Universidad Politécnica
de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Máster Universitario en Inteligencia Artificial

Trabajo Fin de Máster

Predicción de Capacidad de Difusión de Monóxido de Carbono a Largo Plazo en Pacientes de COVID-19 con Redes Bayesianas

Autor: Jorge Angulo Rodríguez
Tutores: Concha Bielza Lozoya y Pedro Larrañaga Múgica

Madrid, Abril 2023

Este Trabajo Fin de Máster se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Máster
Máster Universitario en Inteligencia Artificial Inteligencia Artificial

Predicción de Capacidad de Difusión de Monóxido de Carbono a Largo Plazo en Pacientes de COVID-19 con Redes Bayesianas

Abril 2023

Autor: Jorge Angulo Rodríguez
Tutores: Concha Bielza Lozoya y Pedro Larrañaga Múgica
Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Resumen

Han pasado casi tres años desde el comienzo de la pandemia causada por el COVID-19 y aún sigue sin haber concluido. Entre los pacientes de la enfermedad se ha observado que muchos presentan secuelas que persisten más allá del periodo habitual de recuperación (COVID persistente). Ejemplos comunes de ello son la pérdida de olfato y fatiga. Estas secuelas continúan afectando negativamente a un gran número de pacientes y suponen un reto para la comunidad médica. En este trabajo nos centramos en estudiar la pérdida de capacidad de difusión respiratoria. En concreto, usaremos como métrica *diffusing capacity or transfer factor of the lung for carbon monoxide* (DLCO), que mide la capacidad de difusión de los pulmones para intercambiar monóxido de carbono (CO). Esta medida sirve de representante para conocer la difusión de oxígeno.

Este trabajo, que forma parte del proyecto "*REACT*", estudia el pronóstico a lo largo de un año del valor de DLCO en pacientes de COVID-19 en varios intervalos de tiempo. Un nivel reducido de DLCO es uno de los síntomas más comunes del COVID persistente. Debido a que hablamos de periodos largos de tiempo, es deseable poder predecir, con la mayor exactitud posible, la probabilidad de que el paciente presente un nivel reducido de DLCO meses después de su recuperación.

Para ello, disponemos de datos de pacientes, proporcionados por Oriol Sibila y Rosa Faner, médicos e investigadores del hospital Clínic de Barcelona. Dichos datos nos dan información clínica de los pacientes en varios periodos de tiempo. Nos planteamos con esta información varios problemas de clasificación supervisada en distintos momentos de tiempo, que resolveremos usando redes bayesianas. Los resultados los compararemos con los obtenidos usando regresión logística (modelo muy usado en la literatura médica). Veremos que los modelos de redes bayesianas obtienen mejores resultados de precisión y área bajo la curva ROC que los modelos de regresión logística . Además, nos permitirán presentarlos gráficamente usando GeNIe [1], una herramienta interactiva de BayesFusion, que posibilita realizar consultas diversas y presentar el proceso de inferencia de manera más interpretable.

En este trabajo hemos visto que, con los datos que tenemos, se pueden obtener modelos de clasificación supervisados usando redes bayesianas que nos permiten predecir el nivel del DLCO (normal/anormal) con precisión y AUC en torno a 0.71 y 0.7 respectivamente. En general, vemos que dichos modelos mejoran ligeramente o igualan a los modelos de regresión logística (modelos usados de forma habitual en el contexto médico). Además, en este trabajo vemos que ciertas variables, como DLH, comorbilidades y calcio (entre otras) son particularmente relevantes para la predicción del DLCO.

Abstract

It has been almost three years since the beginning of the COVID-19 pandemic, and it is still ongoing. Among the patients who endured this disease, it has been observed that many of them present sequelae that persist beyond the usual recovery period (Long COVID or Persistent COVID). Common examples of this are the loss of smell and fatigue. These sequelae continue to negatively affect a large number of patients and pose a challenge to the medical community. In this work, we focus on studying the loss of respiratory diffusion capacity. Specifically, we will use DLCO (diffusing capacity of the lungs for carbon monoxide) as a metric, which measures the capacity of the lungs to exchange carbon monoxide (CO). This measure serves as a proxy to know the diffusion of oxygen.

This work, which is part of the "*REACT*" project, studies the prognosis over a year of the DLCO value in COVID-19 patients at various intervals of time. A reduced level of DLCO is one of the most common symptoms of persistent COVID. Because we are talking about long periods of time, it is desirable to be able to predict, with the greatest possible accuracy, the probability that the patient will have a reduced DLCO level months after their recovery.

For this, we have patient data, provided by Oriol Sibila and Rosa Faner, doctors and researchers from Barcelona's Clínic Hospital. This data contains clinical information from the patients at various points in time. With this information, we pose several supervised classification problems, which we will solve using bayesian networks. We will compare the results obtained using logistic regression (a model widely used in medical literature). We will see that, in general, the bayesian network models obtain better accuracy and area under the ROC curve results than the logistic regression models. In addition, they will graphically present the models using GeNIe, an interactive tool form BayesFusion, which allows us to make various queries and present the inference process in a more interpretable way.

In this work, we have seen that, with the available data, we can train supervised classifier models using Bayesian networks that allow us to predict DLCO (normal/abnormal) with an accuracy and AUC of 0.7 and 0.7 respectively. Generally, these models achieve slightly better or equal results when compared to logistic regression models (models regularly used in a medical context). Furthermore, we have seen that some variables, such as DLH, comorbidities, and calcium, are especially relevant in DLCO prediction.

Agradecimientos

Quiero expresar mi agradecimiento a mis tutores Concha Bielza y Pedro Larrañaga, por la oportunidad de realizar este trabajo de fin de máster así como por su ayuda y guía durante su desarrollo. También quiero agradecer a Oriol Sibila y Rosa Faner por su colaboración en este proyecto y por proporcionarme los datos necesarios para su realización. Asimismo, quiero dar las gracias a REACT por financiar este proyecto. Por último, agradecer a mi familia y a Inés todo su apoyo durante estos meses.

Tabla de contenidos

Capítulo 1

Introducción

En este capítulo vamos a describir los objetivos y el contexto en el cual se ha desarrollado el trabajo.

1.1. Descripción del proyecto

1.1.1. Proyecto *REACT*

La crisis sanitaria del COVID-19 llevó a la comunidad investigadora y sanitaria global a centrar sus esfuerzos en investigar la enfermedad. Como parte de la respuesta a dicha pandemia se creó el proyecto *REACT*, un proyecto cuyo objetivo es desarrollar y desplegar infraestructura tecnológica en la Comunidad de Madrid que permita la creación de sistemas inteligentes de ayuda a la decisión para el diagnóstico y apoyo médico y sanitario (durante y después de la pandemia), la gestión hospitalaria, el apoyo asistencial y el apoyo epidemiológico. Este espacio de datos será usado en el contexto de la pandemia de COVID-19, y estará preparado para futuras pandemias. En particular, los sistemas inteligentes consideran aspectos como transparencia, interpretabilidad y explicabilidad de los modelos, como garantía de uso en el dominio sanitario.

1.1.2. Colaboración con Oriol Sibila y Rosa Faner (Hospital Clínico de Barcelona)

Obtener datos de COVID persistente para un trabajo como el presente supone un gran reto. Los datos médicos son datos sensibles y, en general, es necesario colaborar con una institución médica para acceder a los mismos. Para el caso particular del COVID persistente muchos registros médicos no incluyen este diagnóstico, debido a que se trata de un diagnostico nuevo con diversos criterios que se han fijado recientemente.

Además, no se realiza un seguimiento sistemático de pacientes [2], [3] de COVID para verificar si éstos desarrollan COVID persistente en las semanas y meses posteriores al alta. En consecuencia, decidimos buscar investigadores que hubiesen publicado en el área de COVID persistente, ya que estos artículos describen estudios médicos diseñados para estudiar estas secuelas. Esto nos llevó a trabajar con los doctores

Oriol Sibila y Rosa Faner, que han publicado varios artículos [4] sobre secuelas respiratorias en pacientes con COVID-19.

Oriol Sibila es Jefe del Servicio de Neumología del Hospital Clínic de Barcelona y Rosa Faner es investigadora en el Centro de Investigación Biomédica en Red de Enfermedades Respiratorias de Barcelona (CIBERES). Su investigación en el COVID persistente se centra en las secuelas respiratorias de esta enfermedad. Gracias a su colaboración utilizaremos en este proyecto los datos de su estudio, expandiendo de esta forma el trabajo que ellos previamente han realizado.

1.1.3. Cambios en el contexto del proyecto

El objetivo inicial de nuestro trabajo fue estudiar el COVID persistente. Por lo tanto, el trabajo inicial en este proyecto consistió en buscar datos relacionados con esta patología; no obstante, resulta complicado encontrar datos sobre COVID persistente, ya que se trata de una enfermedad nueva cuyos criterios fueron establecidos recientemente por la OMS [5] y, por lo tanto, hay pocos diagnósticos explícitos. En general, los estudios disponibles se basan en el seguimiento de pacientes de COVID con secuelas específicas [6]. Es decir, crear un estudio, hacer un seguimiento de los pacientes y tomar medidas de variables de forma proactiva. Eso hace que los conjuntos de datos disponibles en estos estudios sean más pequeños

Tras intentos fallidos, nos pusimos en contacto con Oriol Sibila y Rosa Faner, porque su trabajo lidiaba con una de las secuelas del COVID, y nos propusieron continuar trabajando con los datos de los que ellos disponían; sobre niveles reducidos de DLCO como secuela pulmonar de COVID. Aunque es importante aclarar que COVID persistente y secuelas del COVID no son lo mismo, sí están estrechamente relacionados [5].

Otro cambio en el contexto cuya mención merece ser comentada en esta sección es referente al lenguaje de programación. Inicialmente `Python` fue seleccionado como lenguaje de programación, ya que se ajustaba a los requisitos iniciales y es un lenguaje que ofrece un amplio ecosistema de librerías que facilitaría trabajar en el contexto de aprendizaje automático. No obstante, tras haber realizado parte del desarrollo usando `Python` y al decidir usar la herramienta GeNIe [1] para mostrar estos modelos fue necesario cambiar, y usar `R`. Esto fue debido a que solo en este lenguaje encontramos librerías para poder exportar los modelos en un formato compatible con GeNIe.

1.1.4. Objetivos

Los objetivos se han establecido en colaboración con Oriol Sibila y Rosa Faner, que han aportado su perspectiva desde su punto de vista como expertos para especificar las preguntas de interés médico. Se plantearon varias lineas de investigación:

1. Como se ha mencionado ya, el objetivo inicial del trabajo era estudiar desde un punto de vista de diagnóstico y con uso de redes bayesianas el COVID persistente. Está línea de investigación no fue posible debido a la falta de datos.
2. Continuar el trabajo de Oriol Sibila, Rosa Faner y colaboradores [4]. El objetivo es continuar el estudio que ellos realizaron sobre DLCO y extenderlo con los datos proporcionados a los 12 meses tras el alta. Asimismo, se plantea realizar

un seguimiento de los pacientes a lo largo del tiempo, describiendo los niveles de DLCO (normal/anormal) y cuáles son las variables para determinar esos valores. Dado que en este trabajo usaremos redes bayesianas, parte de esta línea de investigación es comparar, tanto en precisión como en interpretabilidad, el uso de este modelo con los modelos de regresión logística usados en el contexto médico

3. Se propusieron, tras hablar con los médicos, varios problemas de clasificación supervisada que representan distintas cuestiones de diagnóstico en varios momentos de tiempo. De esta forma, con la información disponible en un cierto momento de tiempo se pretende predecir la evolución de los pacientes. La resolución de dichos problemas producirá un modelo de clasificación que puede servir como herramienta diagnóstica.

Así, establecemos los siguientes objetivos:

1. Resolver los seis problemas de clasificación supervisada que explicaremos más adelante en la sección 2.2.3 usando los datos proporcionados y redes bayesianas.
2. Comparar los resultados obtenidos con un modelo de regresión logística, así como calcular las métricas para evaluar los modelos.
3. Analizar el impacto de las diferentes variables sobre si el DLCO es normal o anormal en un momento concreto del tiempo, e intentar sacar conclusiones generales sobre cuales tienen mayor impacto.
4. Si es posible, crear modelos basados en redes bayesianas que permitan resolver los problemas de clasificación supervisada planteados, con el objetivo de que dichos modelos puedan ayudar en un contexto médico, como ayuda diagnóstica (valorando la interpretabilidad) y para determinar las variables más relevantes.

1.1.5. Estructura del documento

Este documento está organizado en varios capítulos, además de un resumen, una sección de bibliografía.

1. Resumen: da una visión global del trabajo realizado, tanto en español como en inglés.
2. Introducción (este capítulo) incluye:
   a) Descripción del proyecto: explicamos el contexto, los objetivos y decisiones que rodean al trabajo a realizar.
   b) Estado del arte: tiene como objetivo explorar la literatura en la que se ha basado el presente trabajo, tanto fundamentos teóricos como trabajo realizado en líneas similares.
3. Desarrollo del trabajo: en esta sección se explica el trabajo realizado paso a paso.
   a) Introducción al conjunto datos: explicamos las características del conjunto de datos que nos ha sido proporcionado.
   b) Imputación: explicamos el proceso de imputación que hemos seguido.

   c) Discretización: cómo se ha realizado la discretización.

   d) Selección de variables: los métodos que se han usado para realizar la selección de variables.

   e) Plantear los seis problemas que vamos a resolver.

   f) Explicar el proceso general que vamos a seguir para resolver los problemas, así como la presentación de los resultados.

   g) Justificar los modelos que vamos a usar y establecer detalles sobre las librerías usadas en la programación.

4. Resultados: en esta sección damos resolución a los modelos planteados. Para cada uno de los seis modelos presentaremos:

   a) Resolución del problema usando una red bayesiana. Mostraremos las variables seleccionadas para dicho modelo y evaluaremos el rendimiento de dicho modelo en base a las métricas de precisión y área bajo la curva ROC.

   b) Mostraremos el modelo de red bayesiana de forma gráfica usando la herramienta GeNIe [1]. Además, usaremos esta herramienta para realizar consultas (*queries*) al modelo. Dichas consultas nos ayudarán a entender el modelo y cómo diagnosticaría a diferentes pacientes.

   c) Resolución del problema usando regresión logística. Incluiremos las variables que se han seleccionado para cada modelo, precisión y área bajo la curva ROC. Además, mostraremos los parámetros del modelo de regresión logística.

   d) Compararemos las variables seleccionadas en ambos modelos, además de la comparativa de las métricas (y si podemos afirmar que un modelo es mejor que el otro).

5. Conclusiones: evaluaremos si hemos cumplido los objetivos establecidos y posibles líneas de investigación futuras.

1.2. Estado del arte

En esta sección vamos a realizar una breve exploración del estado del arte. Para ello, hemos realizado una búsqueda de trabajos que utilizan técnicas de aprendizaje automático en COVID persistente y secuelas del COVID, publicados hasta la fecha en la que terminó esta parte del trabajo: primavera de 2022. Además, incluiremos libros y artículos que sirven de base teórica para este trabajo, realizando una breve exploración de las diferentes opciones que existen para resolver los problemas que en él se plantean.

1.2.1. Imputación

En esta sección examinamos los enfoques de diferentes autores en el tratamiento de datos incompletos. Nos centramos en particular en técnicas de imputación (completar los datos), aunque no es la única forma de tratarlos que veremos. De este modo, nos centramos en el aspecto de lidiar con datos incompletos, aunque la bibliografía es unánime en la forma más eficaz de lidiar con este tipo de datos para mejorar el

proceso de captura de datos [7]. Es decir, asegurarse que no hay datos incompletos haciendo un mejor seguimiento de pacientes. Sin embargo, esto es algo que queda fuera del alcance del presente trabajo, que se centra en el análisis de los datos.

En la literatura el problema de los datos incompletos no siempre es tratado, y en los estudios médicos no siempre se hace un tratamiento riguroso de esta problemática. Como introducción al problema, vamos a definir una taxonomía común que se usa para categorizar el tipo de *missing values*. Se trata de una taxonomía común en la literatura, pero podemos ver una explicación detallada en Haukoos and Newgard [8]. Esta clasificación se basa en la distribución de los datos incompletos y posibles dependencias con respecto al *dataset*. Consideremos la siguiente notación, sea I un conjunto de variables aleatorias que modelan los datos incompletos. Siguen una distribución de probabilidad a priori desconocida. Asimismo, dado S el *dataset*, podemos dividir este en dos conjuntos: $S = (S_{comp}, S_{incp})$ de datos completos e incompletos [9]. Los tipos de datos incompletos, en función de su distribución, suelen clasificarse en tres categorías:

1. MCAR (*missing completely at random*): si el hecho de que un dato esté incompleto es independiente del resto de los datos observados (completos). Este caso permite tratar los datos incompletos sin tratar de conocer la distribución de los mismos, siendo un enfoque común. No obstante, es una hipótesis poco realista [10]. Este tipo de relación verifica que:

$$P(I|S) = P(I)$$

2. MAR (*missing at random*): en este caso existe dependencia entre dato incompleto y, al menos, una de las variables predictoras, pero no depende del resto de *missing values*.

$$P(I|S) = P(I|S_{comp}) \tag{1.1}$$

3. MNAR (*missing not at random*): en este caso, la distribución de los datos incompletos puede depender del resto de datos incompletos. Por tanto, la ecuación (1.1) no se verifica. Esta es la situación más general, y no asume nada sobre los datos incompletos.

El problema a la hora de seleccionar un método de imputación es que la hipótesis sobre la distribución de los datos incompletos MCAR es la menos restrictiva, sin embargo, múltiples autores han concluido que no es realista, en especial en el contexto de datos médicos. La hipótesis más restrictiva, MNAR, es una situación que puede darse en datos reales, aunque complica la tarea de manejar los datos incompletos. Sin embargo, asumir MAR es el consenso en la bibliografía [9] [11], [12].

Una de las estrategias más sencillas para lidiar con los datos incompletos es eliminar a dichos pacientes o dichas variables. Esta estrategia es sencilla, pero no recomendable. Por un lado, eliminar información puede afectar la capacidad predictiva del modelo si el *dataset* no era muy extenso inicialmente (algo habitual con datos clínicos). Además, si los datos incompletos no siguen un patrón MCAR, se introducirá sesgo en los datos [13].

Otros métodos comunes son los llamados simples o univariantes. Entre estos se realiza la imputación con la media. La media parece un estimador razonable para los

datos incompletos, pero si los datos no son MCAR introduce sesgo. Además, completar con la media no introduce información nueva [14]. Otros métodos de esta familia incluyen *last observation carried fordward*, que consiste en completar con la última observación para ese mismo paciente u otra variante *baseline observation carried fordward*, en la que se usa el valor de base para los valores incompletos. En el contexto médico, estos son los valores de antes del estudio. Estos dos método son usados generalmente para datos de un mismo tipo tomados en varios momentos espaciados en el tiempo. Son populares por su simplicidad e interpretabilidad, pero no solo depende de que no haya cambios en el tiempo, sino que además pueden producir resultados sesgados, incluso con hipótesis de MCAR [12] [11].

Otro grupo de métodos de imputación univariantes son los métodos de regresión (lineal o logística) e interpolación (lineal, *splines*, etc). El objetivo es aproximar el conjunto de datos por una función que nos permita estimar el valor de los datos incompletos. Estos métodos son susceptibles de inducir sesgo bajo hipótesis MAR. Sin embargo, permiten mantener el total del *dataset*, mantener la distribución y desviación típica [14]. Son particularmente sensibles a valores atípicos [15].

Otra familia de métodos popular son los llamados métodos multivariantes o de imputación múltiple (ver figura 1.1). Este tipo de imputación genera varios *datasets* diferentes, proporcionando para cada valor incompleto varios posibles valores con los cuales completarlo. Después se realiza un análisis de cada *dataset* por separado. Finalmente, se combinan estos resultados mediante un proceso de votación para obtener un único *dataset* completo [16] [17]. Este tipo de imputación tiene buenas propiedades, ya que funciona bajo la hipótesis MAR, e incluso puede manejar datos MNAR en ciertos casos [11]. Sin embargo, su aplicabilidad no es universal, en particular, si el número de datos incompletos es elevado (>40 %) su uso no es recomendable [16].

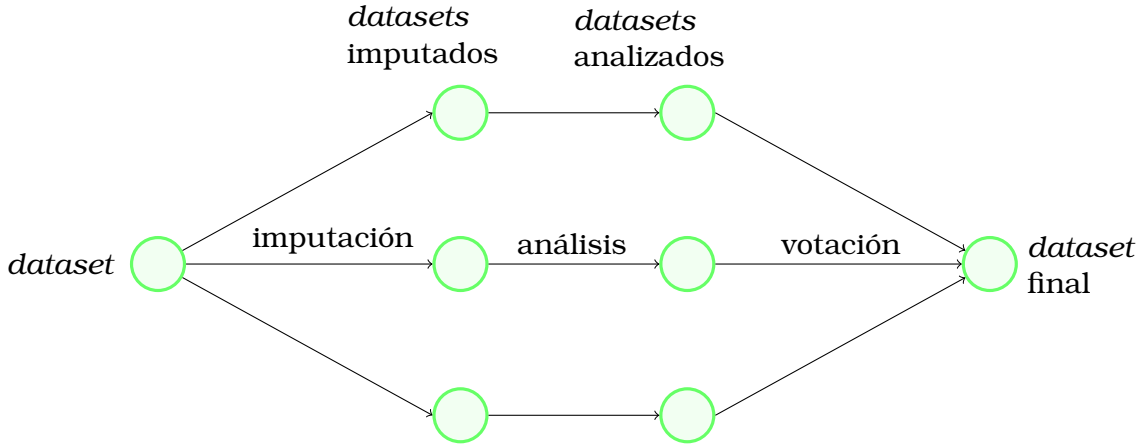

Figura 1.1: Pasos principales de la imputación múltiple, Inspirado por [17]

Otro enfoque para manejar datos incompletos es usar modelos que puedan ser entrenados usando este tipo de datos. En el contexto del COVID este tipo de modelos son frecuentes. En [6], [18] usan XGBoost [19], dicho algoritmo está pensado para se entrenado con datos no completos, y opor tanto puede ser entrenado sin necesitad de imputar. Por tanto, muchos autores que usan este algoritmo no realizan ese proceso.

Si analizamos la bibliografía de trabajos sobre secuelas de COVID o COVID persistente encontramos que la mayoría de los autores no hacen mención a métodos de imputación. Entre los que lo hacen, la mayoría utilizan métodos relativamente simples. En [20] optan por eliminar datos incompletos. En el caso de [21] deciden sustituir por media, mientras que en [22] utilizan regresión lineal (método univariante).

1.2.2. Selección de variables y discretización

Analizamos ahora el proceso de selección de variables, un paso común en modelos de aprendizaje automático. Dicha selección es importante por los siguientes motivos:

- Eliminar variables con baja varianza que no aportan mucha información y pueden dar lugar a problemas en algunos modelos.
- Mejorar la capacidad predictora de los modelos, seleccionando un conjunto "óptimo" de variables para su entrenamiento.
- Hacer los modelos más comprensibles e interpretables, al tener menor número de variables.
- Reducir las dimensiones del conjunto de datos (y agilizar el proceso de entrenamiento).
- Hacer que los modelos sean más útiles desde un punto de vista práctico, pues permiten a los médicos conocer las variables más relevantes para el problema.

Existen múltiples métodos de selección de variables. Por un lado, y relevante en el contexto de la investigación médica y del COVID, encontramos que algunos autores seleccionan las variables mediante la opinión de expertos. Por ejemplo, en [23] verifican que este método produce resultados comparables con otros métodos automáticos.

Entre los métodos de selección de variables basados en los datos, distinguimos cuatro tipos principales [24], [25]:

- Métodos *filter*: son independientes del modelo o algoritmo de aprendizaje, y se centran en evaluar los posibles subconjuntos de variables a seleccionar utilizando diferentes métricas. Debido a que no es necesario entrenar el modelo para usar estos métodos, suelen ser computacionalmente menos costosos y por tanto permiten realizar una búsqueda más exhaustiva. Debido a que no tienen en cuenta el modelo, pueden obviar variables relevantes.
- Métodos *wrapper*: se diferencian de los métodos de *filtering* en que en este caso evaluar un subconjunto de variables implica entrenar un modelo. La puntuación del subconjunto irá ligada a alguna métrica (como precisión) relacionada con un modelo que se ha entrenado con dicho subconjunto de variables. Estos modelos suelen dar mejores resultados, aunque suelen ser computacionalmente más costosos (y por tanto no se puede hacer búsquedas exhaustivas).
- Los modelos embebidos (*embedded*) realizan selección de variables como parte del proceso de entrenamiento del modelo. Un ejemplo prototípico de modelo embebido es el modelo LASSO, un tipo de regresión logística con penalización. Las variables asociadas a coeficientes con valor 0 son eliminadas. Otro ejemplo son los árboles de decisión.

- Modelos híbridos que combinan *filtter* y *wrapper*: la idea es usar un método *filter* para reducir el número inicial de variables con un coste computacional reducido. A continuación, se usa un método *wrapper* sobre el conjunto de variables seleccionadas. De este modo se puede usar el método *wrapper*, más eficaz, pero reduciendo el tiempo de computación (explorar el espacio de subconjuntos puede ser exponencial en el número de variables).

En la bibliografía analizamos los métodos utilizados. Como hemos mencionado anteriormente, mucho autores en el contexto clínico realizan la selección por medio de opinión de expertos. Este es el caso de [26]. Otros autores no hacen mención al proceso de selección de variables. En [6] se realiza también selección manual, relativamente mínima, eliminando variables con poca información.

De los artículos revisados que usan selección de variables basada en datos, [23] utiliza un método *filter* basado en minimizar la correlación de variables de Pearson para seleccionar las variables para el modelo. En [22], se propone un modelo basado en *ensemble* que incorpora selección de variables usando un método *filter* basado en correlación. En [21] usaron un metodo *embedded* basado en regresión logística LASSO.

En general, observamos que la selección de variables no es un problema que la mayoría de los autores que abordan esta cuestión desde un punto de vista médico tratan. Parece que en la mayoría de los casos la selección se hace manualmente o no se considera necesaria.

Consideramos ahora el problema de discretización de variables continuas. Este es un problema que en la bibliografía relacionada no suele ser tratado, ya que la mayor parte de los modelos pueden trabajar con variables continuas. Discretizar supone en general una pérdida de información (que se puede mitigar aumentando el número de particiones). La ventaja es que permite reducir el coste computacional para los algoritmos de predicción y selección de variables [27]. Además, es necesario para ciertos tipos de modelos, como las redes bayesianas discretas.

Podemos hablar de discretización basada en expertos, donde alguien con conocimiento en la materia crea los intervalos para discretizar, o métodos basados en algoritmos. Dentro de los basados en algoritmos, podemos distinguir entre métodos locales y globales. Los métodos locales (univariantes) realizan la discretización de cada variable sin tener en cuenta las demás, mientras que los método globales discretizan todas las variables continuas a la vez.

Algunos ejemplos de discretización local incluyen la partición del dominio en intervalos del mismo tamaño. Otro enfoque es elegir k intervalos de forma que se maximice la entropía [28]. Los métodos locales requieren saber de antemano el número de intervalos en los que se va a partir cada variable [29]. Un ejemplo de método global consiste en realizar *clustering*. Se parte de tantos *clusters* como elementos del *dataset*. Se van juntando los *custers* más cercanos hasta que se cumpla a una condición de parada, entonces se determinan los intervalos en base a los clusters.

1.2.3. Modelos

Vamos a introducir en esta sección la base teórica de los modelos que vamos a utilizar en este trabajo. Los algoritmos de clasificación se pueden separar en supervisados o

no supervisados [30]. Los métodos no supervisados no dependen de la variable clase, mientras que los métodos supervisados si. Todos los algoritmos que presentamos a continuación son supervisados.

1.2.3.1. Modelos de redes bayesianas

Los modelos que vamos a usar son un tipo de red bayesiana [31] con una estructura particular. Las redes bayesianas son una tupla $(D, \boldsymbol{\theta})$, donde D es un grafo acíclico y dirigido (DAG), en el que los nodos son variables aleatorias X_j, $j \in \{1, 2, \ldots, n\}$ y los arcos (X_i, X_j) representan una relación de dependencia probabilística condicional directa entre ambas variables. Que el grafo sea dirigido supone que los arcos tienen dirección, es decir que (X_i, X_j) y (X_j, X_i) son diferentes. Que sea acíclico significa que cualquier camino dirigido que siga los arcos del grafo debe ser finito (no resulta en un bucle o ciclo). Obtenemos:

$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$ representa la distribución de probabilidad condicionada de cada variable, es decir, $\theta_i = P(X_i|\mathbf{Pa}(X_i))$ es la distribución de probabilidad de X_i condicionada a sus padres, es decir, variables X_j para las cuales existe un arco en el DAG que conecta X_j con X_i ($(X_j, X_i) \in D$). En caso que X_i no tenga padres será simplemente la distribución de probabilidad marginal de X_i.

Esta formulación hace que las redes bayesianas factoricen la distribución de probabilidad conjunta usando las independencias condicionadas. Usando la regla de la cadena en la primera y segunda igualdad (en esta varias veces) obtenemos:

$$\begin{aligned} P(X_1, X_2, \ldots, X_n) &= P(X_1|X_2, \ldots, X_n)P(X_2, \ldots, X_n) \\ &= P(X_1)|X_2, \ldots, X_n)P(X_2|X_3, \cdots, X_n) \ldots P(X_{n-1}|X_n) \end{aligned}$$

Es decir, obtenemos que $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})$. Continuamos bajo la hipótesis, común en la práctica, de que podemos encontrar un subconjunto $\mathbf{Pa}(X_i) \subset \{X_i\}_{i=1}^{n}$ tal que X_i es independiente de todas la variables en $\{X_1, X_2 \ldots, X_n\} \setminus \mathbf{Pa}(X_i)$. En consecuencia, obtenemos:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|\mathbf{Pa}(X_i)) \tag{1.2}$$

Esta factorización es de particular utilidad, ya que permite simplificar la distribución de probabilidad conjunta, cuya tabla crece exponencialmente con el número de variables. De esta forma usamos la información sobre independencia condicionada para simplificar su almacenamiento. Además, las redes Bayesinas nos dan una herramienta visual para representar las relaciones que existen entre las variables de un problema. Asimismo, la factorización dada por la ecuación (1.2) simplifica cálculos de inferencia.

El tipo de red bayesiana que vamos a usar se llama **Bayesian network augmented naive Bayes** o BAN [32]. Este es un tipo particular de red, en el cual se hacen ciertas suposiciones a priori. La primera de ellas es que todas las variables tienen a la variable clase C (que debe aparecer) como padre o ancestro. Además, la variable clase no tiene padres. Asimismo, asumimos que cada una de las variables predictoras (aquellas que no son la variable clase) forman una red bayesiana sin restricciones. La estructura global de la red es bastante genérica, dando libertad al modelo en el proceso de aprendizaje (Figura 1.2).

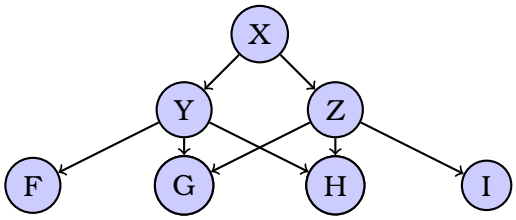

Figura 1.2: Ejemplo de estructura de BAN

1.2.3.2. Regresión logística

La regresión logística es un modelo probabilístico muy conocido y de aplicación general [33] [13]. Lo hemos seleccionado en este trabajo porque su uso es habitual en la medicina [4], y nos va a servir para compararlo con las redes bayesianas.

El modelo logístico permite clasificar una variable clase C a partir de una o más variables predictoras $X_1, \ldots, X_n$. Dado que se trata de un problema de clasificación binario y por tanto la variable clase sigue una distribución de Bernuilli, tenemos que $E(C|\mathbf{X}^{(j)}) = P(C = 1|\mathbf{X}^{(j)})$ (probabilidad de que la observación $X^{(j)}$ se clasifica en la clase $C = 1$). El modelo logístico tiene parámetros $\beta_0, \ldots, \beta_n$, que se estiman a partir de los datos. En base a esto, obtenemos:

$$\pi_j = P(C = 1|\mathbf{X}^{(j)}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1^{(j)} + \ldots + \beta_n X_n^{(j)})}}$$

Esta función logística funciona bien por sus propiedades. Toma valores en el intervalo $[0, 1]$. El proceso de aprendizaje consiste en estimar los parámetros $\{\beta_i\}_{i=0}^n$ a partir del *dataset*. Dicho proceso de aprendizaje suele hacerse mediante estimadores de máxima verosimilitud. No hay una formula cerrada, y se usan métodos numéricos para aproximar los parámetros.

Es común, en el contexto de regresión logística, trabajar con tanto con probabilidad (*probability*) como con *odds*. Este concepto hace referencia al cociente entre la probabilidad de que ocurra un suceso y la de que no ocurra ($p/(1-p)$). Para el modelo que estamos estudiando ahora, al ser de Bernuilli, tenemos que:

$$\text{Odds}(\mathbf{X}^{(j)}) = \frac{P(C = 1|\mathbf{X}^{(j)})}{1 - P(C = 1|\mathbf{X}^{(j)})} = e^{(\beta_0 + \beta_1 X_1^{(j)} + \ldots + \beta_n X_n^{(j)})}$$

En base a esto, podemos definir el concepto de *Odds Ratio*, es decir el cociente de *Odds* de dos variables $\mathbf{X}^{(j)}$ y $\mathbf{X}^{(k)}$ es $\text{OR}(\mathbf{X}^{(j)}, \mathbf{X}^{(k)}) = \text{Odds}(\mathbf{X}^{(j)})/\text{Odds}(\mathbf{X}^{(k)})$. En concreto, si las variables tienen estás características: $\mathbf{X}^{(j)} = (X_1, \ldots, X_{i-1}, 1, X_{i+1}, \ldots, X_n)^t$ y $\mathbf{X}^{(k)} = (X_1, \ldots, X_{i-1}, 0, X_{i+1}, \ldots, X_n)^t$, tenemos que:

$$\text{OR}(\mathbf{X}^{(j)}, \mathbf{X}^{(k)}) = \frac{e^{(\beta_0 + \beta_1 X_1 + \ldots + \beta_{i-1} X_{i-1} + \beta_i \cdot 1 + \beta_{i+1} X_{i+1} + \ldots \beta_n X_n)}}{e^{(\beta_0 + \beta_1 X_1 + \ldots + \beta_{i-1} X_{i-1} + \beta_i \cdot 0 + \beta_{i+1} X_{i+1} + \ldots \beta_n X_n)}} = e^{\beta_i} \tag{1.3}$$

Como hemos mencionado, los parámetros del modelo se estiman usando métodos numéricos. Denotemos por $\{\hat{\beta}_i\}_{i=0}^n$ a las aproximaciones resultantes. Así mismo, dada una n-upla que instancie los valores de las variables predictoras $\mathbf{x}^{(j)} = (x_1^{(j)}, x_2^{(j}), \ldots, x_n^{(j)})^t$ el modelo estimará $P(C|\mathbf{x}^{(j)}) = \hat{\pi}^{(j)} \in [0, 1]$. Además de usar este modelo de regresión

logística para obtener una estimación de la variable clase, es posible dar intervalos de confianza para dicha predicción. El intervalo de confianza (95 %) para $\pi^{(j)}$ es

$$\left(\frac{1}{1+e^{\hat{\beta}_0+x_1^{(j)}\hat{\beta}_1+\ldots+x_n^{(j)}\hat{\beta}_n-1.96\cdot\text{ASE}}}\,,\quad \frac{1}{1+e^{\hat{\beta}_0+x_1^{(j)}\hat{\beta}_1+\ldots+x_n^{(j)}\hat{\beta}_n+1.96\cdot\text{ASE}}}\right),$$

donde

$$\text{ASE} = \sqrt{\text{Var}(\hat{\beta}_0+x_1^{(j)}\hat{\beta}_1+\ldots+x_n^{(j)}\hat{\beta}_n} = \sqrt{(1\ (\mathbf{x}^{(j)})^t)\text{Cov}(\hat{\boldsymbol{\beta}})\begin{pmatrix}1\\ \mathbf{x}^{(j)}\end{pmatrix}}$$

Así mismo, podemos calcular los intervalos de confianza para los coeficientes estimados. El intervalo de confianza (95 %) para $\hat{\beta}_i$ es:

$$\left(\hat{\beta}_i - 1.96 * \text{SE}(\hat{\beta}_i)\,,\quad \hat{\beta}_i + 1.96 * \text{SE}(\hat{\beta}_i)\right)$$

Donde $\text{SE}(\hat{\beta}_i)$ es el error estándar estimado de $\hat{\beta}_i$. Usando la ecuación 1.3, y para $\mathbf{X}^{(j)}$ y $\mathbf{X}^{(k)}$ como en dicha ecuación, el intervalo de confianza del 95 % para el *Odds Ratio* ($OR(\mathbf{X}^{(j)},\mathbf{X}^{(k)})$) es: $\hat{\beta}_i$ es:

$$\left(e^{\hat{\beta}_i-1.96\cdot\text{SE}(\hat{\beta}_i)}\,,\quad e^{\hat{\beta}_i+1.96\cdot\text{SE}(\hat{\beta}_i)}\right)$$

1.2.3.3. Regresión logística multinomial

La regresión logística es un modelo muy flexible, que permite clasificar con variables discretas y variables continuas. No es posible, sin embargo, que la clase que tome 3 o más valores no ordinales usando regresión logística, pues la regresión logística que hemos mostrado solo genera dos regiones de decisión (a ambos lado de la curva definida por la función descrita en el apartado anterior). Para solventar esta limitación podemos usar regresión logística multinomial.

De forma conceptual podemos entender este modelo de la siguiente forma. Sean $k > 2$ valores de la variable clase. Dada la etiqueta $i \in \{1, 2, \ldots, k\}$, podemos crear un modelo que clasifique el *dataset* en elementos que pertenecen a la clase que toma el valor i ($C = i$) y elementos que no tomen el valor i ($C \neq i$). Este proceso lo repetimos para $k-1$ valores. Aplicando estos $k-1$ clasificadores de regresión logística podemos determinar el valor de la variable clase de un elemento del *dataset* a clasificar.

Formalmente lo podemos plantear de la siguiente forma [34]. Consideramos $k > 2$ valores que toma la variable clase C, y una observación $X^{(i)} = (X_1^{(i)}, \ldots, X_n^{(i)})$. Sin pérdida de generalidad, usamos el valor $C = k$ como pivot. Para cada uno de los $k-1$ valores de la variable clase, planteamos un problema de regresión frente al pivot:

$$\begin{aligned}
\frac{P(C=1)}{P(C=k)} &= e^{\beta_0^{(1)}+\beta_1^{(1)}X_1^{(i)}+\ldots+\beta_n^{(1)}X_n^{(i)}}\\
\frac{P(C=2)}{P(C=k)} &= e^{\beta_0^{(2)}+\beta_1^{(2)}X_1^{(i)}+\ldots+\beta_n^{(2)}X_n^{(i)}}\\
&\vdots\\
\frac{P(C=k-1)}{P(C=k)} &= e^{\beta_0^{(k-1)}+\beta_1^{(k-1)}X_1^{(i)}+\ldots+\beta_n^{(k-1)}X_n^{(i)}}
\end{aligned}$$

Despejando $P(C = k)$, y usando que las probabilidades $P(C = j)$, $j = 1, \ldots, k$ deben sumar 1, tenemos que:

$$P(C = k) = 1 - \sum_{j=1}^{k-1} P(C = j) \Rightarrow \frac{\sum_{j=1}^{k-1} P(C = j)}{P(C = k)} = \sum_{j=1}^{k-1} e^{\beta^(j)_0 + \beta_1^{(j)} X_1^{(i)} + \ldots + \beta_n^{(j)} X_n^{(i)}} \Rightarrow$$

$$P(C = K) = \frac{1}{1 + \sum_{j=1}^{K-1} e^{\beta_0^{(j)} + \beta_1^{(j)} X_1^{(i)} + \ldots + \beta_n^{(j)} X_n^{(i)}}}$$

Usando lo anterior, podemos volver a las $k-1$ ecuaciones que planteamos al principio y sustituir $P(C = k)$, con lo que obtenemos:

$$\pi_{C_1} = P(C = 1|X^{(i)}) = \frac{e^{\beta_0^{(1)} + \beta_1^{(1)} X_1^{(i)} + \ldots + \beta_n^{(1)} X_n^{(i)}}}{1 + \sum_{j=1}^{k-1} e^{\beta_0^{(j)} + \beta_1^{(j)} X_1^{(i)} + \ldots + \beta_n^{(j)} X_n^{(i)}}}$$

$$\pi_{C_2} = P(C = 2|X^{(i)}) = \frac{e^{\beta_0^{(2)} + \beta_1^{(2)} X_1^{(i)} + \ldots + \beta_n^{(2)} X_n^{(i)}}}{1 + \sum_{j=1}^{k-1} e^{\beta_0^{(j)} + \beta_1^{(j)} X_1^{(i)} + \ldots + \beta_n^{(j)} X_n^{(i)}}}$$

$$\vdots$$

$$\pi_{C_{k-1}} = P(C = k-1|X^{(i)}) = \frac{e^{\beta_0^{(k-1)} + \beta_1^{(k-1)} X_1^{(i)} + \ldots + \beta_n^{(k-1)} X_n^{(i)}}}{1 + \sum_{j=1}^{k-1} e^{\beta_0^{(j)} + \beta_1^{(j)} X_1^{(i)} + \ldots + \beta_n^{(j)} X_n^{(i)}}}$$

Capítulo 2

Descripción del proceso del trabajo realizado

En este capítulo vamos a explicar el trabajo realizado.

2.1. Introducción al conjunto de datos

El conjunto de datos contiene información de pacientes con PCR positiva, de los cuales se ha realizado un seguimiento en de plazo de un año, con periodos de control al ingreso, 3, 6, 9 y 12 meses. Los datos has sido proporcionados por Oriol Sibila y Rosa Faner, y son datos que han usado en su investigación con pacientes del Hospital Clínic de Barcelona.

La base de datos contiene 605 pacientes y 309 variables, con valores incompletos en muchos casos. Las variables se agrupan temporalmente en los siguientes grupos:

1. Variables atemporales: corresponden a características que no varían a lo largo del tiempo del estudio (obesidad, sexo, fumador, etc).
2. Variables d_0: corresponden a valores tomados en el momento de ingreso del paciente.
3. Variables d_2: corresponden a valores tomados en el segundo día de ingreso.
4. Variables d_{alta}: corresponden a valores tomados en el momento del alta del paciente.
5. Variables "<meses tras el alta_>m", donde meses tras el alta puede ser 3, 6, 9 o 12. Notar que finalmente no se usarán los datos del mes 9 debido a que había menos medidas asociadas con ese periodo.

Podemos examinar los datos de manera preliminar para saber cómo se distribuyen los pacientes con respecto a la variable clase. Consideramos pacientes cuyo porcentaje de DLCO es más de $80\,\%$ como normales (valor 1) y aquellos valores menores como reducido o anormal (valor 0). Esta es la distinción que Oriol Sibila y colaboradores hacen en su artículo [4]. Para los 605 pacientes, en el Cuadro 2.1 tenemos la distribución de los pacientes en el tiempo y categoría en base al porcentaje de DLCO: Los 1 indican que los valores del paciente eran normales, mientras que los 0 indican

3 meses	6 meses	12 meses	Total
0	0	0	236
0	0	1	66
0	1	0	61
0	1	1	37
1	0	0	113
1	0	1	39
1	1	0	22
1	1	1	31

Cuadro 2.1: Distribución de los valores del DLCO por número de pacientes

que eran anormales; el total son los pacientes en dicha categoría. Según los datos obtenidos, dos tercios de los pacientes tuvieron valores anormales tras el alta. Por otra parte, hay un grupo significativo de pacientes que presentaron valores normales y luego pasaron a tener valores anormales, aunque el grupo de pacientes que solo tuvieron valores anormales a los 3 meses y luego no presentan valores anormales es grande.

2.2. Preprocesado de los datos

Como primer paso para poder trabajar con los datos ha sido necesario realizar una limpieza y procesado de los datos. Esto tiene el objetivo de preparar los datos para que puedan ser usados por los modelos que vamos a usar en el trabajo, así como simplificarlos, eliminado información redundante, de baja calidad y complique inecesariamente el proceso de aprendizaje. Este proceso ha seguido los siguientes pasos:

1. Imputación de variables 1: Usando imputación multivariante.
2. Imputación de variables 2: Usando interpolación lineal.
3. Selección de variables eliminando elementos de baja varianza
4. Agrupar en 3 periodos por meses. Eliminar periodos intermedios.
5. Discretización
6. Selección de variables usando *filter*
7. Selección de variables usando *wrapper* (para cada clasificador)

2.2.1. Imputación de variables *missing values*

El *dataset*, tal y como hemos dicho, tiene valores incompletos, es decir, que para ciertos pacientes no se han tomado medidas de algunas variables o se desconoce algún tipo de información. Si bien algunos modelos pueden trabajar con esto, en este trabajo hemos decidido imputar o completar estos valores.

Como hemos analizado en el estado del arte, este es un problema común en la ciencia de datos, y existen múltiples formas de lidiar con este problema. Para el presente conjunto de datos descartamos algunos métodos, como eliminar los datos incompletos. No solo sabemos que este enfoque sesga los resultados excepto bajo la hipótesis

MCAR, sino que además, debido al limitado tamaño del *dataset*, reducir su tamaño afectará a la calidad de los modelos.

Descartamos otros métodos simples, debido a que no nos permiten aprovechar al máximo la información de los datos disponibles, como puede ser completar con la media o *last value carried fordward*. Además, el objetivo es trabajar bajo la hipótesis MAR, lo cual nos lleva a imputación múltiple, y al siguiente esquema de imputación.

1. Primero, separamos en dos grupos los pacientes, aquellos con más del $20\,\%$ de valores incompletos en la tupla de variables asociada a dicho paciente (194 de los 605 pacientes) y aquellos con menos de dicho porcentaje.

2. Para el grupo con menos de un $20\,\%$ de *missing values* usamos imputación multivariable. Hemos descrito con más detalle el funcionamiento de esta técnica en el estado del arte (sección 1.1). El *dataset* imputado será insesgado bajo la hipótesis de que los *missing values* son de tipo MAR. Además, utiliza los valores observados de un paciente para imputar los incompletos.

   En este caso, la imputación se ha realizado usando Python, con `sklearn.impute.IterativeImputer`. Este algoritmo está basado en MICE [17] [35]. Solo lo aplicamos a los pacientes donde el número de *missing values* es bajo, ya que como hemos visto, la imputación múltiple no da buenos resultados en casos con número elevado de datos incompletos.

3. A continuación, juntamos los dos grupos de pacientes. Vamos a usar interpolación lineal (scipy, `interpolate`). En este caso la usamos para pacientes donde el número de *missing values* es elevado. Asumimos que intentar predecir que valor tomaran las variables de estos pacientes usando el anterior método no será muy preciso, así que recurrimos a interpolación lineal. Aceptamos el riesgo de introducir sesgo debido a la necesidad de mantener el tamaño del *dataset* (de por si no muy grande) y aprovechar las propiedades deseables de este método que mencionamos en el estado del arte.

2.2.2. Selección de variables y discretización

Habiendo completado el proceso de imputación continuamos con el pre-procesado de los datos realizando selección de variables.

El primer paso es eliminar aquellas variables con varianza baja. Este proceso nos permite reducir el número de variables de 303 a 163. Las variables de varianza baja pueden dar problemas en los clasificadores y en el caso de las de varianza cero, no aportan información. Además, reducir el número de variables nos ayudará a simplificar el proceso de discretización.

El siguiente paso es agrupar las variables por periodos. Es decir, formar varios *datasets* con las variables de inicio, estáticas y de alta (estas últimas las eliminaremos cuando sea necesario para el modelo) y las variables de un periodo $3m$, $6m$, o $12m$ (como hemos mencionado en la sección anterior, las variables de $9m$ son eliminadas). Lo que queremos es que las variables que estén en el periodo 3 lo estén también en el periodo 6 y 12. De este modo podemos obtener resultados más generales sobre las variables con mayor valor predictor. Esto nos lleva a reducir el número de variables a 77. En los anexos 2 se puede ver un resumen de la composición de dichos *datasets*;

las variables, con los rangos de valore en el caso de las variables discretas y valores en el caso de las continuas.

Muchas de las variables de nuestra base de datos son continuas. Las redes bayesianas pueden trabajar fácilmente con variables continuas si éstas siguen una distribución Gaussiana [36]. Por desgracia, un análisis preliminar de los datos nos muestra que muchas de nuestras variables no siguen una distribución normal. Por tanto, discretizaremos las variables. La discretización es un proceso que conlleva una pérdida de información, pero es necesario para que podamos usar redes bayesians discretas en este problema.

Como hemos mencionado, hay métodos algorítmicos para discretizar variables. Sin embargo, dado que nuestros datos son datos médicos, para los que existen intervalos comunes para discretizar, usaremos este enfoque basado en opinión de expertos. Las variables continuas de nuestra base de datos son variables clínicas. Estas variables tienen un rango de valores dentro de los que se consideran normales (asignaremos valor 0) y pueden estar por encima (1) o por debajo de dicho rango (-1). Notemos que estos valores son una generalización, pueden variar en función de otras variables (sexo, edad, etc.). Incluso pueden variar con las opiniones de los expertos. En este caso, hemos validado los rangos de discretización de las variables con ayuda de Oriol Sibila y Rosa Faner (ver anexos 2).

La discretización de la variable clase que usaremos en la mayoría de los modelos, es decir, el porcentaje de DLCO del paciente en periodo de tiempo concreto, será discretizado de acuerdo con los criterios médicos [4]. Es decir, consideramos que es elevado si supera el 80 % (valor 1 o normal) y bajo (valor 0 o reducido) en el caso contrario.

Llegados aquí, creamos 3 *datasets*. Dichos *datasets* contendrán las variables estáticas, de ingreso y de día dos. Además, cada *dataset* contendrá las variables del periodo correspondiente: 3, 6 o 12 meses. El motivo de no usar las variables de 9 meses es que se realizaron pocas medidas en dicho periodo de tiempo, y por tanto hay pocas variables a los nueve meses. De las variables medidas tras el alta, las que se toman a los 3, 6 y 12 meses. Se ha decidido considerar solo las que los 3 periodos con los que vamos a trabajar. Es decir, dada una variable como ‘ “linfocidos_3m”, consideramos su inclusión solo si las variables “linfocidos_6m” y “linfocidos_12m” también existen.

El siguiente paso consiste en realizar selección de variables mediante un método híbrido, que combina *filter* y *wrapper* (en nuestro caso *minimal-redundancy-maximal-relevance*) [37]. Esto consiste en aplicar primero un método *filter* y después un método *wrapper*. El método *filter* que usaremos consiste en realizar una búsqueda exhaustiva sobre el conjunto potencia de S, $\mathcal{P}(\mathcal{S})$, donde $\mathcal{S}$ es el conjunto de variables. Buscamos un conjunto S^* que maximiza una puntuación (*score*) . En concreto, si C es la variable clase buscamos un conjunto de variables predictoras que verifica que:

$$S^* = \arg\max_S(R(S,C) - r(S,C))$$

donde:

1. $R(S,C) = \frac{1}{|S|}\sum_{X_i \in S} \mathbb{I}(X_i, C)$ representa la **relevancia** ($\mathbb{I}(X_i, C)$ es la información mutua entre las variables X_i y C),
2. $r(S,C) = \frac{1}{|S|^2}\sum_{X_i, X_j \in S} \mathbb{I}(X_i, X_j)$ representa la **redundancia**.

Después de esto, aplicamos un método *wrapper*. En este caso, se aplciará un *wrapper* voraz que funcionará de la siguiente manera:

1. Crear un conjunto $S_0 = \{C\}$ que contenga únicamente la variable clase. Recordemos que S^* era el conjunto seleccionado en la fase *filter*. Denotemos $S_0^* = S^* \setminus S_0$.

2. Añadir a dicho conjunto una variable de las seleccionadas en la fase *filter*, formando $S_1^i = S_0 \cup \{X_i\}, \forall X_i \in S_0^*$. En la siguiente fase entrenamos el correspondiente clasificador y evaluamos su precisión. Nos quedaremos con el conjunto S_1^i que maximiza la precisión del clasificador (lo denotamos S_1).

3. Repetimos este proceso, con $S_j^i = S_{j-1} \cup \{X_i\}, \forall X_i \in S_{j-1}^*$, donde $S_{j-1}^* = S^* \setminus S_{j-1}$. De nuevo, escogeremos el conjunto S_j^i que maximiza la precisión en este paso y denotamos como S_j. En este trabajo hemos hecho una variante. No se aumenta el conjunto a no ser que se mejore la precisión del anterior.

4. La condición de parada es que se alcance un número máximo de iteraciones (paso 3) o que se seleccione un conjunto con 10 variables. Este número es útil en la práctica, y hace más interpretables los resultados.

5. La precisión de los modelos aumenta de forma monótona (en esta variante) según crecen los conjuntos, aunque no existe garantía de que obtengamos el máximo, pudiéndose obtener diferentes resultados en función del orden en el cual se extraigan las variables. En este caso, si no se alcanza un conjunto con 10 variables, se vuelve a comenzar (el orden inicial de variables es aleatorio, y cambia con cada ejecución del algoritmo).

Este proceso de selección de variables de tipo *wrapper* debe ser realizado para cada modelo, consecuentemente cada modelo será entrenado con un conjunto distinto de variables.

2.2.3. Problemas de clasificación

Nos planteamos varios problemas de clasificación usando este conjunto de datos que hemos procesado. Estos problemas los hemos planteado con los médicos y tienen como objetivo darles una herramienta de ayuda al prognóstico en diferentes situaciones.

1. **Predecir DLCO-12 al alta** (Problema 1): Predecir el valor de la DLCO a 12 meses usando las variables de ingreso (y día 2) y estáticas. En este problema se plantea predecir la función pulmonar del paciente usando únicamente los datos disponibles en el momento de ingreso.

2. **Predecir DLCO-12 a los 3 meses** (Problema 2): Predecir con las variables de ingreso, además de las variables disponibles a los 3 meses los valores de DLCO a los 12 meses. Este problema es similar al anterior, pero con más información, ya que se parte de un momento más avanzado en el tiempo.

3. **Predecir evolución al alta** (Problema 3): Predecir, usando únicamente las variables estáticas y de ingreso, la evolución completa del paciente en el tiempo; es decir, se trata de un problema, en el que la variable clase puede tomar 8 valores posibles en función de que combinación de valores (0 ó 1) tome la variable clase en cada uno de los tres periodos (3, 6 y 12).

4. **Predecir DLCO-3 alta** (Problema 4): Predecir, con las variables de ingreso y estáticas el valor de DLCO a los 3 meses. Es un problema similar al problema 1, pero en un plazo de tiempo menor.

5. **Predecir DLCO-6 a los 3 meses** (Problema 5): Predecir, con las variables de ingreso, estáticas y de 3 meses, los valores de DLCO a los 6 meses.

6. **Predecir DLCO-12 a los 6 meses** (Problema 6): Predecir con las variables de ingreso, estáticas, de 3 meses y de 6 meses, el valor del DLCO a los 12 meses. Este es el problema del que más información disponible se tiene.

Como se puede ver, todos estos problemas planteados, si bien similares, plantean problemas de clasificación diferentes que pretende predecir como evolucionará el paciente en un futuro, con la información disponible en un momento concreto del tiempo.

2.2.4. Resolución de los problemas

Explicaremos aquí la metodología general para resolver los problemas planteados. Cada problema lo vamos a resolver usando un modelo de red bayesiana tipo BAN, y un modelo de regresión logística (o regresión logística multinomial).

1. Como hemos mencionado anteriormente, el primer paso consiste en realizar el último paso *wrapper* del proceso de selección de variables. Es decir, usando un método *greedy-wrapper*, seleccionar las variables que se van a usar para entrenar el modelo. Por motivos de interpretabilidad y eficiencia, para este proceso hemos establecido como condiciones de parada alcanzar un conjunto de 10 variables así como un límite de iteraciones.

2. Habiendo seleccionado las variables podemos entrenar el modelo para dicho subconjunto de variables del *dataset*. Para ese conjunto de variables vamos a mostrar cómo se distribuye cada variable para cada valor de la variable clase (menos en el caso del problema 3, donde la variable clase puede tomar 8 valores). Es decir, para cada variable mostramos la media de los valores de dicha variable, para DLCO normal y anormal, así como el intervalo de confianza, p-valor (del test de Student) y rango de valores que toma dicha variable. Además, calculamos para el modelo las métricas de precisión y área bajo la curva ROC.

3. Mostramos los modelos usando GeNIe. Dichos modelos los podemos usar para realizar consultas. Estas consultas consisten en dar al modelo un conjunto de valores para algunas de las variables, y observar los resultados de probabilidad inferidos por este. Esto nos permite entender el modelo mejor, observando el diagnóstico que éste realiza para un paciente para ciertos valores de las variables predictoras o, cuál es el paciente medio dado un cierto valor de la variable clase.

4. Continuamos con la resolución del problema usando regresión logística (o regresión multinomial en el caso del problema 3). Esto comienza, al igual que en el caso del problema anterior, con la selección de variables usando el mismo método *greedy-wrapper*. Presentaremos las variables seleccionadas para este modelo y las compararemos con las seleccionadas para la red bayesiana. Construimos también la tabla de valores medios de cada variable para estas variables, y para cada valor de la variable clase.

5. Finalmente, calculamos las métricas de precisión y AUC para este modelo y la comparamos con las del modelo BAN usando test de hipótesis. De este modo podemos establecer si uno de los dos modelos es mejor que el otro de forma significativa. También mostraremos una tabla con los coeficientes de la regresión logística.
6. Como hemos mencionado, el caso del problema 3 es especial debido a que la variable clase no es binaria. La principal diferencia es que usaremos regresión logística multinomial. Además, algunas de las tablas que mostramos en los otros modelos resultan poco legibles en el caso de la regresión logística, por lo que mostramos otras formas de visualizar los resultados.

2.2.5. Modelos

Como hemos mencionado en los objetivos, el estudio y uso de redes bayesianas aplicadas al COVID fue el motivo principal de la colaboración. Son bien conocidas por la ventajas de este tipo de modelos en lo que respeta a interpretabilidad, cualidad importante en el contexto médico (además de por su versatilidad y potencia). La principal desventaja en este caso es la necesidad de discretizar los datos.

Como hemos mencionado, el tipo de red que vamos a usar es el BAN por su versatilidad de estos modelos en cuanto a la estructura de la red. Así mismo, la variable clase no tiene padres el DAG asociado. Estas suposiciones son razonables en un problema de clasificación supervisada, como el que nos planteamos en este trabajo.

La elección de la regresión logística como modelo viene justificada por múltiples motivos. La principal motivación es la necesidad de un modelo referente para comparar los resultados obtenidos con los modelos de clasificación bayesianos. La regresión logística es un modelo muy conocido, en particular en el contexto médico, por tanto, su uso lo hace atractivo para que sea más familiar a lectores familiarizados con literatura médica. Sirve además para dar continuidad al trabajo de Rosa Faner, Oriol Sibila y colaboradores [4], que usan este tipo de modelos. Para el problema 3, en el cual la variable clase toma más de dos valores no ordinales, es necesario usar regresión logística multinomial.

El aprendizaje de los modelos lo hemos realizado usando R, motivado por la necesidad de poder obtener un formato compatible con GeNIe en los modelos. Para las redes bayesianas, hemos usado la librerías `bnclassify` y `bnlearn` Para las la regresión multinomial, hemos usado la función `multinom`, de la librería de R `nnet`, y para la regresión logística la función `glm` de R. El código usado en este trabajo se puede encontrar `aquí`.

Capítulo 3

Resultados y conclusiones

En este capítulo presentaremos, por un lado, los resultados obtenidos para cada uno de los modelos que resuelven los 6 problemas presentados en el capitulo anterior y, por otro lado, daremos detalles sobre la implementación de los mismos si resulta pertinente.

3.1. Predecir DLCO-12 al alta: BAN

En este problema nos planteamos la estimación del valor del DLCO (normal o anormal) a los 12 meses usando las variables disponibles en el momento de ingreso (problema 1). Este modelo es uno de los más interesantes desde un punto de vista médico, ya que nos permitiría predecir la evolución del paciente a largo plazo usando únicamente la información al momento de ingreso.

El primer paso consiste en aplicar el método *wrapper* a las variables relevantes (obtenidas por *filtering*) para este problema. Las siguientes variables han sido seleccionadas (Obtenidas por *wrapper*):

1. "`fumador`": Variable estática, indica si el paciente fuma.
2. "`SDRA`": Síndrome de distrés respiratorio.
3. "`albumina_d0`": Valor de la albúmina en el momento de ingreso.
4. "`expl_respiratoria_al_ingreso`": Si se ha realizado exploración respiratoria en el momento de ingreso.
5. "`bicarbonato_d0`": Valor del bicarbonato en el momento de ingreso.
6. "`dimero_d2`": Valor del dímero D el segundo día. Se trata de un producto de degradación de la proteína FDP que se usa como marcador de posibles trombosis, entre otras patologías.
7. "`neumonia_organizada`": Si el paciente sufrió neumonía organizada durante el ingreso.
8. "`paquetes_año`": Paquetes de cigarrillos que el paciente fuma al año (por decenas).

En el cuadro 3.1 vemos un resumen de las variable seleccionadas, sus rangos de valores y como se distribuyen con respecto a los dos valores de la variable clase. A

Variable	DLCO<80 % N= 400 (66.01) %	DLCO>80 % N= 206 (33.99 %)	*p*-valor	Valores
fumador	0.765 ± 0.9445	0.961 ± 0.8958	0.01278	[0 1 2]
SDRA	0.2925 ± 0.4555	0.5073 ± 0.5012	$4.323 \cdot 10^{-7}$	[0 1]
albumina_d0	0.0125 ± 0.7096	0.4439 ± 0.6882	$2.415 \cdot 10^{-12}$	[0 -1 1]
expl_respiratoria_al_ingreso	0.355 ± 0.6893	0.3366 ± 0.7201	0.7628	[0 1 2 3 4]
bicarbonato_d0	-0.0325 ± 0.2676	-0.01463 ± 0.2095	0.368	[0 -1 1]
dimero_d2	0.36 ± 0.8557	-0.1512 ± 0.9349	$1.884 \cdot 10^{-10}$	[0 1 -1]
neumonia_organizada	0.3925 ± 0.4889	0.6195 ± 0.4867	$1.005 \cdot 10^{-7}$	[0 1]
paquetes_año	2.76 ± 0.8685	2.463 ± 1.153	0.001306	[0 1 3 2 4]

Cuadro 3.1: Comparación de las variables seleccionadas para cada uno de los valores de la clase.
El p-valor corresponde al test de Student comparando los valores de la misma variable separados en dos conjuntos en base a los valores de la variable clase
Los valores son los posible valore que toma la variable tras discretizar

continuación procedemos a entrenar el modelo. Obtenemos el modelo de la figura 3.1

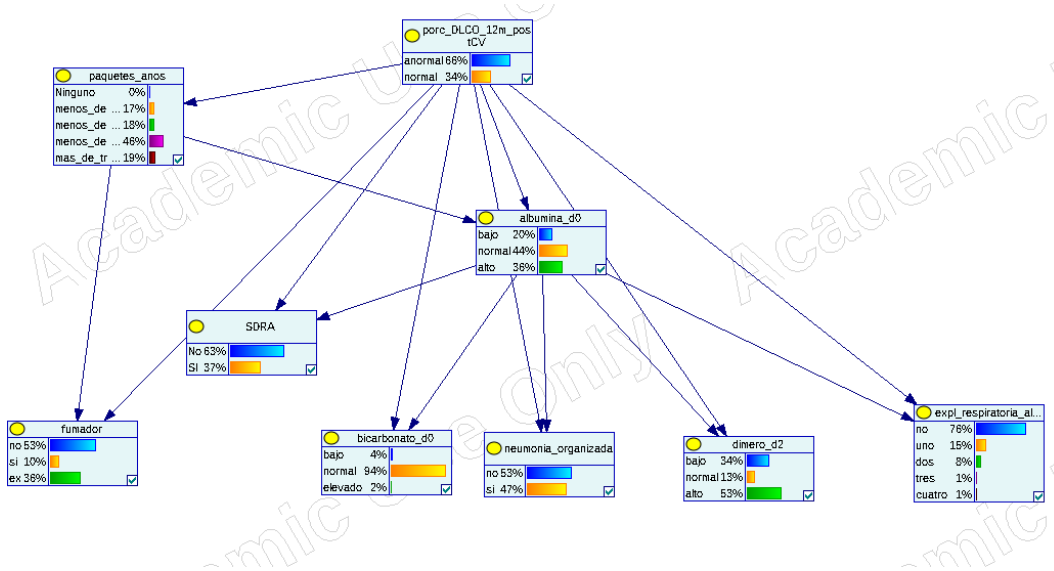

Figura 3.1: Estructura del modelo 1 con GeNIe

Para evaluar el modelo miramos su precisión como clasificador y al área bajo la curva ROC. En este caso obtenemos que el modelo tiene una precisión de 0.7155 y el área bajo la curva ROC es 0.677.

Con este modelo exportado a GeNIE, podemos realizar inferencias con el mismo. A continuación presentamos preguntas que le hacemos al modelo.

1. **Q1** (Figura 3.2): Para un paciente con los valores: ex-fumador, 10 o menos paquetes/año, sin neumonía organizada y no SDRA, el modelo predice que el DL-

CO a 12 meses será inferior al 80 % (anormal) con una probabilidad de 0.81. La probabilidad de DLCO anormal es de 0.66 en el conjunto total de pacientes.

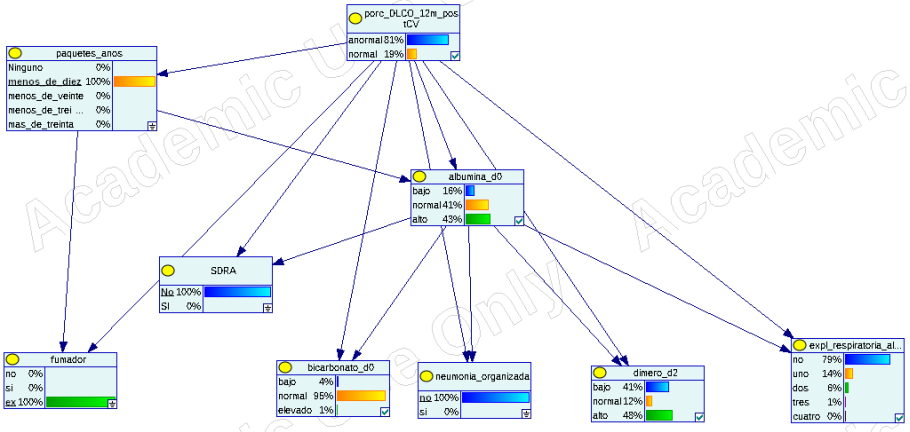

Figura 3.2: Query 1 del modelo 1 con GeNIe

2. **Q2** (Figura 3.3): Nos planteamos como es el paciente prototípico (obtenido por usando MAP) con valor anormal del DLCO a 12 meses
(variable clase porc_DLCO_12m_postCV = 1). Obtenemos los siguientes resultados:

   a) Paquetes/año: entre 20 y 30 decenas con probabilidad del 0.53 frente a un 0.46 en la muestra poblacional.

   b) Albumina d0: Normal con una probabilidad de 0.50 frente a un 0.44 en la muestra poblacional.

   c) SDRA: No, con probabilidad 0.71 frente a 0.63 en la muestra poblacional.

   d) Fumador: No, con probabilidad 0.59 frente a 0.53 en la muestra poblacional.

   e) Bicarbonato: Normal, con probabilidad 0.93 frente a 0.94 en la muestra poblacional.

   f) Neumonía organizada: No, con probabilidad 0.61, frente a 0.53 en la muestra poblacional.

   g) Dímero D día 2: Alto con probabilidad 0.61, frente a 0.53 en la muestra poblacional.

   h) Exploración respiratoria ingreso: No, con probabilidad 0.74, frente a 0.76 en la muestra poblacional.

3. **Q3** (Figura 3.4): Nos planteamos el paciente prototípico con DLCO normal.

   a) Paquetes/año: entre 20 y 30 decenas con probabilidad del 0.31 frente a 0.46 en la muestra poblacional. Con la misma probabilidad (0.31), es que este valor sea de 10 o menos, frente a 0.17 en la muestra.

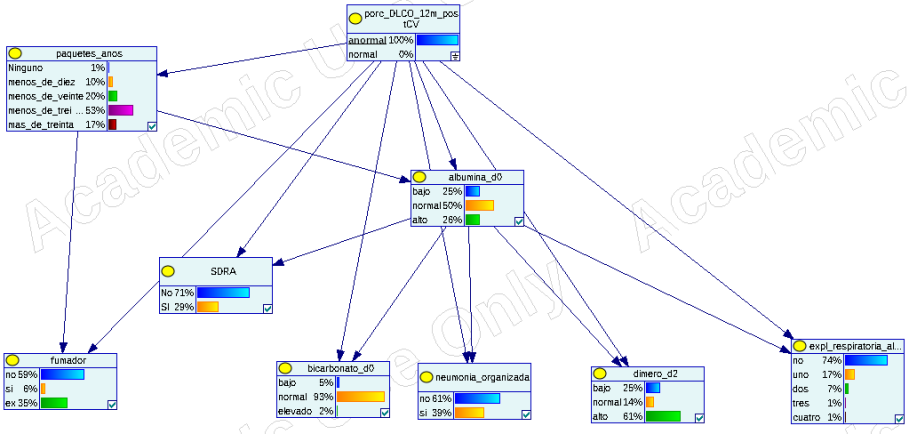

Figura 3.3: *Query 2* del modelo 1 con GeNIe

b) Albumina d0: Alto, con una probabilidad 0.56 frente a 0.36 en la muestra poblacional.

c) SDRA: Si, con probabilidad 0.51 frente a 0.37 en la muestra poblacional.

d) Fumador: No, con probabilidad 0.42 frente a 0.53 en la muestra poblacional.

e) Bicarbonato: Normal, con probabilidad 0.96 frente a 0.94 en la muestra poblacional.

f) Neumonía organizada: Si, con probabilidad 0.61, frente a 0.53 en la muestra poblacional.

g) Dímero D día 2: Bajo con probabilidad 0.52, frente a 0.34 en la muestra poblacional.

h) Exploración respiratoria ingreso: No, con probabilidad 0.78, frente a 0.76 en la muestra poblacional.

4. **Q4** (Figura 3.5) Para un paciente con dímero D alto, fuma entre 20 y 30 decenas de paquetes al año, SDRA no y albumina normal; el modelo predice un valor anormal (0) de DLCO a 12 meses con probabilidad 0.78.

5. **Q5** (Figura 3.6): Para un paciente con dímero D bajo, fuma entre 1 y 10 decenas de paquetes al año, SDRA sí y albúmina alta; el modelo predice un valor normal (1) de DLCO a 12 meses con probabilidad 0.89.

En algunos casos los resultados resultan poco intuitivos. Por ejemplo, vemos que los pacientes con valores normales de DLCO es más probable que sean fumadores que los que tienen DLCO anormal. También vemos que es más probable con un paciente con DLCO normal tenga SDRA que un paciente con DLCO normal. Por ello es importante recordar que las relaciones representadas en la red bayesiana no son necesariamente causales. En otros casos, los resultados parecen ser más realistas. Por ejemplo, los pacientes con DLCO anormal son fuman más paquetes al año.

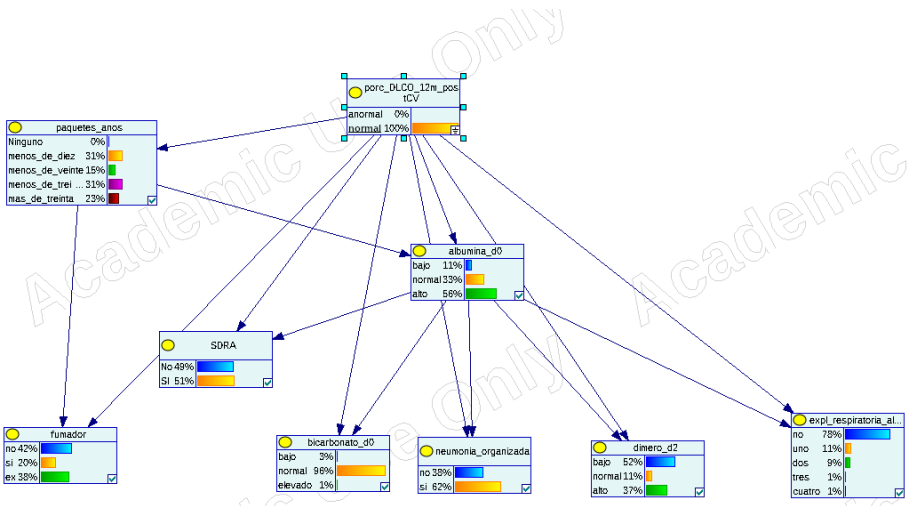

Figura 3.4: *Query* 3 del modelo 1 con GeNIe

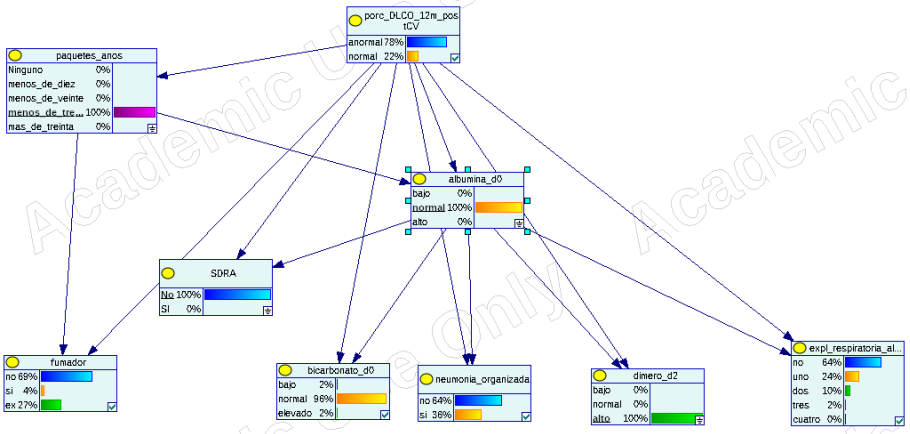

Figura 3.5: *Query* 4 del modelo 1 con GeNIe

3.2. Predecir DLCO-12 al alta: Regresión Logística

Podemos comparar los resultados con la resolución del problema usando regresión logística. Igual que hicimos con el clasificador Bayesiano, realizamos el último paso de la selección de variables con *wrapper*. Estas son las variables seleccionadas (ver cuadro 3.2):

1. "`bilirrubi._total_d0`": Valor de la bilirrubina total en el momento del ingreso.

2. "`leucocitos_total_d0`": Valor absoluto de los leucocitos totales en el momento del ingreso.

3. "`gas_arterial_po2_d0`": Valor de la presión de oxígeno (po2) en la gasometría arterial en el momento del ingreso.

4. "`obesidad_morbida`": Si el IMC del paciente supera 35.

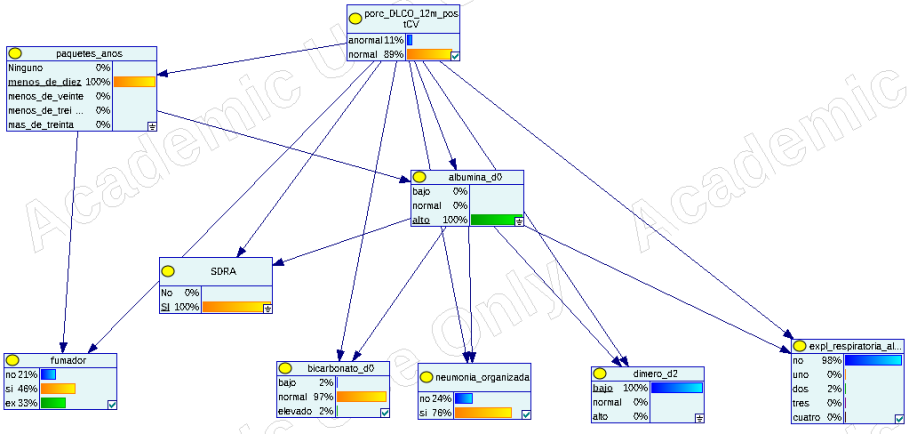

Figura 3.6: *Query* 5 del modelo 1 con GeNIe

5. "`plaquetas_d0`": Valor de las plaquetas en el momento del ingreso.
6. "`albumina_d0`": Valor de la albúmina en el momento del ingreso.
7. "`paquetes_años`": Paquetes de cigarrillos que el paciente fuma al año (por decenas).
8. "`procalcitonina_d2`": Valor de la procalcitonina en el segundo día de ingreso.

En la tabla 3.2 mostramos las variables seleccionadas para este problema, la media y desviación típica para dichas variables de los pacientes con DLCO normal y anormal, los p-valores del test de Student comparando los valore de una variable asociados a un valore de la variable calse frente al otro as los valores que toman.

Variable	**DLCO<80 %** **N= 400 (66,01 %)**	**DLCO>80 %** **N= 206 (33,99 %)**	***p*-valor**	**Valores**
bilirrubi._total_d0	0.535 ± 0.4994	0.6732 ± 0.4702	0.0008813	[0 1]
leucocitos_total_d0	-0.0325 ± 0.4869	-0.02927 ± 0.395	0.93	[0 -1 1]
gas_arterial_po2_d0	-0.0525 ± 0.4419	0.01951 ± 0.3424	0.02743	[0 -1 1]
obesidad_morbida	0.32 ± 0.4724	0.5171 ± 0.5009	$4.177 \cdot 10^{-6}$	[0 1 -1]
plaquetas_d0	-0.3925 ± 0.5236	-0.561 ± 0.5535	0.0003475	[0 -1 1]
albumina_d0	0.0125 ± 0.7096	0.4439 ± 0.6882	$2.415 \cdot 10^{-12}$	[0 -1 1]
paquetes_años	2.76 ± 0.8685	2.463 ± 1.153	0.001306	[0 1 3 2 4]
procalcitonina_d2	0.635 ± 0.482	0.7463 ± 0.4362	0.004348	[0 1]

Cuadro 3.2: Comparación de las variables seleccionadas para cada uno de los valores de la clase.

En el cuadro 3.3 se muestran las variables que se han seleccionado en ambos modelos:

Observamos que las variables seleccionadas cambian de un modelo a otro, aunque vemos que la albúmina y los paquetes al año son seleccionados por ambos modelos. Esto nos puede indicar que estas métricas son particularmente valiosas para predecir

Red Bayesiana	Comunes	R. Logística
fumador	DLCO_12m_postCV	bilirrubi._total_d0
SDRA	albumina_d0	leucocitos_total_d0
expl_respiratoria_al_ingreso	paquetes_año	gas_arterial_po2_d0
bicarbonato_d0		obesidad_morbida
dimero_d2		plaquetas_d0
neumonia_organizada		procalcitonina_d2

Cuadro 3.3: Variables seleccionadas para el modelo 1 (Red Bayesiana (RB) y Regresión Logística (RL))

la variable clase. Con este modelo obtenemos una precisión de 69.70 % y un área bajo la curva ROC de 0.6034. Los obtenidos con la red bayesiana son ligeramente superiores a los obtenidos con regresión logística.

	Red Bayesiana	R. Logística	I.C. 95 %	p-valor
Precisión	0.7155	0.697	(-0.0218, 0.059)	0.359
AUC	0.677	0.6034	**(0.0216, 0.126)**	**0.007**

Cuadro 3.4: Comparativa de los dos modelos para el problema 1
Intervalo de confianza y p-valor obtenidos con el test de Student comparando el conjunto de los resultados de precisión y AUC para cada uno de los k-pliegues de la validación con los del conjunto obtenido por el otro modelo.

Como vemos en el cuadro 3.4, el modelo de red bayesiana es mejor que la regresión logística con diferencia estadísticamente significativa (p-valor 0.007), usando el área bajo la curva ROC como métrica. Aunque la precisión también es mayor, no podemos afirmar que es mejor desde un punto de vista de significatividad estadística.

Finalmente, vemos en el cuadro 3.5 los coeficientes ($\hat{\beta}$) de la regresión logística. Esta tabla muestran los parámetros de la regresión logística. Se puede ver que hay una variedad de variables que se han analizado para determinar el *Odds Ratio* (OR) y el intervalo de confianza de los mismos (CI 95 %). Además, se muestra el p-valor. Estos resultados pueden ser útiles para comprender la relación entre las variables y el resultado de la regresión logística.

Vemos que los resultados son similares, aunque claramente mejore respecto al AUC en el caso del modelo bayesiano. Asimismo, las variables seleccionadas difieren considerablemente, aunque la albúmina y los paquetes al año han sido seleccionados en ambos modelos.

3.3. Predecir DLCO-12 a los 3 meses: BAN

En este problema nos planteamos una cuestión muy similar a la del problema anterior, pero en este caso incluimos las variables medidas a los tres meses. De nuevo el objetivo es predecir el valor del porcentaje de DLCO a 12 meses (problema 2). Presumiblemente este problema es más fácil, ya que tenemos más información, incluido el valor de la variable clase a los 3 meses. Además, en este caso, tenemos acceso a los datos de alta, días de ingreso, días en la UCI, etc. Tras tres meses, todos los pacientes admitidos reciben el alta, dándonos información adicional que podemos usar en este

Variables	**OR**	**CI 95 %**	**p-valor**	β
leucocitos_total_d0(ref -1) valor 0	1.38	$(0.77, 2.47)$	0.877	1.056
leucocitos_total_d0(ref -1) valor 1	0.77	$(0.33, 1.81)$	0.68	$8.132 \cdot 10^{-1}$
bilirrubina_total_d0	1.67	$(1.15, 2.41)$	0.407	$7.94 \cdot 10^{-1}$
gas_arterial_po2_d0 (ref -1) valor 0	2.35	$(1.15, 4.82)$	0.322	1.49
gas_arterial_po2_d0 (ref -1) valor 1	2.18	$(0.85, 5.6)$	0.393	1.62
obesidad_morbida	2.13	$(1.48, 3.06)$	0.225	$5.99 \cdot 10^{-1}$
plaquetas_d0 (ref -1) valor 0	0.47	$(0.33, 0.68)$	0.686	1.141
plaquetas_d0 (ref -1) valor 1	1.16	$(0.38, 3.55)$	0.233	2.47
albumina_d0 (ref -1) valor 0	1.53	$(0.88, 2.68)$	0.06	1.779
albumina_d0 (ref -1) valor 1	4.55	$(2.62, 7.9)$	< 0.001	6.534
paquetes_años (ref 0) valor 1	3722861.83	$(0, \infty)$	0.982	$1.856 \cdot 10^{6}$
paquetes_años (ref 0) valor 2	865172.11	$(0, \infty)$	0.983	$4.654 \cdot 10^{5}$
paquetes_años (ref 0) valor 3	638816.19	$(0, \infty)$	0.983	$4.253 \cdot 10^{5}$
paquetes_años (ref 0) valor 4	1412120	$(0, \infty)$	0.983	$7.482 \cdot 10^{5}$
procalcitonina_d2	1.51	$(1.03, 2.23)$	0.399	$7.985 \cdot 10^{-1}$

Cuadro 3.5: Regresión logística modelo 1 Odds Ratio (OR), intervalos de confianza (CI), p-valor y parámetros (β)

modelo. Las variables seleccionadas por el método *wrapper* son:

1. "porc_DLCO_3m_postCV": Valor de la variable clase a los tres meses.
2. "alt_d0": Valor de la alanina aminotransferasa (ALT) en el momento del ingreso.
3. "tvptep": Sufre tromboembolismo pulmonar (TEP) o trombosis venosa profunda (TVP).
4. "creatinina_d0": Valor de la creatinina en el momento del ingreso.
5. "LDH_d0": Valor de la LDH en el momento del ingreso.
6. "bicarbonato_d0": Valor del bicarbonato en el momento del ingreso.
7. "LDH_d2": Valor de la LDH en el día 2 del ingreso.
8. "neumonia_organizada": Si el paciente sufrió neumonía organizada durante el ingreso.
9. "albumina_d0": Valor de la albumina en el momento de ingreso.

El Cuadro 3.6 nos muestra la media y varianza de las variables seleccionadas para este problema, junto con el p-valor y los valores que toman.

Vemos algunas variables que aparecen en el problema anterior. Desataca la albúmina en ingreso (además del bicarbonato y neumonía organizada). También es importante destacar la elección de la variable clase a los tres meses. Otras variables (tvptep, LDH) no habían aparecido antes, y el clasificador bayesiano presenta una estructura distinta. En este caso el modelo tiene una precisión de **0.7288** y el área bajo la curva ROC es **0.7189**. Esto supone una mejora con respecto al modelo del problema

Variable	DLCO<80 % N= 400 (66.01 %)	DLCO>80 % N= 206 (33.99 %)	*p*-valor	Valores
porc_DLCO_3m_postCV	0.2625 ± 0.4405	0.3415 ± 0.4754	0.04821	[1 0]
alt_d0	0.2275 ± 0.443	0.1854 ± 0.3895	0.2304	[0 1 -1]
tvptep	0.1125 ± 0.3164	0.3805 ± 0.4867	$6.963 \cdot 10^{-12}$	[0 1]
creatinina_d0	-0.01 ± 0.8103	-0.4488 ± 0.7433	$7.961 \cdot 10^{-11}$	[-1 1 0]
LDH_d0	0.6025 ± 0.5001	0.322 ± 0.4787	$5.841 \cdot 10^{-11}$	[0 1 -1]
bicarbonato_d0	-0.0325 ± 0.2676	-0.01463 ± 0.2095	0.368	[0 -1 1]
LDH_d2	0.365 ± 0.8388	-0.1561 ± 0.9156	$3.713 \cdot 10^{-11}$	[0 1 -1]
albumina_d0	0.0125 ± 0.7096	0.4439 ± 0.6882	$2.415 \cdot 10^{-12}$	[0 -1 1]
neumonia_organizada	0.3925 ± 0.4889	0.6195 ± 0.4867	$1.005 \cdot 10^{-7}$	[0 1]

Cuadro 3.6: Comparación de las variables seleccionadas para cada uno de los valores de la clase

anterior, lo cual es razonable al tener más información disponible. Al exportar este modelo a GeNIe obtenemos la estructura de la Figura 3.7.

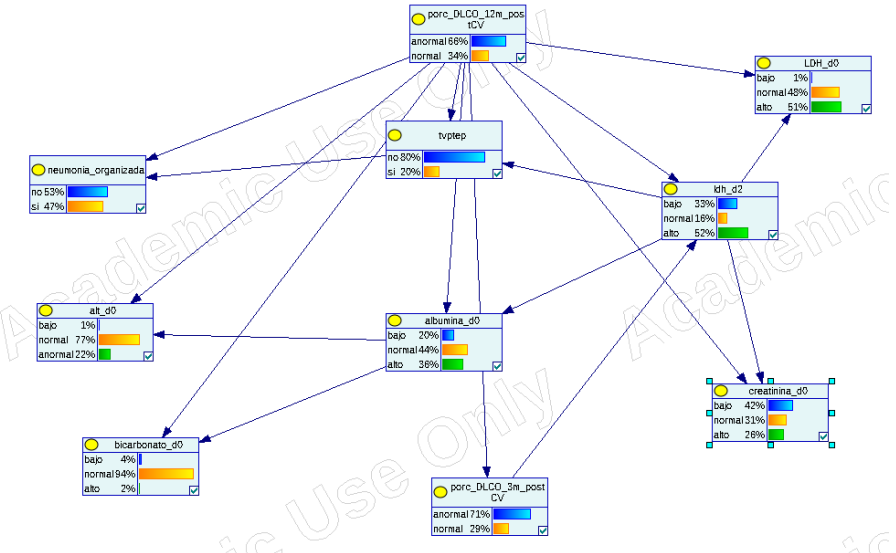

Figura 3.7: Estructura del clasificador obtenido para el problema 2, con GeNIe

Vamos a realizar preguntas al modelo, usando GeNIe, para entenderlo mejor y ver como clasifica.

1. **Q1** (Figura 3.8): Comencemos con un paciente hipotético, que podría esperarse que tenga valores anormales de DLCO a 12 meses (en base a intuición y lo que ya conocemos del modelo 1). Consideramos paciente con DLCO anormal a los 3 meses, LDH_d0, LDH_d2 y creatinina altos y neumonia_organizada no. El modelo predice que el DLCO será anormal a los 12 meses con una probabilidad de 0.89 % (frente a 0.66 base).

2. **Q2** (Figura 3.9): Nos planteamos como es el paciente prototípico con DLCO anormal.

   a) tvptep: No, con probabilidad 0.89 frente a 0.80 en la muestra poblacional.

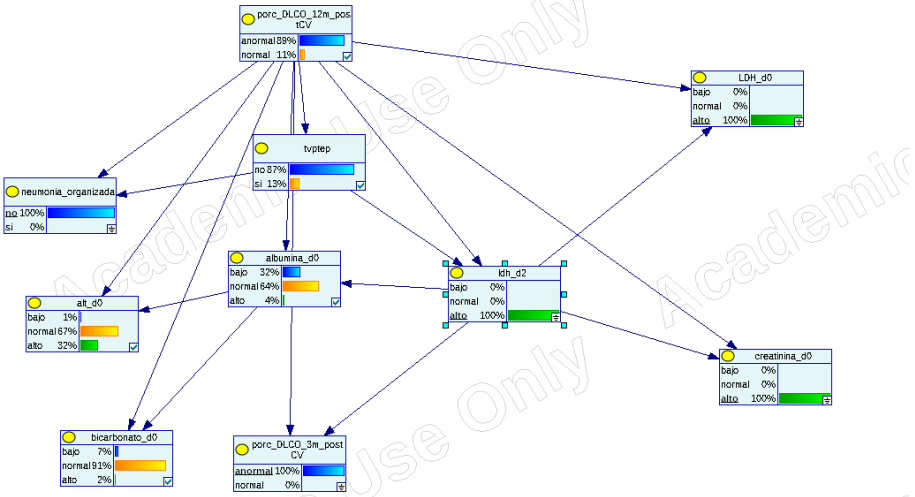

Figura 3.8: *Query 1*, modelo 2 con GeNIe

b) neumonia_organizada: No, con probabilidad de 0.61 frente a 0.53 en la muestra poblacional.

c) LDH_d0: Alto, con probabilidad 0.61 frente a 0.51 en la muestra poblacional.

d) albumina_d0: Normal, con probabilidad 0.50 frente a 0.44 en la muestra poblacional.

e) LDH_d2: Alto, con probabilidad 0.60 frente a 0.52 en la muestra poblacional.

f) alt_d0: Normal con probabilidad 0.75 frente a 0.77 en la muestra poblacional.

g) creatinina_d0: Normal, con probabilidad 0.34 frente a 0.31 en la muestra poblacional.

h) bicarbonato_d0: Normal, con probabilidad 0.93 frente a 0.94 en la muestra poblacional.

i) porc_DLCO_3m_postCV: Anormal, con probabilidad 0.74 frente a 0.71 en la muestra poblacional.

3. **Q3** (Figura 3.10): Nos planteamos como es el paciente prototípico con DLCO normal.

a) tvptep: No, con probabilidad 0.62 frente a 0.80 en la muestra poblacional.

b) neumonia_organizada: Si, con probabilidad 0.61 frente a 0.47 en la muestra poblacional.

c) LDH_d0: Normal, con probabilidad 0.67 frente a 0.48.

d) albumina_d0: Alta, con probabilidad 0.56 frente a 0.36 en la muestra poblacional.

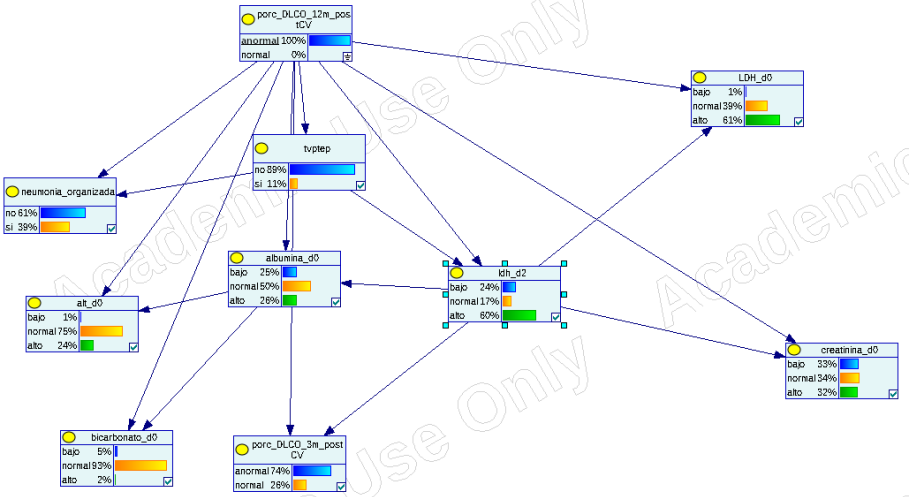

Figura 3.9: *Query 2*, modelo 2 con GeNIe

e) LDH_d2: Bajo, con probabilidad 0.51 frente a 0.33 en la muestra poblacional.

f) alt_d0: Normal con probabilidad 0.81 frente a 0.77 en la muestra poblacional.

g) creatinina_d0: Baja, con probabilidad 0.60 frente a 0.42 en la muestra poblacional.

h) bicarbonato_d0: Normal, con probabilidad 0.96 frente a 0.94 en la muestra poblacional.

i) porc_DLCO_3m_postCV: Anormal, con probabilidad 0.66 frente a 0.71 en la muestra poblacional.

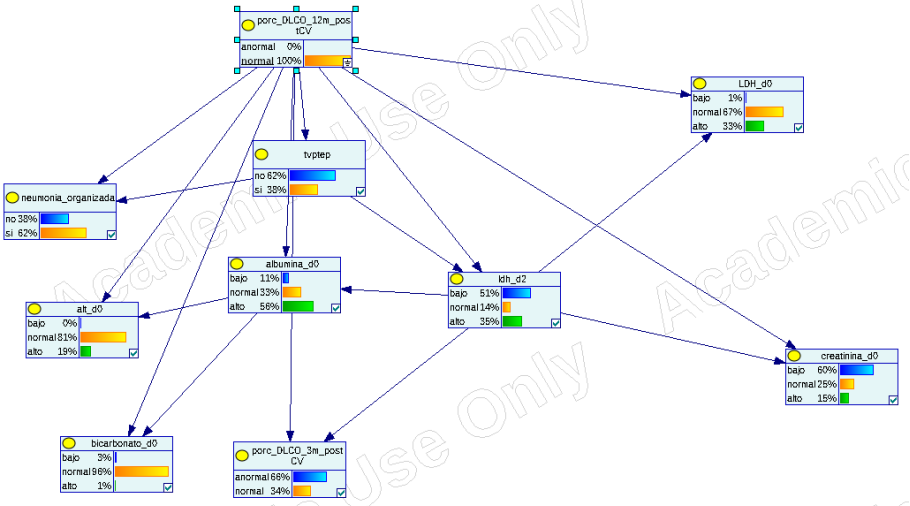

Figura 3.10: query 4, modelo 2 con GeNIe

4. **Q4** (Figura 3.11): Para un paciente sin neumonía organizada, LDH alto (d0 y d2), porc_DLCO_3m_postCV anormal y con tvptep; el modelo predice que el DLCO a 12 meses será anormal con probabilidad de 0.98 (frente a 0.66 en la muestra poblacional).

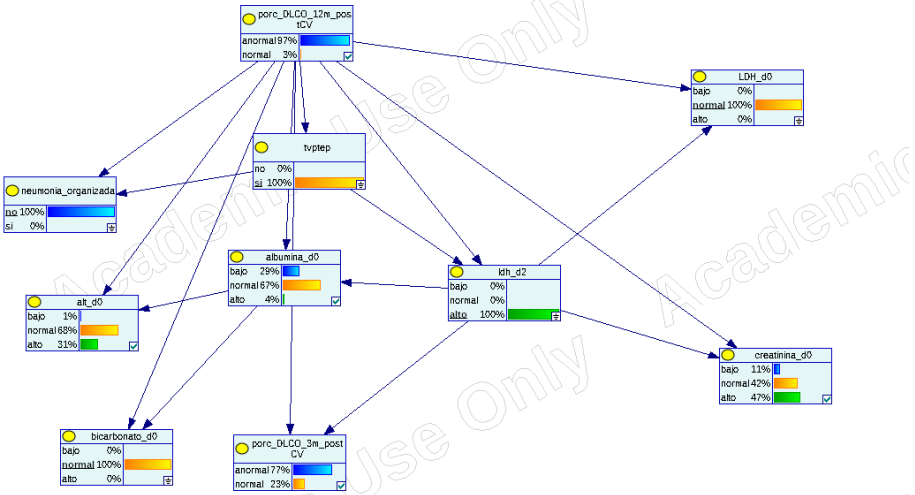

Figura 3.11: *Query 4*, modelo 2 con GeNIe

5. **Q5** (Figura 3.12): Para un paciente con LDH bajo (d0 y d2), sin neumonía organizada, porc_DLCO_3m_postCV normal, y bicarbonato_d0; el modelo predice que el valor del DLCO a los 12 meses será anormal con probabilidad de 0.98, frente a 0.34 en la muestra poblacional.

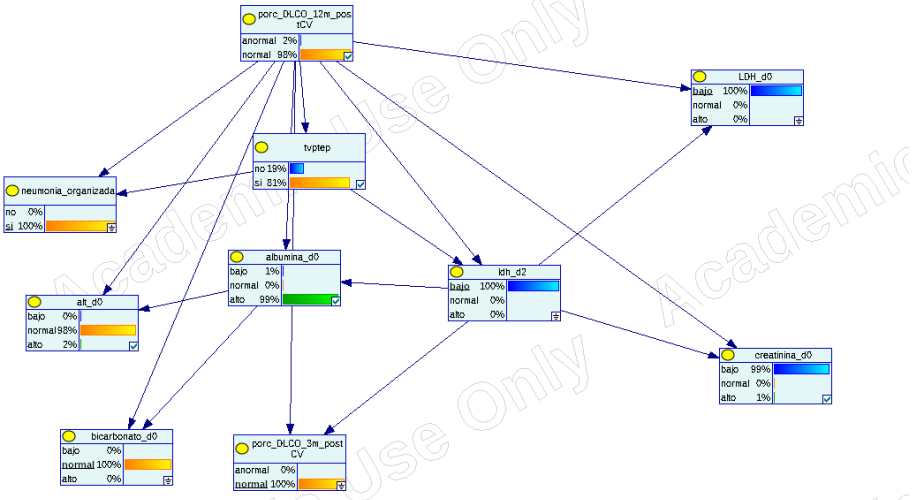

Figura 3.12: *Query 5*, modelo 2 con GeNIe

3.4. Predecir DLCO-12 a los 3 meses: Regresión Logística

Podemos comparar los resultados con la resolución del problema usando regresión logística. Igual que hicimos con el modelo Bayesiano, realizamos el último paso de la selección de variables con *wrapper*, obteniendo las siguientes variables (Cuadro 3.7):

1. "`tvptep`": Sufre tromboembolismo pulmonar (TEP) o trombosis venosa profunda (TVP).
2. "`LDH_d2`": Valor de la LDH en el día 2 del ingreso.
3. "`hematocrito_d0`": Valor del hematócrito en el momento del ingreso.
4. "`dimero_d_al_alta`": Valor del Dímero D en el alta.
5. "`pas_d0`": Presión arterial sistólica (PAS) en el momento del ingreso.
6. "`LDH_d0`": Valor de la LDH en el momento del ingreso.
7. "`procalcitonina_d0`": Valor de la procalcitonina en el momento de ingreso.
8. "`SPD_3m`": Niveles de SPD (ng/ml) por ELISA en plasma a los 3 meses.
9. "`plaquetas_d0`": Valor de las plaquetas en el momento del ingreso.

Podemos observar que ambos modelos tienen varias variables en común (Cuadro 3.8). Vemos también como se incluye información de alta y de los tres meses (pese a que no aparece el DLCO a los 3 meses).

Variable	**DLCO<80 % N= 400 0.6601**	**DLCO>80 % N= 206 (33,99 %)**	***p*-valor**	**Valores**
dimero_d_al_alta	0.5525 ± 0.59	0.2829 ± 0.5313	$2.334 \cdot 10^{-8}$	[0 1 -1]
tvptep	0.1125 ± 0.3164	0.3805 ± 0.4867	$6.963 \cdot 10^{-12}$	[0 1]
SPD_3m	0.535 ± 0.4994	0.3463 ± 0.477	$7.605 \cdot 10^{-6}$	[0 1]
LDH_d0	0.6025 ± 0.5001	0.322 ± 0.4787	$5.841 \cdot 10^{-11}$	[0 1 -1]
pas_d0	0.79 ± 0.4078	0.8195 ± 0.3855	0.3827	[1 0]
LDH_d2	0.365 ± 0.8388	-0.1561 ± 0.9156	$3.713 \cdot 10^{-11}$	[0 1 -1]
hematocrito_d0	-0.315 ± 0.5014	-0.5561 ± 0.4981	$3.462 \cdot 10^{-8}$	[0 -1 1]
plaquetas_d0	-0.3925 ± 0.5236	-0.561 ± 0.5535	0.0003475	[0 -1 1]
procalcitonina_d0	0.2675 ± 0.4432	0.1268 ± 0.3336	$1.469 \cdot 10^{-5}$	[0 1]

Cuadro 3.7: Comparación de las variables seleccionadas para los valores de la clase

Red Bayesiana	Comunes	R. Logística
porc_DLCO_3m_postCV	DLCO_12m_postCV	dimero_d_al_alta
alt_d0	tvptep	SPD_3m
creatinina_d0	LDH_d0	pas_d0
bicarbonato_d0	LDH_d2	hematocrito_d0
albumina_d0		plaquetas_d0
neumonia_organizada		procalcitonina_d0

Cuadro 3.8: Variables seleccionadas para el modelo 2 (RB y RL)

Con este modelo obtenemos una precisión de **0.7585** y un área bajo la curva ROC de **0.6726**. Al igual que con la red bayesiana, al tener más información obtenemos me-

jores resultados. En este caso, el modelo logístico obtiene una precisión ligeramente superior a la red bayesiana, aunque el área bajo la curva ROC es mejor en la red bayesiana. Pero no podemos afirmar que estás diferencias son significativas, véase Cuadro 3.9(con $p < 0.05$). Las características del modelo logístico se muestran en el Cuadro 3.10

	Red bayesiana	R. logística	I.C. 95 %	p-valor
Precisión	0.7288	0.7585	(-0.0873, 0.028)	0.301
AUC	0.7189	0.6726	(-0.0031, 0.0958)	0.0651

Cuadro 3.9: Comparativa de los dos modelos para el problema 2
Intervalo de confianza y p-valor obtenidos con el test de Student comparando el conjunto de los resultados de precisión y AUC para cada uno de los k-pliegues de la validación con los del conjunto obtenido por el otro modelo.

Variables	OR	CI 95 %	p-valor	β
tvptep 1 vs 0	4.67	(3,7.27)	<0.001	3.98
dimero_d_al_alta, valor 0 (ref -1)	1.97	$(0.82, 4.74)$	0.465	1.458
dimero_d_al_alta, valor 1 (ref -1)	0.57	$(0.24, 1.39)$	0.256	$5.694 \cdot 10^{-1}$
plaquetas_d0, valor 0 (ref -1)	0.45	$(0.31, 0.65)$	0.342	1.36
plaquetas_d0, valor 1 (ref -1)	1.16	$(0.38, 3.56)$	0.054	3.646
SPD_3m 1 vs 0	0.5	$(0.35, 0.73)$	0.37	1.289
LDH_d0, valor 0 (ref-1)	1.67	$(0.15, 18.64)$	0.781	$7.016 \cdot 10^{-1}$
LDH_d0, valor 1 (ref-1)	0.59	$(0.05, 6.59)$	0.584	$4.926 \cdot 10^{-1}$
pas_d0 1 vs 0	1.24	$(0.79, 1.95)$	0.34	$7.6753 \cdot 10^{-1}$
LDH_d2, valor 0 (ref -1)	0.42	$(0.24, 0.72)$	0.136	$4.0983 \cdot 10^{-1}$
LDH_d2, valor 1 (ref -1)	0.28	$(0.19, 0.42)$	0.118	$3.7183 \cdot 10^{-1}$
hematocrito_d0, valor 0 (ref -1)	0.45	$(0.31, 0.64)$	0.129	1.8853
hematocrito_d0, valor 1 (ref -1)	0	$(0, \infty)$	0.981	$9.726 \cdot 10^{-1}$
procalcitonina_d0 1 vs 0	0.44	$(0.27, 0.7)$	0.708	$8.966 \cdot 10^{-1}$

Cuadro 3.10: Regresión logística del modelo 2: *odds ratio* (OR), intervalos de confianza (CI), p-valore y parámetros (β)

3.5. Predecir evolución al alta: BAN

El problema 3 se plantea como un problema en el cual se busca predecir los valores de DLCO en los diferentes momentos de tiempo usando las variables conocidas en el momento de ingreso. Es decir, en este caso, la variable clase toma $2^3 = 8$ valores. Para un paciente con DLCO anormal a los 3 y 6 meses, y DLCO normal a los 12 meses, la variable clase tomaría valor $(1, 1, 0)$ o valor $1 \cdot 2^0 + 1 \cdot 2^1 + 0 \cdot 2^2 = 3$. Esta nueva variable clase tendrá el nombre "new_class". En la Figura 3.13 vemos como se distribuyen los pacientes en las 8 clases. El proceso de selección de variables sigue el mismo procedimiento para los modelos anteriores. Para este modelo, las siguientes variables han sido seleccionadas:

1. "`LDH_d2`" : El valor de la lactato deshidrogenasa en el segundo día de ingreso.
2. "`tvptep`" : Sufre tromboembolismo pulmonar (TEP) o trombosis venosa profunda (TVP).

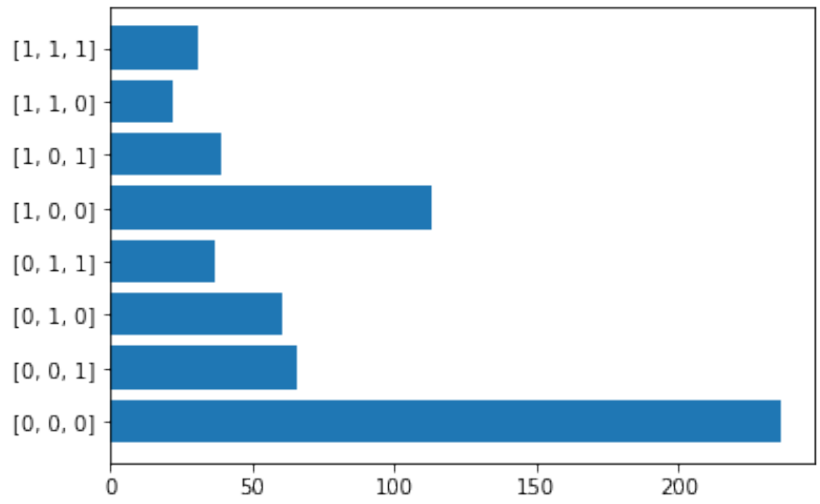

Figura 3.13: Evolución de DLCO de los pacientes a los 3, 6 y 12 meses

3. "`comorbilidades`": Si presenta co-morbilidades (durante o previo al ingreso).
4. "`peso`" : Peso del paciente.
5. "`fosfat_alcali._d0`" : Valor de la fosfatasa alcalina en el momento del ingreso.
6. "`pas_d0`" : Presión arterial sistólica (PAS) en el momento del ingreso.
7. "`consumo_de_alcohol`" : Si el paciente consume alcohol
8. "`bicarbonato_d0`" : Valor del bicarbonato en el momento del ingreso.
9. "`procalcitonina_d2`": Valor de la procalcitonina el segundo día de ingreso.

Vemos variables que han aparecido ya en modelos anteriores, como bicarbonato, procalcitonina, LDH, tvptep y fumador. El modelo tiene una precisión de **0.4592** y el area bajo la curva ROC es **0.6731**. En este caso, la precisión baja. Notamos que el problema es más complicado, teniendo la variable clase 8 posibles valores. Aún así, el área bajo la curva ROC se mantiene relativamente alta.

La Figura 3.14 representa el modelo 3 en GeNIe: A continuación, vamos a hacer consultas al modelo usando GeNIe para entender como funciona.

1. **Q1** (Figura 3.15): Consideramos un paciente que se ha mantenido valores anormales de DLCO a los 3, 6 y 12 meses (configuración $(0, 0, 0)$ de la clase). Vemos en este paciente:
   - La probabilidad de comorbilidades positiva es 0.93 frente a un 0.88 en la muestra poblacional.
   - La probabilidad de LDH_d2 elevado es 0.73 frente a 0.52 en la muestra poblacional.
2. **Q2** (Figura 3.16): Consideramos un paciente que se ha mantenido valores normales de DLCO a los 3, 6 y 12 meses (configuración $(1, 1, 1)$ de la clase). Vemos

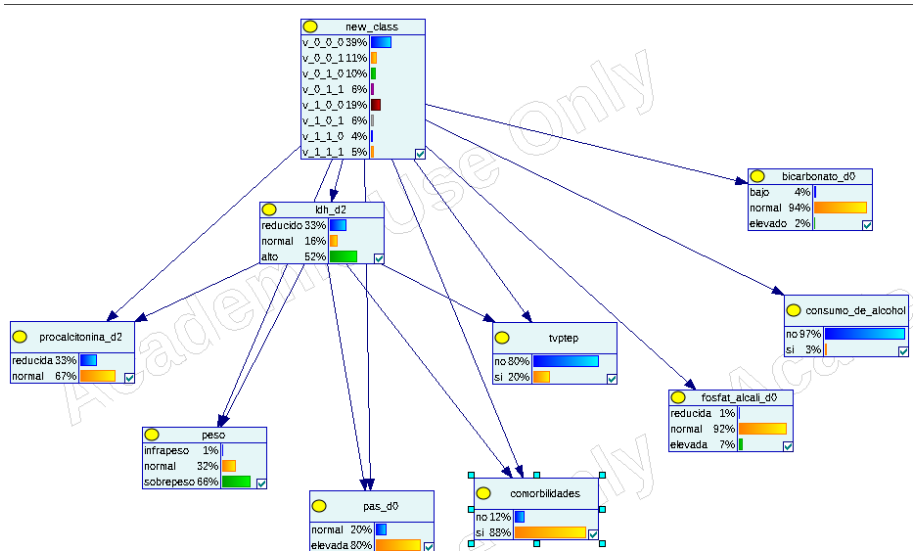

Figura 3.14: Modelo 3 con GeNIe

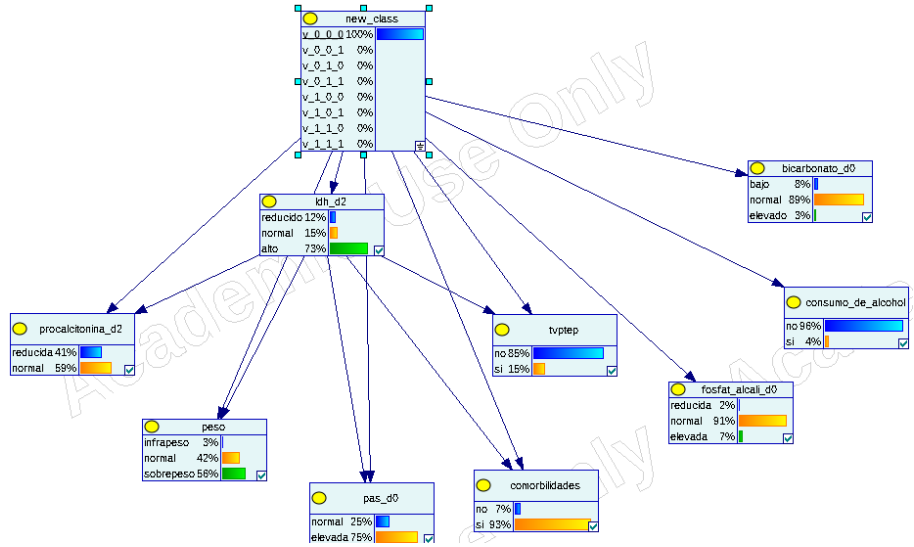

Figura 3.15: *Query 1*, modelo 2 con GeNIe

en este paciente:

- La procalcitonina_d2 reducida con probabilidad 0.58 frente a 0.33 en la muestra poblacional.
- El valor de tvptep es negativo con probabilidad 0.97 frente a 0.80.
- La probabilidad de comorbilidades es 0.68 frente a 0.88 en la muestra poblacional.

3. **Q3** (Figura 3.17): Consideramos un paciente con procalcitonina_d2 reducida, que consume alcohol y con sobrepeso. Para dicho paciente, el modelo predice que tendrá valores anormales en los 3 periodos (clase $(0, 0, 0)$) con probabilidad 0.63 frente a 0.39 en la muestra poblacional.

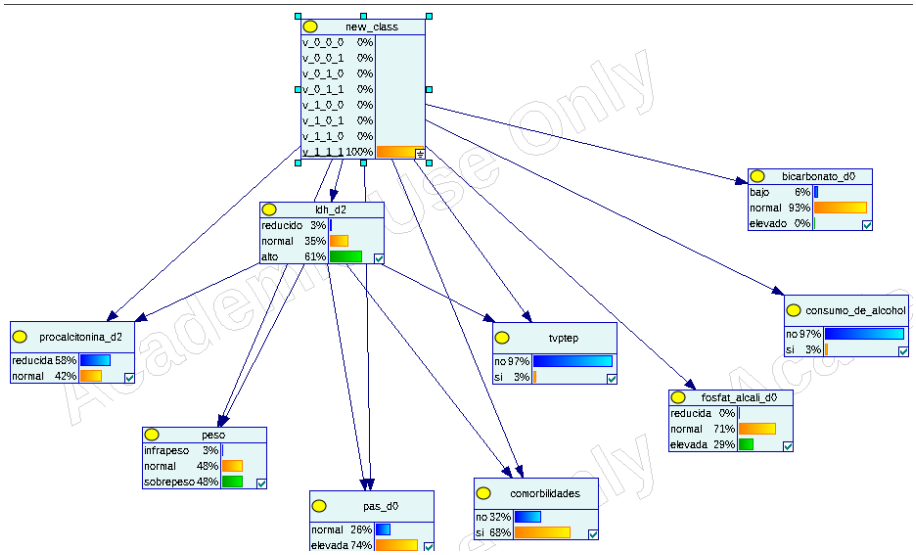

Figura 3.16: *Query 2*, modelo 3 con GeNIe

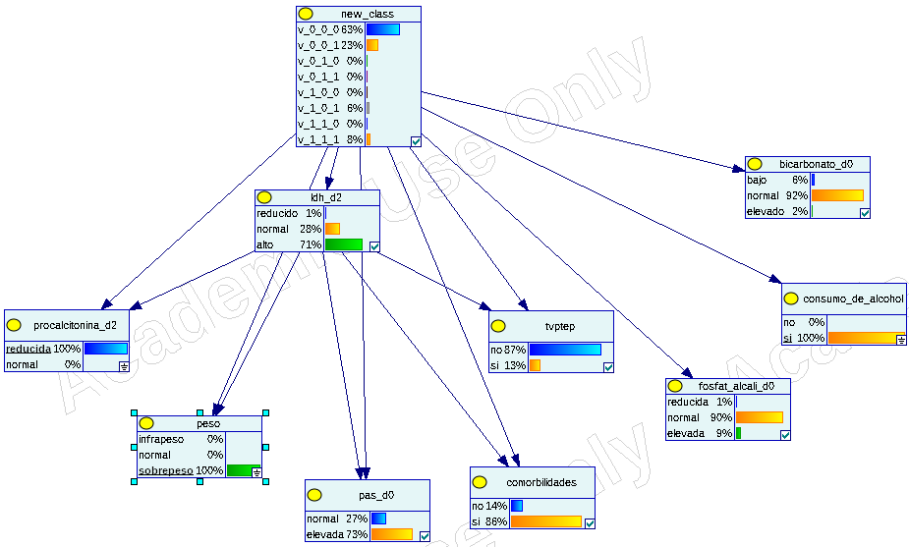

Figura 3.17: *Query 3*, modelo 3 con GeNIe

4. **Q4** (Figura 3.18): Para un paciente con procalcitonina normal, peso normal, pas_d0 normal, sin comorbilidades, tvptep negativo, LDH_d2 normal, bicarbonato_d0 normal y que no consume alcohol, el modelo predice que el paciente pertenece a la clase $(1, 0, 1)$ con probabilidad 0.26 frente a 0.06 en la muestra poblacional.

5. **Q5** (Figura 3.19): Para un paciente sin comorbilidades y LDH_d2 normal, el modelo predice que pertenecerá a la clase $(0, 0, 1)$ con probabilidad 0.4 frente a 0.11 en la muestra poblacional.

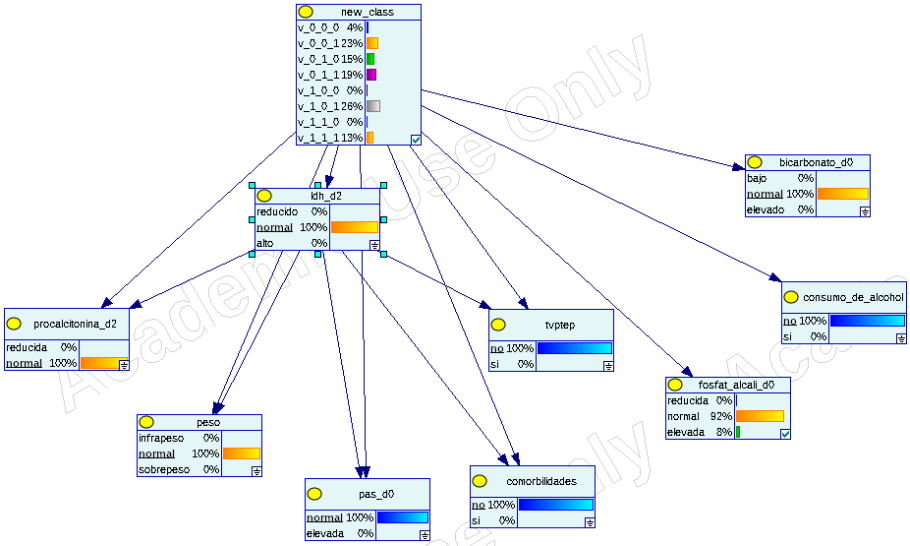

Figura 3.18: query 4, modelo 3 con GeNIe

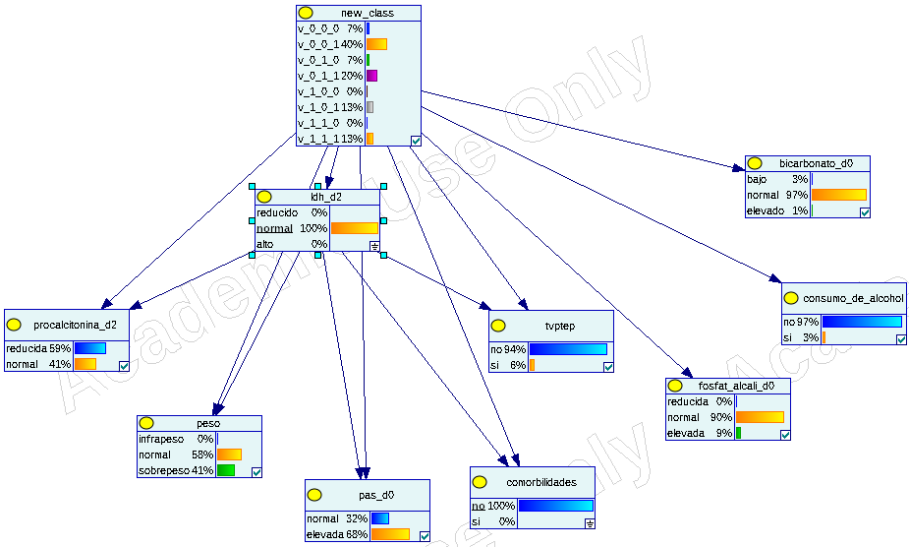

Figura 3.19: *Query 5*, modelo 3 con GeNIe

3.6. Regresión Logística (multinomial)

Vamos a entrenar un modelo logístico que resuelva este problema para comparar los resultados de la red bayesiana. En este caso, debido a que la variable clase toma valores no binarios, debemos usar un modelo de regresión multinomial. Estas son las variables seleccionadas:

1. "`LDH_d2`" : El valor de la lactato deshidrogenasa en el segundo día de ingreso.
2. "`tvptep`" : Sufre tromboembolismo pulmonar (TEP) o trombosis venosa profunda (TVP).
3. "`"fosfat_alcali_d0`" : Valor de la fosfatasa alcalina en el momento del ingreso.
4. "`gas_arterial_po2_d0`" : Valor de la presión de oxígeno (po2) en la gasometría

arterial en el momento del ingreso.

5. "saturacion_basal_d0": Saturación basal en el momento del ingreso.
6. "LDH_d0" : El valor de la lactato deshidrogenasa en el momento del ingreso.
7. "calcio_d0" : Valor del calcio en el momento del ingreso.
8. "fumador" : Consumo de tabaco.

En lo que a selección de variables respecta, vemos que hay un amplia intersección entre las variables, como se contempla en el Cuadro 3.11.

Red bayesiana	**Comunes**	**R. Logistica**
comorbidades	new_class	gas_arterial_po2_d0
peso	LDH_d2	saturacion_basal_d0
pas_d0	tvptep	LDH_d0
consumo_de_alcohol	fosfat_alcali_d0	calcio_d0
bicarbonato_d0		fumador
procalcitonina_d2		

Cuadro 3.11: Variables seleccionadas comunes y no comunes para este problema

	Red Bayesiana	R. Logística	I.C. 95%	p-valor
Precisión	0.4495	0.5314	$(-0.5602, -0.4937)$	$5.007 \cdot 10^{-11}$
AUC	0.683	0.676	$(-0.0268, 0.0393)$	0.693

Cuadro 3.12: Comparativa de los dos modelos para el problema 3
Intervalo de confianza y p-valor obtenidos con el test de Student comparando el conjunto de los resultados de precisión y AUC para cada uno de los k-pliegues de la validación con los del conjunto obtenido por el otro modelo.

Para este problema, vemos en el Cuadro 3.12 que la regresión logística tiene mayor precisión siendo las diferencias estadísticamente significativas. Y el área bajo la curva ROC es similar en los dos modelos.

3.7. Predecir DLCO-3 alta: BAN

Se trata de predecir con las variables de ingreso el DLCO a los 3 meses (si el porcentaje de DLCO a los 3 meses del alta estará o no por encima del 80%). Este problema (problema 4) es sencillo de plantear, y parece razonable esperar buenos resultados de los modelos, ya que se trata de predecir en un espacio de tiempo más corto. Para esto, tenemos disponible la información del paciente en el momento de ingreso. Comenzamos el problema con la selección de variables, cuyo resultado se puede ver en el Cuadro 3.13:

1. "comorbilidades": Si presenta co-morbilidades (durante o previo a ingreso).
2. "alt_d0": Valor de la alanina aminotransferasa (ALT) en el momento del ingreso.
3. "LDH_d2": Valor de la LDH en el segundo día de ingreso.
4. "calcio_d0": Valor del calcio en el momento del ingreso.
5. "procalcitonina_d2": Valor de la procalcitonina en el día 2 del ingreso.

6. "Severity_WHO": Escala de severidad en COVID-19 según la WHO.
7. "LDH_d0": Valor de la LDH en el momento del ingreso.
8. "gas_arterial_fio2_d0": Valor de la Fio2 en la gasometría arterial en el momento del ingreso.
9. "Protei._C_Reactiva_d0": Valor de la proteína C Reactiva en el momento del ingreso.

De nuevo, vemos ciertas variables que han sido seleccionadas en modelos con anteriores: LDH, porcalcitonina, comorbilidades, calcio, gas arterial (en este caso fio2). Además, en este problema aparecen nuevas variables como la serveridad de acuerdo con la Organización Mundial de la Salud y la proteína C-reactiva. Quizá estas variables sean más relevante para este modelo, por ser variables relacionadas la infección y este modelo busca predecir el DLCO en un tiempo más próximo al presente.

Variable	**DLCO<80 % N= 430 (70,96 %)**	**DLCO>80 % N= 176 (29,04 %)**	***p*-valor**	**Valores**
comorbilidades	0.9209 ± 0.2702	0.7657 ± 0.4248	$1.173 \cdot 10^{-5}$	[1 0]
alt_d0	0.1535 ± 0.3859	0.36 ± 0.4814	$8.04 \cdot 10^{-7}$	[-1 0 1]
LDH_d2	0.1256 ± 0.9377	0.3429 ± 0.7784	0.003616	[-1 0 1]
calcio_d0	-0.9395 ± 0.2482	-0.84 ± 0.3677	0.001154	[-1 0 1]
procalcitonina_d2	0.7 ± 0.4588	0.6057 ± 0.4901	0.02965	[1 0]
Severity_WHO	4.698 ± 1.319	4.017 ± 1.28	$1.013 \cdot 10^{-8}$	[1 2 3 4 5 6 7 8]
LDH_d0	0.4907 ± 0.5143	0.5486 ± 0.4991	0.2008	[1 0 -1]
gas_arterial_fio2_d0	-0.6 ± 0.6814	-0.6057 ± 0.6512	0.9231	[-1 0 1]
Protei._C_Reactiva_d0	1.723 ± 0.5914	1.446 ± 0.785	$3.453 \cdot 10^{-5}$	[0 1 2 3]

Cuadro 3.13: Comparación de las variables seleccionadas para los valores de la clase

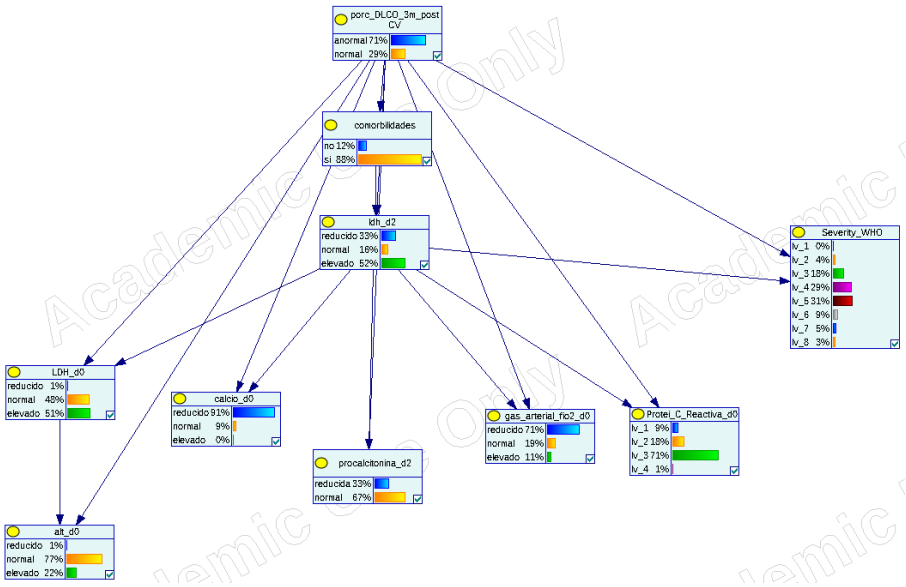

Figura 3.20: Modelo 4 con GeNIe

Con las variables anteriores, el modelo de la Figura 3.20 tiene una **precisión** del

0.7146 y un área bajo la curva ROC de **0.6424**.

Como hemos hecho anteriormente, podemos realizar *queries* al modelo para entenderlo mejor, ver como clasifica a diferentes pacientes.

1. **Q1** (Figura 3.21): Paciente típico con DLCO anormal (valor 0), los cuales componen un 71 % de la población.

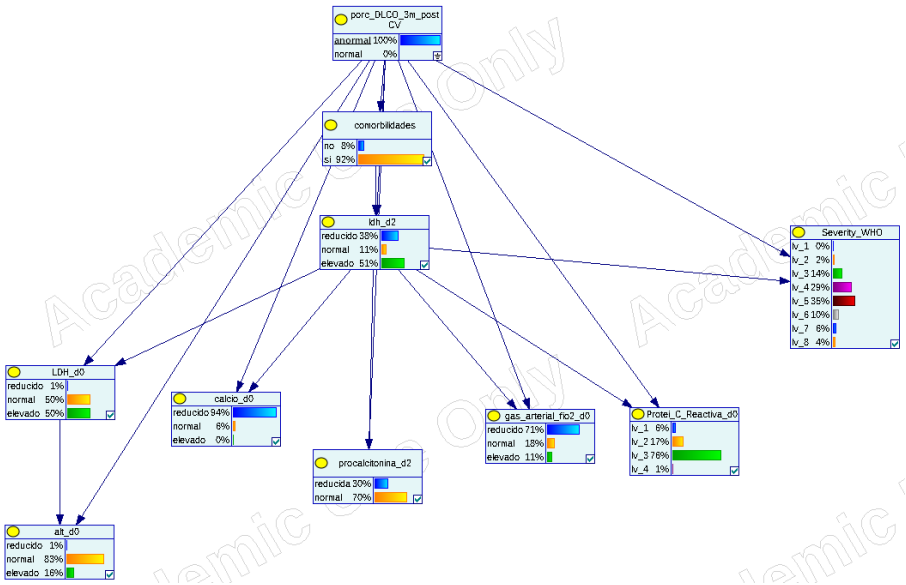

Figura 3.21: *Query 1*, modelo 4 con GeNIe

- Comorbilidades: Sí, con probabilidad 0.92 frente a 0.88 en la muestra poblacional.
- LDH_d2: Elevado, con probabilidad 0.51, frente a 0.52 en la población muestral.
- Calcio_d0: Reducido, con probabilidad 0.94 frente a 0.91 en la población muestral.
- LDH_d0: Elevado, con probabilidad 0.4976 frente a 0.51 en la población muestral.
- alt_d0: Normal con probabilidad 0.88, frente a 0.77 en la población muestral.
- procalcitonina_d2: Normal con probabilidad 0.70 frente a 0.67 en la población muestral.
- gas_arterial_fio2_d0: Reducido con probabilidad 0.71 (igual que la población muestral).
- Proteína C. Reactiva (PCR): Entre 10 y 40 con probabilidad 0.76 frente a 0.71 en la población muestral.
- Severidad WHO: Nivel 5 con probabilidad del 0.35 frente al 0.31 en la población muestral.

2. **Q2** (Figura 3.22): Paciente con DLCO normal (0.29 de la población).

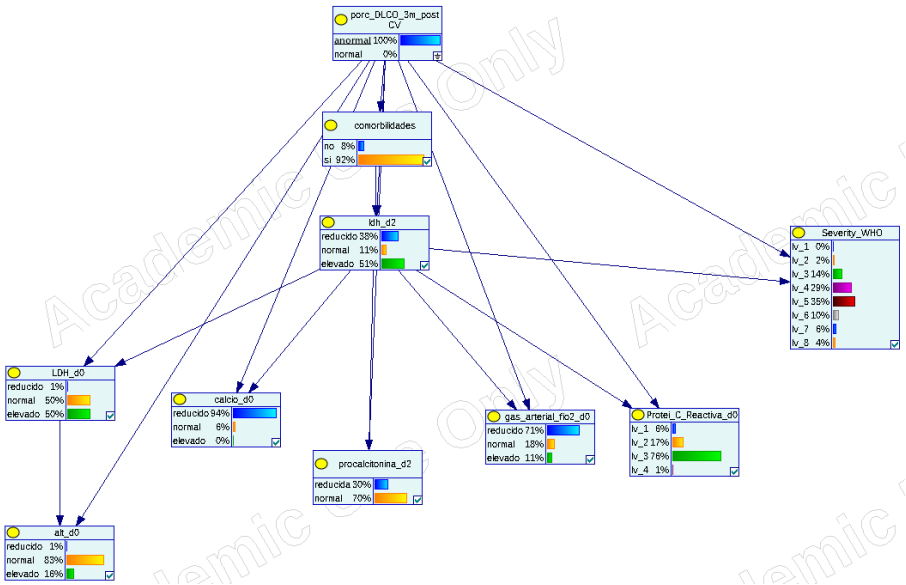

Figura 3.22: *Query 2*, modelo 4 con GeNIe

- Comorbilidades: Sí, con probabilidad 0.77 frente a 0.88 en la muestra poblacional.
- LDH_d2: Elevado, con probabilidad 0.53, frente a 0.52 en la población muestral.
- Calcio_d0: Reducido, con probabilidad 0.84 frente a 0.91 en la población muestral.
- LDH_d0: Elevado, con probabilidad 0.55 frente a 0.51 en la población muestral.
- alt_d0: Normal con probabilidad 0.64, frente a 0.77 en la población muestral.
- procalcitonina_d2: Normal con probabilidad 0.61 frente a 0.67 en la población muestral.
- gas_arterial_fio2_d0: Reducido con probabilidad 0.70 frente a 0.71 en la población muestral.
- Proteína C. Reactiva (PCR): Entre 10 y 40 con probabilidad 0.61 frente al 0.71 % en la población muestral.
- Severidad WHO: Nivel 5 con probabilidad 0.30 frente a 0.31 en la población muestral.

3. **Q3** (Figura 3.23) Consideremos un paciente con nivel de severidad 5, proteína C. reactiva entre 10 y 40, gas arterial fio2 reducido y que presenta comorbilidades. Vemos que el modelo lo clasifica como paciente con DLCO anormal, con probabilidad 0.88 (frente a 0.71 de la población en dicho grupo).

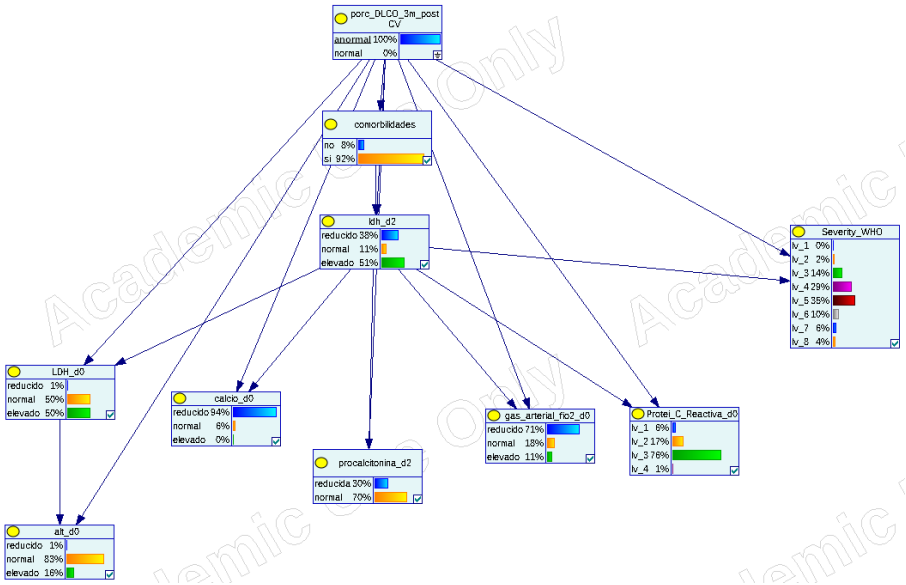

Figura 3.23: *Query 3*, modelo 4 con GeNIe

4. **Q4** (Figura 3.24): Consideramos un paciente con nivel de severidad 3, proteína C reactiva baja (menos de 3), sin comorbilidades, calcio y LDH normales. Para dicho paciente, el modelo predice nivel de DLCO normal con probabilidad 0.96 (frente a 0.29 en la muestra poblacional).

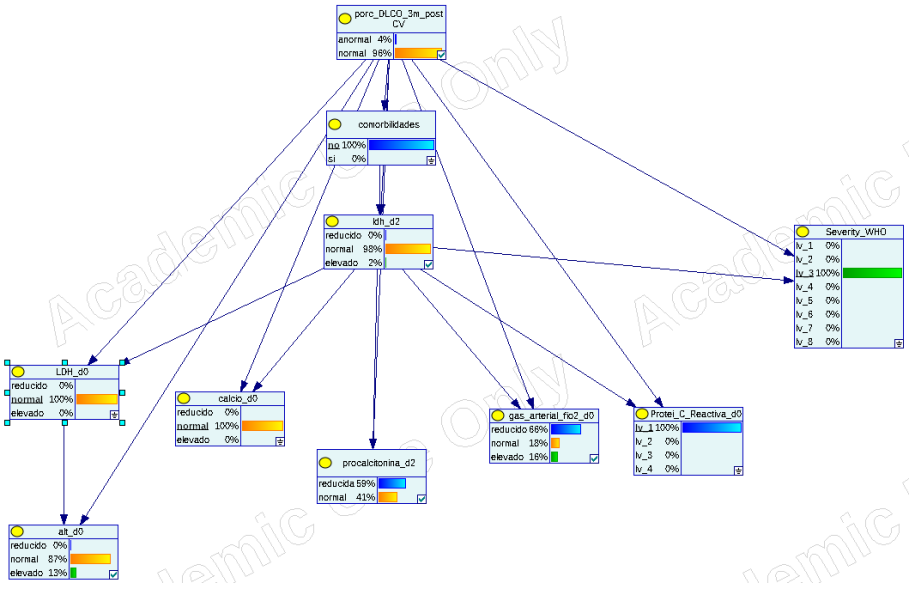

Figura 3.24: *Query 4*, modelo 4 con GeNIe

5. **Q5** (Figura 3.25): Consideramos un paciente con alt_d0 elevado, calcio y LDH (d0 y d2) normales. Para dicho paciente, el modelo predice con probabilidad 0.98 que tendrá DLCO normal a los 3 meses (frente a 0.29 en la población muestral).

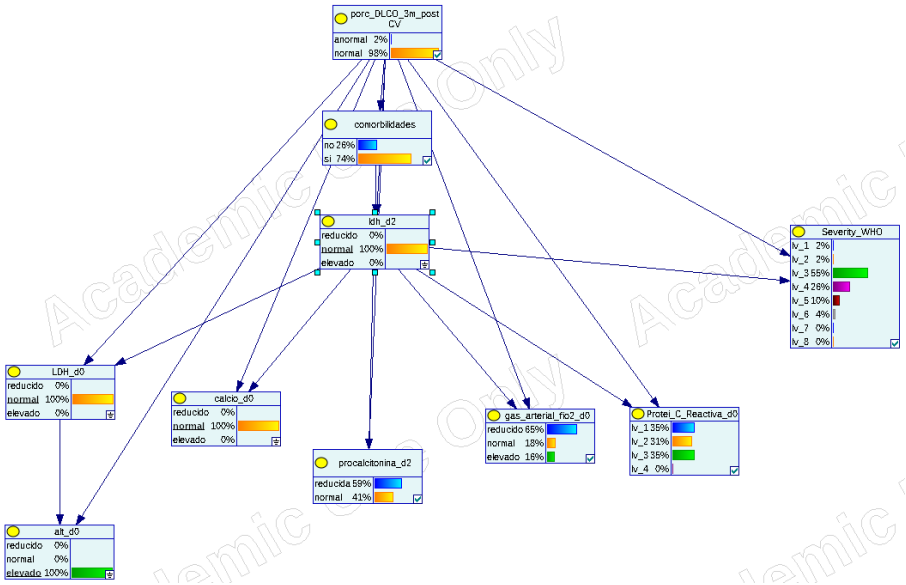

Figura 3.25: *Query 5*, modelo 4 con GeNIe

3.8. Predecir evolución al alta: Regresión logística

Comparamos el modelo anterior con un modelo de regresión logística como referencia. De nuevo, el primer paso es realizar la fase final del proceso de selección de variables para este modelo. Usando el procedimiento habitual, las siguientes variables (véase el Cuadro 3.16) han sido seleccionadas:

1. "`comorbilidades`": Si presenta co-morbilidades (durante o previo a ingreso).
2. "`alt_d0`": Valor de la alanina aminotransferasa (ALT) en el momento del ingreso.
3. "`linfocitos_d0_total`": Valor absoluto de los linfocitos en el momento del ingreso.
4. "`LDH_d2`": Valor de la LDH a los dos días del ingreso.
5. "`hipertension`": Si sufre hipertensión.
6. "`LDH_d0`": Valor de la LDH en el momento del ingreso.
7. "`obesidad_morbida`": Si tiene obesidad mórbida u obesidad G2 (IMC >35).
8. "`frec_respiratoria_ipm_d0`": Frecuencia respiratoria en el momento del ingreso.

Red bayesiana	**Comunes**	**R. logistica**
calcio_d0	porc_DLCO_3m_postCV	linfocitos_d0_total
procalcitonina_d2	comorbidades	hipertension
Severity_WHO	alt_d0	obesidad_morbida
gas_arterial_fio2_d0	LDH_d2	frec_respiratoria_ipm_d0
Protei._C_Reactiva_d0	LDH_d0	

Cuadro 3.14: Comparativa de las variables seleccionadas en ambos modelos

De nuevo, vemos (Cuadro 3.14) que hay variables comunes en los dos modelos, y algunas de ellas se repiten en otros modelos anteriores (LDH, comorbilidades, alt_d0). Aparecen algunas variables nuevas como frec_respiratoria. El modelo resultante tiene una **precisión** de **0.7497** y **AUC** de **0.62**.

	Red bayesiana	R. logística	I.C. 95 %	p-valor
Precisión	0.7147	0.7497	$(-0.0788, -0.0088)$	0.114
AUC	0.642	0.62	$(-0.024, 0.0688)$	0.33

Cuadro 3.15: Comparativa de los dos modelos para el problema 4
Intervalo de confianza y p-valor obtenidos con el test de Student comparando el conjunto de los resultados de precisión y AUC para cada uno de los k-pliegues de la validación con los del conjunto obtenido por el otro modelo.

Vemos que los resultados de la comparación no son concluyentes. Parece que el modelo de regresión logística tiene mayor precisión que la red bayesiana, y la red bayesiana mayor AUC. Sin embargo, las diferencias no son estadísticamente significativas. En el Cuadro 3.17 se muestra más información del modelo logísticol.

Variable	**DLCO<80 % N= 430 (70.96 %)**	**DLCO>80 % N= 176 (29.04 %)**	***p*-valor**	**Valores**
comorbidades	0.9209 ± 0.2702	0.7657 ± 0.4248	$1.173 \cdot 10^{-5}$	[1 0]
alt_d0	0.1535 ± 0.3859	0.36 ± 0.4814	$8.04 \cdot 10^{-7}$	[0 1 -1]
linfocitos_d0_total	-0.3767 ± 0.4947	-0.5029 ± 0.5014	0.005167	[0 -1 1]
LDH_d2	0.1256 ± 0.9377	0.3429 ± 0.7784	0.003616	[1 -1 0]
hipertension	0.5372 ± 0.4992	0.3143 ± 0.4656	$2.976 \cdot 10^{-7}$	[0 1]
LDH_d0	0.4907 ± 0.5143	0.5486 ± 0.4991	0.2008	[1 0 -1]
obesidad_morbida	0.4395 ± 0.5016	0.2571 ± 0.4383	$1.161 \cdot 10^{-5}$	[0 1 -1]
frec_respiratoria_ipm_d0	0.9744 ± 0.1581	0.9486 ± 0.2461	0.1998	[1 0 -1]

Cuadro 3.16: Comparación de las variables seleccionadas para los valores de la clase

3.9. Predecir DLCO-3 alta: BAN

Nos planteamos un nuevo problema (Problema 5), similar al anterior. En este caso, con la información disponible a los 3 meses (junto con las variables de ingreso) tratamos de predecir el valor del DLCO a los 6 meses (si es normal o está por debajo del 80 %). El primer modelo es el modelo de red bayesiana, para el cual realizamos selección de variables siguiendo el procedimiento idéntico a los modelos anteriores. Las siguientes variables han sido seleccionadas (véase el Cuadtro 3.18).

1. "`ast_d0`": Valor de la aspartato aminotransferasa (AST) en el momento del ingreso.
2. "`creatin_kinasa_d0`" : Valor de la creatin kinasa en el momento del ingreso.
3. "`porc_DLCO_3m_postCV`": Si el porcentaje DLCO a los 3 meses de ingreso es mayor que 80 %.
4. "`glicemia_d0`": Valor de la glicemia en el momento del ingreso.
5. "`EPOC_Enfisema`": Si la EPOC que tiene es de tipo enfisematoso.

Variables	OR	CI 95 %	p-valor	β
comorbidades: 1 vs 0	0.31	$(0.18, 0.51)$	0.008	$4.557 \cdot 10^{-1}$
alt_d0: valor 0 (ref -1)	682666.45	$(0, \infty)$	0.989	$8.39 \cdot 10^{6}$
alt_d0: valor 1 (ref -1)	1903292.17	$(0, \infty)$	0.988	$2.235 \cdot 10^{7}$
linfocitos_d0_total: valor 0 (ref -1)	0.66	$(0.46, 0.95)$	0.79	$9.37 \cdot 10^{-1}$
linfocitos_d0_total: valor 1 (ref -1)	0	$(0, \infty)$	0.993	$1.783 \cdot 10^{-7}$
LDH_d2: valor 0 (ref -1)	5.04	$(2.87, 8.86)$	< 0.001	6.75
LDH_d2: valor 1 (ref -1)	1.96	$(1.25, 3.08)$	0.078	2.627
hipertension: 1 vs 0	0.41	$(0.28, 0.59)$	0.002	$5.1065 \cdot 10^{-1}$
LDH_d0: valor 0 (ref -1)	832142.16	$(0, \infty)$	0.989	$1.43 \cdot 10^{7}$
LDH_d0: valor 1 (ref -1)	945983.32	$(0, \infty)$	0.989	$6.723 \cdot 10^{6}$
obesidad_morbida: valor 0 (ref -1)	425346.01	$(0, \infty)$	0.995	$8.93 \cdot 10^{6}$
obesidad_morbida: valor 1 (ref -1)	195921.66	$(0, \infty)$	0.995	$1.29 \cdot 10^{7}$
frec_respiratoria_ipm_d0: valor 0 (ref -1)	0	$(0, \infty)$	0.994	$4.83 \cdot 10^{-8}$
frec_respiratoria_ipm_d0: valor 1 (ref -1)	0	$(0, \infty)$	0.994	$3.15 \cdot 10^{-8}$

Cuadro 3.17: Regresión logística modelo 4: *odds ratio* (OR), intervalos de confianza (CI), p-valores y parámetros (β)

Variable	DLCO<80 % N= 453 (74.75 %)	DLCO>80 % N= 153 (25.25 %)	*p*-valor	Valores
ast_d0	0.4415 ± 0.5273	0.4079 ± 0.4931	0.4756	[0 1 -1]
creatin_kinasa_d0	0.5077 ± 0.5093	0.5987 ± 0.518	0.06109	[0 1 -1]
porc_DLCO_3m_postCV	0.234 ± 0.4238	0.4539 ± 0.4995	$2.065 \cdot 10^{-6}$	[1 0]
glicemia_d0	0.1788 ± 0.4005	0.07237 ± 0.2843	0.0003956	[0 -1 1]
EPOC_Enfisema	0.574 ± 0.8477	0.7434 ± 0.9594	0.05375	[2 0 1]
pcr_al_alta	0.0287 ± 0.1671	0.03289 ± 0.179	0.7995	[1 0]
gas_arterial_po2_d0	-0.0596 ± 0.4295	0.06579 ± 0.3389	0.0002758	[0 -1 1]
sexo	0.3289 ± 0.4703	0.4145 ± 0.4943	0.0628	[1 0]
calcio_d0	-0.9249 ± 0.272	-0.8684 ± 0.3391	0.06373	[-1 0 1]
LDH_al_alta	0.1192 ± 0.3311	0.08553 ± 0.2806	0.2228	[0 1 -1]

Cuadro 3.18: Comparación de las variables seleccionadas para los valores de la clase

6. "`pcr_al_alta`": Valor de la proteína C reactiva al alta.
7. "`gas_arterial_po2_d0`" : Valor de la presión de oxígeno (po2) en la gasometría arterial en el momento del ingreso.
8. "`sexo`": Sexo del paciente.
9. "`calcio_d0`": Valor del calcio en el momento del ingreso.
10. "`LDH_al_alta`": Valor de la LDH en el alta.

Vemos que algunas de las variables seleccionadas han aparecido en modelos previos como LDH, pcr o calcio; aunque vemos también un número elevado de variables que no han aparecido en modelos previos, como sexo, EPOC_Enfisema o creatin_kinasa. Además, observamos que en este modelo tenemos acceso a información de alta y medidas a los tres meses del alta.

El modelo BAN generado con estas variables es el de la Figura 3.26. Tiene una **precisión** de 0.7589 y un **AUC** de **0.7594**.

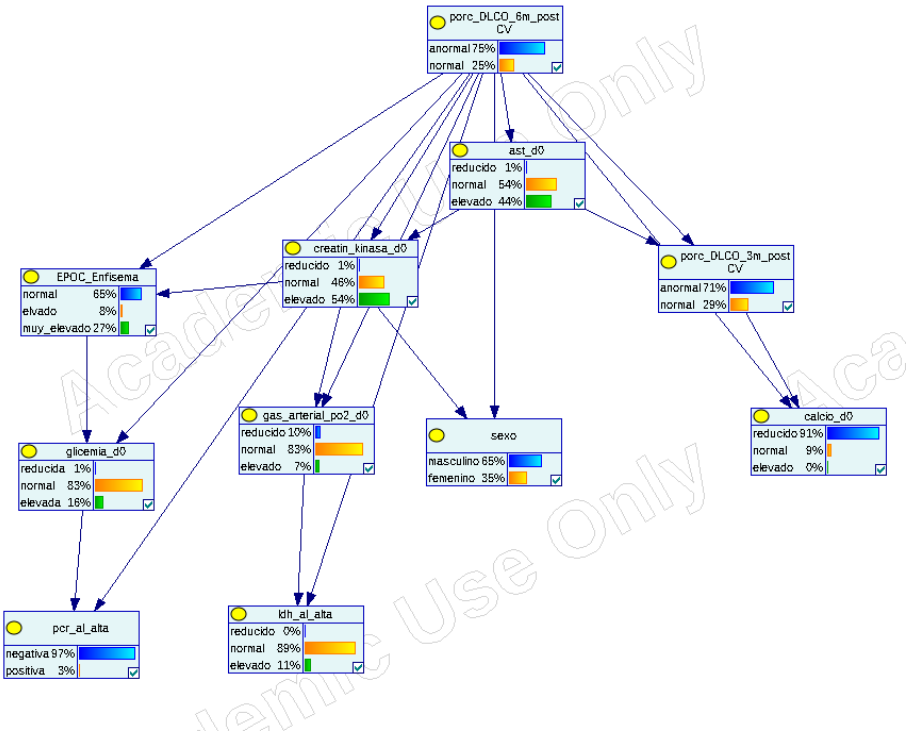

Figura 3.26: Modelo 5 con GeNIe

Vamos a analizar el modelo, haciendo consultas, de manera similar a los problemas anteriores.

1. **Q1** (Figura 3.27): Paciente prototípico con DLCO anormal en mes 6. El 75 % de la muestra poblacional pertenecen a esta categoría.
   - ast_d0: Normal, con probabilidad 0.53 frente a 0.54 en la muestra poblacional.
   - creatin_kinasa_d0: Elevada, con probabilidad 0.51, frente a 0.54 en la población muestral.
   - Calcio_d0: Reducido, con probabilidad 0.93 frente a 0.91 en la población muestral.
   - LDH_al_alta: Normal, con probabilidad 0.88 frente a 0.89 en la población muestral.
   - gas_arterial_po2_d0: Normal con probabilidad 0.81, frente a 0.83 en la población muestral.
   - porc_DLCO_3m_postCV: Anormal con probabilidad 0.77 frente a 0.71 en la población muestral.
   - glicemia_d0: Reducido con probabilidad 0.81 frente a 0.83 en la población muestral.

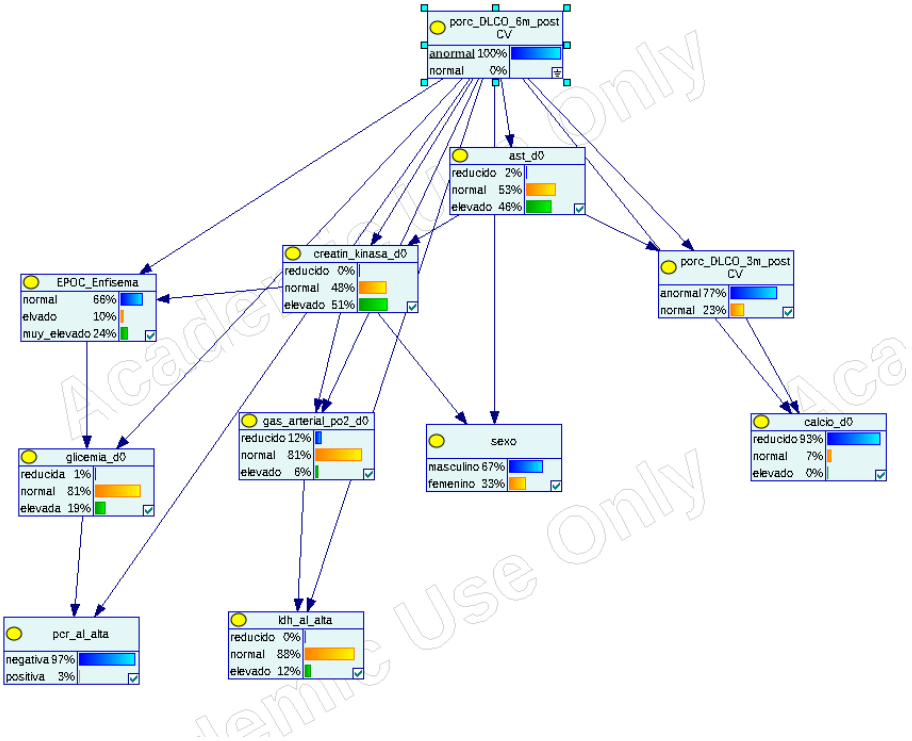

Figura 3.27: *Query 1*, Modelo 5 con GeNIe

- pcr_al_alta: Negativa con probabilidad 0.97 (igual que en la población muestral).
- EPOC_Enfisema: Normal con probabilidad 0.66 frente a 0.65 en la población muestral.
- sexo: Masculino, con probabilidad 0.67, frente a 0.65 en la población muestral.

2. **Q2** (Figura 3.28): Paciente prototípico con DLCO normal en mes 6. El 25 % de la muestra poblacional pertenecen a esta categoría.
   - ast_d0: Normal, con probabilidad 0.59 frente a 0.54 en la muestra poblacional.
   - creatin_kinasa_d0: Elevada, con probabilidad 0.61, frente a 0.54 en la población muestral.
   - Calcio_d0: Reducido, con probabilidad 0.87 frente a 0.91 en la población muestral.
   - LDH_al_alta: Normal, con probabilidad 0.91 frente a 0.89 en la población muestral.
   - gas_arterial_po2_d0: Normal con probabilidad 0.88, frente a 0.83 en la población muestral.
   - porc_DLCO_3m_postCV: Anormal con probabilidad 0.55 frente a 0.71 en la población muestral.

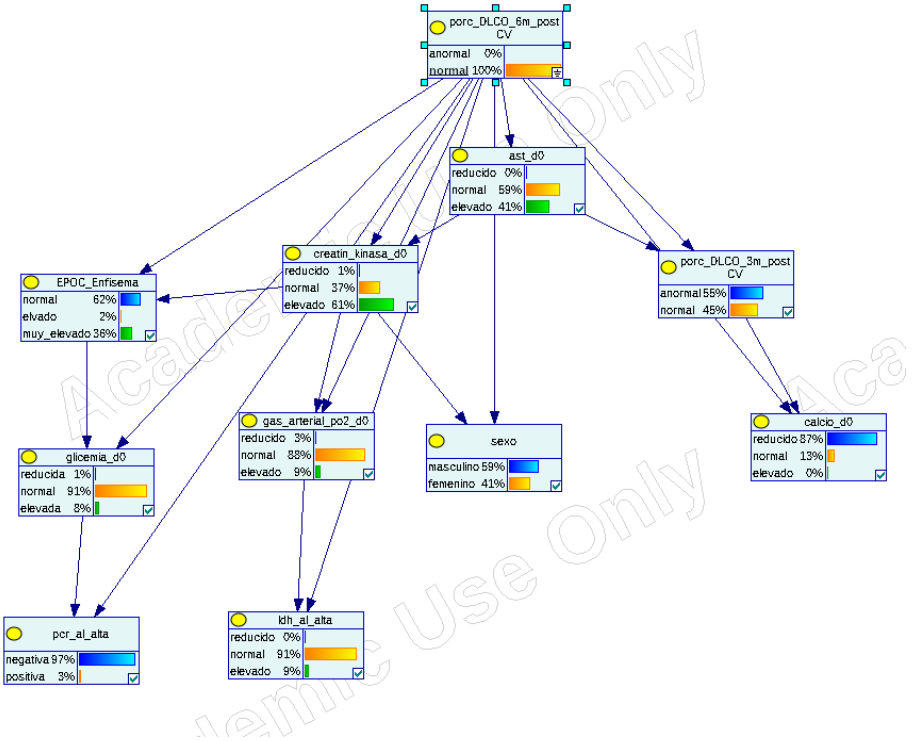

Figura 3.28: *Query 2*, Modelo 5 con GeNIe

- glicemia_d0: Reducido con probabilidad 0.91 frente a 0.83 en la población muestral.
- pcr_al_alta: Negativa con probabilidad 0.97 (igual que en la población muestral).
- EPOC_Enfisema: Normal con probabilidad 0.66 frente a 0.65 en la población muestral.
- Sexo: Masculino, con probabilidad 0.62, frente a 0.65 en la población muestral.

3. **Q3** (Figura 3.29): Consideramos un paciente con calcio normal, ast_d0 normal, creatin_kinasa_d0 elevada, sexo femenino y LDH_al_alta normal. El modelo mostrado en la Figura 3.29 predice que dicho paciente tendrá DLCO a los 6 meses normal con probabilidad 0.59 (frente a 0.25 de la muestra poblacional).

4. **Q4** (Figura 3.30) Consideramos un paciente con DLCO anormal a los 3 meses. El modelo predice que dicho paciente tendrá DLCO normal a los 6 meses con probabilidad 0.81 (frente a 0.75 en la muestra poblacional).

5. **Q4** (Figura 3.31) Consideramos un paciente con sexo masculino, gas_arterial_po2_d0 reducido y EPOC_Enfisema elevado. El modelo predice que dicho paciente tendrá DLCO normal a los 6 meses con probabilidad 0.99 (frente a 0.75 en la muestra poblacional).

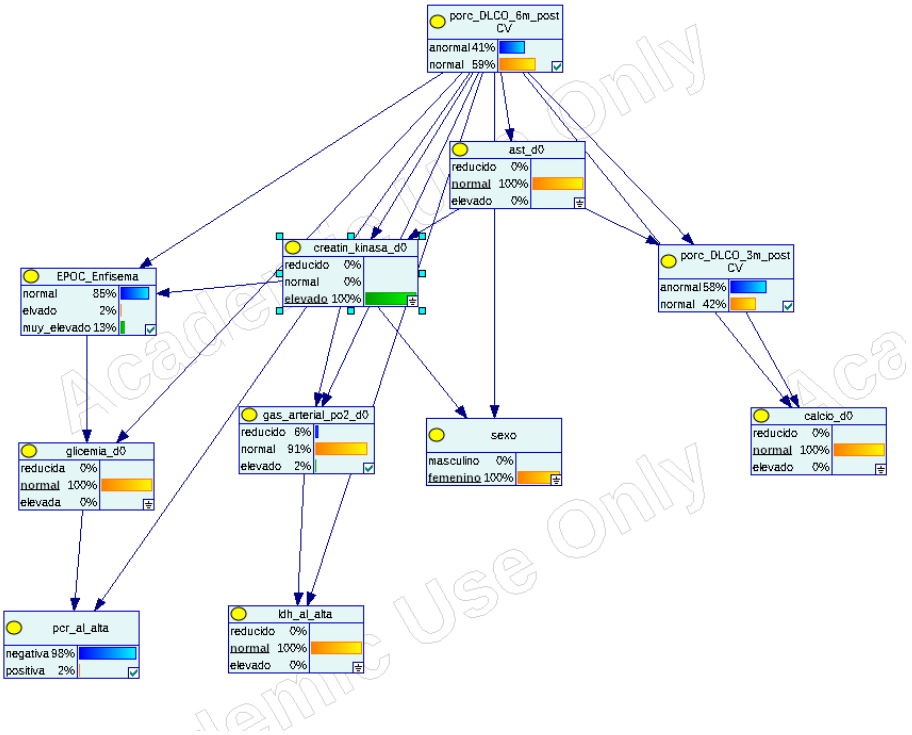

Figura 3.29: *Query 3*, Modelo 5 con GeNIe

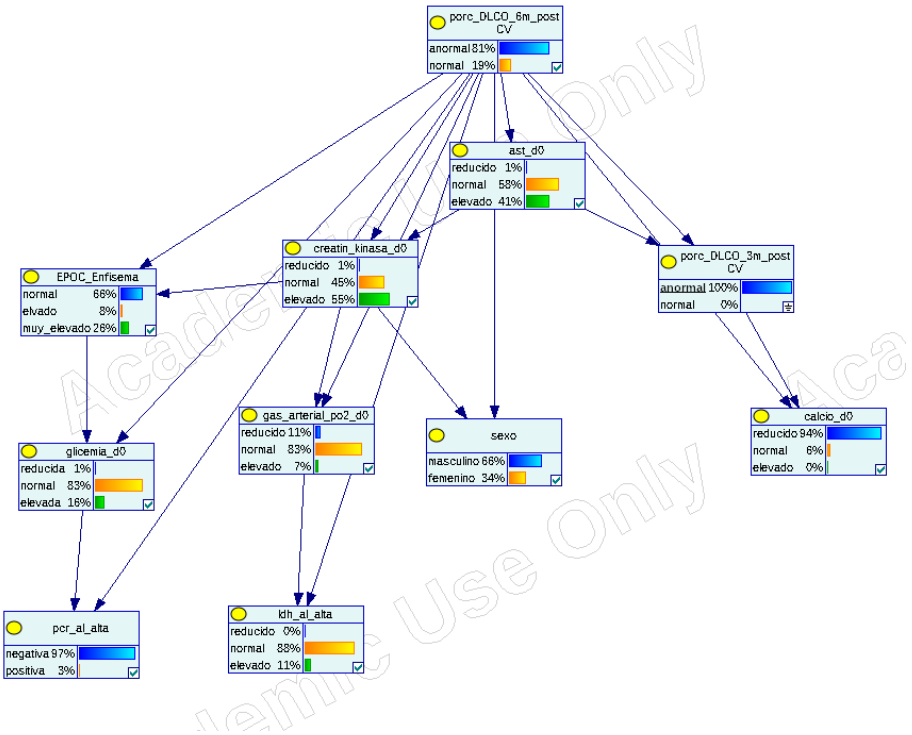

Figura 3.30: *Query 4*, Modelo 5 con GeNIe

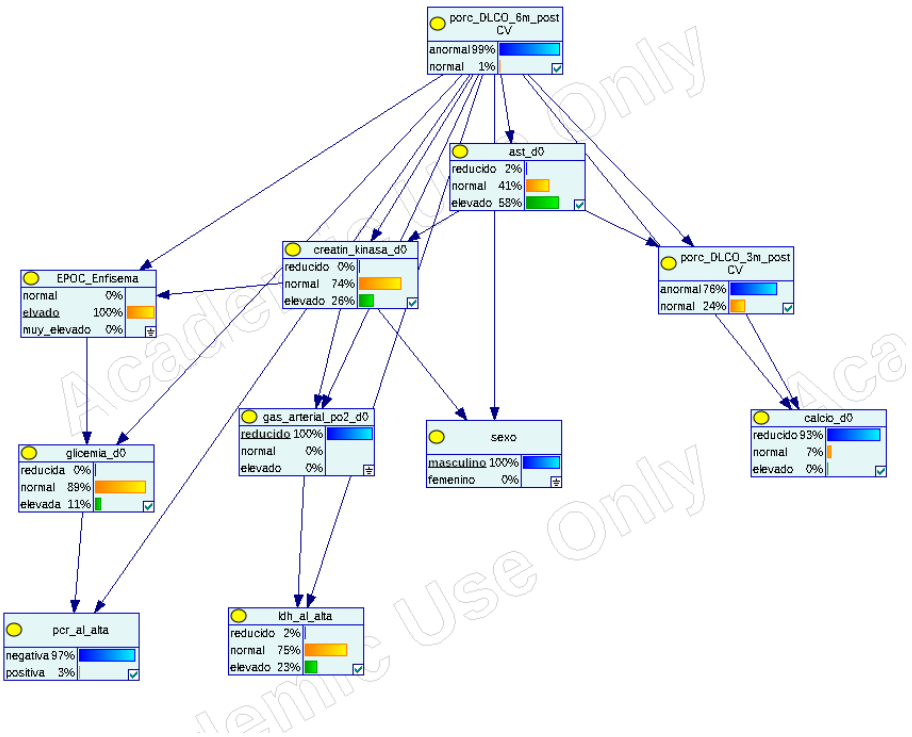

Figura 3.31: *Query 5*, Modelo 5 con GeNIe

3.10. Predecir DLCO-6 a los 3 meses: Regresión logística

Vemos a continuación, usando regresión logística. En este caso, las siguiente variables han sido seleccionadas (Cuadro 3.19):

1. "`consumo_de_alcohol`": Si el paciente consume alcohol.
2. "`creatin_kinasa_d0`": Valor de la creatin kinasa en el momento del ingreso.
3. "`leucocitos_total_d0`": Valor absoluto de los leucocitos totales en el momento del ingreso.
4. "`frec_respiratoria_ipm_d0`": Frecuencia respiratoria en el momento del ingreso.
5. "`LDH_al_alta`": Valor de la LDH en la alta.
6. "`linfocitos_porcentaje_d0`": Porcentaje de linfocitos en el momento del ingreso.
7. "`linfocitos_d0_total`": Valor absoluto de los linfocitos totales en el momento del ingreso
8. "`calcio_d0`": Valor del calcio en el momento del ingreso.

Podemos comparar las variables seleccionadas con las de la red bayesiana en el cuadro 3.20.

Vemos que variables como calcio o LDH aparecen comunes, y son variables que ya han aparecido en otros modelos. Otras variables seleccionadas incluyen leucocitos y

Variable	DLCO<80 % N= 453 (74.75 %)	DLCO>80 % N= 153 (25.25 %)	*p*-valor	Valores
consumo_de_alcohol	0.03311 ± 0.1791	0.006579 ± 0.08111	0.01329	[0 1]
creatin_kinasa_d0	0.5077 ± 0.5093	0.5987 ± 0.518	0.06109	[-1 0 1]
leucocitos_total_d0	-0.006623 ± 0.4727	-0.1053 ± 0.4013	0.01283	[0 -1 1]
frec_respiratoria_ipm_d0	0.9691 ± 0.1856	0.9605 ± 0.1954	0.6361	[-1 0 1]
LDH_al_alta	0.1192 ± 0.3311	0.08553 ± 0.2806	0.2228	[-1 0 1]
linfocitos_porcentaje_d0	-0.5806 ± 0.5073	-0.3816 ± 0.5138	$4.623 \cdot 10^{-5}$	[-1 0 1]
linfocitos_d0_total	-0.4238 ± 0.5036	-0.3816 ± 0.4874	0.3598	[-1 0 1]
calcio_d0	-0.9249 ± 0.272	-0.8684 ± 0.3391	0.06373	[-1 0 1]

Cuadro 3.19: Comparación de las variables seleccionadas para los valores de la clase

Red Bayesiana	Comunes	R. Logistica
ast_d0	porc_DLCO_6m_postCV	consumo_de_alcohol
porc_DLCO_3m_postCV	creatin_kinasa_d0	leucocitos_total_d0
glicemia_d0	calcio_d0	frec_respiratoria_ipm_d0
EPOC_Enfisema	LDH_al_alta	linfocitos_porcentaje_d0
pcr_al_alta		linfocitos_d0_total
gas_arterial_po2_d0		
sexo		

Cuadro 3.20: Comparativa de las variables seleccionadas en ambos modelos

linfocitos, también seleccionadas para otros modelos. La creatin_kinasa ha aparecido en los dos modelos de este problema 5, aunque no aparece en ningún otro modelo.

El modelo de regresión logística que hemos obtenido tiene una **precisión** de **0.7405** y una **AUC** de **0.5026**. El cuadro 3.21 compara los resultados obtenidos con los del modelo anterior.

	Red Bayesiana	R. Logística	I.C. 95 %	p-valor
Precisión	0.7484	0.74	$(-0.0439, 0.05987)$	0.7577
AUC	0.596	0.5026	$(0.056, 0.131)$	$1.673 \cdot 10^{-5}$

Cuadro 3.21: Comparativa de los dos modelos para el problema 5
Intervalo de confianza y p-valor obtenidos con el test de Student comparando el conjunto de los resultados de precisión y AUC para cada uno de los k-pliegues de la validación con los del conjunto obtenido por el otro modelo.

A continuación mostramos los parámetros de este modelo de regresión logística en el cuadro 3.22

3.11. Predecir DLCO-12 a los 6 meses: BAN

Predecir con las variables de ingreso, de los 3 meses y de los 6 meses el valor del DLCO a los 12 meses (Problema 6). Este modelo cuenta con mucha información comparado con los otros para predecir a corto plazo. Cabe esperar que los resultados sean relativamente buenos en comparación con el resto de problemas. Al igual que

Variables	**OR**	**CI 95 %**	**p-valor**	β
consumo_de_alcohol: 1 vs 0	0.2	$(0.03, 1.52)$	0.087	$1.57 \cdot 10^{-1}$
creatin_kinasa_d0 valor 0 (ref -1)	0.25	$(0.03, 1.78)$	0.199	$2.227 \cdot 10^{-1}$
creatin_kinasa_d0: valor 1 (ref -1)	0.4	$(0.06, 2.86)$	0.256	$2.664 \cdot 10^{-1}$
leucocitos_total_d0: valor 0 (ref -1)	0.91	$(0.51, 1.6)$	0.524	1.25
leucocitos_total_d0: valor 1 (ref -1)	0.21	$(0.07, 0.68)$	0.126	$3.787 \cdot 10^{-1}$
frec_respiratoria_ipm_d0: valor 0 (ref -1)	389616.93	$(0, \infty)$	0.991	$7.51 \cdot 10^{6}$
frec_respiratoria_ipm_d0: valor 1 (ref -1)	250468.02	$(0, \infty)$	0.992	$3.643 \cdot 10^{6}$
LDH_al_alta: valor 0 (ref -1)	262478.78	$(0, \infty)$	0.992	$1.771 \cdot 10^{6}$
LDH_al_alta: valor 1 (ref -1)	194808.47	$(0, \infty)$	0.992	$1.71 \cdot 10^{6}$
linfocitos_porcentaje_d0: valor 0 (ref -1)	2.25	$(1.52, 3.32)$	0.004	2.61
linfocitos_porcentaje_d0: valor 1 (ref -1)	3.1	$(0.51, 19.01)$	0.251	3.73
linfocitos_d0_total: valor 0 (ref -1)	1.24	$(0.84, 1.83)$	0.082	$5.621 \cdot 10^{-1}$
linfocitos_d0_total: valor 1 (ref -1)	0	$(0, \infty)$	0.985	$9.066 \cdot 10^{-8}$
calcio_d0: valor 0 (ref -1)	2.07	$(1.14, 3.74)$	0.005	2.572
calcio_d0: valor 1 (ref -1)	0	$(0, \infty)$	0.992	$5.647 \cdot 10^{-7}$

Cuadro 3.22: Regresión logística del modelo 5: *odds ratio* (OR), intervalos de confianza (CI), p-valor y parámetros (β)

con los otros modelos, comenzamos por el proceso de selección de variables. En este caso las variables seleccionadas son:

1. "`SPA_6m`": Niveles de SPA (ng/ml) por ELISA en plasma a los 6 meses.
2. "`calcio_d0`": Valor del calcio en el momento del ingreso.
3. "`pcr_al_alta`": Valor de la proteína C reactiva al alta.
4. "`pcr_d2`": Valor de la proteína C reactiva segundo día del ingreso.
5. "`fosfat_alcali._d0`": Valor de la fosfatasa alcalina en el momento del ingreso.
6. "`SPD_6m`": Niveles de SPD (ng/ml) por ELISA en plasma a los 6 meses.
7. "`porc_DLCO_3m_postCV`": Si el porcentaje de DLCO a los 3 meses del alta es superior al 80 %.
8. "`comorbilidades`": Si presenta co-morbilidades (durante o previo a ingreso).
9. "`porc_DLCO_6m_postCV`": Si el porcentaje de DLCO a los 6 meses del alta es superior al 80 %.
10. "`LDH_d2`": Valor de la LDH en el día 2 del ingreso.
11. "`dias_ingreso_UCI`": Días de ingreso en UCI.

Notamos que se han seleccionado el DLCO a 3 y 6 meses y días en la UCI (algo que parece lógico). También aparecen otras variables que ya hemos visto en otros problemas como pcr, calcio, LDH, fostata alcalina. Finalemente, aparecen SPA y SPD como marcadores importantes a los 6 meses.

Con estas variables obtenemos el modelo de la Figura 3.32.

Variable	DLCO<80 % N= 400 (66.01 %)	DLCO>80 % N= 206 (33.99 %)	*p*-valor	Valores
SPA_6m	3.145 ± 0.8978	3.61 ± 1.487	$5.215 \cdot 10^{-5}$	[0 1 2 3 4 5]
porc_DLCO_3m_postCV	0.2625 ± 0.4405	0.3415 ± 0.4754	0.04821	[1 0]
ferritina_d0	0.6425 ± 0.52	0.4146 ± 0.5133	$4.126 \cdot 10^{-7}$	[1 0 -1]
comorbidades	0.875 ± 0.3311	0.878 ± 0.328	0.9142	[1 0]
hemoglobina_d0	-0.0975 ± 0.3053	-0.03902 ± 0.1941	0.004337	[0 -1 1]
sodio_d0	-0.0725 ± 0.3041	-0.03902 ± 0.2179	0.1203	[0 -1 1]
tvptep	0.1125 ± 0.3164	0.3805 ± 0.4867	$6.963 \cdot 10^{-12}$	[0 1]
dimero_d_d0	0.51 ± 0.5619	0.2829 ± 0.483	$3.427 \cdot 10^{-7}$	[0 1 -1]
dolor_pleuritico	0.2775 ± 0.4483	0.5512 ± 0.4986	$1.328 \cdot 10^{-10}$	[1 0]
plaquetas_d0	-0.3925 ± 0.5236	-0.561 ± 0.5535	0.0003475	[0 -1 1]

Cuadro 3.23: Comparación de las variables seleccionadas para los valores de la variable clase

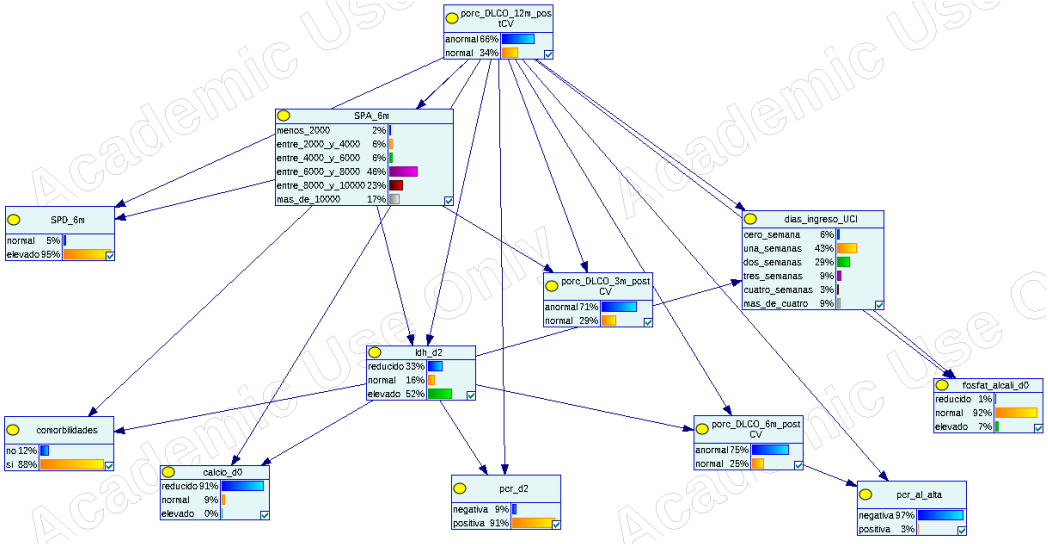

Figura 3.32: Modelo 6 con GeNIe

Evaluamos este modelo, y obtenemos que la **precisión** es **0.7595** y el **AUC** es **0.7476**. Vamos a realizar consultas al modelo para entender el problema en detalle.

1. **Q1** (Figura 3.33): Consideramos el paciente prototípico con DLCO anormal (66 % de la muestra poblacional).
   - SPA_6m: Entre 6000ng/ml y 8000ng/ml, con probabilidad 0.58 frente a 0.46 en la muestra poblacional.
   - SPD_6m: Elevado, con probabilidad 0.95, igual que la muestra poblacional.
   - LDH_d2: Elevado, con probabilidad 0.60 frente a 0.52 de la muestra poblacional.
   - porc_DLCO_3m_postCV: Enormal, con probabilidad 0.74 frente a 0.71 de la muestra poblacional.
   - dias_ingreso_UCI: Dos semanas, con probabilidad 0.34 frente a 0.29 de la

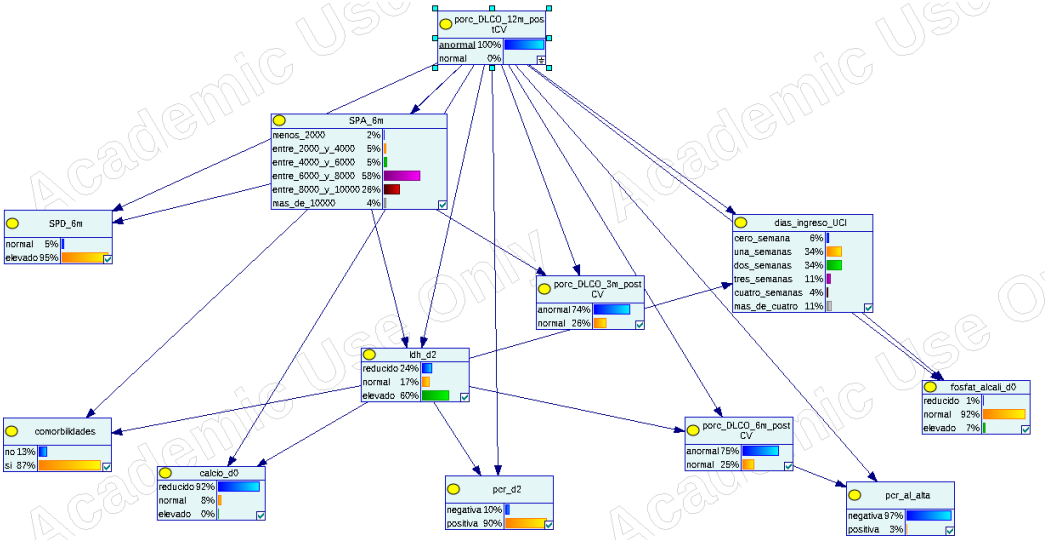

Figura 3.33: *Query 1*, Modelo 6 con GeNIe

muestra poblacional.

- porc_DLCO_6m_postCV: Anormal, con probabilidad 0.75, igual que la muestra poblacional.
- comorbilidades: Sí, con probabilidad de 0.87 frente al 0.88 de la muestra poblacional.
- calcio_d0: Reducido, con probabilidad de 0.92 frente al 0.91 la muestra poblacional.
- pcr_d2: Positiva, con probabilidad de 0.90 frente al 0.91 la muestra poblacional.
- pcr_al_alta: Negativa, con probabilidad 0.97, igual que la muestra poblacional.
- fosfat_alcali_d0: Normal, con probabilidad 0.92, igual que la muestra poblacional.

2. **Q2** (Figura 3.34): Consideramos el paciente prototípico con DLCO normal (34 % de la muestra poblacional)
   - SPA_6m: Más de 10000ng/ml, con probabilidad 0.41 frente a 0.17 en la muestra poblacional.
   - SPD_6m: Elevado, con probabilidad 0.95, igual que la muestra poblacional.
   - LDH_d2: Reducido, con probabilidad 0.51 frente a 0.35 la muestra poblacional.
   - porc_DLCO_3m_postCV: Anormal, con probabilidad 0.74 frente al 0.66 la muestra poblacional.
   - dias_ingreso_UCI: Una semana, con probabilidad 0.61 frente al 0.43 la muestra poblacional.

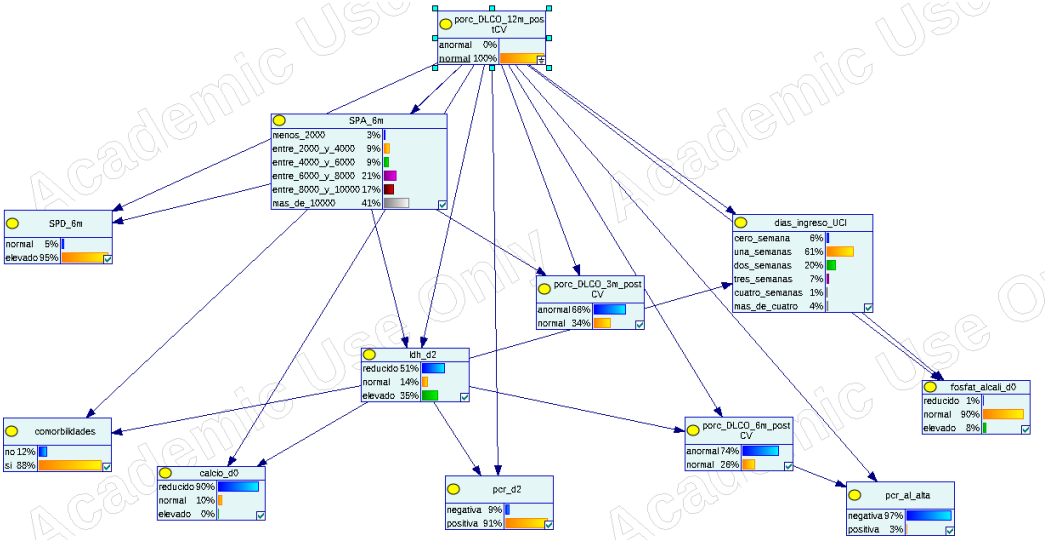

Figura 3.34: *Query 2*, Modelo 6 con GeNIe

- porc_DLCO_6m_postCV: Anormal, con probabilidad 0.75, igual que la muestra poblacional.
- comorbididades: Si, con probabilidad 0.88, igual que la muestra poblacional.
- calcio_d0: Reducido, con probabilidad 0.91, igual que la muestra poblacional.
- pcr_d2: Positiva, con probabilidad 0.91, igual que la muestra poblacional.
- pcr_al_alta: Negativa, con probabilidad 0.97, igual que la muestra poblacional.
- fosfat_alcali_d0: Normal, con probabilidad 0.90 frente a 0.92 la muestra poblacional.

3. **Q3** (Figura 3.35): Consideramos un paciente hipotético con LDH_d2 elevado, porc_DLCO_6m_postCV anormal, SPA_6m entre 6000ng/ml y 8000ng/ml y dos semanas de ingreso en la UCI. Para esa consulta, el modelo predice que el DLCO será anormal con probabilidad de 0.92. En la Figura vemos las variables seleccionadas

4. **Q4** (Figura 3.36): Consideramos un paciente con pcr_al_alta negativa, porc_DLCO_6m_postCV anormal,
SPA_6m entre 6000ng/ml y 8000ng/ml. El modelo predice DLCO anormal con probabilidad 0.92.

5. **Q5** (Figura 3.37): Consideramos un paciente con SPA_6m de más de 10000ng/ml, pcr_d2 positiva, SPD_6m elevado y con comorbilidades. Para dicho paciente el modelo predice que el DLCO a los 12 meses será anormal con probabilidad 1.

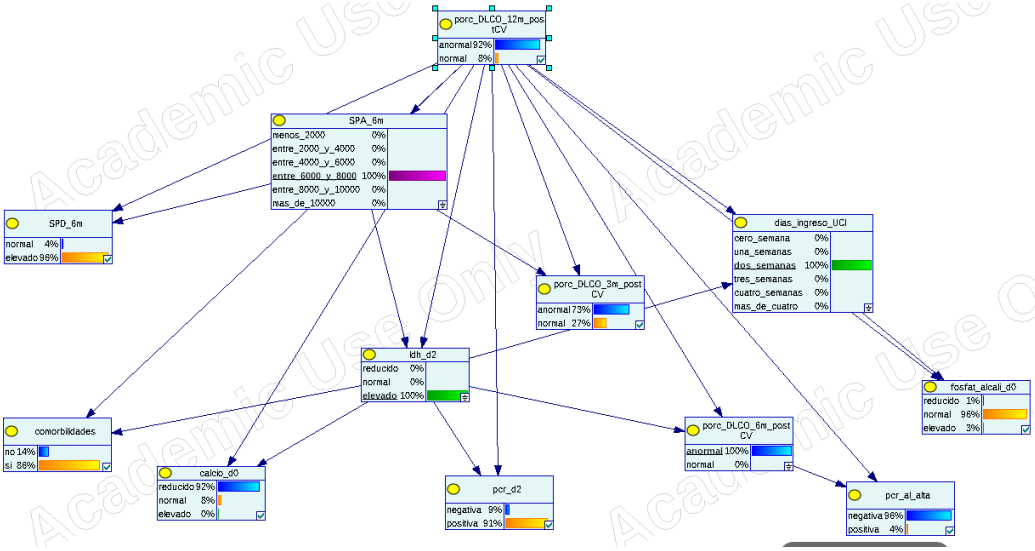

Figura 3.35: *Query 3*, Modelo 6 con GeNIe

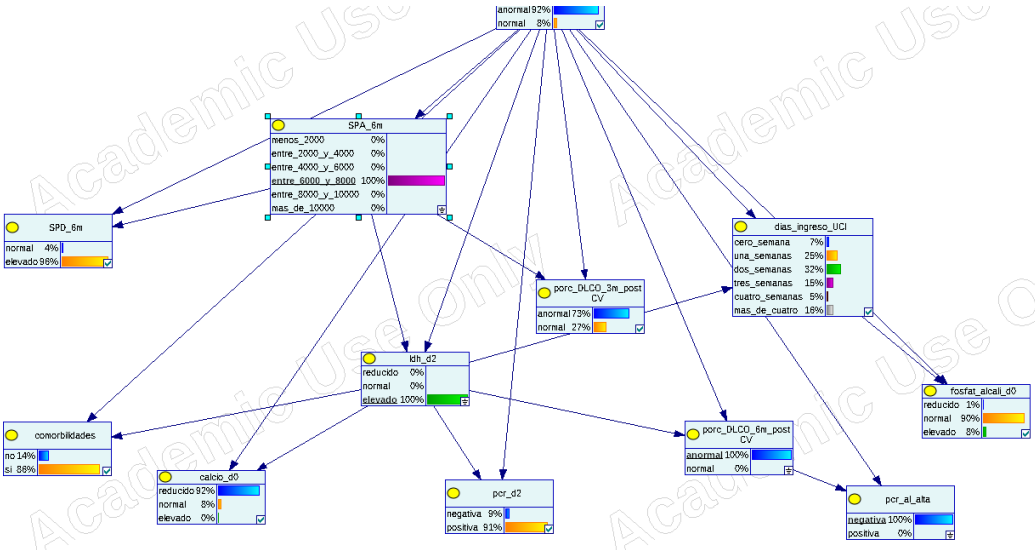

Figura 3.36: *Query 4*, Modelo 6 con GeNIe

3.12. Predecir DLCO-12 a los 6 meses: Regresión logística

Continuamos con el modelo de regresión logística, aplicada al problema presente. El proceso de selección de variables genera el siguiente resultado:

1. "`SPA_6m`": Niveles de SPA (ng/ml) por ELISA en plasma a los 6 meses.
2. "`porc_DLCO_3m_postCV`": Si el porcentaje de DLCO a los 3 meses del alta es superior al 80 %.
3. "`ferritina_d0`": Valor de la ferritina en el momento del ingreso.
4. "`comorbilidades`": Si presenta co-morbilidades (durante o previo a ingreso).
5. "`hemoglobina_d0`": Valor de la hemoglobina en el momento del ingreso.

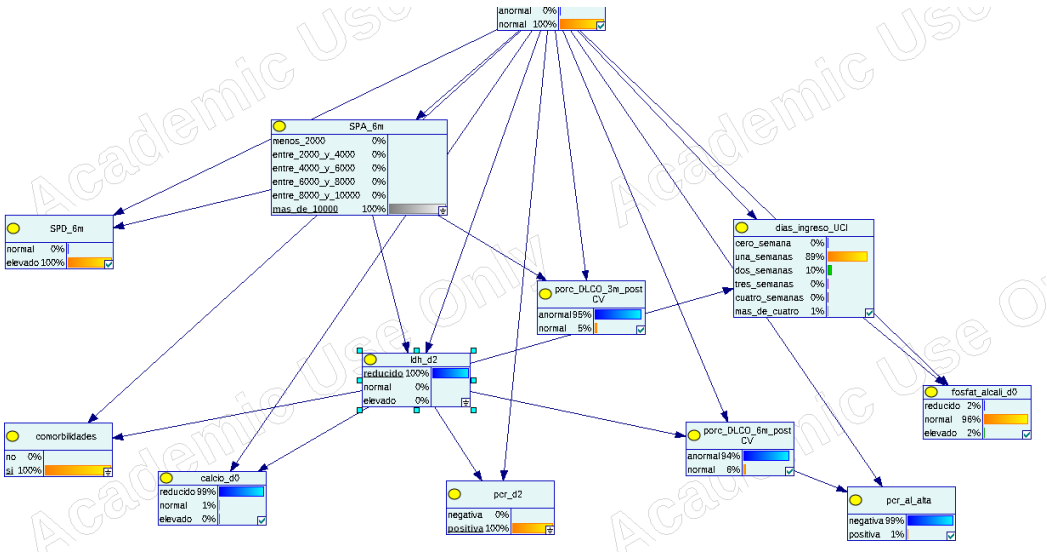

Figura 3.37: *Query 5*, Modelo 6 con GeNIe

Variable	**DLCO<80 %** **N= 400 66.01 %**	**DLCO>80 %** **N= 206 33.99 %**	***p*-valor**	**Valores**
SPA_6m	3.145 ± 0.8978	3.61 ± 1.487	$5.215 \cdot 10^{-5}$	[1 4 3 5 2 0]
calcio_d0	-0.9175 ± 0.2844	-0.8976 ± 0.304	0.4357	[-1 0 1]
pcr_al_alta	0.03 ± 0.1708	0.02927 ± 0.169	0.96	[1 0]
pcr_d2	0.9025 ± 0.297	0.9122 ± 0.2837	0.6956	[1 0]
fosfat_alcali._d0	0.055 ± 0.2686	0.06829 ± 0.3055	0.5984	[0 -1 1]
SPD_6m	0.9525 ± 0.213	0.9463 ± 0.2259	0.7465	[1 0]
porc_DLCO_3m_postCV	0.2625 ± 0.4405	0.3415 ± 0.4754	0.04821	[1 0]
comorbidades	0.875 ± 0.3311	0.878 ± 0.328	0.9142	[1 0]
porc_DLCO_6m_postCV	0.2475 ± 0.4321	0.2585 ± 0.4389	0.7687	[0 1]
LDH_d2	0.365 ± 0.8388	-0.1561 ± 0.9156	$3.713 \cdot 10^{-11}$	[0 1 -1]
dias_ingreso_UCI	3.065 ± 1.369	2.502 ± 1.06	$3.943 \cdot 10^{-8}$	[5 1 2 3 4 6]

Cuadro 3.24: Comparación de las variables seleccionadas para los valores de la variable clase

6. "`sodio_d0`": Valor del sodio en el momento del ingreso.
7. "`tvptep`": Sufre tromboembolismo pulmonar (TEP) o trombosis venosa profunda (TVP).
8. "`dimero_d_d0`": Valor del Dímero D en el momento del ingreso.
9. "`dolor_pleuritico`": Si tuvo dolor pleurítico antes del ingreso.
10. "`plaquetas_d0`": Valor de las plaquetas en el momento del ingreso.

Como vemos, las variables incluidas son de ingreso, alta, 3 y 6 meses. Vemos que algunas de las variables son comunes porc_DLCO_3m_postCV, SPA_6m, comorbidades. Otras no aparecen en el modelo anterior, pero las hemos visto en otros problemas como tvptep, ferritina etc.

Comparáramos las variables seleccionadas en ambos modelos en el Cuadro 3.25. El

Red Bayesiana	**Comunes**	**R. Logistica**
calcio_d0	porc_DLCO_12m_postCV	ferritina_d0
pcr_al_alta	SPA_6m	hemoglobina_d0
pcr_d2	porc_DLCO_3m_postCV	sodio_d0
fosfat_alcali._d0	comorbidades	tvptep
SPD_6m		dimero_d_d0
porc_DLCO_6m_postCV		dolor_pleuritico
LDH_d2		plaquetas_d0
dias_ingreso_UCI		

Cuadro 3.25: Comparativa de las variables seleccionadas en ambos modelos

modelo de regresión logística que hemos obtenido tiene una **precisón** de **0.7582** y una **AUC** de **0.6897**. Podemos comparar los resultados obtenidos con el modelo de red bayesiana en el cuadro 3.26: Vemos que los valores de precisión y AUC son altos

	Red Bayesiana	R. Logística	I.C. 95 %	p-valor
Precisión	75.84 %	75.82 %	$(-0.05, 0.0516)$	0.978
AUC	0.7594	0.6897	$(0.00486, 0.1346)$	0.0355

Cuadro 3.26: Comparativa de los dos modelos para el problema 6
Intervalo de confianza y p-valor obtenidos con el test de Student comparando el conjunto de los resultados de precisión y AUC para cada uno de los k-pliegues de la validación con los del conjunto obtenido por el otro modelo.

en los dos casos, comparados con los otros problemas. Esto es algo que podríamos estadísticamente esperar debido a tener más información. Además, vemos que no hay diferencias significativas entre ambos modelos en cuanto a la precisión, aunque se constata que el modelo de red bayesiana es mejor que la regresión logística en área bajo la curva ROC.

A continuación mostramos los parámetros de este modelo de regresión logística en el Cuadro 3.27.

Variables	**OR**	**CI 95 %**	**p-valor**	β
SPA_6m: valor 1 (ref 0)	0.94	$(0.27, 3.26)$	0.97	1.025848
SPA_6m: valor 2 (ref 0)	0.89	$(0.26, 3.08)$	0.812	1.172692
SPA_6m: valor 3 (ref 0)	0.18	$(0.06, 0.55)$	0.009	$2.114018 \cdot 10^{-1}$
SPA_6m: valor 4 (ref 0)	0.3	$(0.1, 0.92)$	0.043	$2.766536 \cdot 10^{-1}$
SPA_6m: valor 5 (ref 0)	4.76	$(1.48, 15.37)$	0.063	3.551931
porc_DLCO_3m_postCV: 1 vs 0	1.45	$(1, 2.1)$	< 0.001	2.419924
ferritina_d0: valor 0 (ref -1)	3.49	$(0.73, 16.78)$	0.275	3.123955
ferritina_d0: valor 1 (ref -1)	1.29	$(0.27, 6.18)$	0.368	2.410972
comorbilidades: 1 vs 0	1.09	$(0.64, 1.84)$	0.727	$8.955125 \cdot 10^{-1}$
hemoglobina_d0: valor 0 (ref -1)	2.65	$(1.21, 5.78)$	0.202	1.826027
hemoglobina_d0: valor 1 (ref -1)	0	$(0, \infty)$	0.984	$2.686853 \cdot 10^{-5}$
sodio_d0: valor 0 (ref -1)	1.91	$(0.89, 4.07)$	0.605	$7.923254 \cdot 10^{-1}$
sodio_d0: valor 1 (ref -1)	0.73	$(0.08, 7.1)$	0.953	$9.247458 \cdot 10^{-1}$
tvptep: 1 vs 0	4.86	$(3.18, 7.45)$	0.029	2.030959
dimero_d_d0: valor 0 (ref -1)	3.43	$(0.96, 12.27)$	0.365	2.061648
dimero_d_d0: valor 1 (ref -1)	1.21	$(0.34, 4.4)$	0.798	1.230932
dolor_pleuritico: 1 vs 0	3.11	$(2.18, 4.45)$	0.225	1.575866
plaquetas_d0: valor 0 (ref-1)	0.47	$(0.33, 0.68)$	0.352	1.328912
plaquetas_d0: valor 1 (ref-1)	1.21	$(0.4, 3.71)$	0.022	4.753187

Cuadro 3.27: Regresión logística Modelo 6: *odds ratio* (OR), Intervalos de confianza (CI) y parámetros (β)

Capítulo 4

Conclusiones y trabajo futuro

En este trabajo de fin de máster planteamos seis problemas de clasificación supervisada que hemos resuelto utilizando redes bayesianas. Hemos comparado los resultados alcanzados con aquellos obtenidos con regresión logística (un método muy empleado en la literatura médica) y hemos logrado resultados modestos pero positivos. En general, los valores de precisión, y especialmente de AUC (métrica más relevante en estos problemas) son buenos. Vemos que los modelos bayesianos obtienen mejores resultados respecto de esta métrica, revindicando la elección de este tipo de modelos, cuyo estudio era uno de los objetivos de este proyecto. Además, hemos estudiado el problema de predecir los niveles de DLCO en diferentes plazos, examinando qué variables son más relevantes en su predicción. Esta tarea ha sido facilitada gracias a la posibilidad de poder consultar y ver gráficamente los modelos usando GeNIe.

Se ha realizado mucha investigación de aprendizaje automático relacionada con el COVID desde el comienzo de la pandemia, no habiendo sido objeto de la misma profundidad de estudio el COVID persistente y las secuelas por parte de este área de conocimiento (casi toda la investigación es desde el punto de vista médico). En este sentido, hemos visto que la literatura era escasa y no hemos encontrado ningún trabajo que haya usado redes bayesianas para afrontarlo.

El trabajo ha cumplido el objetivo de continuar la investigación de Oriol Sibila y Rosa Faner, continuando con los datos de 12 meses. Además, hemos expandido la metodología, con un proceso más extenso desde el punto de vista del área de aprendizaje automático, usando redes bayesianas, un modelo poco común en el contexto médico. El trabajo también hace más énfasis en métricas que evalúan la capacidad preditiva del modelo, algo que es más común en el contexto del aprendizaje automático. Si examinamos el trabajo desde el contexto del aprendizaje automático, vemos que se ha resuelto un problema real de investigación. Además, hemos incorporado el punto de vista de médicos, que son los expertos en el área que este trabajo pretende resolver. Hemos mantenido, de forma consciente, el uso de la regresión logística como modelo de aprendizaje automático e incorporado el mismo tipo de tablas que Oriol Sibila y Rosa Faner usaron en su trabajo como punto de comparación.

En general, los resultados no son suficientemente buenos como para usarlos como única herramienta diagnóstica. Estamos hablando de modelos cuya precisión es algo superior al 70 % y cuyo AUC es cercano a 0.7. Estos resultados son, sin embargo, positivos, y los modelos pueden servir de apoyo en el proceso de diagnóstico. En este

sentido, podría ser de mayor utilidad tener los modelos BAN en una herramienta como GeNIe, para mostrar de forma más transparente los resultados y facilitar la introducción de los datos del paciente. Sin embargo, esto requiere un cierto de nivel de familiaridad con las redes bayesianas, así como acceso a una herramienta visual como GeNIe.

En el Cuadro 4.1 podemos ver resumidas las métricas para cada modelo. Vemos que que en general los modelos de regresión logística tienen mejor precisión (en torno al 74 %) frente a los modelos BAN (en torno al 71 %). Sin embargo, vemos que los modelos BAN tienen mejor AUC en casi todos los casos.

	Precisión		Mejor Modelo	AUC		Mejor Modelo
Problema	RB	RL		RB	RL	
1	75.84 %	75.82 %	EQ	0.7594	0.6897	RB
2	72.88 %	75.85 %	RL	0.7189	0.6726	RB
3	44.95 %	53.14 %	RL	0.6830	0.6760	EQ
4	71.47 %	74.97 %	RL	0.6420	0.6201	RB
5	74.84 %	74.05 %	EQ	0.5960	0.5026	RB
6	75.84 %	75.82 %	EQ	0.7594	0.6897	RB
Media	71.12 %	73.99 %		0.7070	0.6509	

Cuadro 4.1: Comparativa de los dos modelos para los seis problemas
EQ: No hay diferencias significativas
RB: Red bayesiana mejor con p<0.05 %
RL: Red bayesiana mejor con p<0.05

No solo los modelos son herramientas diagnósticas; las variables obtenidas en el proceso de selección dan información a los médicos. Podemos analizar las variables predictoras. Consideramos las que han sido seleccionadas más veces en los 12 modelos (6 de regresión logística y 6 clasificadores bayesianos) obviando en este caso el momento en el que se tomó la medida. No incluimos tampoco las variables de DLCO. Vemos en la Figura 4.1 que ciertos indicadores aparecen repetidos varias veces. Notablemente, ldh (aunque hay algunos modelos que incluían esta variable en varios momentos del tiempo).

En este caso, las variables seleccionadas más que aportar una perspectiva nueva a este área de investigación, confirman lo que los médicos ya sabían. De acuerdo con Oriol Sibila, estos resultados son consistentes con el trabajo que han realizado ellos y otros investigadores en este área. En este sentido, el presente trabajo sirve para replicar resultados obtenidos en este área de investigación. Asimismo, que se hayan obtenido resultados consistentes con la investigación existente en este área valida el trabajo realizado.

Hemos visto que DLCO anormal en un periodo no es necesariamente indicador de DLCO anormal en el futuro, siendo necesario examinar más información de los pacientes. También hemos visto que la discretización supone una pérdida de mucha información. Variables como proteína C-reactiva o tvptep pierden información al ser discretizadas. Una futura línea para continuar el presente trabajo sería trabajar con variables continuas o con un método de discretización no basado en expertos. Otra mejora en este aspecto es discretizar todas las variables, no solo las que pasaron el primer proceso de selección de variables, con el objetivo de que el proceso final de

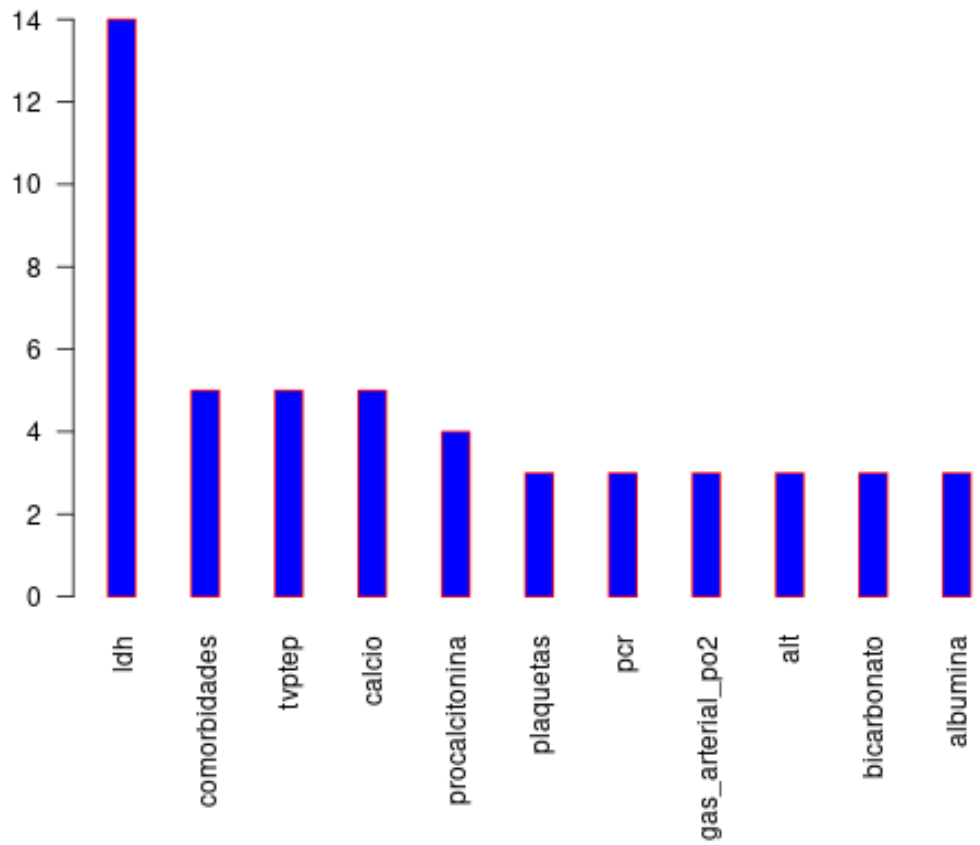

Figura 4.1: Variables más veces seleccionadas en los 12 modelos (3 veces o más)

selección de variables parta de una mayor cantidad de variables.

Sobre expandir el trabajo, hay más posibles mejoras del proceso que se pueden perseguir además de cambiar el proceso de discretización. Una opción es usar otros algoritmos además de la regresión logística para realizar la comparación. Otra opción es usar redes causales y razonamientos contra-fácticos.

Una de las primeras líneas en las que se podría continuar este trabajo es ayudando a los médicos a integrar esta herramienta en un entorno donde la puedan consultar, preferiblemente online, aunque eso está limitado por las herramientas que pueden interpretar los modelos. GeNIe es un programa, cuyo uso solo es gratuito en el entorno académico y supone una barrera de coste y entrenamiento que limita su aplicabilidad. Una posible solución sería crear una herramienta (o adaptar una existente) que permita interactuar con el modelo, aunque sea de forma sencilla.

Oriol Sibila y Rosa Faner tienen datos nuevos disponibles, que podrían permitir continuar la linea de investigación con un *dataset* ampliado y refinando las técnicas utilizadas en el presente trabajo. Los datos que tienen disponibles corresponden a

nuevas variables tomadas sobre los mismos pacientes. Esto puede llevarse por dos líneas. Por un lado se puede usar esta nueva información para ver si es posible mejorar los modelos desarrollados para resolver los problemas 1-6 en . Por otro lado, se puede investigar la capacidad de estas variables para predecir el nivel de DLCO considerando horizontes de tiempo que superan los 12 meses.

Finalmente, sería deseable trabajar con Rosa Faner y Oriol Sibila en estudiar la interpretabilidad de los modelos de red bayesiana comparados con los modelos de regresión logística (comúnmente utilizada en el contexto médico). Además de eso, validar la utilidad de los modelos y su aplicabilidad.

Bibliografía

[1] Bayesfusion LLC. *GeNIe Modeler Academic*. Ver. 4.0. Dic. de 2022. URL: https://www.bayesfusion.com/genie/.

[2] Mandeep Garg et al. "The conundrum of 'long-COVID-19: a narrative review". En: *International journal of general medicine* (2021), págs. 2491-2506.

[3] Rachel R Deer et al. "Characterizing long COVID: deep phenotype of a complex condition". En: *EBioMedicine* 74 (2021).

[4] Oriol Sibila et al. "Elevated plasma levels of epithelial and endothelial cell markers in COVID-19 survivors with reduced lung diffusing capacity six months after hospital discharge". En: *Respiratory Research* 23.1 (2022), págs. 1-10.

[5] World Health Organization. *A clinical case definition of post COVID-19 condition by a Delphi consensus, 6 October 2021*. Inf. téc. 2021.

[6] Emily R Pfaff et al. "Identifying who has long COVID in the USA: a machine learning approach using N3C data". En: *The Lancet Digital Health* (2022), e532-e541.

[7] Roderick J. Little et al. "The prevention and treatment of missing data in clinical trials". En: *New England Journal of Medicine* 367.14 (2012), págs. 1355-1360.

[8] Jason S. Haukoos y Craig D. Newgard. "Advanced statistics: missing data in clinical research—part 1: An introduction and conceptual framework". En: *Academic Emergency Medicine* 14.7 (2007), págs. 662-668.

[9] Joseph L Schafer y John W Graham. "Missing data: Our view of the state of the art." En: *Psychological Methods* 7.2 (2002), págs. 147-177.

[10] Donald B Rubin. "Inference and missing data". En: *Biometrika* 63.3 (1976), págs. 581-592.

[11] Ane B. Pedersen et al. "Missing data and multiple imputation in clinical epidemiological research". En: *Clinical Epidemiology* 9 (mar. de 2017), págs. 157-166.

[12] Stef Van Buuren. *Flexible Imputation of Missing Data*. CRC press, 2018.

[13] David W. Hosmer y Stanley Lemeshow. *Applied Logistic Regression*. 2nd Edition. 2000.

[14] Kang Hyun. "The prevention and handling of the missing data". En: *Korean Journal of Anesthesiology* 64.5 (2013), págs. 402-406.

[15] Christophe Genolini, Hélène Jacqmin-Gadda et al. "Copy mean: A new method to impute intermittent missing values in longitudinal studies". En: *Open Journal of Statistics* 3.04 (2013), págs. 26-34.

[16] Janus Christian Jakobsen et al. "When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts". En: *BMC Medical Research Methodology* 17.1 (2017), págs. 1-10.

[17] Stef van Buuren y Karin Groothuis-Oudshoorn. "MICE: Multivariate imputation by chained equations in R". En: *Journal of Statistical Software* 45.3 (2011), 1–67.

[18] Yazeed Zoabi, Shira Deri-Rozov y Noam Shomron. "Machine learning-based prediction of COVID-19 diagnosis based on symptoms". En: *npj Digital Medicine* 4.1 (2021), págs. 1-5.

[19] Tianqi Chen et al. "Xgboost: Extreme gradient boosting". En: *R package version 0.4-2* 1.4 (2015), págs. 1-4.

[20] Changyong Huang et al. "6-month consequences of COVID-19 in patients discharged from hospital: A cohort study". En: *Lancet* 397.10270 (2021), págs. 220-232.

[21] Jessica González et al. "Pulmonary function and radiologic features in survivors of critical COVID-19: A 3-month prospective cohort". En: *Chest* 160.1 (2021), págs. 187-198.

[22] Aditya Gupta, Vibha Jain y Amritpal Singh. "Stacking ensemble-based intelligent machine learning model for predicting post-COVID-19 complications". En: *New Generation Computing* 40.4 (2022), págs. 987-1007.

[23] Rohan P Joshi et al. "A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results". En: *Journal of Clinical Virology* 129 (2020).

[24] Vipin Kumar y Sonajharia Minz. "Feature selection: A literature review". En: *SmartCR* 4.3 (2014), págs. 211-229.

[25] Jundong Li et al. "Feature selection: A data perspective". En: *ACM Computing Surveys (CSUR)* 50.6 (2017), págs. 1-45.

[26] Wei Qin et al. "Diffusion capacity abnormalities for carbon monoxide in patients with COVID-19 at 3-month follow-up". En: *European Respiratory Journal* 58.1 (2021).

[27] Topi Talvitie, Ralf Eggeling y Mikko Koivisto. "Learning Bayesian networks with local structure, mixed variables, and exact algorithms". En: *International Journal of Approximate Reasoning* 115 (2019), págs. 69-95. ISSN: 0888-613X.

[28] Andrew KC Wong y David KY Chiu. "Synthesizing statistical knowledge from incomplete mixed-mode data". En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1987), págs. 796-805.

[29] Michal R Chmielewski y Jerzy W Grzymala-Busse. "Global discretization of continuous attributes as preprocessing for machine learning". En: *International Journal of Approximate Reasoning* 15.4 (1996), págs. 319-331.

[30] Rajashree Dash, Rajib Lochan Paramguru y Rasmita Dash. "Comparative analysis of supervised and unsupervised discretization techniques". En: *International Journal of Advances in Science and Technology* (2011), págs. 29-37.

[31] Judea Pearl. "*Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*." En: (1988).

[32] Kazuo J Ezawa, Moninder Singh y Steven W Norton. "Learning goal oriented Bayesian networks for telecommunications risk management". En: *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*. 1996, págs. 139-147.

[33] David G Kleinbaum y Mitchell Klein. *Logistic Regression*. Springer, 2002.

[34] Scott Menard. *Applied Logistic Regression Analysis*. Sage, 2002. Cap. 5, págs. 91-93.

[35] S. F. Buck. "A method of estimation of missing values in multivariate data suitable for Use with an electronic computer." En: *Journal of the Royal Statistical Society. Series B (Methodological)* vol. 22, no. 2 (1960), págs. 302-306.

[36] Ross D Shachter y C Robert Kenley. "Gaussian influence diagrams". En: *Management science* 35.5 (1989), págs. 527-550.

[37] Hanchuan Peng, Fuhui Long y C. Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy". En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (2005), págs. 1226-1238.

Anexo

A continuación consideramos los *datasets* completos (sin discretizar) y antes de realizar el proceso de selección de variables *filter-wrapper*.

Variables	**Rangos y Valores**
edad	[19.0, 89.0]
sexo	{0, 1}
peso	[45.0, 135.0]
comorbidades	{0, 1, 41}
EPOC_Enfisema	{0, 1, 2}
hipertension	{0, 1}
obesidad_morbida	{0, 1}
fumador	[0.0, 2.0]
paquetes_años	[0.0, 120.0]
consumo_de_alcohol	{0, 1, 9}
dias_ingreso_hospital	[-9.0, 136.0]
disnea	{0, 1}
dolor_pleuritico	{0, 1}
expectoracion	{0, 1, 2, 3, 4}
temperatura	[34.2999999999999, 40.0]
pas_d0	[70.0, 191.0]
pad_d0	[40.0, 113.0]
saturacion_basal_d0	[24.0, 100.0]
frec_cardiaca_bpm_d0	[21.0, 151.0]
frec_respiratoria_ipm_d0	[12.0, 42.0]
glasgow_d0	{2, 6, 13, 14, 15}
expl_respiratoria_al_ingreso	{0, 1, 2, 3, 4}
Protei._C_Reactiva_d0	[0.1, 269.0]
creatinina_d0	[0.1, 297.0]
ast_d0	[8.0, 1261.0]
alt_d0	[7.0, 750.0]
bilirrubi._total_d0	[0.1, 259.0]
fosfat_alcali._d0	[1.0, 1079.0]
LDH_d0	[129.0, 1314.0]
proteinas_totales_d0	[6.0, 89.0]
albumina_d0	[30.0, 48.0]
sodio_d0	[109.0, 151.0]
procalcitonina_d0	[0.0, 784.0]
ferritina_d0	[17.0, 4149.0]

leucocitos_total_d0	[1010.0, 270000.0]
hemoglobina_d0	[32.2, 182.0]
hematocrito_d0	[0.4, 54.0]
plaquetas_d0	[24000.0, 2480000.0]
neutrófilos_total_d0	[700.0, 103000.0]
neutrofilos_porcentaje_d0	[0.888888888888889, 8564.35643564357]
dimero_d_d0	[100.0, 55000.0]
linfocitos_d0_total	[100.0, 16430.0]
linfocitos_porcentaje_d0	[0.98438560760353, 171.145833333333]
creatin_kinasa_d0	[30.0, 613.0]
urea_d0	[18.6, 141.0]
glicemia_d0	[3.8, 580.0]
troponina_d0	[0.0, 9775.1]
calcio_d0	[1.7, 10.5]
magnesio_d0	[1.0, 44229.0]
gas_arterial_fio2_d0	[21.0, 100.0]
gas_arterial_ph_d0	[7.176, 7464.0]
gas_arterial_po2_d0	[30.0, 277.399999999999]
gas_arterial_pco2_d0	[22.7, 51.9]
bicarbonato_d0	[13.6, 34.0]
SDRA	[0.0, 1.0]
neumonia_organizada	{0, 1}
comp_respiratoria	{0, 1}
Severity_WHO	[1.0, 8.0]
dias_ingreso_UCI	[0.0, 109.0]
pcr_d2	[0.13, 645.0]
pcr_al_alta	[0.0, 8.96]
ferritina_d2	[23.0, 6880.0]
ferritina_alta	[56.0, 2524.0]
ldh_d2	[143.0, 1139.0]
ldh_al_alta	[119.0, 956.0]
procalcitonina_d2	[0.0, 999.0]
procalcitonina_al_alta	[0.0, 251.0]
linfos_totales_d2	[100.0, 33100.0]
linfos_totales_alta	[400.0, 37900.0]
dimero_d2	[100.0, 13300.0]
dimero_d_al_alta	[0.0, 13600.0]
tvptep	{0, 1}
Sintomas_3m	{0, 1}
disnea_3m	{0, 1}
SPD_3m	[63.43, 174.3]
SPA_3m	[392.63, 3396.94]

Cuadro 2: Variables seleccionadas 3 meses

Variables	**Rangos y Valores**
edad	[19.0, 89.0]

sexo	{0, 1}
peso	[45.0, 135.0]
comorbidades	{0, 1, 41}
EPOC_Enfisema	{0, 1, 2}
hipertension	{0, 1}
obesidad_morbida	{0, 1}
fumador	[0.0, 2.0]
paquetes_años	[0.0, 120.0]
consumo_de_alcohol	{0, 1, 9}
dias_ingreso_hospital	[-9.0, 136.0]
disnea	{0, 1}
dolor_pleuritico	{0, 1}
expectoracion	{0, 1, 2, 3, 4}
temperatura	[34.2999999999999, 40.0]
pas_d0	[70.0, 191.0]
pad_d0	[40.0, 113.0]
saturacion_basal_d0	[24.0, 100.0]
frec_cardiaca_bpm_d0	[21.0, 151.0]
frec_respiratoria_ipm_d0	[12.0, 42.0]
glasgow_d0	{2, 6, 13, 14, 15}
expl_respiratoria_al_ingreso	{0, 1, 2, 3, 4}
Protei._C_Reactiva_d0	[0.1, 269.0]
creatinina_d0	[0.1, 297.0]
ast_d0	[8.0, 1261.0]
alt_d0	[7.0, 750.0]
bilirrubi._total_d0	[0.1, 259.0]
fosfat_alcali._d0	[1.0, 1079.0]
LDH_d0	[129.0, 1314.0]
proteinas_totales_d0	[6.0, 89.0]
albumina_d0	[30.0, 48.0]
sodio_d0	[109.0, 151.0]
procalcitonina_d0	[0.0, 784.0]
ferritina_d0	[17.0, 4149.0]
leucocitos_total_d0	[1010.0, 270000.0]
hemoglobina_d0	[32.2, 182.0]
hematocrito_d0	[0.4, 54.0]
plaquetas_d0	[24000.0, 2480000.0]
neutrófilos_total_d0	[700.0, 103000.0]
neutrofilos_porcentaje_d0	[0.888888888888889, 8564.35643564357]
dimero_d_d0	[100.0, 55000.0]
linfocitos_d0_total	[100.0, 16430.0]
linfocitos_porcentaje_d0	[0.98438560760353, 171.145833333333]
creatin_kinasa_d0	[30.0, 613.0]
urea_d0	[18.6, 141.0]
glicemia_d0	[3.8, 580.0]
troponina_d0	[0.0, 9775.1]
calcio_d0	[1.7, 10.5]
magnesio_d0	[1.0, 44229.0]
gas_arterial_fio2_d0	[21.0, 100.0]

gas_arterial_ph_d0	[7.176, 7464.0]
gas_arterial_po2_d0	[30.0, 277.399999999999]
gas_arterial_pco2_d0	[22.7, 51.9]
bicarbonato_d0	[13.6, 34.0]
SDRA	[0.0, 1.0]
neumonia_organizada	{0, 1}
comp_respiratoria	{0, 1}
Severity_WHO	[1.0, 8.0]
dias_ingreso_UCI	[0.0, 109.0]
pcr_d2	[0.13, 645.0]
pcr_al_alta	[0.0, 8.96]
ferritina_d2	[23.0, 6880.0]
ferritina_alta	[56.0, 2524.0]
ldh_d2	[143.0, 1139.0]
ldh_al_alta	[119.0, 956.0]
procalcitonina_d2	[0.0, 999.0]
procalcitonina_al_alta	[0.0, 251.0]
linfos_totales_d2	[100.0, 33100.0]
linfos_totales_alta	[400.0, 37900.0]
dimero_d2	[100.0, 13300.0]
dimero_d_al_alta	[0.0, 13600.0]
tvptep	{0, 1}
Sintomas_6m	[0.0, 1.0]
disnea_6m	[0.0, 1.0]
SPD_6m	[18.2, 701.01]
SPA_6m	[289.97, 13000.64]

Cuadro 3: Variables seleccionadas 6 meses

Variables	**Rangos y Valores**
edad	[19.0, 89.0]
sexo	{0, 1}
peso	[45.0, 135.0]
comorbidades	{0, 1, 41}
EPOC_Enfisema	{0, 1, 2}
hipertension	{0, 1}
obesidad_morbida	{0, 1}
fumador	[0.0, 2.0]
paquetes_años	[0.0, 120.0]
consumo_de_alcohol	{0, 1, 9}
dias_ingreso_hospital	[-9.0, 136.0]
disnea	{0, 1}
dolor_pleuritico	{0, 1}
expectoracion	{0, 1, 2, 3, 4}
temperatura	[34.2999999999999, 40.0]
pas_d0	[70.0, 191.0]
pad_d0	[40.0, 113.0]

saturacion_basal_d0	[24.0, 100.0]
frec_cardiaca_bpm_d0	[21.0, 151.0]
frec_respiratoria_ipm_d0	[12.0, 42.0]
glasgow_d0	{2, 6, 13, 14, 15}
expl_respiratoria_al_ingreso	{0, 1, 2, 3, 4}
Protei._C_Reactiva_d0	[0.1, 269.0]
creatinina_d0	[0.1, 297.0]
ast_d0	[8.0, 1261.0]
alt_d0	[7.0, 750.0]
bilirrubi._total_d0	[0.1, 259.0]
fosfat_alcali._d0	[1.0, 1079.0]
LDH_d0	[129.0, 1314.0]
proteinas_totales_d0	[6.0, 89.0]
albumina_d0	[30.0, 48.0]
sodio_d0	[109.0, 151.0]
procalcitonina_d0	[0.0, 784.0]
ferritina_d0	[17.0, 4149.0]
leucocitos_total_d0	[1010.0, 270000.0]
hemoglobina_d0	[32.2, 182.0]
hematocrito_d0	[0.4, 54.0]
plaquetas_d0	[24000.0, 2480000.0]
neutrófilos_total_d0	[700.0, 103000.0]
neutrofilos_porcentaje_d0	[0.888888888888889, 8564.35643564357]
dimero_d_d0	[100.0, 55000.0]
linfocitos_d0_total	[100.0, 16430.0]
linfocitos_porcentaje_d0	[0.98438560760353, 171.145833333333]
creatin_kinasa_d0	[30.0, 613.0]
urea_d0	[18.6, 141.0]
glicemia_d0	[3.8, 580.0]
troponina_d0	[0.0, 9775.1]
calcio_d0	[1.7, 10.5]
magnesio_d0	[1.0, 44229.0]
gas_arterial_fio2_d0	[21.0, 100.0]
gas_arterial_ph_d0	[7.176, 7464.0]
gas_arterial_po2_d0	[30.0, 277.399999999999]
gas_arterial_pco2_d0	[22.7, 51.9]
bicarbonato_d0	[13.6, 34.0]
SDRA	[0.0, 1.0]
neumonia_organizada	{0, 1}
comp_respiratoria	{0, 1}
Severity_WHO	[1.0, 8.0]
dias_ingreso_UCI	[0.0, 109.0]
pcr_d2	[0.13, 645.0]
pcr_al_alta	[0.0, 8.96]
ferritina_d2	[23.0, 6880.0]
ferritina_alta	[56.0, 2524.0]
ldh_d2	[143.0, 1139.0]
ldh_al_alta	[119.0, 956.0]
procalcitonina_d2	[0.0, 999.0]

procalcitonina_al_alta	[0.0, 251.0]
linfos_totales_d2	[100.0, 33100.0]
linfos_totales_alta	[400.0, 37900.0]
dimero_d2	[100.0, 13300.0]
dimero_d_al_alta	[0.0, 13600.0]
tvptep	{0, 1}
Sintomas_12m	[0.0, 1.0]
disnea_12m	[0.0, 1.0]
SPA_12m	[2008.63, 44483.77]
SLPI_12m	[11.3, 374.99]
SPD_12m	[12.31, 341.05]

Cuadro 4: Variables seleccionadas 12 meses