde la segunda familia está asociada a dos posibles factorizaciones distintas, junto a ellas encontramos una posible representación de cada factorización a la que llamaremos red Bayesiana. Se puede observar que a factorizaciones distintas le corresponden redes distintas. La red se forma tomando como nodos las variables de la factorización correspondiente. Estas factorizaciones se obtienen a partir de la fórmula producto, mediante la cual vamos a poder escribir cualquier función de probabilidad como productos de condicionadas. Traduciendo las probabilidades condicionadas en arcos, vamos a completar la red. Así, vemos que si en la factorización nos aparece una variable A condicionada a una variable B, en la red aparecerá un arco cuyo nodo inicial será B y su nodo final será A. Una vez obtenida nuestra red podremos estudiar las propiedades de ella, una de las que nos interesa es la independencia condicionada, la cual vamos a estudiar a continuación.

Tomemos el segundo caso de la imagen, la función de probabilidad que tenemos es:

$$P(A,B,C) = P(B|C)P(C|A)P(A)$$

queremos ver si A y B son independientes condicionalmente dado C, es decir, queremos ver si se verifica la fórmula de independencia condicionada

$$P(AB|C) = P(A|C)P(B|C) \tag{1.2}$$

Por la fórmula de Bayes tenemos que

$$P(AB|C) = \frac{P(ABC)}{P(C)}$$

y a su vez sustituyendo *P(ABC)* por su factorización, concluimos que

$$\frac{P(B|C)P(C|A)P(A)}{P(C)} = P(A|C)P(B|C)$$

Por lo tanto, deducimos que A y B son independiente condicionalmente dado C.

En el tercer caso, ocurre algo similar. Tenemos que la función de probabilidad es *P(A,B,C)= P(A|C)P(B|C)P(C)*. Veamos si A y B son independientes condionalmente dado C, es decir si se cumple la ecuación (1.2). Partamos de *P(A,B|C)*, usamos la fórmula de Bayes y sustituimos por la factorización que nos dan y nos queda que

$$P(AB|C) = \frac{P(ABC)}{P(C)} = \frac{P(A|C)P(B|C)P(C)}{P(C)}$$

Luego obtenemos que

$$P(AB|C) = P(A|C)P(B|C)$$

por tanto deducimos que A y B son independientes condicionalmente dado C.

Sin embargo con el primer caso no ocurre lo mismo, por lo que no podemos garantizar la independencia condicionada de A y B. Pero tampoco podemos afirmar que no sean independientes condicionalmente dado C, ya que podría darse el caso en el que A, B, C sean independientes dando esta factorización. Esto es así, porque la familia de probabilidad conjunta que factoriza como *P(A,B,C)=P(C|A,B)P(A)P(B)* incluye a la familia de probabilidades conjuntas que factoriza como *P(A,B,C)=P(A)P(B)P(C)* ya que introducir un arco en el grafo nos aumenta la familia de factorizaciones existentes de acuerdo a esta red.

La independencia condicionada de A y B dado C en la segunda y tercera red se puede asociar a que el único camino que une A con B pasa por C, siendo C un nodo de arcos divergentes o de arcos en series.

En lo que sigue vamos a usar cierta terminología específica para un grafo, comencemos con un término esencial.

Definición 1.1.1 (Camino no dirigido) *Sea* G=(V,A) *se define un camino no dirigido entre* v_{i1} y v_{ik} *a una secuencia de nodos* v_{i1},v_{i2},..,v_{ik} *de forma que existe el arco* $(v_{ij},v_{ij+1}) \in A$ *o* $(v_{ij+1},v_{ij}) \in A$

También utilizaremos la siguiente notación para los nodos.

Definición 1.1.2 *Se define un nodo de arcos convergentes a un nodo tal que dado un camino dirigido solo posee arcos incidentes. Similarmente, un nodo que solo tiene arcos salientes, se le llama nodo de arcos divergentes.*
Se dice que un nodo es un nodo de arcos en serie si dado un camino dirigido,

posee un solo arco incidente y un solo arco saliente.

Se llama v-estructura a las conexiones entre tres nodos en las que los dos nodos no adyacentes no son independientes condicionalmente dado un tercero. Es decir, una conexión convergente donde no existen más arcos.

Para formalizar las propiedades topológicas inducidas por la red que hemos visto, será de ayuda definir el concepto de D-separación.

Definición 1.1.3 (D-SEPARACIÓN) *Si A,B,C son tres subconjuntos disjuntos de nodos en un DAG G, entonces C se dice que d-separa A de B, denotado* $A \perp\!\!\!\perp_G B|C$*, si a lo largo de cada secuencia de arcos entre un nodo de A y un nodo de B hay un nodo v que satisface una de las siguientes dos condiciones:*

1. *v tiene arcos convergentes y ninguno de v o sus descendentes están en C.*

2. *v está en C y no tiene arcos convergentes.*

Por tanto, en términos de d-separación tendríamos que en los casos 2 y 3 de la figura 1.1 , C d-separa A de B mientras que en el caso 1 C no d-separa A de B.

En general, si inicialmente nos dan una red lo ideal sería que la d-separación fuese equivalente a la independencia condicional, sin embargo esto no es siempre cierto por lo que surge el concepto de mapa de independencia.

Definición 1.1.4 (MAPAS) *Un grafo G es un mapa de independencia (I-map) de la estructura dependiente de la probabilidad P de X si existe una correspondencia biunívoca entre las variables aleatoria* X_i *y los nodos V de G, de modo que para todos los subconjuntos disjuntos A,B,C*

$$A \perp\!\!\!\perp_P B|C \Longleftarrow A \perp\!\!\!\perp_G B|C$$

Análogamente, G es un mapa de dependencia (D-map) de P si tenemos

$$A \perp\!\!\!\perp_P B|C \Longrightarrow A \perp\!\!\!\perp_G B|C.$$

G se dice que es un mapa perfecto de P si es I-map y D-map, es decir,

$$A \perp\!\!\!\perp_P B|C \Longleftrightarrow A \perp\!\!\!\perp_G B|C$$

y en este caso, P se dice que es isomorfo a G.

1.1.2. Estructuras de red equivalentes

A partir de la figura 1.1, podemos observar que la conexión en serie y la conexión divergente están asociadas a factorizaciones equivalentes; cada una se puede obtener de la otra aplicando repetidas veces el teorema de Bayes, como vamos a ver a continuación.
Supongamos cierta la primera factorización

$$P(A, B, C) = P(B|C)P(C|A)P(A)$$

queremos llegar a partir de ella a

$$P(A, B, C) = P(A|C)P(B|C)P(C)$$

Usando la fórmula producto tenemos

$$P(A, B, C) = P(C)P(B|C)P(A|B, C) \tag{1.3}$$

con esto obtenemos dos de los tres términos necesarios, el término que no coincide con lo que queremos es *P(A|B,C)* intentemos transformarlo utilizando de nuevo la formula producto

$$P(A|B, C) = \frac{P(A, B, C)}{P(B|C)P(C)} \tag{1.4}$$

despejando ahora *P(B|C)* de la factorización de la que partimos tenemos

$$P(B|C) = \frac{P(A, B, C)}{P(C|A)P(A)}$$

lo sustituimos en la ecuación (1.4) donde obtenemos

$$P(A|B, C) = \frac{P(A)P(C|A)}{P(C)}$$

y sabemos que

$$P(A)P(C|A) = P(A, C) \text{ y también que } \frac{P(A, C)}{P(C)} = P(A|C)$$

Luego obtenemos que

$$P(A|B, C) = P(A|C)$$

sustituyendo esto en la ecuación (1.3), obtenemos la segunda factorización como queríamos demostrar.

Tales estructuras probabilísticamente equivalentes se conocen como estructuras equivalentes de Markov. Dado que la equivalencia es simétrica, reflexiva y transitiva, cada conjunto de estructuras equivalentes forma una

clase de equivalencia.

Generalizando lo visto en el ejemplo de la figura 1.1, se puede mostrar que los únicos arcos cuya dirección es necesaria para identificar una clase de equivalencia son los que pertenecen al menos a una v-estructura. Las clases de equivalencia suelen estar representadas por grafos acíclicos dirigidos (DAGs), en los que sólo se dirigen los arcos pertenecientes a v-estructuras y los que introducen v-estructuras o ciclos adicionales. Tales arcos se llaman obligados, ya que su dirección está determinada por la clase de equivalencia, aunque no son parte de ninguna v-estructura. Cambiar la dirección de cualquier otro arco no obligado da otra red de la misma clase de equivalencia, siempre y cuando no introduzca ninguna nueva v-estructura o ningún nuevo ciclo.

1.1.3. Envolvente de Markov

Otra propiedad fundamental que está estrechamente relacionada con la definición de mapa de independencia y d-separación es la envolvente de Markov. Esencialmente representa el conjunto de nodos que d-separa completamente un nodo dado del resto del grafo.

Definición 1.1.5 *La envolvente de Markov de un nodo $A \in V$ es el mínimo subconjunto S de V tal que*

$$A \perp\!\!\!\perp_D V - S - A | S$$

En cualquier red Bayesiana, la envolvente de Markov de un nodo A es el conjunto de los padres de A, los hijos de A y todos los otros nodos que compartan hijos con A (ver apéndice A) .

La envolvente de Markov facilita la comparación de la redes Bayesianas con modelos gráficos basados en grafos no dirigidos, conocidos como redes de Markov. Un DAG se puede transformar en el grafo no dirigido asociado a la red de Markov correspondiente siguiendo los pasos que veremos a continuación.

1. Se conectan los nodos no adyacentes en cada v-estructura con un arco no dirigido. Esto es equivalente a añadir un arco no dirigido entre cualquier nodo de la envolvente de Markov y el nodo en el que se centra dicha envolvente.

2. Se ignora la dirección de los otros arcos. Esto efectivamente reemplaza los arcos dirigidos.

La transformación anterior se llama moralización ya que "se casan"padres no adyacentes que comparte un hijo común. El grafo resultante se llama grafo moral.

1.2. Modelado de redes Bayesianas

La tarea de adecuar una red Bayesiana a un modelo probabilístico suele llamarse aprendizaje, un término tomado de la teoría de sistemas expertos y la inteligencia artificial. Se realiza en dos etapas diferentes, que corresponden a la selección de modelos gráficos y técnicas de estimación de parámetros en modelos estadísticos clásicos.

El primer paso se llama aprendizaje de la estructura y consiste en identificar la estructura de grafo de la red Bayesiana. Idealmente, debería ser el I-map mínimo de la estructura de dependencia de los datos o, en su defecto, debería al menos resultar en una distribución lo más cercana posible a la correcta en el espacio de probabilidad. Se proponen varios algoritmos para el aprendizaje de la estructura. A pesar de la variedad de antecedentes teóricos y terminología, se clasifican en tres grandes categorías: algoritmos basados en restricciones, basados en puntuaciones y algoritmos híbridos. Como alternativa, la estructura de la red se puede construir manualmente a partir del conocimiento de un experto y de la información previa disponible sobre los datos.

El segundo paso se llama aprendizaje paramétrico. Como su nombre indica, implementa la estimación de los parámetros de la distribución conjunta. Esta tarea se puede realizar eficientemente mediante estimaciones de máxima verosimilitud sobre los parámetros de las distribuciones locales que implica la estructura obtenida en el paso anterior.

1.2.1. Algoritmos de aprendizaje de estructura basados en restricciones

Los algoritmos de aprendizaje de estructura basados en restricciones se basan en el trabajo original de Pearl en mapas y su aplicación a modelos de grafos causales. Su algoritmo de causalidad inductiva (IC) proporciona un marco para el aprendizaje de la estructura de las redes Bayesianas mediante pruebas de independencia condicional. Los detalles del IC se describen a continuación.

Algoritmo 1 Algoritmo de causalidad inductiva

1: Para cada par de variables A y B en V, buscar el conjunto $S_{AB} \subset \mathrm{V}$ (incluyendo $S=\emptyset$) de tal manera que A y B son independientes dado S_{AB} y $A,B \notin \mathrm{S}_{AB}$. Si no hay tal conjunto, coloque un arco no dirigido entre A y B.

2: Para cada par de variables no adyacentes A y B con un vecino común C, comprobar si $C \in \mathrm{S}_{AB}$. Si esto no es cierto, establezca la dirección del arco A-C y C-B como $A \rightarrow C$ y $C \leftarrow B$, denotamos - al eje existente entre dos nodos.

3: Establecer la dirección de los arcos que aún no están dirigidos aplicando recursivamente las dos siguientes reglas:

   1. Si A es adyacente a B y hay un camino dirigido entre A y B entonces establecer la dirección de A -B como $A \rightarrow B$.
   2. Si A y B no son adyacentes pero $A \rightarrow C$ y C-B, entonces cambiar el anterior por $C \rightarrow B$.

4: Devuelve el grafo acíclico dirigido resultante.

El primer paso identifica qué pares de variables están conectadas por un arco, independientemente de su dirección. Estas variables no pueden ser independientes dado ningún otro subconjunto de variables, porque no se pueden d-separar. Este paso también se puede ver como un procedimiento de selección hacia atrás a partir del modelo saturado con un grafo completo y la poda basada en pruebas estadísticas para la independencia condicional.

El segundo paso trata de identificar las v-estructuras entre todos los pares de nodos no adyacentes A y B con un vecino común C. Por definición, las v-estructuras son las únicas conexiones fundamentales en las que los dos nodos no adyacentes no son independientes condicionalmente dado un tercero. Por lo tanto, si no hay subconjuntos de nodos que contiengan a C y d-separa A y B, los tres nodos son parte de una v-estructura centrada en C. Esta condición se puede verificar mediante la realización de una prueba de independencia condicional para A y B con cada subconjunto de sus vecinos comunes que incluye a C. Al final de la segunda estapa, se conocen tanto la estructura del grafo como las v-estructuras de la red, por lo que la clase de equivalencia a la que pertenece la red Bayesiana se identifica de manera única.

El tercer y último paso del algoritmo IC identifica los arcos obligados y los orienta recursivamente para obtener el DAG describiendo la clase de equivalencia identificada por los pasos anteriores.

Un problema importante del algoritmo IC es que los dos primeros pasos

pueden no ser aplicables en problemas reales debido al número exponencial de posibles relaciones de independencia condicional. Esto ha llevado al desarrollo de algoritmos mejorados, como los siguientes:

- PC: la primera aplicación práctica del algoritmo *IC* (Spirtes et al., 2001 [2]), un procedimiento de selección hacia atrás a partir del grafo saturado.
- Grow-Shrink (GS): basado en el algoritmo de envolvente de Markov de Grow-Shrink (Margaritis, 2003 [3]), un sencillo enfoque de detección de envolventes de Markov.
- IAMB: basado en el algoritmo global de la asociación incremental de envolventes de Markov (Tsamardinos et al.,2003 [4]), un esquema de selección en dos fases basado en una selección hacia adelante seguida de una hacia atrás.
- Fast-IAMB: una variante de IAMB que utiliza la selección especulativa paso a paso para reducir el número de pruebas de independencia condicional (Yaramakala y Margaritis, 2005 [5]).
- Inter-IAMB: otra variante de IAMB que utiliza la selección por pasos hacia adelante (Tsamardinos et al., 2003 [4]) para evitar falsos positivos en la fase de detección general de Markov.

Todos estos algoritmos, con la excepción de PC, primero aprenden la envolvente de Markov de cada nodo de la red. Este paso preliminar simplifica enormemente la identificación de los vecinos de cada nodo, ya que la búsqueda se puede limitar a la envolvente de Markov. Como resultado, el número de pruebas de independencia condicional realizadas por el algoritmo de aprendizaje y su complejidad global se reducen significativamente.

1.2.2. Algoritmos de aprendizaje de estructura basados en puntuaciones

Los algoritmos de aprendizaje de la estructura basados en la puntuación (también conocidos como algoritmos de búsqueda y puntuación) representan la aplicación de técnicas generales de optimización heurística al problema de aprender la estructura de una red Bayesiana. A cada red candidata se le asigna una puntuación de red que refleja su bondad de ajuste, que el algoritmo luego intenta maximizar. Algunos ejemplos de esta clase de algoritmos son los siguientes:

- Algoritmos de búsqueda codiciosos (greedy) como el de escalada con reinicios aleatorios o búsqueda tabú. Estos algoritmos exploran el espacio de búsqueda a partir de una estructura de red (generalmente el

grafo vacío) y añade, borra o invierte un arco cada vez hasta que la puntuación ya no se pueda mejorar.

Algoritmo 2 Algoritmo de escalada

1: Elija una estructura de red G sobre V, generalmente(pero no necesariamente) vacía.
2: Calcule la puntuación de G, denotada como $Score_G$= *Score(G).*
3: Establezca $maxscore$=$Score_G$.
4: Repita los siguientes pasos siempre y cuando la puntuación máxima ($maxscore$) aumente:
   1. Por cada posible adición, eliminación o inversión de un arco que no dé una red cíclica:
      a) Calcule la puntuación de la red modificada $G*$, $score_{G*}$= $score(G*)$;
      b) Si $score_{G*}$>$Score_G$, tomar G=G* y $Score_G$=$Score_{G*}$.
   2. Actualice maxscore con el nuevo valor de $Score_G$.
5: Devuelve el grafo acíclico dirigido G.

- Algoritmos genéticos, que imitan la evolución natural a través de la selección iterativa de los modelos "más aptos" y la hibridación de sus características. En este caso el espacio de búsqueda se explora a través del crossover (que combina la estructura de dos redes) y la mutación (que introduce alteraciones aleatorias) de operadores estocásticos.

- Recocido simulado. Este algoritmo realiza una búsqueda local estocástica al aceptar cambios que aumentan la puntuación de la red y, al mismo tiempo, permiten cambios que la disminuyen con una probabilidad inversamente proporcional a la disminución de la puntuación.

1.2.3. Algoritmos de aprendizaje de estructura híbrida

Los algoritmos híbridos de aprendizaje de estructuras combinan algoritmos basados en restricciones y puntuación para compensar sus debilidades y producir estructuras de red fiables en una amplia variedad de situaciones. Los dos miembros más conocidos de esta familia son el algoritmo del candidado disperso (SC) de Friedman et al. (1999b) [6] y el algoritmo de escalada Min-Max (MMHC) de Tsamardinos et al. (2006) [7]. El primero corresponde al Algoritmo 3.

Algoritmo 3 Algoritmo del candidato disperso

1: Elija una estructura de red G sobre V, generalmente (pero no necesariamente) vacía.
2: Repita los siguientes pasos hasta la convergencia:
   1. Restringir: selecciona un conjunto C_i de padres candidatos para cada nodo $X_i \in$V, que debe estar incluido en el conjunto de los padres de X_i en G.
   2. Maximizar: encuentra la estructura de red $G*$ que maximiza la puntuación $Score(G*)$ entre las redes en las que los padres de cada nodo X_i están incluidos en el conjunto C_i correspondiente.
   3. Establecer $G=G*$.
3: Devuelve el grafo acíclico dirigido G.

Ambos algoritmos se basan en dos pasos llamados restringir y maximizar. En el primer paso, el conjunto de candidatos para los padres de cada nodo X_i se reduce de todo el conjunto de nodos V a un conjunto más pequeño $C_i \subset V$ de nodos, cuyo comportamiento se ha demostrado que está relacionado de alguna manera con el de X_i. Esto a su vez resulta en un espacio de búsqueda más pequeño. El segundo paso busca la red que maximiza una función de puntuación dada sujeta a las restricciones impuestas por los conjuntos C_i.

En el algoritmo del candidato disperso estos dos pasos se aplican iterativamente hasta que no hay cambio en la red o ninguna red mejora la puntuación de la red; la elección de la heurística utilizada para realizarlos se deja a la implementación. Por otra parte, en el algoritmo MMHC, restringir y maximizar se realizan sólo una vez; la heurística de Min-Max Padres e Hijos (MMPC) se utiliza para aprender el conjunto candidato C_i y una búsqueda codiciosa de escalada para encontrar la red óptima.

1.2.4. Elección de distribuciones, independencia condicional. Pruebas y puntuación de red.

En principio, hay muchas opciones posibles para las funciones de distribución conjunta y local, dependiendo de la naturaleza de los datos y los objetivos del análisis. Sin embargo, nos vamos a centrar principalmente en dos casos:

- Variables multinomiales: se utiliza para conjuntos de datos discretos/categóricos y a menudo se refiere como el caso discreto. Tanto las distribuciones conjuntas como las locales son multinomiales, y estas últimas se representan como tablas condicionadas de probabilidad (CPT). Las redes Bayesianas correspondientes se conocen como redes Bayesianas discretas.

- Variables normales multivariantes: esta representación se utiliza para conjuntos de datos continuos y por lo tanto se conoce como el caso continuo. La distribución conjunta es normal multivariante, mientras que las distribuciones locales son variables aleatorias normales univariantes vinculadas por restricciones lineales. Las distribuciones locales son, de hecho, modelos lineales en los que los padres desempeñan el papel de variables explicativas. Estas redes Bayesianas se llaman redes gaussianas Bayesianas.

Otros supuestos de distribuciones requieren algoritmos de aprendizaje especiales o presentan varias limitaciones debido a la dificultad de especificar las funciones de distribución en forma cerrada. Por ejemplo, los modelos de datos mixtos imponen restricciones a la elección de los padres para los nodos.

La elección de un conjunto particular de distribuciones conjuntas y locales también determina qué pruebas de independencia condicional y qué puntuaciones de red se pueden utilizar para aprender la estructura de la red Bayesiana.

Las pruebas de independencia condicional y puntuaciones de red para datos discretos son funciones de las tablas condicionadas de probabilidad implicadas por la estructura gráfica de la red a través de las frecuencias observadas $\{n_{ijk}, i=1,...,R, j=1,...,C, k=1,...,L\}$ para las variables aleatorias X e Y y todas las configuraciones de las variables de condicionamiento Z. Dos pruebas comunes de independencia condicional son las siguientes:

- Información mutua, una medida de distancia de la información teórica definida como

$$MI(X,Y|Z) = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{L}\frac{n_{ijk}}{n}log\frac{n_{ijk}n_{++k}}{n_{i+k}n_{+jk}}$$

  Es proporcional al test de razón de log-verosimilitudes G^2 (difieren por un factor $2n$, donde n es el tamaño de la muestra), y está relacionado con la desviación de los modelos contrastados.

- La clásica prueba X^2 de Pearson para tablas de contingencia,

$$X^2(X,Y|Z) = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{L}\frac{(n_{ijk}-m_{ijk})^2}{m_{ijk}}, \text{ donde } m_{ijk} = \frac{n_{i+k}n_{+jk}}{n_{++k}}$$

En ambos casos, la hipótesis nula de independencia puede contrastarse utilizando la distribución asintótica ${\chi^2}_{(R-1)(C-1)L}$. Otras opciones posibles son la prueba exacta de Fisher y el estimador de contracción para la información

mutua.

Las puntuaciones de red más comunes son las siguientes:

- La puntuación equivalente de Dirichlet-Bayesiano (BDe), la densidad posterior asociada a una distribución uniforme a priori tanto en el espacio de las estructuras de red como en los parámetros de cada distribución local.
- El criterio de información Bayesiana (BIC), una puntuación de probabilidad penalizada definida como

$$BIC = \sum_{i=1}^{n} logP_{Xi}(X_i|\pi_{Xi}) - \frac{d}{2}logn,$$

  donde d es el número de parámetros de la distribución conjunta. BIC converge asintóticamente a la densidad a posteriori BDe

Las funciones de puntuación asignan la misma puntuación a redes pertenecientes a la misma clase de equivalencia. También se pueden descomponer en las componentes asociadas a cada nodo, lo que es una ventaja computacional significativa cuando se aprende la estructura de la red (las únicas partes de la puntuación que deben recalcularse son las que difieren entre las redes que se comparan).

1.2.5. Aprendizaje paramétrico

Una vez que la estructura de la red se ha aprendido de los datos, la tarea de estimar y actualizar los parámetros de la distribución conjunta se simplifica en gran medida por la aplicación de la propiedad de Markov.

Las distribuciones locales en la práctica implican sólo un pequeño número de variables. Además, su dimensión no suele escalar con el tamaño de X y a menudo se asume que está limitado por una constante al calcular la complejidad computacional de los algoritmos. Esto a su vez alivia la maldición de la dimensionalidad, porque cada distribución local tiene un número comparativamente pequeño de parámetros para estimar de la muestra y porque las estimaciones son más exactas debido a la mejor relación entre el tamaño del espacio de parámetros y el tamaño de la muestra. Hay dos enfoques principales para la estimación de esos parámetros: uno basado en estimación de máxima verosimilitud y el otro basado en la estimación Bayesiana.

El número de parámetros necesarios para identificar de manera única la distribución conjunta, que es la suma del número de parámetros de las

distribuciones locales, se reduce también porque las relaciones de independencia condicional codificadas en la estructura de la red hace constante una gran parte del espacio de parámetros. Por ejemplo, en las redes Bayesianas de Gauss, los coeficientes de correlación parcial que implican variables independientes (condicionales) son iguales a cero por definición, y las frecuencias conjuntas se factorizan en distribuciones marginales multinomiales.

Sin embargo, la estimación de parámetros sigue siendo problemática en muchas situaciones. Por ejemplo, es cada vez más común tener tamaños de muestra mucho menores que el número de variables incluidas en el modelo. Esto es típico de los conjuntos de datos biológicos de alto rendimiento, como los microarrays, que tienen unas pocas, diez o cien observaciones y miles de genes. En este contexto, que se llama "n pequeño, p grande", las estimaciones tienen una gran variabilidad a menos que se tenga especial cuidado tanto en la estructura como en los aprendizajes de parámetros.

1.2.6. Discretización

Una forma sencilla de aprender redes Bayesianas a partir de datos mixtos es convertir todas las variables continuas en variables discretas y luego aplicar las técnicas descritas en las secciones anteriores. Este enfoque, que se llama discretización o binning, evita completamente el problema de definir un modelo probabilístico para los datos. La discretización también se puede aplicar para tratar con datos continuos cuando una o más variables presentan desviaciones severas de la normalidad (sesgo, colas, etc.).

Los intervalos en los que las variables serán discretizadas se pueden elegir de una de las siguientes maneras:

- Uso de conocimientos previos sobre los datos. Las cotas de los intervalos se definen, para cada variable, para corresponder a escenarios del mundo real significativamente diferentes, como la concentración de un contaminante particular (ausente, peligroso, letal) o clases de edad (niños, adultos, ancianos).
- Uso de la heurística antes de aprender la estructura de red.
- Elegir el número de intervalos y sus cotas para equilibrar la exactitud y la pérdida de información, repetirlo cada vez para una variable antes de que se haya aprendido la estructura de la red.
- Realizar el aprendizaje y la discretización de forma iterativa hasta que no se haga ninguna mejora.

Estas estrategias difieren entre sí en el equilibrio que hacen entre la exactitud de la representación discreta de los datos originales y la eficiencia computacional de la transformación.

Capítulo 2

Algoritmos de inferencia en Redes Bayesianas

Las redes Bayesianas, como otros modelos estadísticos, pueden ser utilizadas para responder preguntas sobre la naturaleza de los datos que van más allá de la mera descripción de la muestra observada. Las técnicas basadas en nuevas evidencias que se utilizan para obtener esas respuestas, se conocen en general como inferencia. Para las redes Bayesianas, el proceso de responder a estas preguntas también se conoce como razonamiento probabilístico o actualización de creencias, mientras que las preguntas mismas se llaman consultas.

2.1. Razonamiento probabilístico y evidencias

En la práctica, el razonamiento probabilístico sobre las redes Bayesianas tiene sus raíces en la estadística Bayesianas y se centra en el cálculo de probabilidades o densidades a posteriori. Por ejemplo, supongamos que hemos aprendido una red Bayesiana B con la estructura G y parámetros Θ. Posteriormente, queremos investigar los efectos de una nueva evidencia E sobre la distribución de X usando el conocimiento codificado en B, es decir, para investigar la distribución a posteriori $P(X|E,B)=P(X|E,G,\Theta)$.

Los enfoques utilizados para este tipo de análisis varían en función de la naturaleza de E y de la naturaleza de la información que nos interesa. Las dos evidencias más comunes son las siguientes:

- Evidencia sólida, una observación de una o más variables en la red. En otras palabras,

$$E = \{X_{i_1} = e_i, X_{i_2} = e_2, ..., X_{i_k} = e_k\}, i_1, ..., i_k \in \{1, ..., n\}$$

- Evidencia suave, una nueva distribución para una o más variables en la red. Dado que tanto la estructura de la red como las hipótesis de

distribución se consideran fijas, las evidencias no concluyentes suelen especificarse como un nuevo conjunto de parámetros,

$$E = \{X_{i1} \sim (\Theta_{Xi_1}), X_{i2} \sim (\Theta_{Xi_2}), ..., X_{i_k} \sim (\Theta_{Xi_k})\}$$

Esta nueva distribución puede ser, por ejemplo, la hipótesis nula en un problema de contraste de hipótesis.

En lo que respecta a las consultas, nos centraremos en las consultas de probabilidad condicional (CPQ) y máximo a posteriori (MAP), también conocidas como la explicación más verosímil. Esta explicación representaría el principio de la navaja de Ockham, según el cual: "La explicación más probable suele ser la correcta". Ambos se aplican principalmente a la evidencia sólida, aunque se pueden usar en combinación con evidencia suave.

Las consultas de probabilidad condicional se refieren a la distribución de un subconjunto de variables Q=$\{X_{j_1}$,...,$X_{j_l}\}$ dado alguna evidencia sólida E en otro conjunto X_{i_1},...,X_{i_k} de variables en X. Los dos conjuntos de variables se suponen generalmente disjuntos. En redes Bayesianas discretas, esta distribución se calcula como la probabilidad a posteriori

$$CPQ(Q|E,B) = P(Q|E,G,\Theta) = P(X_{j_1}, ..., X_{j_l}|E,G,\Theta)$$

que es la distribución marginal de la probabilidad a posteriori de Q, es decir,

$$P(Q|E,G,\Theta) = \int P(X|E,G,\Theta)d(X\backslash Q) \tag{2.1}$$

En las redes Bayesianas de Gauss, del mismo modo,

$$CPQ(Q|E,B) = f(Q|E,G,\Theta) = \int f(X|E,G,\Theta)d(X\backslash Q)$$

Esta clase de consultas tiene muchas aplicaciones útiles debido a su versatilidad. Por ejemplo, se pueden utilizar consultas condicionales de probabilidad para evaluar la interacción entre dos conjuntos de factores de diseños experimentales para un tratamiento de interés dado. Mientras que el último (es decir, el rasgo) sería considerado como la evidencia sólida E, el primero jugaría el papel del conjunto de variables de consultas Q. Otro ejemplo común sería evaluar las probabilidades de un resultado desfavorable Q para diferentes conjuntos de evidencias sólidas E_1,E_2,...,E_m, esto precisamente es lo que haremos en el capítulo siguiente en un caso experimental.

La explicación más verosímil se refieren a encontrar la configuración q^* de las variables en Q que tienen la mayor probabilidad a posteriori,

$$MAP(Q|E,B) = q^* = \underset{q}{argmax} P(Q = q|E,G,\Theta)$$

o la máxima densidad a posteriori

$$MAP(Q|E,B) = q^* = \underset{q}{argmax} f(Q = q|E, G, \Theta)$$

en redes Bayesianas de Gauss. Las aplicaciones de este tipo de consultas se dividen en dos categorías: imputando los datos faltantes de la evidencia sólida parcialmente observada, donde las variables en Q que no se observan deben ser imputadas de las en E, o comparando q^* con los valores observados para las variables en Q para la evidencia sólida completamente observada.

Tanto las consultas condicionales de probabilidad como la explicación más verosímil también pueden utilizarse con evidencias suaves, aunque con interpretaciones diferentes. Por ejemplo, cuando E codifica evidencias sólidas no es estocástico sino un valor observado. En este caso, $P(Q{=}q|E,G,\Theta)$ no es estocástico. Sin embargo, cuando E codifica evidencias suaves sigue siendo una variable aleatoria, y a su vez $P(Q{=}q|E,G,\Theta)$ también es estocástico. Por lo tanto, los resultados de las consultas descritas en esta sección deben evaluarse de acuerdo con la naturaleza de las evidencias en las que se basan.

2.2. Algoritmos para el razonamiento probabilístico: Inferencia exacta y aproximada

La estimación de las probabilidades y densidades a posteriori mostrada en la sección anterior es un problema fundamental en la evaluación de consultas. Las consultas que implican probabilidades muy pequeñas o redes grandes son particularmente problemáticas incluso con los mejores algoritmos debido a retos computacionales y probabilísticos. En el peor de los casos, su complejidad computacional es exponencial en el número de variables.

Los algoritmos para el razonamiento probabilístico pueden caracterizarse como exactos o aproximados. Ambos se basan en las propiedades fundamentales de las redes Bayesianas para evitar la maldición de la dimensionalidad a través del uso de cómputos locales, es decir, utilizando sólo distribuciones locales.

Por ejemplo, la marginalización en la ecuación (2.1) puede ser reescrita

como:

$$P(Q|E,G,\Theta) = \int P(X|E,G,\Theta)d(X\backslash Q) =$$
$$= \int \Big[\sum_{i=1}^{p} P(X_i|E,\pi_{Xi},\Theta_{X_i})\Big] d(X\backslash Q) =$$
$$= \prod_{i:X_i\in Q} \int P(X_i|E,\pi_{Xi},\Theta_{X_i})dX_i$$

La correspondencia entre la d-separación y la independencia condicional también puede utilizarse para reducir aún más la dimensión del problema. Por definición, las variables que son d-separadas de Q por E no pueden influir en el resultado de la consulta. Por lo tanto, pueden ser completamente ignoradas en el cálculo de las probabilidades a posteriori.

Los algoritmos de inferencia exacta combinan aplicaciones repetidas del teorema de Bayes con cálculos locales para obtener valores exactos de *P(Q|E,G,Θ)*. Sin embargo, su viabilidad se limita a redes pequeñas o muy simples, como árboles o poliárboles.

Los dos algoritmos de inferencia exacta más conocidos son: el de eliminación de variables y el de actualizaciones de creencias basadas en árboles de unión. Ambos se derivaron originalmente para redes discretas y más tarde se han extendido a las redes continuas o mixtas. La eliminación de variables utiliza directamente la estructura de la red Bayesiana, especificando la secuencia óptima de operaciones en las distribuciones locales y cómo almacenar los resultados intermedios para evitar cálculos innecesarios. Por otra parte, las actualizaciones de creencias también se pueden realizar transformando primero la red Bayesiana en un árbol de unión.
Como se ilustra en el Algoritmo 4, un árbol de unión es una transformación del grafo moral de B en el que los nodos originales se agrupan para reducir cualquier estructura de red en un árbol.

Algoritmo 4 Algoritmo de agrupamiento en árboles de unión

1: Moralizar: cree el grafo moral de la red Bayesiana B.
2: Triangular: divida cada ciclo que abarca 4 o más nodos en subciclos de exactamente 3 nodos agregando arcos al grafo moral, obteniendo así un grafo triangulado.
3: Cliques: identifique los cliques del grafo triangulado, es decir, subconjuntos máximos de nodos en los que cada elemento es adyacente a todos los demás.
4: Árbol de unión: cree un árbol en el que cada clique es un nodo, y los cliques adyacentes están unidos por arcos.
5: Reparametrizar: use los parámetros de las distribuciones locales de B para calcular los conjuntos de parámetros de los nodos compuestos del árbol de unión.

Posteriormente, las actualizaciones de creencias se pueden realizar eficientemente usando el algoritmo de Kim y Pearl.
Los algoritmos de inferencia aproximada utilizan simulaciones de Monte Carlo para tomar muestras de las distribuciones locales y así estimar $P(Q|E,G,\Theta)$. En particular, generan un gran número de muestras a partir de B y estiman las probabilidades condicionales pertinentes ponderando las muestras que incluyen tanto a E como $Q{=}q$ frente a las que incluyen únicamente a E. En el área de la computación, estas muestras aleatorias son a menudo llamadas partículas, y los algoritmos que hacen uso de ellos se conocen como filtros de partículas o métodos basados en partículas.

En la literatura podemos encontrar diversos enfoques que se han desarrollado para describir procesos de muestreo aleatorio o para estimar los parámetros respecto a la muestra obtenida. Esto ha resultado en varios algoritmos aproximados. Dentro de las técnicas más utilizadas podemos encontrar el muestreo de aceptación y rechazo o diversas variantes. La combinación más simple de estos métodos de muestreo y ponderación se conoce como muestreo hacia adelante o lógico. Esto se describe en el Algoritmo 5.

Algoritmo 5 Algoritmo de muestreo lógico

1: Ordene las variables en X de acuerdo con el orden topológico implicado por G, digamos $X_{(1)} < X_{(2)} <...< X_{(p)}$.
2: Para un número adecuado de muestras $x^* = (\ {x^*}_1,..,\ {x^*}_p)$.
   1. para $i{=}1,..p$, genera ${x^*}_{(i)}$ a partir de $X_{(1)}|\Pi_{X_{(1)}}$
   2. si E está incluido en x, tomar $n_E{=}n_E{+}1$
   3. si tanto $Q = q$ como E están incluidos en x, tomar $n_{E,q}{=}n_{E,q}{+}1$
3: Estimar $P(Q|E,G,\Theta)$ con $n_{E,q}$ / n_E

En el punto 1 podemos encontrar esa ordenación de las componentes de X debido a que las redes Bayesianas con las que trabajamos son grafos acíclicos, es decir, no tienen ciclos.

El muestreo lógico combina el método de aceptación y rechazo con pesos uniformes, contando esencialmente la proporción de muestras generadas incluyendo E que también incluyen $Q = q$. Claramente, tal algoritmo puede ser muy ineficiente si $P(E)$ es pequeño, porque la mayoría de las muestras aleatorias serán descartadas sin contribuir a la estimación de $P(Q|E,G,\Theta)$. Sin embargo, su simplicidad facilita la implementación y es muy general en sus posibles aplicaciones ya que permite especificaciones muy complejas en E y Q tanto para $MAP(Q|E,B)$ como para $CPQ(Q|E,B)$. En el otro extremo del espectro, algoritmos complejos aproximados, como el esquema de muestreo de importancia adaptativa, pueden estimar probabilidades condicionales tan pequeñas como 10^{-41}. También funcionan mejor en grandes redes. Sin embargo, sus supuestos a menudo los restringen a datos discretos y pueden requerir la especificación de parámetros de ajuste no triviales.

Capítulo 3

Aplicación experimental en el contexto del COVID-19

Los datos usados para nuestro estudio son tomados del *"Informe nº 22. Situación de COVID-19 en España a 13 de abril de 2020. Equipo COVID-19"* elaborado por la Red Nacional de Vigilancia Epidemiológica (ver Apéndice C).

Tomando como base el informe anterior vamos a modelar una red Bayesiana. En primer lugar, a partir del informe, seleccionamos un conjunto de variables relativas a características del enfermo, síntomas, enfermedades de riesgo y situación clínica . Al elegir estas variables y los valores que toman debemos hacerlo con la precaución de que las probabilidades condicionadas que vayamos a introducir en la red debemos extraerlas de los datos cuantitativos de dicho informe. Por tanto, estamos muy limitados a la hora de construir un modelo realista de red Bayesiana.
Del mismo modo, la topología de la red no puede ser muy densa ya que no podemos extraer toda la información cuantitativa que necesitaríamos para introducirla. Por ello, solo hemos añadido algunas de las relaciones que deberían tenerse en cuenta. Aun así, utilizando el modelo obtenido, al que llamaremos *red diana*, ejemplificaremos las diferentes metodologías que se han descrito en los capítulo anteriores, en particular, veremos como a partir de la red podemos deducir independencias condicionales, también veremos como funcionan los algoritmos de aprendizaje y daremos respuestas mas verosímiles a una evidencia.

3.1. Descripción de las variables y modelado de la red Bayesiana

Las variables que usaremos en nuestro estudio serán:

- *Edad* (Edad), dentro de esta variable distinguiremos los siguientes grupos de edad de acuerdo a la información cuantitativa que aparece en las tablas del informe:
  - <2, codificado con el valor 1.
  - 2-4, codificado con el valor 2.
  - 5-14, codificado con el valor 3.
  - 15-29, codificado con el valor 4.
  - 30-39,codificado con el valor 5.
  - 40-49, codificado con el valor 6.
  - 50-59, codificado con el valor 7.
  - 60-69, codificado con el valor 8.
  - 70-79, codificado con el valor 9.
  - $\geq$ 80, codificado con el valor 10.
- *Sexo* (Sex), distinguiremos entre
  - *Mujer*, codificado con el valor 1.
  - *Hombre*, codificado con el valor 2.

Todas las variables siguientes están relacionadas con los síntomas del paciente y tomarán el valor 0 si ese síntoma no se padece y el valor 1 si se padece.

- *Fiebre* (Fie)
- *Tos* (Tos)
- *Dolordegarganta* (DG)
- *Disnea* (Dis)
- *Escalofríos* (Esc)
- *Vómitos* (Vom)
- *Diarrea* (Diar)
- *Síndrome de distrés respiratorio agudo* (SDRA)
- *Fallo renal agudo* (FRA)
- *Otros síntomas respiratorios* (OSR)

A continuación, las variables relacionadas con las enfermedades y factores de riesgos, tomarán el valor 1 si el paciente padece dicha enfermedad o 0 si no la padece.

- *Neumonía* (Neum)
- *Enfermedad cardiovascular* (EnCar)
- *Enfermedad respiratoria* (EnRes)
- *Diabetes* (Diab)
- *Hipertensión arterial* (HipArt)

Las últimas variables corresponden a la situación clínica del paciente, estas son:

- *Hospitalización* (Hosp), si se requiere de hospitalización esta variable tomará el valor 1 si no se requiere será 0.
- *Admisión en UCI* (UCI), si es necesario la admisión en UCI tendremos que vale 1 esta variable, sino 0.
- *Defunción* (Def), si muere el paciente esta variable tomará el valor 1 y 0 en otro caso.

Una vez definidas las variables de la red, es decir los nodos, vamos a introducir la información topológica. Hay que tener en cuenta que cada arco que se introduzca supone tener que introducir una información cuantitativa que la deduciremos de las tablas numéricas que aparecen en el informe, por lo que estamos muy limitados a la hora de introducir arcos.
En base a las tablas de este informe, hemos decidido introducir los siguientes arcos:

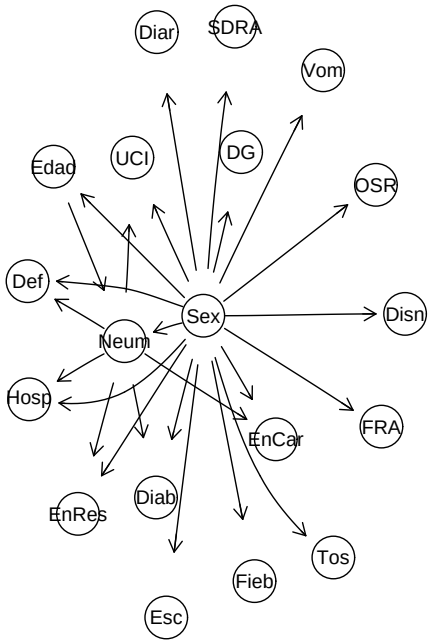


Figura 3.1: Red Diana.

Obtenemos una red con 19 nodos, los cuales son las variables descritas anteriormente. Esta red consta de 25 arcos que representan las probabilidades condicionadas existentes entre las variables. Se ve que existen dos nodos, *Sexo* y *Neumonía*, que tienen dependencia directa con otras muchas variables. La razón de esto, es que en el informe, estas variables aparecen en varias de las tablas numéricas a partir de las cuales se puede estimar probabilidades condicionadas.

Como se puede ver en la red que hemos modelado, existen síntomas como *tos* o *fiebre* que no influyen, por ejemplo, en la *hospitalización* conociendo el sexo del paciente. Esto se debe a que el único camino que une *tos* con *hospitalización* es a través de *sexo* y este nodo no es de aristas convergentes. Esto es una limitación del modelo que usamos pero no tenemos la suficiente información cuantitativa para introducir una red más densa. Aunque lo ideal sería trabajar con un modelo más denso, este será un modelo muy simple que se aproxima a la distribución del vector aleatorio que estamos estudiando.

Finalmente, para introducir en nuestra red las probabilidades condicio-

nadas necesarias nos hemos encontrado básicamente con dos formas distintas de hacerlo. La primera y la más simple, se refiere a las variables que solo dependen del nodo *sexo*. Estas se relacionan con los síntomas y la *edad* y las hemos podido tomar directamente de la Tabla 2 del informe.

Características		N	Total N (%)	Mujeres N (%)	Hombres N (%)	p-valor
Sexo		113368		59196 (52,2)	54172 (47,8)	
Edad	Mediana (RIC)[2]	112982	60 (46-75)	58 (44-75)	62 (49-76)	<0,001
Grupo de edad (años)	<2	112982	168 (0,1)	69 (0,1)	99 (0,2)	
	2-4		64 (0,1)	32 (0,1)	32 (0,1)	
	5-14		303 (0,3)	153 (0,3)	150 (0,3)	
	15-29		6155 (5,4)	3975 (6,7)	2174 (4,0)	
	30-39		10764 (9,5)	6518 (11,0)	4245 (7,9)	
	40-49		16915 (15,0)	9369 (15,9)	7544 (14,0)	
	50-59		21057 (18,6)	11369 (19,3)	9682 (17,9)	
	60-69		18801 (16,6)	8671 (14,7)	10127 (18,8)	
	70-79		17912 (15,9)	7507 (12,7)	10402 (19,3)	
	≥80		20843 (18,4)	11348 (19,2)	9494 (17,6)	<0,001
Síntomas[1]	Fiebre o reciente historia de fiebre	10665	7876 (73,8)	3724 (68,7)	4143 (79,1)	<0,001
	Tos	10411	7751 (74,5)	3925 (74,1)	3817 (74,8)	0,422
	Dolor de garganta	8812	2285 (25,9)	1376 (30,1)	904 (21,3)	<0,001
	Disnea	13059	6263 (48,0)	2873 (44,4)	3387 (51,5)	<0,001
	Escalofríos	7785	2666 (34,2)	1339 (33,6)	1324 (34,9)	0,244
	Vómitos	8875	830 (9,4)	504 (11,0)	325 (7,6)	<0,001
	Diarrea	9146	2755 (30,1)	1510 (32,0)	1243 (28,2)	<0,001
	Neumonía (radiológica o clínica)	60495	36017 (59,5)	14901 (51,2)	21112 (67,3)	<0,001
	Síndrome de distrés respiratorio agudo	33237	2296 (6,9)	847 (5,1)	1446 (8,7)	<0,001
	Otros síntomas resp.	35737	5676 (15,9)	2502 (13,8)	3172 (18,1)	<0,001
	Fallo renal agudo	37998	1732 (4,6)	601 (3,1)	1130 (6,0)	<0,001

Figura 3.2: Tabla 2 del informe RENAVE

Por otro lado, para las variables relacionadas con las enfermedades y la situación clínica hemos tenido que desagregar los datos manualmente de la siguiente forma, tomemos de ejemplo la probabilidad de hospitalización condicionado a sexo y padecer o no neumonía.
Para esto, tomaremos la Tabla 5 del informe. Observamos el número total de caso con neumonía (33075) y sin neumonía (5771). Vemos en la parte superior el porcentaje de mujeres con neumonía (41 %), de hombre con neumonía (59 %), mujeres sin neumonía (58 %) y hombres sin neumonía (42 %). Con estos datos, tenemos que sacar la probabilidad de ser hospitalizado sabiendo que se padece neumonía siendo mujer u hombre y la de ser hospitalizado sabiendo que no se padece neumonía siendo mujer u hombre. Tomemos el caso de mujer con neumonía, para sacar esta probabilidad tomamos el número total de casos con neumonía (33075), multiplicamos este número por 0.41, es decir calculamos el 41 % para saber cuantos de esos casos son mujeres y aproximamos a un número natural. Haciendo esto obtenemos que 13561 mujeres son hospitalizadas padeciendo neumonía. Finalmente para obtener la probabilidad, dividimos el número de mujeres hospitalizadas (13561) en-

tre el número total de mujeres con neumonía(14901) y obtenemos 0.91. Para obtener la probabilidad de mujeres no hospitalizadas padeciendo neumonía tomamos el complementario, con lo que tenemos 0.09. Hacemos lo mismo con los datos de los hombres y después repetimos el mismo proceso con los datos de no padecer neumonía y obtenemos todas las probabilidades.

Introducimos los datos en el software especializado *R* (https://www.r-project.org/) usando el siguiente comando

```
#PROB Hosp|Sex,Neum
PHosp = c(0.91, 0.09, 0.92, 0.08, 0.235, 0.765, 0.27, 0.73)
dim(PHosp) = c(2, 2, 2)
dimnames(PHosp) = list("Hosp" = c("Si", "No"),
   "Sex" =  c("Mujer", "Hombre"),
"Neum" = c("Si", "No"))
PHosp
```

El resto de las probabilidades condicionadas se han sacado siguiente un proceso similar. Una vez que sacamos todas las probabilidades, las introducimos en la red como se indica a continuación:

```
net = model2network("[Sex][Edad|Sex][Neum|Edad:Sex][Fieb|Sex]
  [Tos|Sex][DG|Sex][Disn|Sex][Esc|Sex][Diar|Sex][Vom|Sex]
  [FRA|Sex][OSR|Sex][SDRA|Sex][Hosp|Sex:Neum][UCI|Sex:Neum]
  [Def|Edad:Sex][EnCar|Sex:Neum][EnRes|Sex:Neum][Diab|Sex:Neum]")
net

redD = custom.fit(net, dist = list(Sex = PSex, Edad = PEdadSex,
Neum = PNeumSexEdad, Fieb = PFieb, Tos = PTos,DG =PDG,
Disn = PDisn, Esc = PEsc, Diar = PDiar, FRA = PFRA,
    OSR = POSR,SDRA = PSDRA, Hosp= PHosp, UCI = PUCI,
    Def = PDef, EnCar = PEnCar, EnRes= PEnRes,
Diab= PDiab, Vom = PVom), ordinal = c("Edad","Sex"))
redD
```

3.2. Aprendizaje de la red

En esta sección vamos a ver como los métodos de aprendizaje funcionan y para eso hemos diseñado el siguiente experimento.
Tomando de referencia la red modelada anteriormente, a la que llamamos *red diana*, vamos a utilizar *R* para aprenderla.

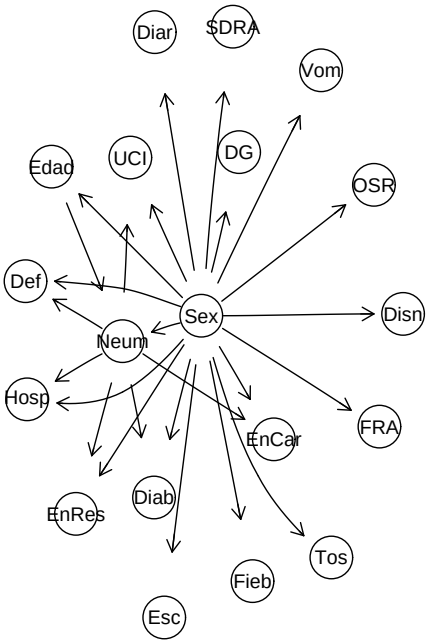


Figura 3.3: Red Diana.

En la siguiente representación podemos observar con mayor claridad como existe una estructura en la que hay dos nodos que jerárquicamente influyen en otro subconjunto de nodos.

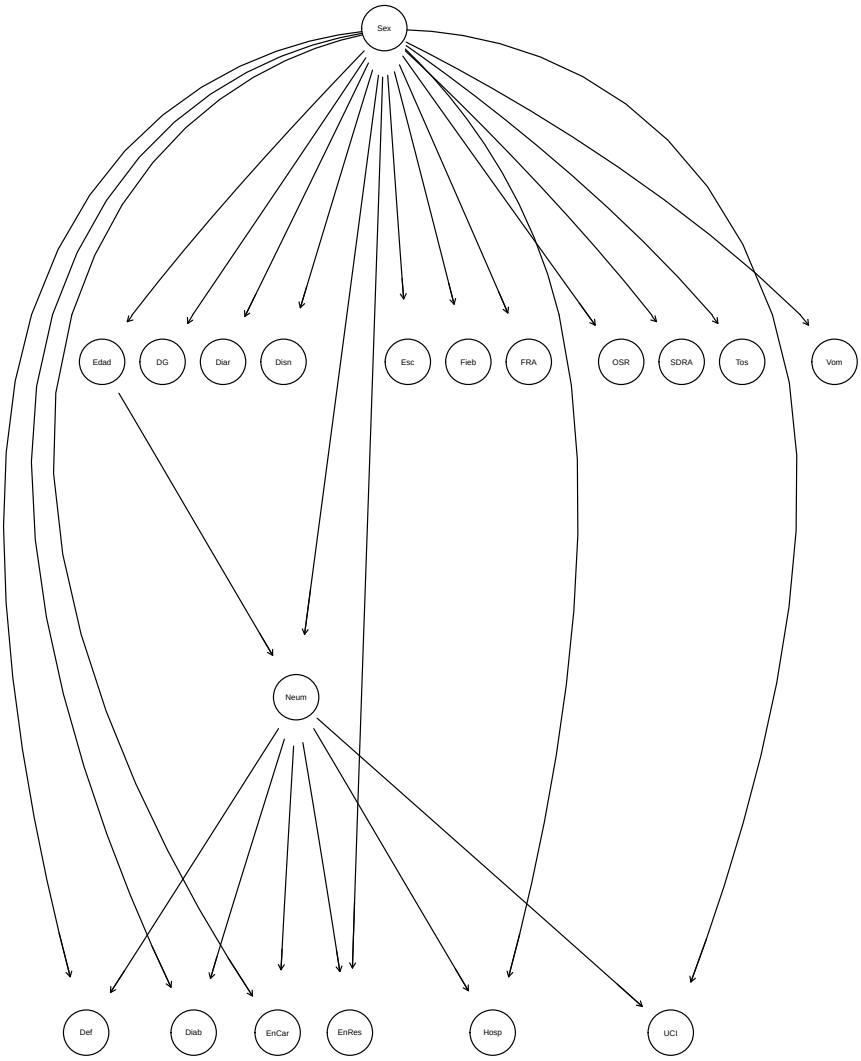


Figura 3.4: Red Diana.

Como estamos trabajado con una red definida directamente mediante probabilidades condicionadas de la cual no tenemos datos por individuos, para aprender la red primero tendremos que hacer una simulación de la que

podría ser nuestra muestra real de datos.
Usaremos para realizar esto las librerías,

```
library(bnlearn)
library(Rgraphviz)
```

Con el siguiente comando obtendríamos la red aprendida.

```
par(mfrow = c(1,2))
graphviz.plot(net, layout = "fdp", main = "RED DIANA")
sim = rbn(redD, n, redD)
sim
graphviz.plot(gs(sim), layout = "fdp", main = "RED APRENDIDA")
```

La simulación (sim) se realiza con el comando rbn al cual le damos nuestra *red diana*, el número de simulaciones que queremos que nos haga (n) y de donde tiene que tomar dichos datos. Con $gs(sim)$ convertimos en grafo nuestra simulación y después la pintamos. La *red aprendida* se parecerá más a la *red diana* cuanto mayor sea n. Vamos a ver esto en las siguientes gráficas.

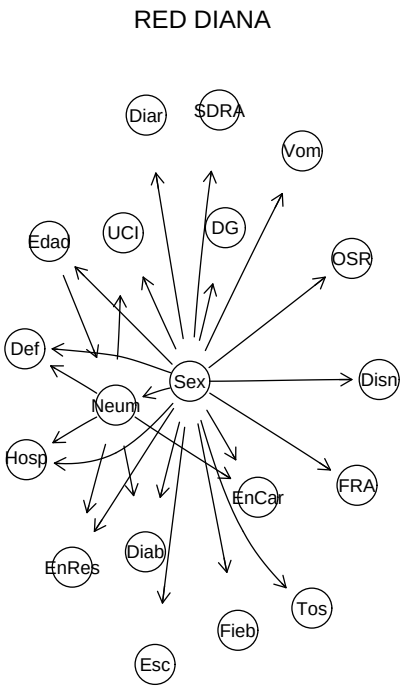


Figura 3.5: Red diana.

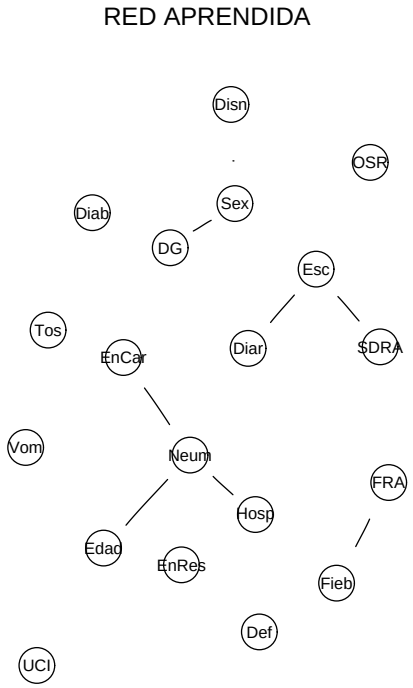


Figura 3.6: Red aprendida para n=1000.

Al hacer la simulación con un número del tamaño de 1000, R nos dibuja la red pero nos devuelve el siguiente error.

```
> graphviz.plot(gs(sim), layout = "fdp", main = "RED APRENDIDA")
Warning message:
In check.data(x, allow.missing = TRUE) :
variable Edad has levels that are not observed in the data.
```

En él dice que hay niveles de la variable *Edad* que no aparecen en la simulación. Esto se debe a que hay grupos de edades con una probabilidad tan pequeña que en 1000 simulaciones no aparecen, como bien dijimos en el capítulo anterior el hecho de tener una evidencia con una probabilidad muy pequeña supone un problema y podría hacer que los métodos necesitasen muchísimas simulaciones. La *red aprendida* añade arcos para hacer frente a la simulación, por lo tanto si la simulación es pequeña no hacen faltan más arcos para explicar las dependencias que establecen los datos. Esto ocurre en este caso, solo obtenemos 6 arcos y ninguno dirigido. Aumentemos las simulaciones y veamos que ocurre entonces.

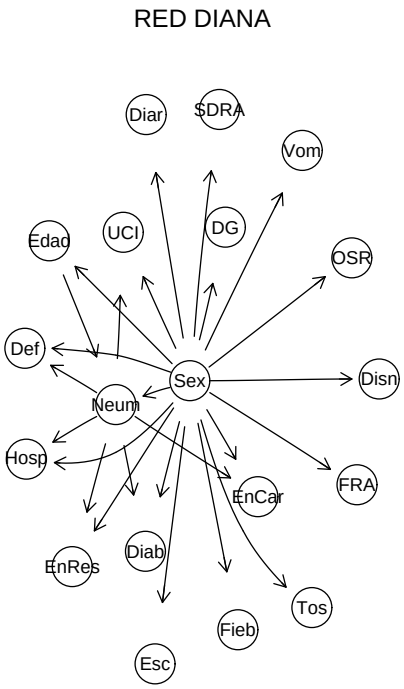


Figura 3.7: Red diana.

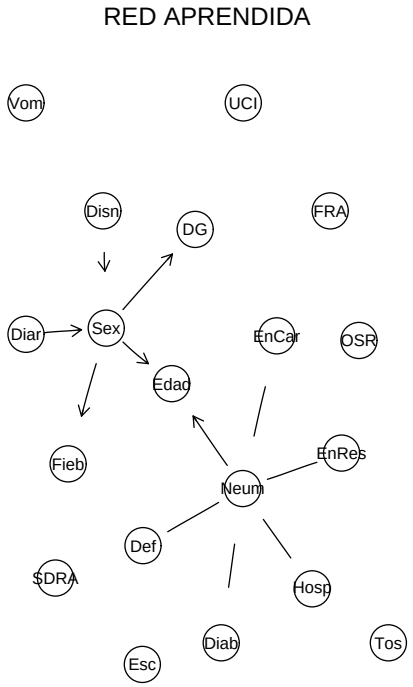


Figura 3.8: Red aprendida para n=100000.

En esta red creada a partir de 100000 simulaciones obtenemos ya más arcos y algunos de ellos ya dirigidos, ya no nos aparece ningún error por lo que podemos suponer que aparecen todos los grupos de *edad* con los que trabajamos. Pero seguimos teniendo muchas variables libres y bastantes arcos no dirigidos.

Aumentamos las simulaciones hasta 100 millones y obtenemos una red bastante parecida a la de partida, hay que tener en cuenta que hoy en día muchas de las aplicaciones donde aparecen las redes Bayesianas obtienen los datos de los clientes a partir de la información suministrada a través de Internet por lo cual se tiene una cantidad masiva de información.

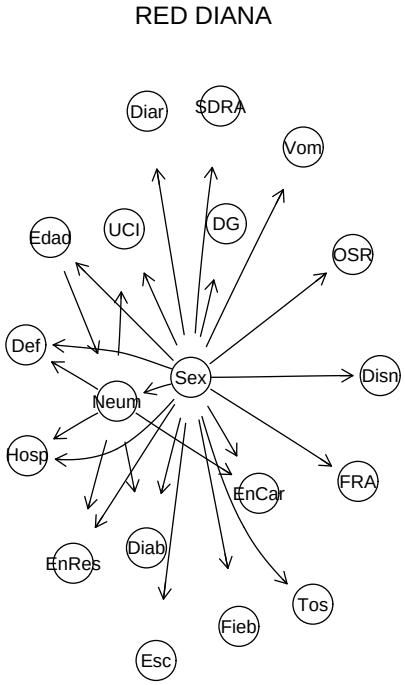


Figura 3.9: Red diana.

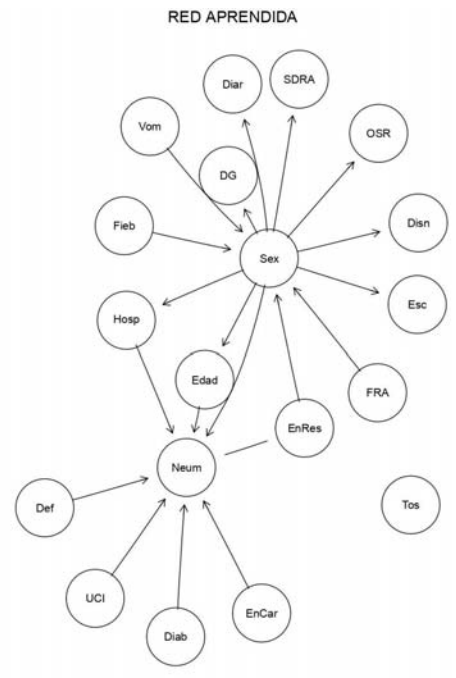


Figura 3.10: Red aprendida para n=100 millones.

En primer lugar, podemos diferenciar ya claramente dos conjuntos de nodos en los cuales el centro es *Sexo* y *Neumonia*, al igual que pasa en la *red diana*. Respecto a los arcos, obtenemos 20 arcos de los 25 que posee la original y solo uno de ellos no dirigido. Además todos los nodos tienen al menos un padre exceptuando la variable *Tos* que queda libre. La falta de algunos arcos se puede deber a que en la *red diana* hayamos añadidos arcos que no sean necesarios para la información cuantitativa que tenemos. Por ejemplo, en la variable *Tos* se asume una dependencia en función de *Sexo* pero los datos obtenidos en la simulación no establecen esa dependencia, luego ese arco no aparece en la *red aprendida*.
Respecto al sentido de los arcos, observamos que no es el mismo en la *red diana* que en la *red aprendida* esto es debido a que se aprende la clase de equivalencia de los arcos, que es la misma sea el sentido que sea siempre y cuando no se generen v-estructuras.

3.3. Gráficas

En esta sección se compararán las gráficas obtenidas mediante la simulación con las gráficas reales que aparecen en el informe *RENAVE* (Apéndice C). El objetivo que se persigue es el de validar experimentalmente la red diana que se ha modelado a través de los datos del informe. Para ello se resumirá por medio de histogramas de frecuencias la base de datos simulada a partir de la red diana y se contrastarán los resultados reales.
Usaremos las siguientes librerías de *R*.

```
library(ggplot2)
library(tidyr)
library(car)
```

Veamos la distribución de la población con *COVID-19* según *Sexo* y *Edad*.

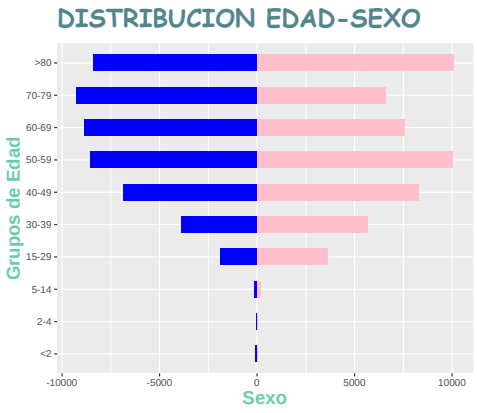


Figura 3.11: Distribución de contagiados según Sexo y Edad en la simulación.

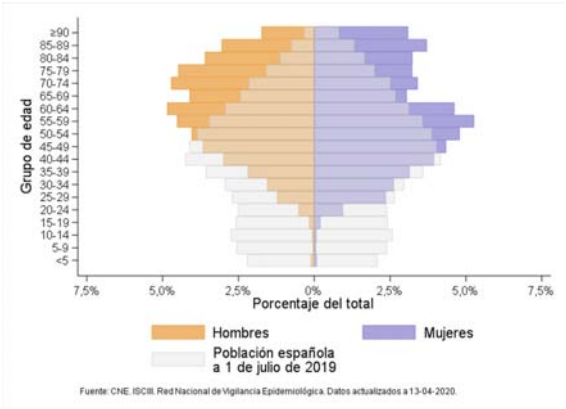


Figura 3.12: Distribución de contagiados según Sexo y Edad en el informe RENAVE.

En la primera comparación podemos observar que la *Figura* 3.11, creada a partir de la simulación, sigue una distribución muy parecida a la de la *Figura* 3.12. Cabe destacar que algunas diferencias son provocadas por los distintos grupos de edad tomados para el estudio, en nuestra figura tomamos solo 10 subconjuntos de *Edad* mientras que en la figura del informe *RENAVE* toma 19 subconjuntos. Aún con estas diferencias se puede ver que la gráfica correspondiente a los hombres en ambos casos sigue una distribución unimodal mientras que la de las mujeres sigue una bimodal con los picos más altos en las edades de 50-59 y $\geq$80.

En la siguiente gráfica tomaremos la distribución de los *No hospitalizados* según *Sexo* y *Edad*.

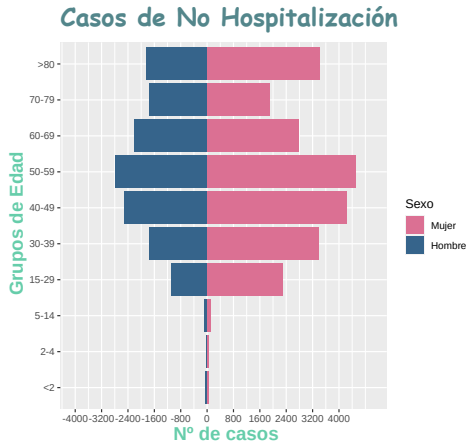


Figura 3.13: Nº de casos de no hospitalizados agrupados por Sexo y Edad en la simulación.

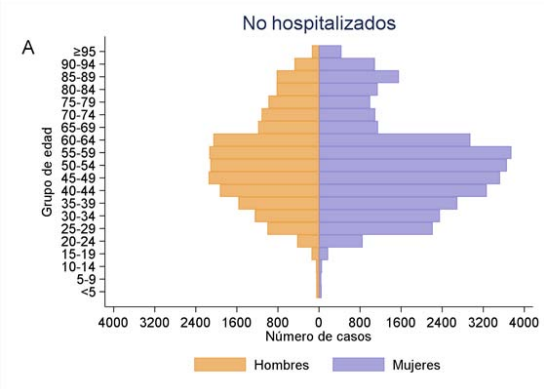


Figura 3.14: Nº de casos de no hospitalizados agrupados por Sexo y Edad en el informe RENAVE.

Al igual que en las gráficas anteriores, nos encontramos con diferencias en el número de franjas de edad tomadas, lo cual nos produce ciertas discrepancias en el número de casos obtenidos en cada grupo de edad. Sin embargo obtenemos distribuciones muy parecidas, en la gráfica correspondiente a los hombres observamos que ambas siguen una distribución unimodal alcanzando el máximo en la franja de edad correspondiente a 50-59 años. En el caso de las mujeres, obtenemos una distribución bimodal, en los dos casos, alcanzando los dos picos en las franjas de 50-59 y $\geq$80.

Por último veremos los casos de defunciones según el grupo de edad.

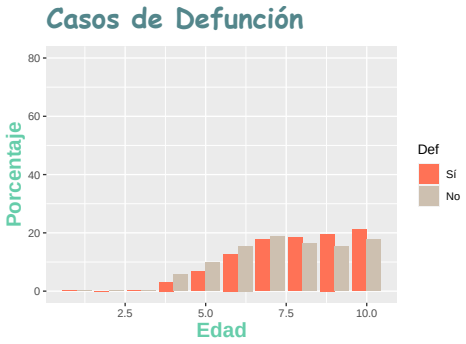


Figura 3.15: Porcentaje de muertes según la edad en la simulación.

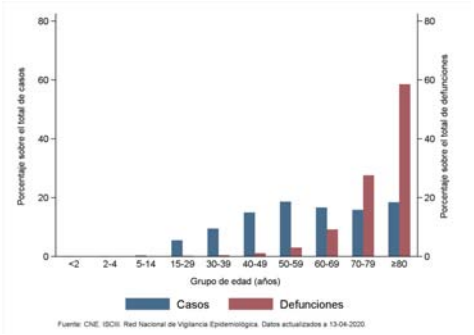


Figura 3.16: Porcentaje de muertes según la edad en el informe RENAVE.

En primer lugar, aclaramos que en el informe cuando se refiere a casos realmente se trata de los casos de *COVID-19* que no son defunciones por lo tanto tiene sentido que comparemos ambas gráficas ya que están re-

presentando lo mismo. Comparando ambas figuras, podemos ver que hay discrepancias evidentes.

Si nos quedamos solo con el porcentaje de *No defunciones* de la *Figura* 3.15 y el porcentaje de *casos* en la *Figura* 3.16, observamos que se parecen bastante. Ambas toman el máximo en la franja de edad de 50-59 años con un poco menos del 20 %, después la gráfica decrece en los siguientes grupos de edad, para volver a crecer en la franja de mayores de 80 años donde el porcentaje es muy próximo al máximo.
Sin embargo, si nos fijamos en los porcentajes de muertes de ambas figuras podemos observar ciertas diferencias. Aunque ambas son crecientes y toman sus máximos en la edad de mayores de 80, nuestra simulación nos da una menor concentración de muertes en las franjas de edades mayores mientras que en las edades más pequeñas nos muestra una mayor concentración comparándola con la gráfica del informe *RENAVE*. Esto nos hace pensar que hay algún fallo en nuestra *red diana* ya que no se disponía de los datos desagregados para poder estimar de un modo más realista las probabilidades condicionadas.

Para solventar este error, hacemos un cambio en la *red diana*, sustituimos el arco entre los nodos *defunción* y *neumonía* por el arco dirigido entre el nodo *edad* y el nodo *defunción*. Con este cambio reforzamos la dependencia existente entre *edad* y *defunción* ya que en la estructura anterior esta dependencia se eliminaba al conocer si el paciente padecía o no neumonía.
El cambio no afecta al experimento de aprendizaje inicial puesto que lo que se comprueba en ese caso es cómo aprende de una red diana. Este tipo de rectificación de la red del modelo suele ocurrir en la práctica cuando se observa discrepancias de este tipo ya que en muchos casos en el modelo se incluyen estimaciones subjetivas de las probabilidades condicionadas.
Con este cambio obtenemos la siguiente gráfica.

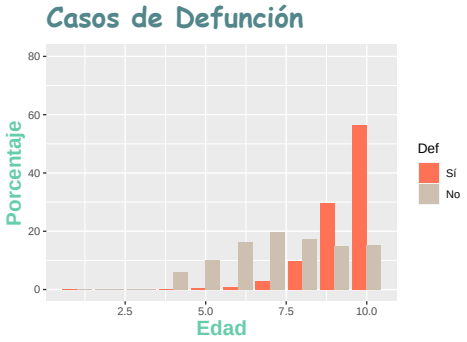


Figura 3.17: Porcentaje de muertes según la edad en la simulación.

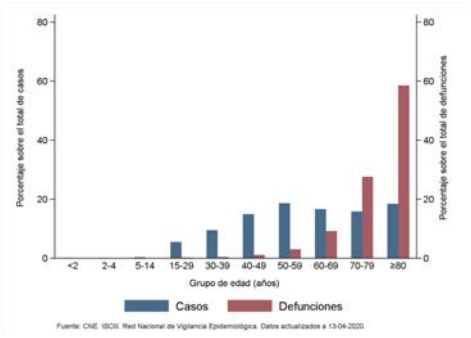


Figura 3.18: Porcentaje de muertes según la edad en el informe RENAVE.

Cambiando un único arco en la *red Diana* hemos podido conseguir que la

simulación con la que trabajamos cambie en el número de muertes según el rango de edad y obtengamos una gráfica prácticamente igual a la del informe *RENAVE*. Donde vemos que el número de muertes es prácticamente cero en los grupos de edad hasta los 30 años y empiezan a aumentar hasta que en mayores de 80 años obtenemos un 60 % del total de las muertes. Por otro lado, la gráfica de las no defunciones no varía apenas respecto a la anterior que si se asemejaba bastante a la gráfica del informe. Por lo que obtenemos una muy buena aproximación en el número de muertes.

Como conclusión, podemos decir que nuestra simulación explica bastante bien la realidad que queríamos estudiar. Por lo tanto, hemos visto que sin tener los datos por individuos, mediante una *red Bayesiana* podemos simularlos y obtener una muy buena aproximación de ellos. En la práctica este proceso de modelado se puede completar con la realización de contrastes de bondad de ajuste entre la simulación y los datos reales lo cual permite además contar con el nivel de significación del modelo usado.

3.4. Explicación más verosímil

En la sección anterior hemos visto que a partir de la *red Diana* podemos simular los datos y trabajar con ellos. En esta sección, vamos a dar la explicación más probable de un suceso a partir de nuestra *red diana*. Para ello usaremos la siguiente librería de R :

```
library(gRain)
```

Queremos saber cuál será la respuesta más probable a partir de unas evidencias.
Supongamos que queremos saber las implicaciones en cuanto al sexo del paciente del hecho de pertenecer al intervalo de edad 50-59 no padeciendo neumonía. Tendríamos de evidencias el hecho de tener entre 50-59 años y el de no padecer neumonía. El siguiente comando nos daría la respuesta más verosímil:

```
> particles = cpdist(redD, nodes = "Sex",
                + evidence = ( Neum == "No")&(Edad == "50-59"))
> prop.table(table(particles))
particles
Mujer    Hombre
0.6189904 0.3810096
> names(which.max(prop.table(table(particles))))
[1] "Mujer"
```

Nos dice que lo más probable es que ese individuo sea una *Mujer*. Si comparamos estas probabilidades con las probabilidades marginales de la variable *Sexo*.

```
> PSex
Mujer Hombre
[1,] 0.522  0.478
```

Observamos que la evidencia aumenta en aproximadamente un 20 % la probabilidad de ser mujer respecto a la que teníamos inicialmente.

Otro ejemplo sería el siguiente. Queremos saber a qué grupo de *edad* corresponde un *Hombre* sabiendo que padece el *COVID-19* y no fallece.

```
> particles2 = cpdist(redD, nodes = "Edad",
                  + evidence = (Def == "No")&(Sex == "Hombre"))
> prop.table(table(particles2))

> names(which.max(prop.table(table(particles2))))
[1] "50-59"
```

Nos dice que lo más probable es que sea un hombre de 50-59, ya que es el hecho de mayor probabilidad.

También podemos obtener la probabilidad de que ocurra un suceso concreto dada una evidencia. Por ejemplo, supongamos que queremos comparar la probabilidad de *defunción* siendo *Hombre o Mujer*.

```
> cpquery(redD,event=(Sex == "Hombre"),evidence = (Def == "Si"))
[1] 0.6530612
> cpquery(redD,event=(Sex == "Mujer"),evidence = (Def == "Si"))
[1] 0.3621838
```

Obtenemos que un hombre tiene una probabilidad de morir de 0.653 mientras que en el sexo mujer obtenemos una probabilidad de 0.362. Con esto podemos decir, que un hombre con *COVID-19* tiene mayor probabilidad de morir que una mujer.

Queremos comparar ahora la probabilidad de morir en dos grupos de edad, jóvenes entre 15-29 años frente a mayores de más de 80 años.

```
> cpquery(redD,event=(Edad == "15-29"),evidence=(Def == "Si"))
[1] 0.001298701
> cpquery(redD,event=(Edad == ">80"),evidence=(Def == "Si"))
[1] 0.5669291
```

Esto nos dice que un joven tiene una probabilidad casi nula de morir mientras que un anciano tiene una probabilidad de 0.5669 de morir por la enfermedad. Por lo cual, se puede afirma que la probabilidad de defunción está relacionada con la edad y los mayores tienen mayor probabilidad.

Con esto, hemos obtenido un método por el cuál obtenemos respuestas a nuestras preguntas.

Apéndice A

Conceptos de Teoría de Grafos

Esta sección se dedica a aclarar términos sobre la teoría de grafos que serán usados a lo largo del trabajo.

Sea G =(V,A) donde V es el conjunto de nodos y A $\subset$ V $\times$ V es el conjunto de arcos. Se dice que G es un grafo dirigido si todos sus arcos tienen un sentido definido. Un arco a=(b,c) se considera dirigido desde b hacia c donde b se denomina nodo inicial y c se denomina nodo final. Un arco no dirigido entre un nodo A y un nodo B, es decir un arco por el cual se puede ir de A hacia B o de B hacia A, se llamará eje. Dos nodos son adyacentes si están conectados por una arco.
Se llama camino dirigido a una secuencia de nodos dentro de un grafo tal que existe un arco dirigido entre dos nodos adyacentes, es decir el nodo final de un arco es el nodo inicial del siguiente.
Un ciclo es un camino cerrado, es decir el primer y el último nodo del camino coinciden. Un grafo que solo pasa una vez por cada nodo y no posee ciclos, se llama grafo acíclico.

Hay que mencionar las relaciones entre los nodos dentro de un grafo, para ello introducimos la imagen A,1

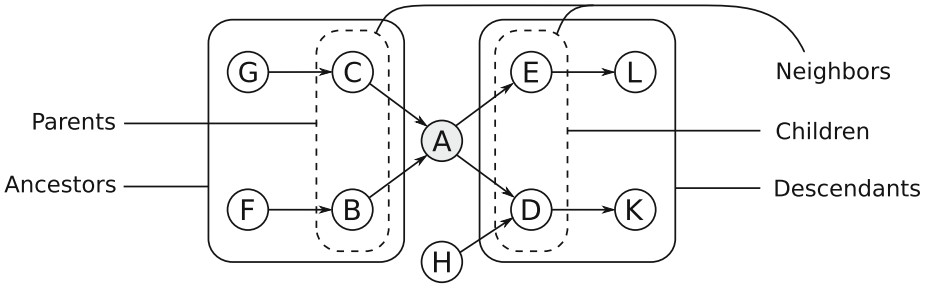


Figura A.1: Padres, hijos, antecesores, descendientes y vecinos de un nodo A en un grafo dirigido.

Fijamos un nodo A, respecto de dicho nodo podemos comenzar por identificar en el resto de nodos dos grupos, nodos antecesores y nodos descendientes. Se dice que un nodo es antecesor de A si desde ese nodo puedo llegar hasta A siguiendo un camino dirigido. Llamamos descendiente de A a un nodo al cual llego mediante un camino que parte de A.
Dentro de los antecesores cabe destacar a los padres de A, que son los nodos antecesores de A que además son adyacentes, es decir son los más próximos dentro de los antecesores.
Dentro de los descendientes encontramos a los hijos de A, que son los nodos descendientes de A que además son adyacentes, es decir son los más próximos a A dentro de los descendientes.
El conjunto de padres e hijos de A se llama conjunto de vecinos o vecindad.

Apéndice B

Scripts del software R

```
##PAQUETES NECESARIOS
install.packages("bnlearn")
library(bnlearn)
install.packages('ggplot2', dependencies = TRUE)
library(ggplot2)
install.packages("tidyverse")
library(tidyr)
install.packages("carData")
library(car)
install.packages("fdth")
library(fdth)
#install.packages("BiocManager")
#BiocManager::install("Rgraphviz")
library(Rgraphviz)
install.packages("gRbase")
install.packages("gRain")
library(gRain)
```

```
##REPRESENTAR MI RED MODELO

#INTRODUCIMOS PROBABILIDADES CONDICIONADAS

#Calculamos las probabilidades de los padres
#que encontramos en el grafo (Sex,Edad,Neum)
PSex = matrix(c(0.522, 0.478), ncol = 2,
```

```
               + dimnames = list(NULL, c("Mujer", "Hombre")))
PSex
PEdad = matrix(c(0.001, 0.001, 0.003, 0.055, 0.095,
                   0.15, 0.186, 0.166, 0.159, 0.184 ),
               + ncol =10 , dimnames = list(NULL,
               + c("<2", "2-4", "5-14", "15-29",
                        "30-39", "40-49", "50-59",
                         "60-69", "70-79", ">=80")))
PEdad
PNeum= matrix(c(0.6, 0.4), ncol = 2,
            + dimnames = list(NULL, c("Si", "No")))
PNeum

#PROB FIEBRE|SEX
PFieb = c(0.687, 0.313, 0.791, 0.209)
dim(PFieb) = c(2,2)
dimnames(PFieb) = list("Fieb" = c("Si", "No"),
                       +"Sex" =  c("Mujer", "Hombre"))
PFieb

#PROB TOS|SEX
PTos= c(0.741, 0.259, 0.748, 0.252)
dim(PTos) = c(2, 2)
dimnames(PTos) = list("Tos" = c("Si", "No"),
                      +"Sex" =  c("Mujer", "Hombre"))
PTos

#PROB DG|SEX
PDG = c(0.301, 0.699, 0.213, 0.787)
dim(PDG) = c(2, 2)
dimnames(PDG) = list("DG" = c("Si", "No"),
                  +   "Sex" =  c("Mujer", "Hombre"))
PDG

#PROB Disn|SEX
PDisn = c(0.444, 0.556, 0.515, 0.485)
dim(PDisn) = c(2, 2)
dimnames(PDisn) = list("Disn" = c("Si", "No"),
                       + "Sex" =  c("Mujer", "Hombre"))
PDisn

#PROB Esc|SEX
PEsc = c(0.336, 0.664, 0.349, 0.651)
dim(PEsc) = c(2, 2)
```

```
dimnames(PEsc) = list("Esc" = c("Si", "No"),
                      + "Sex" =  c("Mujer", "Hombre"))
PEsc

#PROB Vom|SEX
PVom = c(0.110, 0.89, 0.076, 0.924)
dim(PVom) = c(2, 2)
dimnames(PVom) = list("Vom" = c("Si", "No"),
                      + "Sex" =  c("Mujer", "Hombre"))
PVom

#PROB Diarr|SEX
PDiar = c(0.320, 0.68, 0.282, 0.718)
dim(PDiar) = c(2, 2)
dimnames(PDiar) = list("Diar" = c("Si", "No"),
                      +  "Sex" =  c("Mujer", "Hombre"))
PDiar

#PROB SDRA|SEX
PSDRA = c(0.051, 0.949, 0.087, 0.913)
dim(PSDRA) = c(2, 2)
dimnames(PSDRA) = list("SDRA" = c("Si", "No"),
                       + "Sex" =  c("Mujer", "Hombre"))
PSDRA

#PROB OSR|SEX
POSR = c(0.138, 0.862, 0.181, 0.819)
dim(POSR) = c(2, 2)
dimnames(POSR) = list("OSR" = c("Si", "No"),
                     + "Sex" =  c("Mujer", "Hombre"))
POSR

#PROB FRA|SEX
PFRA = c(0.031, 0.969, 0.06, 0.94)
dim(PFRA) = c(2, 2)
dimnames(PFRA) = list("FRA" = c("Si", "No"),
                      + "Sex" =  c("Mujer", "Hombre"))
PFRA

#PROB EDAD|Sex
PEdadSex = c(0.001, 0.001, 0.003, 0.067, 0.11, 0.159, 0.193,
             0.147, 0.127, 0.192, 0.002, 0.001, 0.003, 0.04,
              0.079, 0.14, 0.179, 0.188, 0.194, 0.174)
dim(PEdadSex) = c(10, 2)
```

```
dimnames(PEdadSex) = list("Edad" = c("<2", "2-4","5-14","15-29",
                  "30-39","40-49","50-59","60-69","70-79",">80"),
                        + "Sex" =  c("Mujer", "Hombre"))
PEdadSex

#PROB Hosp|Sex,Neum
PHosp = c(0.91, 0.09, 0.92, 0.08, 0.235, 0.765, 0.27, 0.73)
dim(PHosp) = c(2, 2, 2)
dimnames(PHosp) = list("Hosp" = c("Si", "No"),
                     + "Sex" =  c("Mujer", "Hombre"),
"Neum" = c("Si", "No"))
PHosp

#PROB UCI|Sex,Neum
PUCI = c(0.081, 0.919, 0.083, 0.917, 0.012, 0.988, 0.012, 0.988)
dim(PUCI) = c(2, 2, 2)
dimnames(PUCI) = list("UCI" = c("Si", "No"),
                    + "Sex" =  c("Mujer", "Hombre"),
"Neum" = c("Si", "No"))
PUCI

#PROB DEF|Sex,Edad
PDef = c(0.022, 0.978, 0, 1, 0, 1, 0.001, 0.999, 0.002, 0.998, 0.004,
          0.996, 0.008, 0.992, 0.03, 0.97, 0.1, 0.9, 0.18, 0.82, 0.025,
          0.975, 0, 1, 0, 1, 0.003, 0.997, 0.004, 0.996, 0.007, 0.993,
          0.016, 0.984, 0.056, 0.944, 0.17, 0.83, 0.3, 0.7)
dim(PDef) = c(2, 10, 2)
dimnames(PDef) = list("Def" = c("Si", "No"),
                + "Edad" = c("<2", "2-4","5-14","15-29","30-39","40-49",
                     "50-59","60-69","70-79",">80"),
                + "Sex" = c("Mujer", "Hombre"))
PDef

#PROB EnCar|Sex,Neum
PEnCar = c(0.347, 0.653, 0.352, 0.648, 0.174, 0.826, 0.175, 0.825)
dim(PEnCar) = c(2, 2, 2)
dimnames(PEnCar) = list("EnCar" = c("Si", "No"),
                      + "Sex" =  c("Mujer", "Hombre"),
"Neum" = c("Si", "No"))
PEnCar

#PROB EnRes|Sex,Neum
PEnRes = c(0.099, 0.901, 0.101, 0.899, 0.05, 0.95, 0.057, 0.943)
dim(PEnRes) = c(2, 2, 2)
```

```
dimnames(PEnRes) = list("EnRes" = c("Si", "No"),
                      + "Sex" =  c("Mujer", "Hombre"),
                      + "Neum" = c("Si", "No"))
PEnRes

#PROB Diab|Sex,Neum
PDiab = c(0.174, 0.826, 0.177, 0.823, 0.078, 0.922, 0.079, 0.921)
dim(PDiab) = c(2, 2, 2)
dimnames(PDiab) = list("Diab" = c("Si", "No"),
                      + "Sex" =  c("Mujer", "Hombre"),
"Neum" = c("Si", "No"))
PDiab

#PROB Neum|Sex,Edad
PNeumSexEdad = c(0.127, 0.873, 0.143, 0.857, 0.151, 0.849, 0.174,
                 0.826, 0.268, 0.732, 0.381, 0.619, 0.461, 0.539,
                 0.598, 0.402, 0.697, 0.303, 0.635, 0.365, 0.222,
                 0.778, 0.277, 0.723, 0.25, 0.75, 0.295, 0.705,
                 0.421, 0.579, 0.552, 0.448, 0.629, 0.371, 0.747,
                 0.253, 0.82, 0.18, 0.776, 0.224)
dim(PNeumSexEdad) = c(2, 10, 2)
dimnames(PNeumSexEdad) = list("Neum" = c("Si", "No"),
                            + "Edad" = c("<2", "2-4","5-14","15-29",
                      "30-39","40-49","50-59","60-69","70-79",">80"),
                            + "Sex" = c("Mujer", "Hombre"))
PNeumSexEdad


##USANDO LAS PROBABILIDADES DEFINIDAS ANTERIORMENTE,
# DEFINIMOS NUESTRA RED Y LA DIBUJAMOS
net=model2network("[Sex][Edad|Sex][Neum|Edad:Sex][Fieb|Sex]
               + [Tos|Sex][DG|Sex][Disn|Sex][Esc|Sex][Diar|Sex]
               + [Vom|Sex][FRA|Sex][OSR|Sex][SDRA|Sex]
               + [Hosp|Sex:Neum][UCI|Sex:Neum][Def|Edad:Sex]
               + [EnCar|Sex:Neum][EnRes|Sex:Neum][Diab|Sex:Neum]")
net
graphviz.plot(net, layout = "fdp", main = "RED DIANA")
redD = custom.fit(net, dist = list(Sex = PSex, Edad = PEdadSex,
                      +  Neum = PNeumSexEdad, Fieb = PFieb,
                      +  Tos = PTos,DG =PDG, Disn = PDisn,
                      +  Esc = PEsc, Diar = PDiar, FRA = PFRA,
                      +  OSR = POSR,SDRA = PSDRA, Hosp= PHosp,
                      +  UCI = PUCI, Def = PDef, EnCar = PEnCar,
                      +  EnRes= PEnRes, Diab= PDiab, Vom = PVom),
```

```
                  + ordinal = c("Edad","Sex"))
redD



##A PARTIR DE AHORA REALIZAREMOS UNA SIMULACIÓN DE TAMAÑO n
# Y OBTENDREMOS LA RED APRENDIDA.
n=100000
par(mfrow = c(1,2))
graphviz.plot(net, layout = "fdp", main = "RED DIANA")
sim = rbn(redD, n, redD)
sim
graphviz.plot(gs(sim), layout = "fdp", main = "RED APRENDIDA")


##AHORA REALIZAREMOS LAS GRAFICAS MÁS SIGNIFICATIVAS
#PARA COMPARARLAS CON EL INFORME A PARTIR DE NUESTRA SIMULACIÓN

attach(sim)


#GRAFICA POBLACIONAL, CASOS DE COVID-19 SEGÚN SEXO Y EDAD
Tit = element_text(family="Comic Sans MS",
size=rel(2),
vjust=2,
face="bold",
color="cadetblue4",
lineheight=1.5)
ggplot(sim, aes(x = Edad,
y = recode(as.numeric(Sex),"1=1;2=-1"),
fill = Sex)) +
geom_col(data = subset(sim, Sex == "Hombre"),
width = 0.5, fill = "blue") +
geom_col(data = subset(sim, Sex == "Mujer"),
width = 0.5, fill = "pink") +
coord_flip() +
theme (plot.title = Tit)+
theme (axis.title = element_text(face="bold", colour="aquamarine3",
                                 size=rel(1.5)))+
ggtitle("DISTRIBUCIÓN EDAD-SEXO") +
xlab("Grupos de Edad") + ylab("Sexo")

## GRAFICA SITUACION CLINICA
```

```
#CASOS DE NO HOSPITALIZADOS SEGÚN SEXO Y EDAD
Tit = element_text(family="Comic Sans MS",
size=rel(2),
vjust=2,
face="bold",
color="cadetblue4",
lineheight=1.5)


ggplot(sim, aes(x=Edad, fill =Sex,
y=ifelse(Sex=="Hombre",-(as.numeric(Hosp =="No")),
          as.numeric(Hosp=="No")))) +
geom_bar(stat="identity")+
coord_flip()+
labs(title="Casos de No Hospitalización",
x= " Grupos de Edad",
y=" N° de casos") +
scale_fill_manual(values=c("palevioletred", "steelblue4"),
"Sexo",
labels=c("Mujer","Hombre")) +
theme (plot.title = Tit)+
theme (axis.title = element_text(face="bold",
          colour="aquamarine3", size=rel(1.5)))+
scale_y_continuous(limit = c(-4000,5000), breaks=seq(-4000, 4000, 800))

#GRAFICA DEL PORCENTAJE DE MUERTES POR COVID-19 SEGÚN LA EDAD
#COMPARADA CON NO MUERTE SEGUN LA EDAD.

ggplot(sim, aes(as.numeric(Edad),fill=Def)) +
geom_bar(aes(y=c(..count..[..group..==1]/sum(..count..[..group..==1]),
..count..[..group..==2]/sum(..count..[..group..==2]))*100),

position="dodge") +
labs(title="Casos de Defunción",
x= "Edad",
y="Porcentaje") +
scale_fill_manual(values=c("coral1", "antiquewhite3"),
"Def",
labels=c("Sí","No")) +
theme (plot.title = Tit)+
theme (axis.title = element_text(face="bold",
        colour="aquamarine3", size=rel(1.5)))+
scale_y_continuous(limit=c(0,80))
```

```
#RESPUESTA MÁS VEROSIMIL

#CASO MÁS PROBABLE DE UNA VARIABLE DADAS DOS EVIDENCIAS

#queremos predecir el sexo de un individuo entre 50-59 años
#que padece Covid-19 y no padece de neumoia
particles = cpdist(redD, nodes = "Sex",
            evidence = ( Neum == "No")&(Edad == "50-59"))
prop.table(table(particles))
names(which.max(prop.table(table(particles))))

#queremos predecir la edad de un hombre con covid y
#que no muere
particles2 = cpdist(redD, nodes = "Edad",
             evidence = (Def == "No")&(Sex == "Hombre"))
prop.table(table(particles2))
names(which.max(prop.table(table(particles2))))

#PROBABILIDAD DE UN SUCESO CONCRETO

#queremos comparar la probabilidad de defunción según el Sexo
cpquery(redD,event = (Sex == "Hombre") ,evidence = (Def == "Si") )
cpquery(redD,event = (Sex == "Mujer")  ,evidence = (Def == "Si") )

#queremos comparar la probabilidad de defunción según la edad
cpquery(redD,event = (Edad == "15-29") ,evidence = (Def == "Si") )
cpquery(redD,event = (Edad == ">80")  ,evidence = (Def == "Si") )
```

Apéndice C

Informe de la Red Nacional de Vigilancia Epidemiológica

Informe sobre la situación de COVID-19 en España

Informe COVID-19 nº 22. 13 de abril de 2020

Contenido

Introducción

En diciembre de 2019 surgió un agrupamiento de casos de neumonía en la ciudad de Wuhan (provincia de Hubei, China), con una exposición común a un mercado mayorista de marisco, pescado y animales vivos. El 7 de enero de 2020, las autoridades chinas identificaron como agente causante del brote un nuevo virus de la familia *Coronaviridae* que posteriormente fue denominado SARS-CoV-2[1]. La secuencia genética fue compartida por las autoridades chinas el 12 de enero. La enfermedad causada por este nuevo virus se ha denominado por consenso internacional COVID-19. El Comité de Emergencias del Reglamento Sanitario Internacional (RSI, 2005) declaró el brote como una Emergencia de Salud Pública de Importancia Internacional (ESPII) en su reunión del 30 de enero de 2020. Posteriormente, la OMS lo reconoció como una pandemia global el 11 de marzo de 2020.

En España, las comunidades autónomas (CCAA) notifican diariamente al Ministerio de Sanidad las cifras de casos confirmados acumulados de COVID-19: total de casos, casos en profesionales sanitarios, hospitalizaciones, ingresos en UCI, fallecidos y casos recuperados.

Al mismo tiempo, las CCAA completan, según acceden a la información, la encuesta individualizada para cada uno de dichos casos. La encuesta incluye información clínico-epidemiológica consensuada y aprobada por la Ponencia de Alertas y Planes de Preparación y Respuesta y la Red Nacional de Vigilancia Epidemiológica (RENAVE), y la notifican mediante la plataforma informática SiViES (Sistema para la Vigilancia en España), que gestiona el Centro Nacional de Epidemiología.

Este informe contiene información de los casos de COVID-19 notificados al Centro Nacional de Epidemiología a través de la plataforma SiViES hasta la extracción de datos (12:00 h del 13 de abril de 2020): 113.407 casos que suponen el 67% de los 169.496 totales en España cuantificados hasta el día 12 de abril de 2020 (21:00 h). Su objetivo es obtener una información detallada sobre las características clínicas y epidemiológicas de los casos de COVID-19 y los factores que pueden estar asociados a una mayor gravedad. Los resultados deben confirmarse con posteriores actualizaciones de COVID-19 en SiViES.

[1] https://www.who.int/dg/speeches/detail/who-director-general-s-remarks-at-the-media-briefing-on-2019-ncov-on-11-february-2020

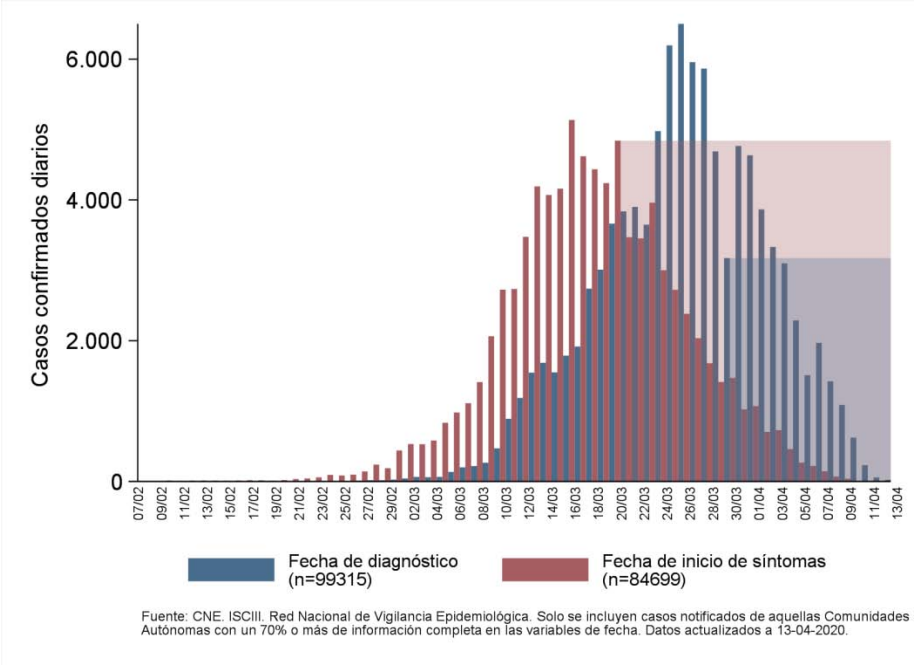
Casos notificados de COVID-19 en España

Tabla 1. Distribución por CCAA. Casos de COVID-19 notificados a la RENAVE

CCAA	Casos1	Casos notificados a la RENAVE2	%
Andalucía	10187	7035	69
Aragón	4187	3152	75
Asturias	1958	841	43
Baleares	1550	1447	93
Canarias	1944	1917	99
Cantabria	1777	1770	100
Castilla La Mancha	14054	8036	57
Castilla y León	12628	10058	80
Cataluña	34726	1928	6
Comunitat Valenciana	9060	8905	98
Extremadura	2658	2409	91
Galicia	7494	6550	87
Madrid	47146	45277	96
Murcia	1463	1275	87
Navarra	4092	3409	83
País Vasco	11018	6830	62
La Rioja	3358	2389	71
Ceuta	95	96	100
Melilla	101	83	82
Total	169496	113407	67

[1]Notificación agregada de casos de COVID-19 hasta las 21:00h del 5 de abril de 2020
[2]Extracción de datos de SiViES a las 12:00h del 6 de abril de 2020

Figura 1. Curva epidémica por fecha de inicio de síntomas[1] y fecha de diagnóstico[1]. Casos de COVID-19 notificados a la RENAVE



[1]Los datos de los recuadros sombreados pueden estar afectados por retraso en la notificación o diagnóstico

Características demográficas, clínicas y epidemiológicas

Tabla 2. Características demográficas, clínicas y antecedentes epidemiológicos de riesgo. Casos de COVID-19 notificados a la RENAVE[1] (N=113407)

Características		N	Total N (%)	Mujeres N (%)	Hombres N (%)	p-valor
Sexo		113368		59196 (52,2)	54172 (47,8)	
Edad	Mediana (RIC)[2]	112982	60 (46-75)	58 (44-75)	62 (49-76)	<0,001
Grupo de edad (años)	<2	112982	168 (0,1)	69 (0,1)	99 (0,2)	
	2-4		64 (0,1)	32 (0,1)	32 (0,1)	
	5-14		303 (0,3)	153 (0,3)	150 (0,3)	
	15-29		6155 (5,4)	3975 (6,7)	2174 (4,0)	
	30-39		10764 (9,5)	6518 (11,0)	4245 (7,9)	
	40-49		16915 (15,0)	9369 (15,9)	7544 (14,0)	
	50-59		21057 (18,6)	11369 (19,3)	9682 (17,9)	
	60-69		18801 (16,6)	8671 (14,7)	10127 (18,8)	
	70-79		17912 (15,9)	7507 (12,7)	10402 (19,3)	
	≥80		20843 (18,4)	11348 (19,2)	9494 (17,6)	<0,001
Síntomas[1]	Fiebre o reciente historia de fiebre	10665	7876 (73,8)	3724 (68,7)	4143 (79,1)	<0,001
	Tos	10411	7751 (74,5)	3925 (74,1)	3817 (74,8)	0,422
	Dolor de garganta	8812	2285 (25,9)	1376 (30,1)	904 (21,3)	<0,001
	Disnea	13059	6263 (48,0)	2873 (44,4)	3387 (51,5)	<0,001
	Escalofríos	7785	2666 (34,2)	1339 (33,6)	1324 (34,9)	0,244
	Vómitos	8875	830 (9,4)	504 (11,0)	325 (7,6)	<0,001
	Diarrea	9146	2755 (30,1)	1510 (32,0)	1243 (28,2)	<0,001
	Neumonía (radiológica o clínica)	60495	36017 (59,5)	14901 (51,2)	21112 (67,3)	<0,001
	Síndrome de distrés respiratorio agudo	33237	2296 (6,9)	847 (5,1)	1446 (8,7)	<0,001
	Otros síntomas resp.	35737	5676 (15,9)	2502 (13,8)	3172 (18,1)	<0,001
	Fallo renal agudo	37998	1732 (4,6)	601 (3,1)	1130 (6,0)	<0,001
Enfermedades y factores de riesgo[1]	Una o más	73046	48868 (66,9)	23199 (63,0)	25660 (70,8)	<0,001
	Enfermedad cardiovascular	63821	21123 (33,1)	8925 (27,8)	12193 (38,4)	<0,001
	Enfermedad respiratoria	63821	6409 (10,0)	2446 (7,6)	3961 (12,5)	<0,001
	Diabetes	63821	10509 (16,5)	4214 (13,1)	6292 (19,8)	<0,001
	Hipertensión arterial*	63821	8990 (14,1)	4244 (13,2)	4746 (15,0)	<0,001
Hospitalización		106437	51853 (48,7)	21825 (39,4)	30014 (58,8)	<0,001
Ventilación mecánica		31068	2169 (7,0)	627 (4,2)	1538 (9,5)	<0,001
Admisión UCI[3]		79677	4070 (5,1)	1172 (2,9)	2894 (7,4)	<0,001
Defunción		113407	8644 (7,6)	3319 (5,6)	5325 (9,8)	<0,001
Contacto estrecho con casos COVID-19 probable o confirmado		2675	1490 (55,7)	914 (62,1)	572 (47,8)	<0,001
Contacto con persona con infección respiratoria aguda		2922	2156 (73,8)	1328 (77,8)	827 (68,2)	<0,001
Profesional sanitario		88994	16446 (18,5)	12134 (26,0)	4310 (10,2)	<0,001
Visita a centro sanitario		996	156 (15,7)	95 (18,4)	61 (12,9)	0,017

[1]Porcentaje sobre los casos de COVID-19 de los que se dispone de información. [2]RIC: rango intercuartil. [3]UCI: Unidad de cuidados intensivos. *La información sobre hipertensión arterial se recoge a partir del 18/03/2020. Datos actualizados a 13-04-2020.

Tabla 3. Caracterización temporal desde el inicio de los síntomas. Casos de COVID-19 notificados a la RENAVE[1]

	N	Total Mediana (RIC)[1]	Mujeres Mediana (RIC)[1]	Hombres Mediana (RIC)[1]	p-valor
Inicio síntomas hasta diagnóstico (días)	74398	5 (2- 9)	5 (2- 9)	6 (3- 9)	<0,001
Inicio síntomas hasta notificación CCAA (días)	83551	6 (3-10)	6 (2-10)	6 (3-10)	<0,001
Inicio síntomas hasta hospitalización (días)	42642	6 (3- 9)	6 (3- 9)	6 (3- 9)	0,912
Inicio síntomas hasta ingreso en UCI (días)	3135	8 (5-11)	8 (5-11)	8 (5-10)	0,466
Inicio síntomas hasta defunción (días)	6155	10 (6-14)	9 (5-13)	10 (6-14)	<0,001

[1]RIC: rango intercuartil. Datos actualizados a 13-04-2020.

Figura 2. Distribución por edad y sexo. Casos de COVID-19 notificados a la RENAVE y población española

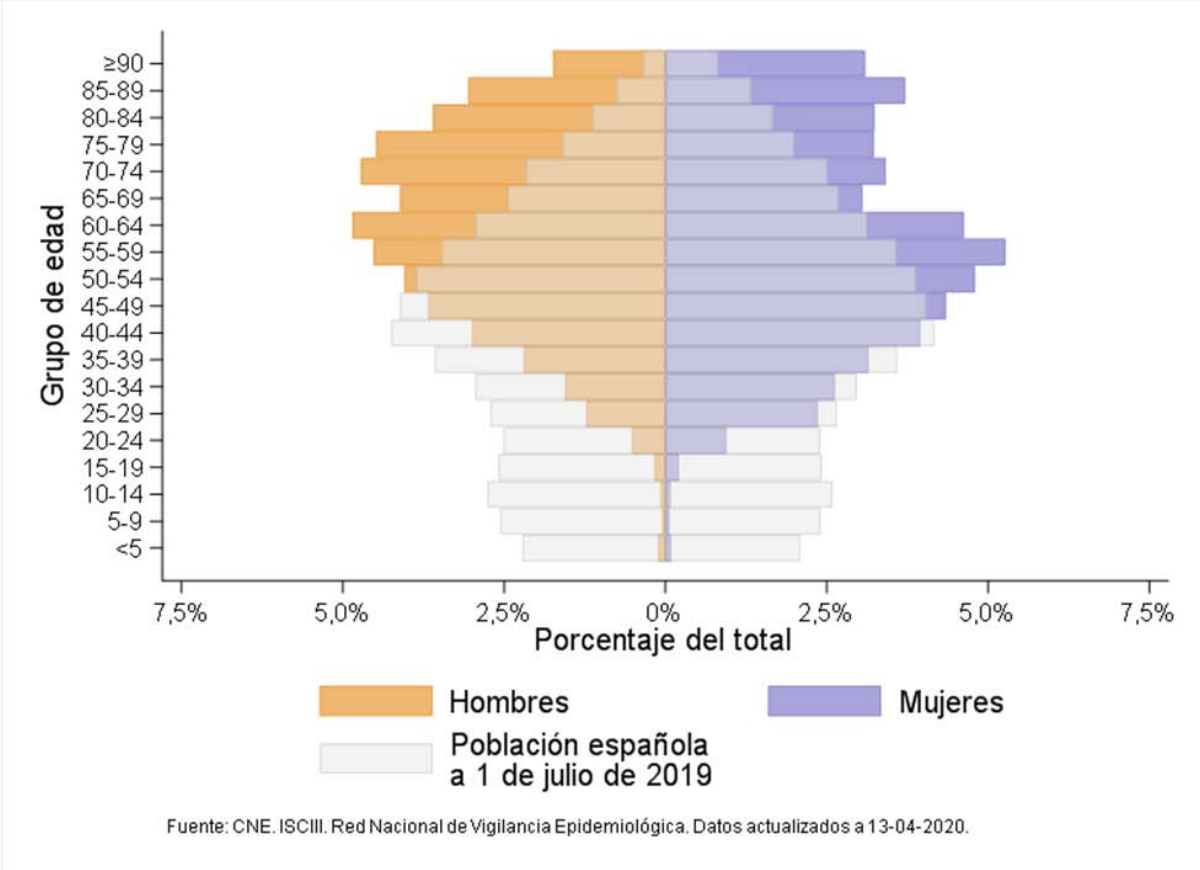


Figura 3. Distribución por grupos de edad y situación clínica. Casos de COVID-19 notificados a la RENAVE

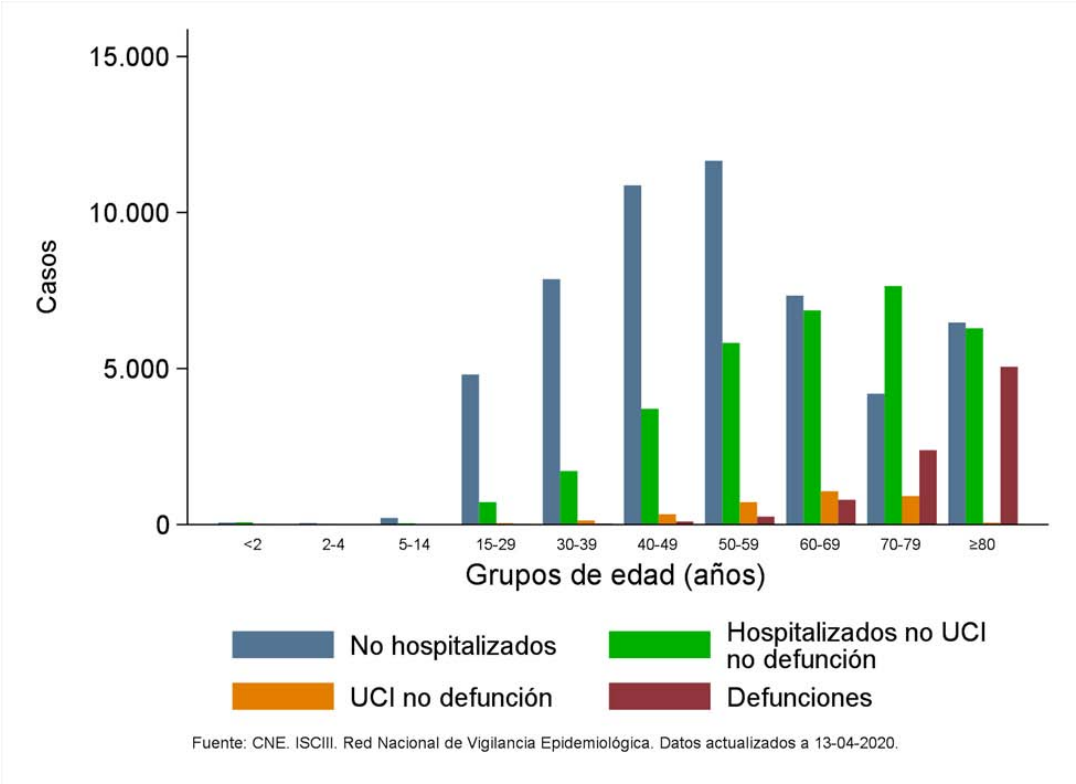
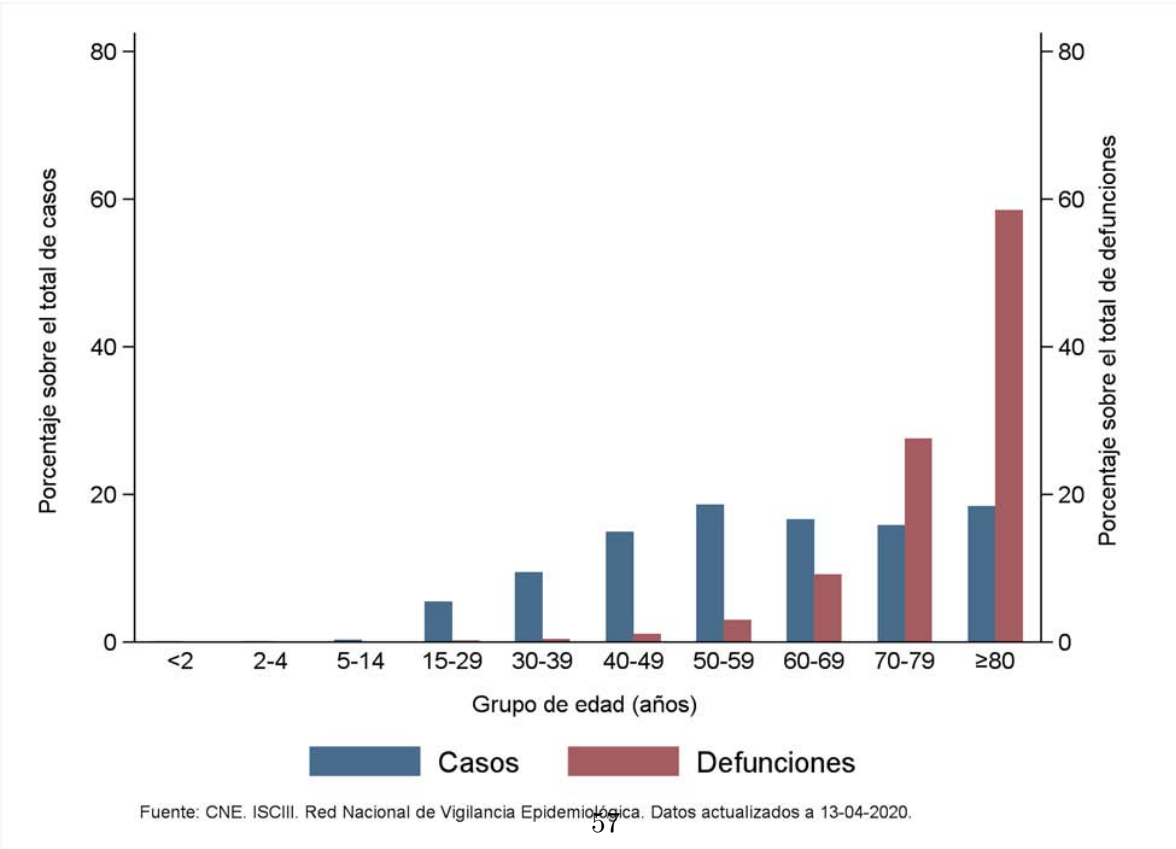


Figura 4. Porcentaje de casos y defunciones por grupo de edad. Casos de COVID-19 notificados a la RENAVE



Características clínicas y gravedad

Tabla 4.1. Número de casos por grupos de edad y situación clínica. Casos de COVID-19 notificados a la RENAVE, total

Grupo de edad (años)	Casos totales N	Hospitalizados N (%)	UCI N (%)	Defunciones N (%)
<2	168	102 (60,7)	14 (8,3)	4 (2,4)
2-4	64	17 (26,6)	2 (3,1)	0 (0,0)
5-14	303	54 (17,8)	4 (1,3)	0 (0,0)
15-29	6155	916 (14,9)	46 (0,7)	14 (0,2)
30-39	10764	2213 (20,6)	142 (1,3)	32 (0,3)
40-49	16915	5000 (29,6)	366 (2,2)	97 (0,6)
50-59	21057	8113 (38,5)	797 (3,8)	257 (1,2)
60-69	18801	10448 (55,6)	1295 (6,9)	795 (4,2)
70-79	17912	12618 (70,4)	1259 (7,0)	2386 (13,3)
≥80	20843	12345 (59,2)	140 (0,7)	5058 (24,3)
Total	**113407**	**51853 (45,7)**	**4070 (3,6)**	**8644 (7,6)**

Datos actualizados a 13-04-2020.

Tabla 4.2. Número de casos por grupos de edad y situación clínica. Casos de COVID-19 notificados a la RENAVE, mujeres

Grupo de edad (años)	Casos totales N	Hospitalizados N (%)	UCI N (%)	Defunciones N (%)
<2	69	41 (59,4)	5 (7,2)	2 (2,9)
2-4	32	9 (28,1)	1 (3,1)	0 (0,0)
5-14	153	30 (19,6)	1 (0,7)	0 (0,0)
15-29	3975	472 (11,9)	19 (0,5)	6 (0,2)
30-39	6518	1032 (15,8)	47 (0,7)	14 (0,2)
40-49	9369	1980 (21,1)	96 (1,0)	35 (0,4)
50-59	11369	3223 (28,3)	224 (2,0)	79 (0,7)
60-69	8671	4135 (47,7)	358 (4,1)	241 (2,8)
70-79	7507	4969 (66,2)	362 (4,8)	701 (9,3)
≥80	11348	5922 (52,2)	59 (0,5)	2241 (19,7)
Total	**59196**	**21825 (36,9)**	**1172 (2,0)**	**3319 (5,6)**

Datos actualizados a 13-04-2020.

Tabla 4.3. Número de casos por grupos de edad y situación clínica. Casos de COVID-19 notificados a la RENAVE, hombres

Grupo de edad (años)	Casos totales N	Hospitalizados N (%)	UCI N (%)	Defunciones N (%)
<2	99	61 (61,6)	9 (9,1)	2 (2,0)
2-4	32	8 (25,0)	1 (3,1)	0 (0,0)
5-14	150	24 (16,0)	3 (2,0)	0 (0,0)
15-29	2174	439 (20,2)	27 (1,2)	8 (0,4)
30-39	4245	1180 (27,8)	95 (2,2)	18 (0,4)
40-49	7544	3020 (40,0)	270 (3,6)	62 (0,8)
50-59	9682	4888 (50,5)	571 (5,9)	178 (1,8)
60-69	10127	6311 (62,3)	937 (9,3)	554 (5,5)
70-79	10402	7646 (73,5)	895 (8,6)	1685 (16,2)
≥80	9494	6422 (67,6)	81 (0,9)	2817 (29,7)
Total	**54172**	**30014 (55,4)**	**2894 (5,3)**	**5325 (9,8)**

Datos actualizados a 13-04-2020.

Tabla 5. Características de los casos según presencia de neumonía[1]. Casos de COVID-19 notificados a la RENAVE

Características[1]		Con neumonía N (%)	Sin neumonía N (%)	p-valor
Total		36017 (60)	24478 (40)	
Sexo	Mujeres	14901 (41)	14211 (58)	
	Hombres	21112 (59)	10256 (42)	<0,001
Edad, mediana (RIC)[2]		67 (54-78)	54 (41-68)	<0,001
Grupo de edad (años)	<2	17 (0)	83 (0)	
	2-4	8 (0)	31 (0)	
	5-14	26 (0)	107 (0)	
	15-29	612 (2)	2055 (8)	
	30-39	1727 (5)	3336 (14)	
	40-49	3956 (11)	4527 (19)	
	50-59	6107 (17)	5051 (21)	
	60-69	7359 (20)	3493 (14)	
	70-79	8589 (24)	2637 (11)	
	≥80	7610 (21)	3083 (13)	<0,001
Enfermedades y factores de riesgo	Una o más	24791 (76)	12884 (58)	<0,001
Enfermedad cardiovascular	Sí	12628 (42)	4282 (23)	<0,001
Enfermedad respiratoria	Sí	3596 (12)	1406 (8)	<0,001
Diabetes	Sí	6337 (21)	1934 (11)	<0,001
Otra	Sí	14885 (41)	6083 (25)	<0,001
Hospitalización	Sí	33075 (92)	5771 (27)	<0,001
Ventilación mecánica	Sí	1856 (11)	159 (2)	<0,001
Admisión UCI	Sí	2977 (10)	300 (2)	<0,001
Defunción	Sí	5501 (15)	875 (4)	<0,001

[1]Análisis sobre los casos de COVID-19 de los que se dispone de información sobre la presencia o ausencia de neumonía. [2]Rango Intercuartílico. Datos actualizados a 13-04-2020.

Tabla 6. Características según hospitalización en Unidad de Cuidados Intensivos (UCI)[1]. Casos de COVID-19 notificados a la RENAVE

Características[1]		Hospitalizados UCI N (%)	Hospitalizados no UCI N (%)	p-valor
Total		4059 (10)	38457 (90)	
Sexo	Mujeres	1168 (29)	16571 (43)	
	Hombres	2887 (71)	21879 (57)	<0,001
Edad, mediana (RIC)[2]		65 (56-72)	69 (55-80)	<0,001
Grupo de edad (años)	<2	14 (0)	76 (0)	
	2-4	2 (0)	11 (0)	
	5-14	4 (0)	40 (0)	
	15-29	46 (1)	725 (2)	
	30-39	142 (4)	1729 (4)	
	40-49	364 (9)	3759 (10)	
	50-59	795 (20)	5947 (15)	
	60-69	1292 (32)	7309 (19)	
	70-79	1256 (31)	9200 (24)	
	≥80	139 (3)	9647 (25)	<0,001
Enfermedades y factores de riesgo	Una o más	2711 (81)	24131 (78)	<0,001
Enfermedad cardiovascular	Sí	1385 (47)	13486 (46)	0,129
Enfermedad respiratoria	Sí	373 (13)	4013 (14)	0,175
Diabetes	Sí	743 (25)	6608 (22)	<0,001
Otra	Sí	1346 (33)	15146 (39)	<0,001
Neumonía (radiológica o clínica)	Sí	2974 (91)	25472 (85)	<0,001
Síndrome de distrés respiratorio agudo	Sí	853 (47)	1072 (8)	<0,001
Defunción	Sí	770 (19)	5556 (14)	<0,001

[1]Porcentaje sobre los casos de COVID-19 de los que se dispone de información sobre las variables señaladas. [2]Rango Intercuartílico. Datos actualizados a 13-04-2020.

Tabla 7. Características según defunción[1] .Casos de COVID-19 notificados a la RENAVE

Características[1]		Defunción N (%)	No defunción N (%)	p-valor
Total		8644 (8)	104763 (92)	
Sexo	Mujeres	3319 (38)	55877 (53)	
	Hombres	5325 (62)	48847 (47)	<0,001
Edad, mediana (RIC)[2]		82 (75-87)	58 (45-73)	<0,001
Grupo de edad (años)	<2	4 (0)	164 (0)	
	2-4	0 (0)	64 (0)	
	5-14	0 (0)	303 (0)	
	15-29	14 (0)	6141 (6)	
	30-39	32 (0)	10732 (10)	
	40-49	97 (1)	16818 (16)	
	50-59	257 (3)	20800 (20)	
	60-69	795 (9)	18006 (17)	
	70-79	2386 (28)	15526 (15)	
	≥80	5058 (59)	15785 (15)	<0,001
Profesional sanitario	Sí	30 (0)	16416 (20)	<0,001
Hospitalización	Sí	7521 (89)	44332 (45)	<0,001
Admisión UCI	Sí	772 (11)	3298 (5)	<0,001
Enfermedades y factores de riesgo	Una o más	6971 (95)	41897 (64)	<0,001
Enfermedad cardiovascular	Sí	4455 (67)	16668 (29)	<0,001
Enfermedad respiratoria	Sí	1241 (19)	5168 (9)	<0,001
Diabetes	Sí	2220 (34)	8289 (14)	<0,001
Otra	Sí	4179 (48)	22184 (21)	<0,001
Neumonía (radiológica o clínica)	Sí	5501 (86)	30516 (56)	<0,001
Síndrome de distrés respiratorio agudo	Sí	779 (27)	1517 (5)	<0,001
Ventilación mecánica	Sí	534 (15)	1635 (6)	<0,001

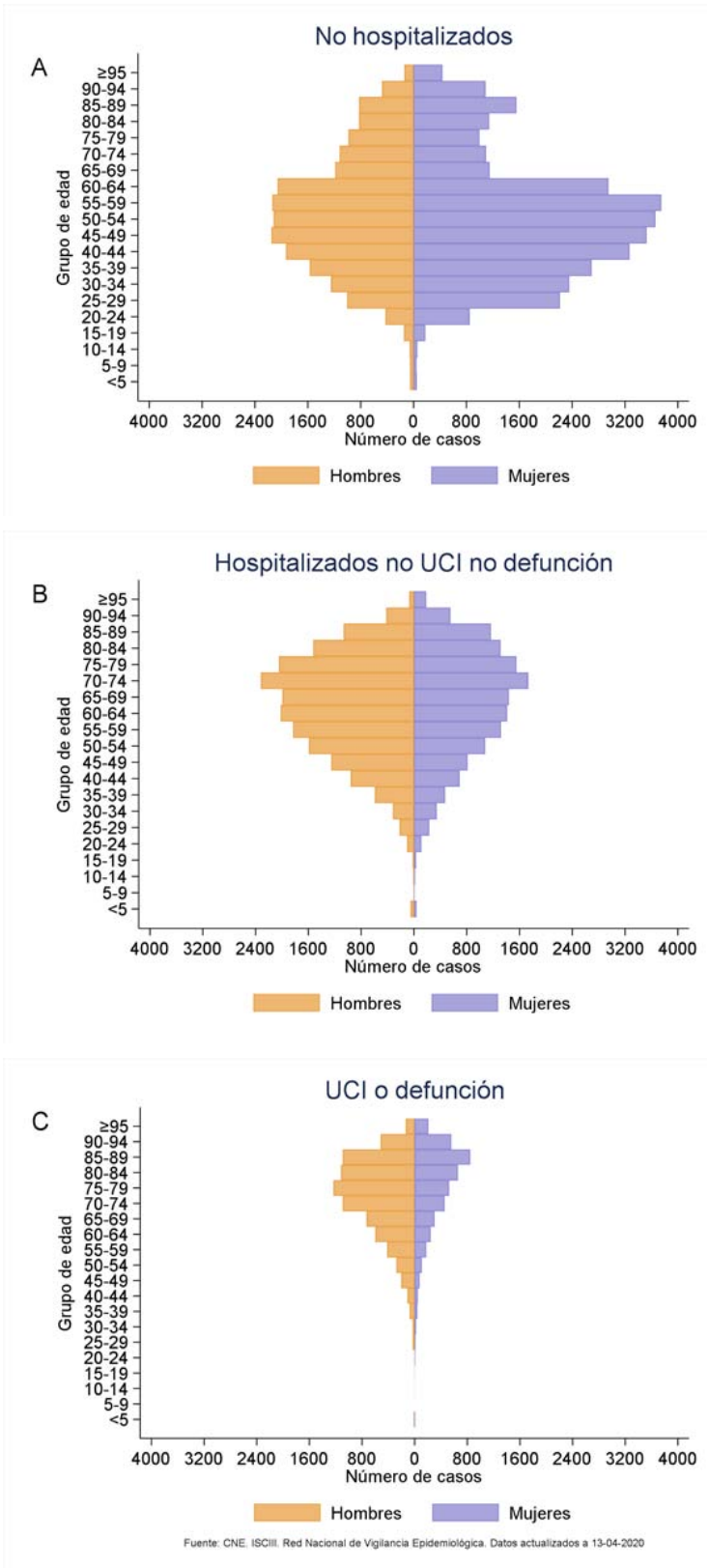
[1]Porcentaje sobre los casos de COVID-19 de los que se dispone de información sobre las variables señaladas. [2]Rango Intercuartílico. Datos actualizados a 13-04-2020.

Tabla 8. Características según nivel de gravedad[1] .Casos de COVID-19 notificados a la RENAVE

Características[1]		No hospitalizados N (%)	Hospitalizados no UCI, no defunción N (%)	UCI no defunción N (%)	Defunción N (%)	p-valor
Total		53630 (54)	32901 (33)	3289 (3)	8644 (9)	
Sexo	Mujeres	33070 (62)	14491 (44)	950 (29)	3319 (38)	
	Hombres	20552 (38)	18403 (56)	2335 (71)	5325 (62)	
Edad, mediana (RIC)[2]		52 (53-77)	66 (40-64)	63 (40-64)	82 (75-87)	<0,001
Grupo de edad (años)	<2	59 (0)	73 (0)	13 (0)	4 (0)	
	2-4	42 (0)	11 (0)	2 (0)	0 (0)	
	5-14	214 (0)	40 (0)	4 (0)	0 (0)	
	15-29	4811 (9)	720 (2)	42 (1)	14 (0)	
	30-39	7864 (15)	1717 (5)	129 (4)	32 (0)	
	40-49	10873 (20)	3709 (11)	333 (10)	97 (1)	
	50-59	11661 (22)	5820 (18)	717 (22)	257 (3)	
	60-69	7339 (14)	6865 (21)	1071 (33)	795 (9)	
	70-79	4192 (8)	7644 (23)	912 (28)	2386 (28)	
	≥80	6477 (12)	6288 (19)	62 (2)	5058 (59)	<0,001
Enfermedades y factores de riesgo	Una o más	12304 (45)	19352 (75)	2071 (78)	6971 (95)	<0,001
Enfermedad cardiovascular	Sí	3884 (16)	10174 (41)	988 (43)	4455 (67)	<0,001
Enfermedad respiratoria	Sí	1373 (6)	3052 (12)	258 (11)	1241 (19)	<0,001
Diabetes	Sí	1777 (7)	5017 (20)	531 (23)	2220 (34)	<0,001
Otra	Sí	6748 (13)	12128 (37)	1002 (30)	4179 (48)	<0,001
Neumonía (radiológica o clínica)	Sí	2747 (15)	21347 (84)	2350 (91)	5501 (86)	<0,001
Ventilación mecánica	Sí	7 (0)	222 (1)	1389 (77)	534 (15)	<0,001

[1]Análisis sobre los casos de COVID-19 de los que se dispone de información sobre la gravedad. [2]RIC: rango intercuartil. Datos actualizados a 13-04-2020.

Figura 5. Distribución por sexo y edad según el nivel de gravedad. Casos de COVID-19 notificados a la RENAVE



Novedades respecto al informe anterior

- Aumenta ligeramente el porcentaje de mujeres respecto a análisis previos, hasta alcanzar el 52,2%.
- Disminuye ligeramente el porcentaje de hospitalizados, del 51,5% al 48,7%
- A su vez, aumentan ligeramente los casos con neumonía: del 57,5 al 59,5%
- Disminuye el porcentaje de pacientes con ventilación mecánica: de un 7,9 a un 7%
- Aumenta la letalidad en los casos de COVID-19 notificados a la REANVE del 5,3 al 7,6%
- La hipertensión arterial también está mas presente de forma significativa en hombres que en mujeres.
- Disminuye de forma importante el % de sanitarios respecto al total de casos notificados: de un 26% a un 18,5% actualmente.
- La prevalencia de enfermedad de base es de 45% en los casos no hospitalizados, 75% en los casos hospitalizados, 78% en los ingresados en UCI y 95% en los fallecidos

Principales resultados

- Desde el inicio de la alerta por SARS-CoV-2 se han notificado 169.496 casos de COVID-19 en España, de los que se ha recibido información en SiVies de 113.407 casos (67%) hasta las 12:00 h del 13 de abril de 2020.
- El 52,2% de los casos de COVID-19 son mujeres y la mediana de edad de los casos es 60 años, siendo mayor en hombres que en mujeres (62 vs 58 años). Los síntomas más frecuentes que se refieren son fiebre, tos, disnea y escalofríos; un 40% presentó clínica digestiva (diarrea o vómitos). Los hombres presentan una mayor prevalencia de fiebre y disnea, mientras que el dolor de garganta y la clínica digestiva son significativamente más frecuentes en mujeres. Un 48,7% de los casos notificados a SiViES han sido hospitalizados, 57% han desarrollado neumonía, un 5,1% han sido admitidos en UCI y un 7,6% han fallecido. Los hombres presentan una mayor prevalencia de neumonía, enfermedades de base (cardiovascular, respiratoria, diabetes, hipertensión) y un mayor porcentaje de hospitalización, admisión en UCI, ventilación mecánica y letalidad que las mujeres. Se estima que un 7% de pacientes necesitan ventilación mecánica, 9,5% en hombres y 4,2% en mujeres.
- La distribución por sexo y grupo de edad indica que los casos de COVID-19, con respecto a la distribución de la población española, están sobrerrepresentados entre los mayores de 50 años, tanto en hombres como en mujeres. En las mujeres llama la atención de forma más acusada entre los 45 y 65 años, y en los hombres a partir de los 60 años. Por el contrario, la presentación de casos de COVID-19 en menores de 25 años de ambos sexos es muy baja.
- Un 18,5% de los casos notificados a SiViES son trabajadores sanitarios (dato calculado sobre los casos que tenían información sobre esta variable), siendo significativamente mayor este porcentaje de trabajadores sanitarios entre las mujeres que entre los hombres (26 vs 10,2%).
- En un análisis específico sobre neumonía se observa que los pacientes con neumonía son significativamente mayores que los que no presentan neumonía (67 vs 54 años, respectivamente). Los hombres, las personas mayores de 60 años y las que presentan enfermedad de base (especialmente enfermedad cardiovascular y diabetes) están más representados entre los pacientes que presentan neumonía. Como es de esperar, el porcentaje de hospitalización, ventilación mecánica, admisión en UCI y defunción es significativamente mayor en los casos con neumonía.

- Los pacientes ingresados en UCI son significativamente más jóvenes que los hospitalizados sin ingreso en UCI (edad mediana 65 vs 69 años), siendo el porcentaje de pacientes mayores de 80 años

con ingreso en UCI del 3% frente al 25% en el grupo de hospitalizados sin ingreso en UCI. Entre los ingresados en UCI, frente a los hospitalizados sin ingreso en UCI, están más representados los hombres y existe una mayor prevalencia de enfermedades de base, neumonía y otras complicaciones respiratorias.

- En un análisis específico sobre defunción se observa que los pacientes fallecidos, frente a los no fallecidos, son significativamente mayores (edad mediana 82 vs 58 años), los hombres están más representados, presentan más frecuentemente enfermedades de base, neumonía y otras complicaciones respiratorias, y han sido hospitalizados e ingresados en UCI con mayor frecuencia.
- En una escala de gravedad de 1) casos no hospitalizados, 2) casos hospitalizados (no UCI, no defunción), 3) casos admitidos en UCI no fallecidos y 4)fallecidos, se observa que el porcentaje de pacientes de mayores de 70 años aumenta de 20% en pacientes no hospitalizados a 87% en pacientes fallecidos. A medida que aumenta la gravedad se observa también un mayor porcentaje de hombres y de pacientes con enfermedad de base. La prevalencia de enfermedad de base es de 45% en los casos no hospitalizados, 75% en los casos hospitalizados, 78% en los ingresados en UCI y 95% en los fallecidos. En definitiva, los hombres, los pacientes de mayor edad y con enfermedades de base y factores de riesgo están más representados a medida que aumenta la gravedad.
- En la pirámides de distribución de casos por sexo y edad en función de la gravedad se observa un predominio de mujeres en los casos no hospitalizados y de hombres en los que requieren hospitalización. El número de casos hospitalizados y no hospitalizados, por debajo de los 25 años, es bajo en ambos sexos. La mayoría de casos hospitalizados sin ingreso en UCI ni defunción, así como los casos con ingreso en UCI o defunción, se dan entre hombres, superando ampliamente a la proporción de mujeres a partir de los 45 años, y de forma más acusada a partir de los 70 años.

Nota metodológica

Desde el principio de la pandemia de COVID-19, la vigilancia de los casos de esta enfermedad en España se basa en la notificación universal de todos los casos confirmados de COVID-19 que se identifican en cada CCAA. Las CCAA notifican al nivel central esta información de dos formas diferentes:

- Una notificación diaria del número de casos agregados de COVID-19 por CCAA al Centro de Coordinación de Alertas y Emergencias (CCAES) del Ministerio de Sanidad. Los datos contienen información sobre casos totales, casos en profesionales sanitarios, hospitalizados, admisiones en UCI, fallecidos y recuperados.
- Una notificación individualizada de casos de COVID-19 a la RENAVE a través de la plataforma informática vía Web SiViES que gestiona el Centro Nacional de Epidemiología (CNE). Esta información procede de la encuesta epidemiológica de caso que cada CA cumplimenta ante la identificación de un caso de COVID-19 y contiene datos demográficos, epidemiológicos y clínicos de los casos de COVID-19 identificados en España. Para conseguir una información completa de cada caso, la CA debe realizar sucesivas actualizaciones de la información de la encuesta porque no siempre toda la información está disponible desde la identificación del caso, o precisa de una actualización según cambia la evolución clínica del paciente. Las actualizaciones de la información de los casos las realizan las CCAA, según disponen de ellas, en la plataforma SiVies mediante un determinado soporte electrónico. Para ello, las CCAA reciben el apoyo permanente del equipo SiVies del CNE. Sin embargo, la situación de intensísima carga de trabajo en todas la Unidades de Salud Pública de las CCAA implica un esfuerzo muy importante para lograr la notificación individualizada a la RENAVE, especialmente, su actualización para completar la información de la encuesta epidemiológica de caso.

Mientras que la notificación agregada se acerca más a la realidad de la evolución de la pandemia de COVID-19 en España, la información de la RENAVE es todavía incompleta, si bien puede ofrecer una información más precisa sobre las características epidemiológicas y clínicas o los factores de riesgo y enfermedades de base que podrían estar asociados a los casos de COVID-19 identificados en España.

Los datos agregados de COVID-19 se pueden consultar en: https://covid19.isciii.es/

La información individualizada debe ser consolidada con sucesivas actualizaciones para evitar interpretaciones erróneas durante su análisis.

Es importante resaltar que todos los resultados son provisionales y deben interpretarse con precaución, porque se ofrece información de cada variable con la información disponible para cada una de ellas. En el caso de "defunción", hemos considerado como "no defunción" los casos sin información en esta variable, y deben ser confirmados en posteriores análisis.

Vigilancia de los excesos de mortalidad por todas las causas. MoMo

MoMo es un sistema de vigilancia de los excesos de mortalidad por todas las causas que proporciona información sobre el impacto en la mortalidad de la población de todo evento que pueda suponer una amenaza para la Salud Pública. Sus resultados pueden apoyar las evaluaciones de riesgo de dichos eventos y contribuir a guiar adecuadamente la respuesta de salud pública y el desarrollo de políticas de control.

Mediante el sistema MoMo, el Centro Nacional de Epidemiología estima los excesos de mortalidad por todas las causas por sexo y grupos de edad y a nivel nacional y por CCAA. Los resultados son estimaciones de exceso de defunciones por todas las causas, es decir, son excesos de defunciones de carácter inespecífico que no se pueden atribuir directamente a una causa determinada.

Por otra parte, en el momento actual observamos un retraso en la notificación de defunciones en los registros civiles de varias CCAA, por lo que los resultados deben interpretarse con precaución, especialmente en los últimos días y siempre deben ser confirmados con el análisis de los próximos días.

Las estimaciones de MoMo pueden consultarse aquí: Informes MoMo 2020

Bibliografía

[1] Radhakrishnan Nagarajan, Marco Scurari, Sophie Lèbre (2013), *Bayesian Networks in R with Application in Systems Biology*, Springer

[2] SpirtesP, GlymourC, Scheines R (2001), *Causation, prediction, and search, 2nd edn*, MIT Press, Cambridge

[3] Margaritis D (2003), *Learning Bayesian network model structure from data*, PhD thesis, School of Computer Science, CArnegie-Mellon University, Pittsburgh, PA, available as Technical Report CMU-CS-03-153

[4] Tsamardinos I, Aliferis CF, Statnikov A (2003), *Algorithms for large scale MArkov Blanket discovery*, In: PToceedings of the 16th international Florida artificial intelligence research society conference, AAAI Press, 376-381

[5] Yaramakala S, Margaritis D (2005), *Speculative Markov blanket discovery for optimal feature selection*, In: ICDM '05: Proceedings of the 5th IEEE international conference on data mining, IEEE Computer Society, 809-812

[6] Friedman J, Pe'er D, Nachman I (1999), *Learning Bayesian network structure from massive datset: the "Sparce Candidate" algorithm*, In: Proceedings of 15th conference on uncertainty in artificial intelligence (UAI), Morgan Kaufmann, 206-215

[7] Tsamardinos I, Brown LEm Aliferis CF (2006), *The max-min hill-cliimbing Bayesian network structure learning algorithm*, Machine Learn 65(1):31-78