Mejora en la clasificación de pacientes mediante técnicas de Machine Learning: Aplicación a un problema neurológico a partir de la obtención de 14 biomarcadores

Manuel Quintana Luque
Master en Bioinformática y Bioestadística
Área 5. Subárea 1: Estadística y bioinformática

Consultor: Santiago Pérez Hoyos
Profesor responsable de la asignatura: Alexandre Sánchez Pla

Fecha Entrega: 2 de enero de 2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Mejora en la clasificación de pacientes mediante técnicas de Machine Learning: Aplicación a un problema neurológico a partir de la obtención de 14 biomarcadores*
Nombre del autor:	*Manuel Quintana Luque*
Nombre del consultor/a:	*Santiago Pérez Hoyos*
Nombre del PRA:	*Alexandre Sánchez Pla*
Fecha de entrega (mm/aaaa):	01/2019
Titulación::	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Àrea 5. Subárea 1: Estadística y bioinform.*
Idioma del trabajo:	*Castellano*
Palabras clave	*Machine learning, biomarcador, ictus*

Resumen del Trabajo (máximo 250 palabras):

El objetivo de este trabajo es conocer si la aplicación de técnicas de Machine Learning es útil para mejorar la clasificación de pacientes con ictus a partir de los valores de una serie de biomarcadores. Para ello, se han utilizado los pacientes con sospecha de ictus de un estudio publicado en el cual se utilizaron modelos de regresión logística sin éxito.

Se han aplicado técnicas de Machine Learning a los datos de entrenamiento (n=541) y se ha evaluado el rendimiento de los algoritmos obtenidos en una muestra de validación (n=766), obteniendo la capacidad diagnóstica de los modelos mediante matrices de confusión.

El mejor algoritmo para clasificar ictus/mimic se ha obtenido mediante un Random Forest entrenado con 10-fold crossvalidation, consiguiendo una precisión del 86.7% en la muestra de validación. El mejor algoritmo para clasificar ictus isquémico/hemorrágico se ha obtenido mediante una red neuronal artificial entrenada con 3-fold crossvalidation, con una precisión del 86.8% en la muestra de validación. No se han mejorado los resultados de los modelos de regresión logística en la clasificación de ictus isquémico/ hemorrágico, pero sí en la de ictus/mimic. No obstante, no se han alcanzado las precisiones del 90% esperadas al inicio del estudio.

En conclusión, las capacidades diagnósticas de los algoritmos obtenidos mediante técnicas de Machine Learning no son muy superiores a los obtenidos mediante regresión logística. Así, hasta que no se hallen otros marcadores más potentes, no es posible clasificar más precozmente a estos pacientes, siendo necesaria la obtención de pruebas complementarias para ello.

Abstract (in English, 250 words or less):

The aim of this work is to know if the use of machine learning techniques with biomarkers is useful to improve the classification of stroke patients. For this purpose, patients with suspected stroke have been obtained from a published study in which logistic regression models were used without success.

Machine Learning techniques have been applied to the training data (n = 541) and the performance of the algorithms has been evaluated in a validation sample (n = 766), obtaining the diagnostic ability of the models through the use of confusion matrices.

The best algorithm to classify stroke/mimic was obtained by a Random Forest trained with 10-fold crossvalidation, obtaining an accuracy of 86.7% in the validation sample. The best algorithm to classify ischemic/hemorrhagic stroke was obtained by an artificial neural network trained with 3-fold crossvalidation, with an accuracy of 86.8% obtained in the validation sample.

The results of the logistic regression models have not been improved by machine learning techniques in the classification of ischemic/hemorrhagic stroke, but an improvement was achieved in the stroke/mimic classification. However, the 90% of accuracy expected at the beginning of the study has not been reached.

In conclusion, the diagnostic abilities of the algorithms obtained by Machine Learning techniques are not much higher than those obtained by logistic regression. Thus, until the obtention of new potent biomarkers, it is not possible to classify these patients earlier, being still necessary the obtention of complementary medical tests.

Índice

Lista de figuras

Lista de tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

En el ámbito de la salud, es necesario diagnosticar bien a los pacientes para ofrecerles el tratamiento adecuado y prevenir secuelas que pueden llegar a ser irreversibles [1,2]. Hay enfermedades como el ictus, en las cuales la rapidez en las que se les administra el tratamiento es clave para su pronóstico final [3-5], por lo que es necesario un rápido diagnóstico de esta enfermedad.

Mediante la obtención de biomarcadores es posible obtener buenas capacidades predictivas para clasificar a los pacientes [6-8], aunque hay casos en los que no se alcanzan los resultados esperados tras aplicar los métodos estadísticos más tradicionales [9-11].

En un estudio publicado en Stroke por Bustamante A, et al [9] se observó, mediante modelos de regresión logística, que los biomarcadores no aportaban más a la clasificación de pacientes de lo que lo hacían las variables clínicas. Para realizar esos modelos, los biomarcadores se categorizaron optimizando su asociación con cada una de las variables de clasificación. Uno de los inconvenientes de este procedimiento es que, aún obteniendo un punto de corte óptimo de cada biomarcador, no implica que la combinación entre los diferentes biomarcadores sea óptima para obtener una buena clasificación, perdiendo información que podría ser útil con puntos de corte diferentes o, incluso, dividiendo cada biomarcador en diferentes tramos, por lo que el modelo final obtenido puede estar muy lejos del modelo de predicción óptimo.

Las técnicas de Machine Learning han resultado ser muy útiles para encontrar algoritmos con altas capacidades predictivas en múltiples estudios de clasificación y predicción [12-19], por lo que pensamos que la aplicación de estas técnicas al problema anterior podría ser una mejor opción para hallar algoritmos que nos permitan clasificar a los pacientes de una forma más precisa, obteniendo así un diagnóstico precoz más fiable.

Así, el propósito de este trabajo es explorar si la aplicación de técnicas de Machine Learning es útil para mejorar la clasificación de pacientes con ictus a partir de los valores de una serie de biomarcadores. Para ello, se analizan los pacientes con sospecha de ictus del estudio de Bustamante A, et al [9] con el objetivo de clasificarlos según el diagnóstico final de ictus o mimic (cualquier diagnóstico que no sea ictus) y, entre los mismos ictus, diferenciar los que son isquémicos de los hemorrágicos.

Como resultado, se esperan obtener algoritmos con capacidades diagnósticas para clasificar a los pacientes superiores al 90%.

1.2 Objetivos del Trabajo

Objetivo general:

- Mejorar la clasificación de pacientes con ictus mediante técnicas de Machine Learning

Objetivos específicos:

- Obtener el mejor modelo o algoritmo para predecir el diagnóstico de ictus vs mimic a través de técnicas de Machine Learning
- Obtener el mejor modelo para diferenciar los ictus isquémicos de los hemorrágicos a través de técnicas de Machine Learning
- Determinar la capacidad diagnóstica de los algoritmos en un grupo de validación
- Determinar si la capacidad diagnóstica de los algoritmos mejoran los modelos ya publicados con los mismos pacientes y obtenidos mediante regresión logística

1.3 Enfoque y método seguido

Para la consecución de los objetivos, se aplica cada una de las técnicas de Machine Learning a unos datos de entrenamiento para obtener una buena

validación en los datos de prueba o validación. Para ello, se tratan de ajustar los parámetros de cada uno de los modelos de forma adecuada y con sentido. El entrenamiento de algoritmos con parámetros no apropiados podría hacernos descartar alguna técnica valiosa para el estudio.

La estrategia que se ha seguido en este estudio, tanto para clasificar ictus vs mimics, como ictus isquémico vs hemorrágico, en cada una de las técnicas es la siguiente:

1. Transformar los datos acorde al algoritmo que se utiliza.
2. Seleccionar los parámetros más adecuados para entrenar el algoritmo según el tipo de datos de nuestro estudio y el tamaño muestral.
3. Entrenar el algoritmo en la muestra de entrenamiento.
4. Testar el algoritmo en la muestra de validación.
5. Volver a entrenar el algoritmo modificando alguno de los parámetros y comprobar si hay mejoría en la clasificación de los pacientes tras probarlo en la muestra de validación. En caso necesario, volver a repetir este paso hasta que se considere que se ha hallado el mejor modelo posible.
6. Analizar cuál es la técnica más adecuada según los resultados obtenidos.
7. Contrastar con los resultados hallados en los modelos de regresión logística por el estudio previo.

1.4 Planificación del Trabajo

Para el desarrollo del trabajo, se ha obtenido una base de datos en SPSS de los autores del artículo de Bustamante A, et al. [9], donde se han dispuesto las variables clínicas y biomarcadores ya transformados de todos los pacientes con sospecha de ictus. También se ha dispuesto de varios archivos en Excel con los datos crudos de los biomarcadores.

Para el procesamiento de datos se ha utilizado el software R (Version 3.4.4) mediante la plataforma RStudio (Version 1.1.414).

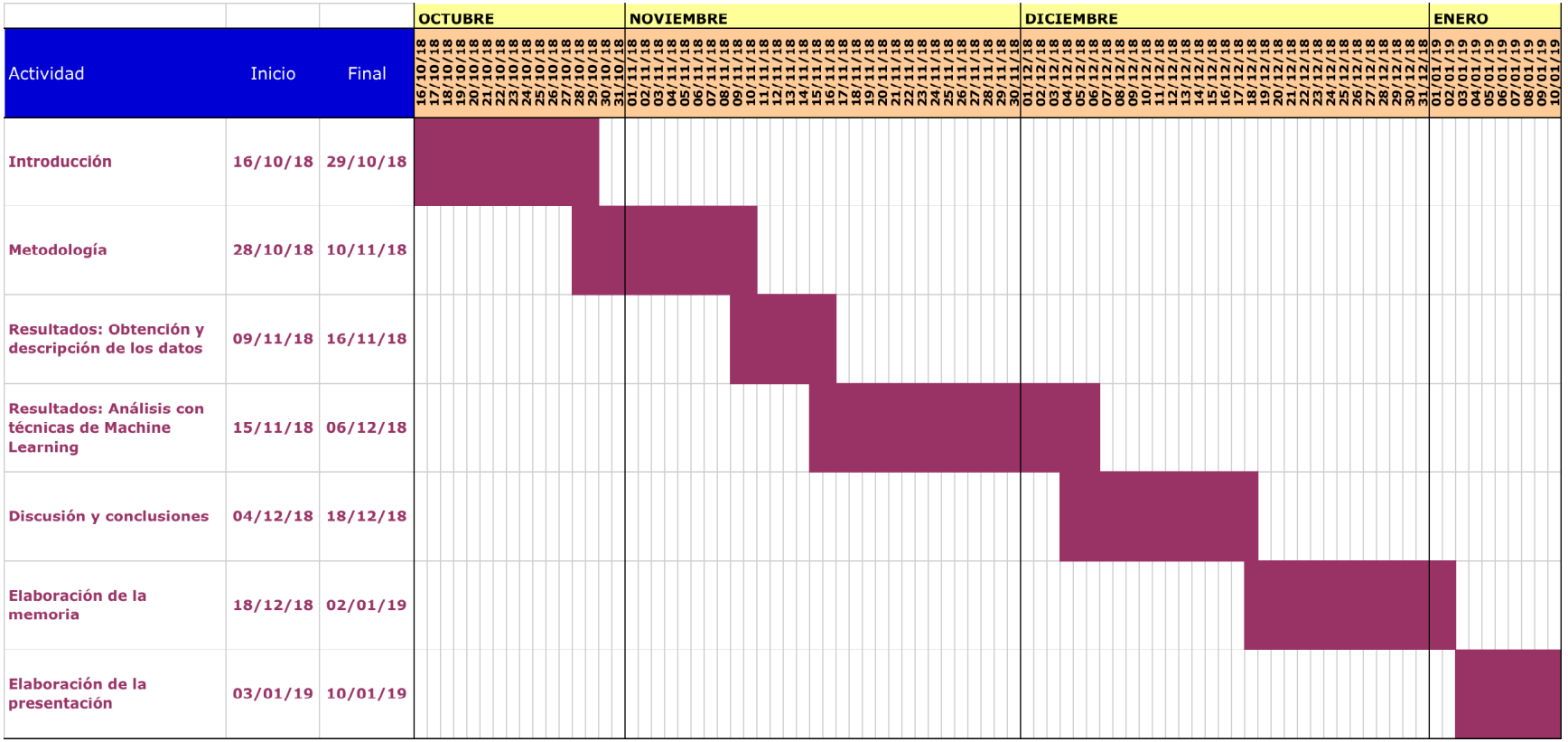
Las tareas planificadas para la realización del trabajo han sido las siguientes:

- Elaboración de la introducción (Situación actual del tema, necesidades, soluciones, aportaciones, objetivos, estrategia, descripción del procedimiento para llevar a cabo el estudio)

- Descripción de la metodología (Origen de los datos, tipo de pacientes, criterios de inclusión, descripción de variables, descripción de métodos de Machine Learning, softwares y paquetes utilizados)
- Obtención de resultados (Importación datos a R, detección de errores, reorganización base de datos, análisis descriptivo, análisis con técnicas de Machine Learning)
- Discusión y conclusiones
- Elaboración de la memoria memoria
- Preparación de presentación virtual

En el siguiente diagrama de Gantt se puede observar la planificación temporal de las tareas que se han realizado en el transcurso del TFM:

Figura 1. Diagrama de Gannt. Planificación temporal de las tareas.



1.5 Breve sumario de productos obtenidos

- El mejor algoritmo para clasificar ictus/mimic se ha obtenido mediante un **Random Forest entrenado con un 10-fold crossvalidation**, consiguiendo una precisión del **86.7%** en la muestra de validación.
- El mejor algoritmo para clasificar ictus isquémico/hemorrágico se ha obtenido mediante una **red neuronal artificial entrenada con 3-fold crossvalidation**, con una precisión del **86.8%** en la muestra de validación.
- Las precisiones de las técnicas en la muestra de validación son algo superiores a las obtenidas mediante modelos de regresión logística, aunque todas están por debajo de las capacidades diagnósticas del 90% esperadas a priori.

1.6 Breve descripción de los otros capítulos de la memoria

En el capítulo 2 de esta memoria se muestra la metodología utilizada para la obtención de los resultados del estudio. En la primera parte de este capítulo se explica como se obtuvo la muestra del estudio, los criterios de inclusión y exclusión, y la descripción de las variables principales y variables clínicas recogidas. En la segunda parte del capítulo, se detallan los biomarcadores analizados y como se procesaron las muestras para obtenerlos. Seguidamente, se enumeran las técnicas de Machine Learning utilizadas, describiendo detalladamente cada una de ellas. Finalmente, se explica como se han analizado los datos y los softwares y paquetes que se han empleado para obtener el análisis. En el capítulo 3 se muestran los resultados obtenidos tras el análisis de datos, desde el descriptivo de la muestra hasta la obtención de los diferentes algoritmos, para acabar comparando los resultados con los del estudio ya publicado. En el capítulo 4 se muestran las conclusiones del estudio y una discusión en base a los resultados obtenidos. Finalmente, en los últimos capítulos se muestra un glosario, las referencias y los anexos.

2. Metodología

2.1 Muestra del estudio

Los datos utilizados para este trabajo provienen del estudio multicéntrico Stroke-Chip [9], en el cual participaron 6 hospitales de Cataluña (Vall d'Hebron, Germans Trias i Pujol y Bellvitge en Barcelona, Josep Trueta en Gerona, Joan XXIII en Tarragona y Verge de la Cinta en Tortosa) entre Agosto de 2012 y Noviembre de 2013. En este estudio se reclutaron de forma prospectiva pacientes con sospecha de ictus en urgencias con los siguientes criterios de inclusión:

- Edad > 18 años
- Tiempo desde el inicio de los síntomas hasta la obtención de muestra sanguínea < 6 horas
- Recogida de sangre previa al tratamiento trombolítico
- Consentimiento informado firmado por paciente o familiar

Fueron excluidos aquellos pacientes en los que no se pudo obtener la muestra de sangre y aquellos en los que no se pudo establecer un diagnóstico claro 1 mes después del episodio de urgencias.
Tras la inclusión, se recogieron los datos clínicos de los pacientes.

Variables principales

Las variables dependientes del estudio y, por tanto, las que utilizamos como variables respuesta para entrenar los algoritmos son las siguientes:

- Diagnóstico (**Ictus vs. Mimic**): El ictus está definido como un trastorno brusco de la circulación cerebral que altera la función de una determinada región del cerebro [20]. En este estudio, el diagnóstico de ictus se realizó por neurólogos especialistas y se confirmaron mediante neuroimagen. Para descartar el ictus en el resto de pacientes, se realizaron las pruebas auxiliares pertinentes en cada caso, y se les acabó diagnosticando como 'mimics'.

- Tipo de Ictus (**Isquémico vs. Hemorrágico**): Un ictus es isquémico cuando hay una disminución importante del flujo sanguíneo que recibe una parte de nuestro cerebro producida generalmente por un coágulo (trombo) que bloquea el paso de sangre a una parte del cerebro. El ictus es hemorrágico cuando se produce la rotura de un vaso cerebral, dejando salir su contenido sanguíneo, dañando a una o varias regiones del cerebro [20].

Variables clínicas

- Demográficas: edad, sexo
- Factores de riesgo: tabaquismo, alcoholismo, hipertensión arterial, diabetes mellitus, dislipemia, fibrilación auricular, coronariopatía, ictus previo
- Discapacidad previa: escala modificada de Rankin (mRS)
- Grado de severidad: escala NIHSS
- Exploración: tensión arterial sistólica, tensión arterial diastólica, glucemia

2.2 Biomarcadores

En el estudio original [9] se analizaron 21 biomarcadores que fueron seleccionados préviamente por tener alguna relación con el ictus, ya fuera por resultados previos del grupo o por información encontrada en la literatura. En un primer corte de datos (n=541, *interim analysis*), se descartaron 7 al no estar nada asociados con ninguna de las variables principales. El resto de pacientes se utilizaron como muestra de validación de los modelos construidos en el *interim analysis*.

Para este estudio se analizan los 14 biomarcadores determinados en la muestra global del estudio, los cuales se detallan a continuación:

- NT-proBNP: N-Terminal Pro-B-Type Natriuretic Peptide
- IGFBP-3: insulin-like growth factor-binding protein-3
- TNF-R1: tumor necrosis factor receptor-1
- GroA: growth-related oncogene-α
- FasL: Fas ligand

- IL-6: interleukin-6
- D-dimer
- vWF: von Willebrand factor
- VAP-1: vascular adhesion protein-1
- Endostatin
- S100B
- Hsc70: heat shock 70 kDA protein-8
- Apo CIII: apolipoprotein CIII
- NCAM: neuron cell adhesion molecule

Procesado de muestras

Las muestras fueron recogidas en tubos EDTA, centrifugados a 1500 g durante 15 minutos a 4ºC, y las alicuotas de plasma se almacenaron a -80ºC. La determinación de los biomarcadores se realizó mediante diferentes inmunoensayos. A continuación se detalla el material, dilución y unidades para la obtención de cada biomarcador:

Biomarcador	**Fabricante**	**Referencia**	**Dilución**	**LIC**	**Unidades**
NT-proBNP	Roche	4,842,464	1/1	5	pg/mL
IGFBP-3, TNF-R1	Aushon 2-plex	85,214	1/25	48.4, 2.34	pg/mL
GroA, FasL, IL-6	Aushon 3-plex	85,723	1/2	0.39, 1.56, 0.20	pg/mL
D-dimer	Stago	947	1/42	10	ng/mL
vWF	Stago	942	1/102	1	%
VAP-1	eBioscience	BMS259TEN	1/1000	0.019	ng/mL
Endostatin	R&D	RYD-DNST0	1/50	0.023	ng/mL
Caspase-3	eBioscience	BMS2012INTS	1/3	0.12	ng/mL
Hsc70	USCNK/Cloud-Clone Corp	E93063Hu/ SED063Hu	1/1	0.134	ng/mL
S100B	Cusabio	CSB-E08065h	1/1	1.17	pg/mL
NCAM	Abnova	KA2003	1/150	10	pg/mL
Apo-CIII	Abnova	KA0465	1/2500	0.002	μg/mL

LIC: límite inferior de cuantificación

Tabla 1. Procesado de muestras de cada biomarcador analizado

2.3 Técnicas de Machine Learning aplicadas

2.3.1 Algoritmo k-NN

Es una de las técnicas de machine learning más sencillas, produciendo un algoritmo de clasificación que se basa en asignar el grupo de un nuevo individuo a la clase del *k* vecino más próximo [21].

El algoritmo k-NN es simple y efectivo, y tiene como principales ventajas que no necesita hacer suposiciones sobre la distribución de los datos y que la fase de entrenamiento del algoritmo es rápida. Por contra, este algoritmo tiene algunas debilidades, ya que no se produce un modelo - por lo que la capacidad es limitada para entender como están relacionadas las características con la respuesta-, la fase de clasificación es lenta y tanto las variables nominales como los datos perdidos requieren de un procesamiento adicional. Además, también se requiere de la selección de un valor de *k* apropiado para obtener un buen modelo.

2.3.2 Naïve Bayes

El algoritmo de Naïve Bayes consiste básicamente en aplicar el teorema de Bayes a problemas de clasificación [21]. Este algoritmo asume que todas las características o variables de la base de datos son igualmente importantes e independientes. No obstante, si las variables no cumplen estas asunciones de independencia, el algoritmo rinde igualmente bastante bien. Como el método Naive-Bayes trabaja con tablas de frecuencia, las variables sobre las que debe construir el algoritmo han de ser categóricas. En caso de tener alguna variable o característica numérica, debemos categorizarla. Para clasificar los datos, el algoritmo calcula las probabilidades de las clases de clasificación, asumiendo que cada una de las características van a ocurrir de forma independiente. Finalmente, para calcular la probabilidad por la cual se regirá el algoritmo para clasificar cada uno de los casos se obtiene la probabilidad obtenida para una de las clases y se divide por la suma de probabilidades total.

El algoritmo Naive-Bayes es simple, rápido y muy efectivo. Además rinde bien con ruido y datos perdidos y requiere relativamente de pocos ejemplos para el entrenamiento, rindiendo también bien en muestras muy grandes. Por el contrario, tenemos que es un algoritmo que se basa en una suposición a menudo errónea de que las variables tengan igual importancia y sean independientes. Además, puede no ser apropiado para bases de datos con muchas variables numéricas.

2.3.3 Redes neuronales artificiales

Una red neuronal artificial modela la relación entre un grupo de señales 'input' (las características o variables de la muestra), y una señal 'output (la/s variable/s respuesta), usando una red de nodos intermedios para ajustar el algoritmo de aprendizaje [21]. Este método que relaciona el 'input' con el 'output' es muy complejo, actuando como una caja negra, ya que los procesos intermedios que se utilizan son muy difíciles de interpretar.

Las redes neuronales suelen constar de las siguientes características:

- Función de activación: Transforma las señales 'input' combinadas de una neurona en una señal 'output' para ser transmitida a la red.
- Topología de la red: Describe el número o nodos en el modelo, así como el número de capas y la manera de cómo están conectadas.
- Algoritmo de entrenamiento: Especifica como son los pesos de las conexiones para inhibir o excitar las neuronas en proporción a cada señal 'input'.

Una de las ventajas de las redes neuronales artificiales es que pueden ser adaptadas tanto a problemas de clasificación como numéricos. Además, hace pocas suposiciones sobre las relaciones que hay entre los datos y pueden modelar patrones más complejos que cualquier otro algoritmo. Por otro lado, se producen algoritmos que son computacionalmente intensivos y lentos para entrenar, sobretodo si la topología de la red es compleja, con unos modelos resultantes que son muy difíciles de interpretar, además de ser propensos a sobreajustar los datos de entrenamiento.

2.3.4 Support Vector Machines

El algoritmo Support Vector Machines (SVM) crea un hiperplano que divide el espacio para crear particiones que sean homogeneas o similares en cada una de las clases, o sea, una superficie que crea un límite entre puntos de datos que, de forma multidimensional, representan los casos y los valores de sus diferentes características[21]. Es un método muy potente, ya que es capaz de modelar relaciones altamente complejas (viene a ser una mezcla de algoritmos k-NN y modelos de regresión lineal). Los algortimos SVM se pueden adaptar para ser usados en cualquier tipo de tarea de aprendizaje, incluyendo tanto tareas de clasificación como de predicción numérica.

Dos parámetros a tener en cuenta en este algoritmo es el valor de coste C y el kernel. El parámetro C te permite ajustar el márgen de error de los casos que caen en el lado equivocado del hiperplano. El kernel te ofrece modelar el hiperplano según diferentes funciones para intentar ajustar el algoritmo SVM de manera que podamos clasificar los casos de una forma más precisa. Los algoritmos SVM con kernels no lineales son muy potentes. Las funciones más comunes de kernel son las siguientes: lineal kernel, polynomial kernel, sigmoid kernel y Gaussian RBF kernel.

Una de las ventajas de estos algoritmos es que, al igual que las redes neuronales, pueden ser usados tanto para problemas de predicción numéricos como de clasificación. Además, estos algoritmos no son propensos a sobreajustar los datos de entrenamiento y son más sencillos de utilizar que las redes neuronales. Como debilidades tenemos que son algoritmos lentos para entrenar cuando tenemos un gran número de variables, son difíciles de interpretar y, si queremos encontrar el mejor modelo, hay que probar con varias combinaciones de parámetros.

2.3.5 Árboles de clasificación

Los árboles de clasificación utilizan una estructura de árbol para modelar las relaciones entre las diferentes características y la variable respuesta [21]. El árbol empieza con un ‘nodo raíz’ y, mediante ‘nodos de decisión’ basados en las diferentes características, el árbol se va ramificando o dividiendo para

optimizar la predicción con la variable respuesta, hasta encontrar un grupo homogéneo donde ninguna decisión mejora la predicción. Estos últimos nodos se denominan 'nodos terminales'. Los árboles de decisión se construyen usando partición recursiva, dividiendo los datos en subgrupos, que a su vez se dividen repetidamente en subgrupos más pequeños hasta que el proceso determina que los datos de los diferentes subgrupos son suficientemente homogenos como para clasificar bien los datos. Se para de dividir cuando todos o casi todos los casos tienen la misma clase, no hay características que se asocien con la variables respuesta y, por tanto, cualquier división es inútil, o el árbol ha crecido a un tamaño predefinido.
Uno de los árboles de decisión más conocidos es el producido por el algoritmo C5.0. El algoritmo C5.0 puede funcionar tan bien como cualquier otro modelo de machine learning avanzado, con la ventaja de que su estructura final es mucho más fácil de interpretar. Otras ventajas que tiene este algoritmo es que tiene un proceso de aprendizaje muy automatizado, pudiéndose manejar bien tanto con variables numéricas como nominales, así como con datos perdidos, excluyendo características que no son importantes para la predicción de la clase. Como desventajas obtenemos que se puede producir fácilmente un sobre o infraajuste del modelo, que pequeños cambios en los datos de entrenamiento pueden condicionar grandes cambios en la lógica de decisión y que son modelos que a menudo están sesgados hacia particiones en características con un gran número de niveles.

2.3.6 Random Forests

Los *random forests* se basan en conjuntos de árboles de decisión, combinando los principios básicos del ensacado con la selección aleatoria de características para añadir diversidad adicional a los modelos de árbol de clasificación [21]. Tras generar el conjunto de árboles, el modelo usa un voto para combinar las predicciones de los árboles. Como el agrupamiento usa sólo una porción aleatoria pequeña de todo el conjunto de características, *random forests* puede manejar bases de datos muy grandes. Sus tasas de error para la mayoría de las tareas de aprendizaje son similares a las de

cualquier otro método. Además, los random forests tienden a ser más fáciles de utilizar y menos propensos al sobreajuste.

Algunas de las fortalezas de estos algoritmos es que pueden manejar ruido o datos perdidos tan bien como las características categóricas o continuas, que seleccionan sólo las características más importantes y que pueden ser usados en bases de datos con un número extremadamente grande de características o casos. Como debilidades tenemos que, a diferencia de los árboles de decisión, el modelo no es fácilmente interpretable y que puede requerir de bastante tiempo para afinar o tunear el modelo a los datos.

2.4 Tratamiento de los datos y softwares empleados para el análisis

Biomarcadores

Para normalizar los biomarcadores obtenidos se ha realizado una transformación logarítmica de éstos y se ha dividido por el valor obtenido de la muestra control de cada placa. Debido a la alta variabilidad entre placas presentadas en la obtención de algunos biomarcadores, se ha decidido además estandarizar todos los valores por placa mediante Z-scores.

Análisis de datos

Para el análisis estadístico, se ha importado la base de datos de un archivo SPSS y se ha utilizado el software R (Version 3.4.4) mediante la plataforma RStudio (Version 1.1.414).

Para empezar, se ha realizado un análisis descriptivo y se han comparado los grupos de clasificación con las variables clínicas y biomarcadores. Para ello, se ha utilizado el test de la ji-cuadrado en la comparación con variables categóricas y el test de la t de Student o la variante de Welch (no igualdad de varianzas entre grupos) en la comparación con variables continuas. La comparación realizada con escalas numéricas se ha realizado mediante el test de la U de Mann-Whitney.

Se ha considerado estadísticamente significativo un p-valor inferior a 0.05 en todas las comparaciones.

Para entrenar los algoritmos se ha utilizado la muestra del 'interim analysis' del estudio previo. El resto de pacientes se ha utilizado para evaluar el rendimiento de los modelos.

En cada una de las técnicas de Machine learning se han utilizado principalmente las siguientes funciones y paquetes de R:

-Algoritmo k-NN: Se ha entrenado mediante la función *knn* del paquete "class". Se han utilizado varios valores de *k* para intentar conseguir el modelo óptimo.

-Naive Bayes: Se ha entrenado mediante la función *naiveBayes* del paquete "e1071". Como los algoritmos de Naive Bayes sólo se pueden entrenar con variables categóricas, se han categorizado las variables numéricas para poderlas incluir en los modelos. Para ello, se han realizado curvas ROC para obtener las sensibilidades y especificidades de cada punto de corte en cada clasificación y se ha determinado el punto de corte óptimo mediante el valor máximo del índice de Youden, obtenido a partir de la siguiente fórmula: J= sensibilidad+especificidad-1 [22]. Se ha intentado mejorar el algoritmo realizando cambios en el parámetro de *Laplace*.

-Redes Neuronales Artificiales: Se han entrenado con la función *neuralnet* del paquete "neuralnet". Se ha probado con diferentes nodos ocultos y se ha realizado una validación cruzada para intentar mejorar el modelos. También se ha utilizado la función *plotnet* del paquete "NeuralNetTools" para obtener una visualización diferente de la red.

-Support Vector Machine (SVM): Se ha utilizado la función *ksvm* del paquete "kernlab" para entrenar los datos. Se han obtenido probado con diferentes funciones para entrenar los algoritmos (métodos 'svmLinear', 'svmRadial', 'svmPoly´) y se han realizado validaciones cruzadas para intentar mejorar los modelos.

-Árboles de decisión: Se han entrenado mediante la función *C5.0* del paquete "C50". Se ha intentado mejorar la precisión entrenando el algoritmo con el parámetro 'trials', al cual le indicábamos el número de árboles de decisión para usar en el equipo 'boosted', o sea, empezando cada árbol con una partición diferente.

-Random Forest: Se ha utilizado la función *randomForest* del paquete "randomForest", en la cual se tenía que especificar el número de árboles que

contenía, en nuestro caso 1000 (ntree=1000). Finalmente se realizó una validación cruzada para mejorar la precisión del modelo.

Mediante la función *ConfusionMatrix* del paquete “caret”, se ha evaluado el rendimiento de los modelos, obteniendo su precisión diagnóstica (*Accuracy*) y concordancia entre la predicción de los modelos y los valores reales (*Kappa*). Los valores del índice *Kappa* van de 0 (concordancia nula) a 1 (concordancia perfecta). Además, mediante el test de McNemar se ha evaluado si la predicción tiende a clasificar a los pacientes hacia un grupo u otro.

3. Resultados

Los resultados se estructuran de la siguiente forma, para cada una de las variables principales (ictus/mimic y tipo de ictus: isquémico/hemorrágico):

1. Se realiza un análisis descriptivo de la muestra
2. Se muestran las asociaciones de las variables clínicas con el grupo de clasificación
3. Se muestra la distribución de cada uno de los biomarcadores por grupo de clasificación y se evaluan diferencias significativas
4. Se aplican las técnicas de clasificación sólo en los biomarcadores: De esta forma determinamos la fuerza que podrían tener los biomarcadores por sí solos antes de realizar la exploración médica o neurológica.
5. Se aplican las técnicas de clasificación sólo a las variables clínicas: Ésto nos permite saber como clasificaríamos a los pacientes si no tuviéramos biomarcadores, para comprobar más adelante cuál es el valor añadido de éstos y saber si vale la pena determinar esos biomarcadores (con lo que comporta de gasto de tiempo y dinero)
6. Se aplican las técnicas de clasificación a las variables clínicas + biomarcadores: Para obtener la mejor precisión posible y conocer si ha valido la pena la obtención de los marcadores.
7. Comparación del rendimiento de los diferentes algoritmos para cada grupo de variables analizadas.
8. Comparación con resultados de análisis de regresión logística publicado.

3.1 Clasificación Ictus-Mimics

3.1.1 Análisis descriptivo

Del total de 1307 pacientes, 717 son hombres y 590 mujeres. La edad media es de 70,5 años y los factores de riesgo más frecuentes son la hipertensión arterial (n=940), la dislipemia (n=615), la fibrilación auricular (n=378) y la diabetes mellitus (n=329). Además, 226 son fumadores, 236 tienen ictus previo y 188 tienen discapacidad previa. Las medias de tensión arterial sistólica y diastólica son 157,1 y 84,2 mmHg respectivamente, el nivel medio de glucemia es de 135 mg/dl y la puntuación mediana de la escala NIHSS de 6 (0-42).
La variable respuesta, de la cual nos interesa predecir su clasificación, es el diagnóstico final, del cual obtenemos que 1115 son ictus y 192 son mimics, que representan un 85.31% y un 14.69% de la muestra respectivamente.

3.1.2 Relación variables clínicas con grupos de clasificación

Los ictus son mucho mayores que los mimics, con una diferencia en la media de edad de casi 10 años (72.0 vs 62.3) y un p-valor que es estadísticamente muy significativo ($p<0.001$, t-test de Welch).

Hay mayor porcentaje de hombres entre los ictus (56,2%) que entre los mímics (46,9%). El test de la ji-cuadrado de Pearson nos revela que la mayor proporción de hombres mostrada en los ictus con respecto a los mimics es estadísticamente significativa ($p=0.016$) para un nivel de significación del 5%.

La mayoría de los factores de riesgo predominan más en los ictus que en los mímics. Los pacientes diagnosticados de ictus presentan significativamente más alcoholismo ($p=0.033$), hipertensión arterial ($p<0.001$), dislipemia ($p=0.001$) y fibrilación auricular ($p<0.001$). Por otro lado, entre los mimics habían más pacientes que presentaron ictus previos de forma significativa ($p=0.007$).

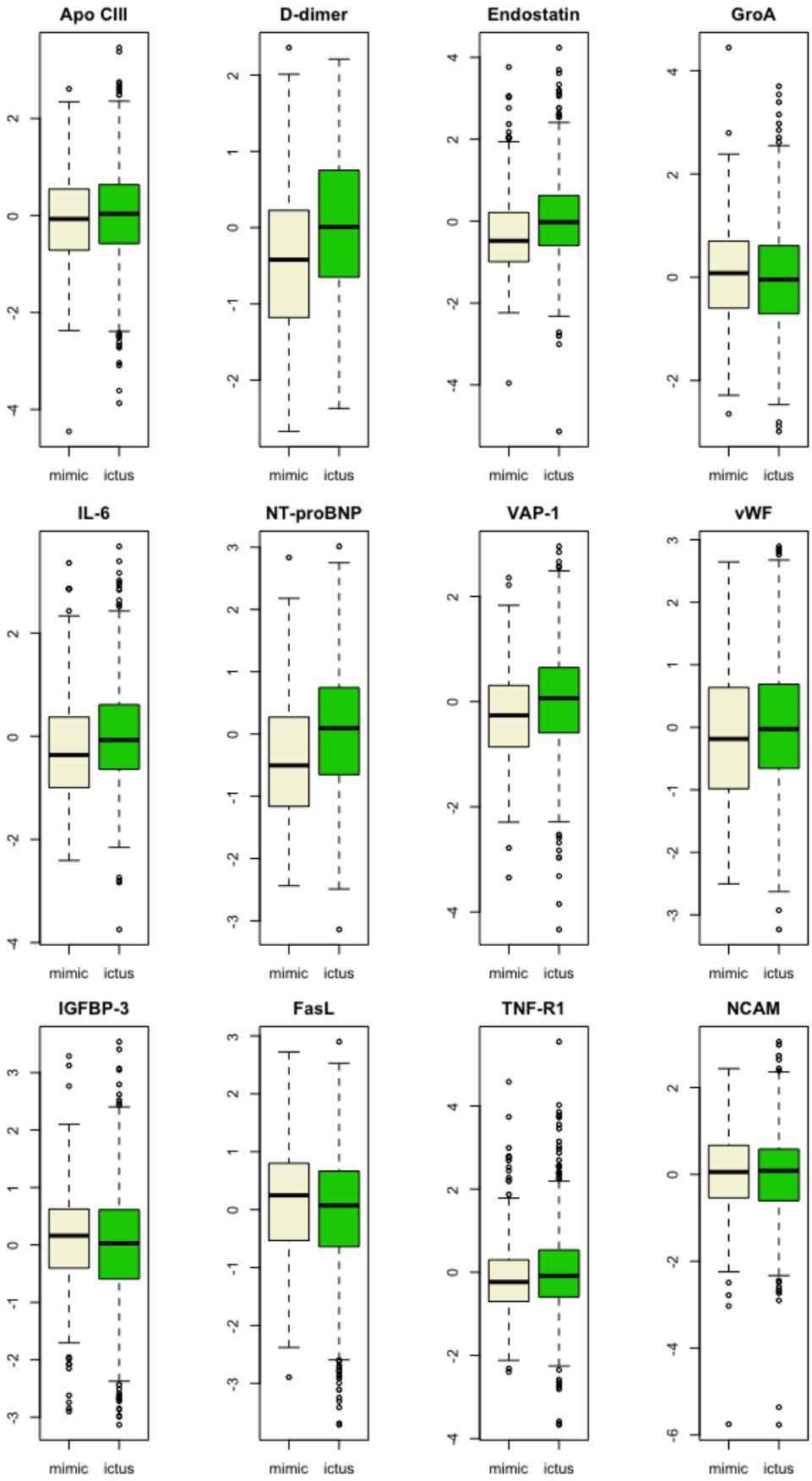
La discapacidad previa medida está medida por la escala mRS (modified Rankin Scale) con puntuaciones que van de 0 (paciente totalmente independiente) a 6 (éxitus). En este caso, todos los pacientes están inicialmente vivos, por lo que el mRS previo más alto que obtenemos es de 5, o sea, con un nivel de dependencia funcional máximo. No hay diferencias significativas entre los 2 grupos (p=0.733, U de Mann-Whitney). Para determinar si un paciente es funcionalmente dependiente o no, la escala mRS se suele dividir en puntuaciones <=2 (independiente) y <2 (dependiente). Tampoco hay diferencias significativas según esta categorizaciónn (p=0.624, ji-cuadrado de Pearson).

Los niveles de tensión arterial sistólica son más elevados entre los ictus. Las diferencias en la presión arterial sistólica son muy amplias, con una media que está incrementada en más de 12 mmHg en los ictus con respecto a los mimics (158.9 vs 146.0, p<0.001, t-test de Welch). Los ictus también muestran unos valores significativamente superiores en la presión arterial diastólica (p=0.012, t-test de Welch). En la glucemia no hay diferencias significativas entre los 2 grupos (p=0.155, t-test de Welch).

El grado de severidad del evento a la llegada a Urgencias se mide con la escala NIHSS. Esta escala puede llegar a puntuar de 0 a 42 puntos, considerando 0 un paciente que prácticamente no tiene síntomas de ictus y puntuaciones mayores de 30 a pacientes que muestran una gravedad muy acusada del ictus. Hay mucha diferencia en esta variable, mostrando más gravedad aquellos pacientes que acaban catalogándose de ictus. El test no paramétrico de la U de Mann Whitney nos revela que las diferencias en la severidad son muy significativas (p<0.001).

3.1.3 Distribución biomarcadores por grupo de clasificación

A continuación se muestra la distribución de los pacientes según su clasificación de ictus o mimics:


Apo CIII
D-dimer
Endostatin
GroA
IL-6
NT-proBNP
VAP-1
vWF
IGFBP-3
FasL
TNF-R1
NCAM
mimic
ictus

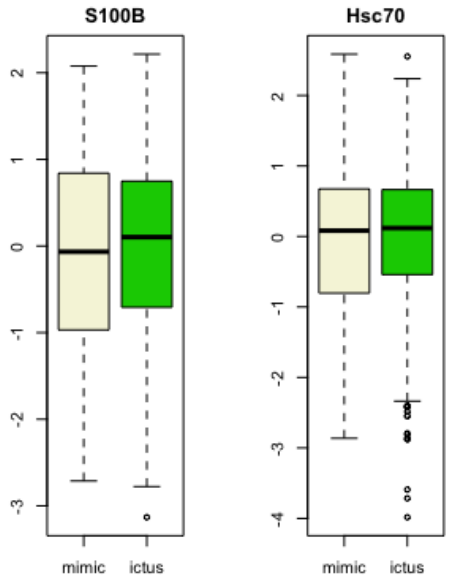


Figura 2. Distribución de los biomarcadores según grupo de clasificación (ictus/mimic)

Los marcadores que fueron significativamente más elevados en los pacientes con ictus con respecto a los mimics fueron: D-Dimer ($p<0.001$, t de Student), Endostatin ($p<0.001$, t-test de Welch), IL-6 ($p=0.004$, t-test de Welch), NT-proBNP ($p<0.001$, t de Student), VAP-1 ($p<0.001$, t de Student) y VWF ($p=0.041$, t-test de Welch). Por contra son los mimics los que tienen valores más altos de FasL ($p=0.011$).

3.1.4 Técnicas clasificación en biomarcadores

Se dividen las muestras en entrenamiento y prueba. Se utilizan los casos del Interim analysis para el training y el resto para la validación, como en el estudio original, por lo que hay 541 casos en el grupo de entrenamiento y 766 en el grupo de prueba.

Técnica 1: k-Nearest Neighbour

Como el grupo de entrenamiento incluye 541 casos, se utiliza primero la raíz cuadrada de éstos para establecer el valor de k, el cual es el valor más

estandarizado para especificar el número de vecinos que entran en votación para clasificar cada caso en un grupo u otro. Así, se entrena el algoritmo con una k=23, obteniendo una precisión global en la muestra de validación del 85.1%. El índice de concordancia Kappa es 0 y el test de McNemar es significativo, debido a que la predicción tiende a dar más diagnósticos de ictus. De hecho, el algoritmo clasifica a todos los pacientes como ictus, por lo que no es bueno para diagnosticar a estos pacientes. Así, se intenta mejorar el modelo modificando el valor de k en el algoritmo.

En la siguiente tabla se muestra un resumen de los resultados obtenidos probando con diferentes valores de k. Se puede observar que no hay otros valores de k que mejoren la predicción de nuestro algoritmo inicial, por lo que este algoritmo no nos clasifica bien a los pacientes, tendiendo a clasificar a los mimics como ictus. Sólo se alcanza un pobre índice máximo de Kappa de 0.058 con un valor de k igual 5, por lo que la concordancia entre los valores reales y los que predice el modelo también son muy débiles.

Valores k	% clasificados correctamente	Índice Kappa
1	79.5%	0.03
3	82.9%	0.01
5	84.6%	0.06
11	85.12%	0.01
16	85.12%	0
21	85.12%	0
26	85.12%	0

Tabla 2. Rendimiento de los algoritmos k-NN según el valor de k (biomarcadores en clasificación ictus/mimic)

Técnica 2: Naive Bayes

El algoritmo de Naive-Bayes sólo trabaja con variables categóricas, por lo que se han de categorizar todas las variables numéricas. En lugar de hacer una categorización arbitraria de ellas, se han realizado curvas ROC para determinar los puntos de corte óptimos de cada una de las variables mediante el valor que maximice la sensibilidad y especificidad para predecir ictus en la muestra de entrenamiento. Para establecer los puntos de corte se

ha calculado el valor máximo del estadístico de Youden.

El algoritmo de Naive Bayes muestra el siguiente rendimiento en la muestra de validación:

```
## Confusion Matrix and Statistics
##
##           Actual
## Predicción mimic ictus
##      mimic    23    65
##      ictus    91   587
```

Este algoritmo no rinde demasiado bien en la muestra de validación, con una precisión global de solo el 79.63%. La sensibilidad es alta, aunque tenemos una baja especificidad y un número elevado de falsos negativos, que son ictus clasificados como mimics.

Para mejorar el modelo, se podría probar de categorizar las variables numéricas de otra manera y ver si obtenemos un mejor algoritmo. No obstante, ya se ha categorizado por los puntos de corte óptimos de cada biomarcador, así que no es de esperar que se pueda mejorar la predicción con otros puntos de corte.

Otra forma de intentar mejorar el modelo es darle un valor al estimador de Laplace para comprobar si se puede mejorar la predicción. Este estimador añade un número a cada una de las frecuencias para asegurar que cada característica tenga una probabilidad superior a cero de ocurrir en cada clase de la variable respuesta. Si se entrena el algoritmo con un valor de *Laplace=1*, se obtiene el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##           Actual
## Predicción mimic ictus
##      mimic    22    67
##      ictus    92   585
```

Se observa que no hay una mejora en la predicción del modelo. La precisión global (79.24%) es igual que con *Laplace=0*, por lo que este estimador no aportaría más beneficio al algoritmo. Se ha probado con otros valores de Laplace, pero la precisión del modelo no varía.

Técnica 3: Artificial Neural Network

Para las redes neuronales artificiales, se utilizan los datos con los biomarcadores sin categorizar ya estandarizados de la muestra de entrenamiento y la muestra de prueba. Se entrena primero el algoritmo con **un nodo oculto** y se obtiene la siguiente red neuronal:

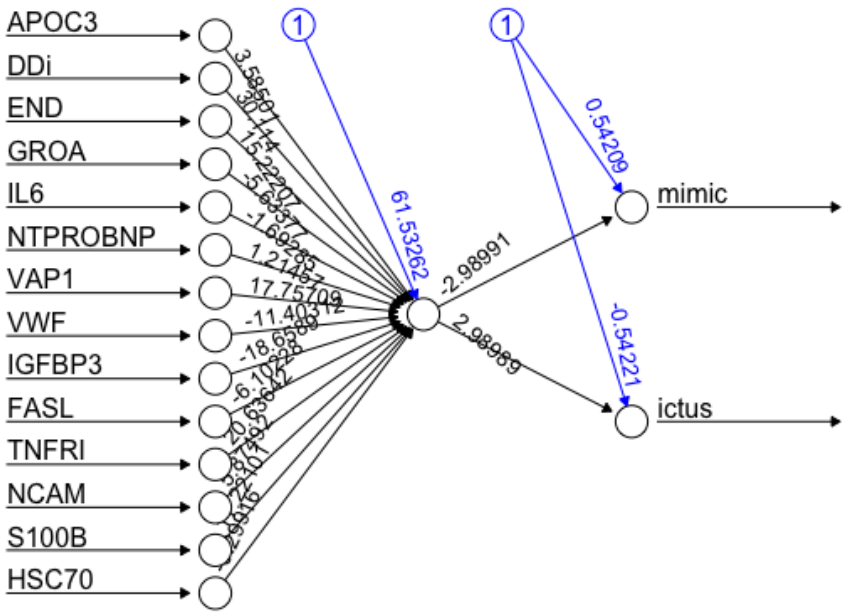


Figura 3. Red neuronal con un nodo oculto (biomarcadores en clasificación ictus/mimic)

En este modelo se han realizado un total de 5525 pasos con un error de estimación igual a 51.05. Los biomarcadores que tienen un mayor peso para predecir las clases de la variable respuesta son el D-dimer, la Endostatin, el VAP-1, el IGFBP3, el TNFR-1 y el NCAM. A continuación se muestra el rendimiento de la red en los datos de validación:

```
## Confusion Matrix and Statistics
##
##
## prediction mimic ictus
##      mimic    30    74
##      ictus    84   578
```

La precisión global del modelo es del 79.37% y el índice de concordancia

kappa es 0.16. No hay muy buena precisión y hay muchos falsos negativos.

El entrenamiento de un modelo neuronal artificial con **dos nodos ocultos** da como resultado la siguiente red:

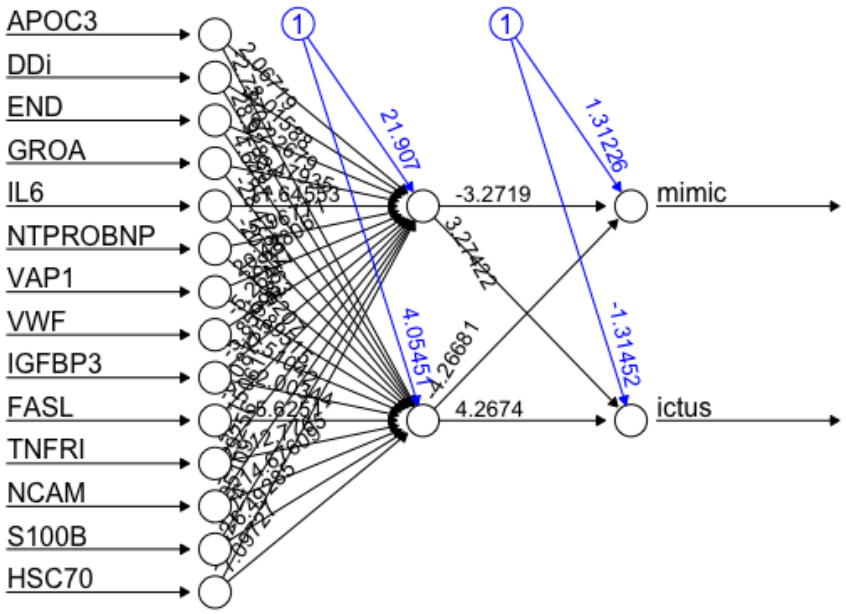


Figura 4. Red neuronal con dos nodos ocultos (biomarcadores en clasificación ictus/mimic)

En este modelo se han realizado un total de 2131 pasos con un error de estimación igual a 42.86, menor que cuando se utiliza sólo un nodo oculto. El resultado del rendimiento del modelo con 2 nodos ocultos es el siguiente:

```
## Confusion Matrix and Statistics
##
##
## prediction2 mimic ictus
##       mimic    31    73
##       ictus    83   579
```

La precisión global de este modelo es del 79.63% y el índice de concordancia Kappa es 0.17. No hay una mejora sustancial. Se ha probado con más nodos ocultos pero empeora el rendimiento del modelo.

Para intentar mejorarlo, se entrena el modelo con **dos nodos ocultos** mediante un **3-fold crossvalidation** y se obtiene el siguiente rendimiento en la muestra de validación:

```
## Confusion Matrix and Statistics
##
##             Bio_test_labels
## prediction3f mimic ictus
##        mimic    19    37
##        ictus    95   615
```

Se observa que la precisión global del nuevo modelo es del 82.77%, algo superior al modelo anterior. La sensibilidad es alta, pero la especificidad y el kappa son bajos (0.14). Por tanto, el modelo neuronal de 2 nodos ocultos obtenido con validación cruzada rinde algo mejor que el de 2 nodos sin validación cruzada, aunque el rendimiento general no acaba de ser bueno.

Técnica 4: Support Vector Machine (SVM)

Primero, se obtiene el algoritmo SVM entrenando los datos con una **función lineal** (kernel=”vanilladot”), obteniendo el siguiente rendimiento en la muestra de validación:

```
## Confusion Matrix and Statistics
##
##                   Bio_test_labels
## lineal_predictions mimic ictus
##              mimic     0     0
##              ictus   114   652
```

La precisión global del modelo SVM utilizando la función lineal es del 85.12%. Se aprecia que todos los pacientes los clasifica como ictus.

Otra de las funciones más utilizadas para obtener un buen rendimiento es la ***Gaussian RBF*** **(**función kernel de base radial gausiana), pero obtenemos exactamente el mismo rendimiento que con el algoritmo anterior.

Para intentar mejorar el modelo, se entrena primero el **modelo lineal con 3-fold crossvalidation** y se evalúa el rendimiento, obteniendo exactamente el mismo resultado que cuando no se realiza validación cruzada. Lo mismo pasa cuando se entrena el modelo con la función **RBF con 3-fold crossvalidation.**

Se prueba por último un 3-fold crossvalidation con un **Kernel polinomial** y se evalúa el rendimiento:

```
## Confusion Matrix and Statistics
##
##                     Bio_test_labels
## prediction3fsvmPoly mimic ictus
##               mimic    11    21
##               ictus   103   631
```

Se observa que el kernel polinomial ya clasifica a algunos mimics, con una precisión global del 83.81% y un índice de concordancia de 0.09. No obstante es insuficiente para el rendimiento que se espera obtener.

Técnica 5: Arbol de Decisión

El entrenamiento de los datos mediante un árbol de decisión nos da el siguiente resultado:

```
## Decision tree:
##
## NTPROBNP_men1.04 = > -1.04: ictus (460.8/52.9)
## NTPROBNP_men1.04 = <= -1.04:
## :...TNFRI_men0.499 = > -0.499: ictus (44.8/7.7)
##     TNFRI_men0.499 = <= -0.499:
##     :...VAP1_men0.0836 = <= -0.0836: mimic (22.5/7)
##         VAP1_men0.0836 = > -0.0836: ictus (12.9/2)
##
## Evaluation on training data (541 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       4   69(12.8%)   <<
##
##     (a)   (b)    <-classified as
##    ----  ----
##      15    63    (a): class mimic
##       6   457    (b): class ictus
##
##
##  Attribute usage:
##
##   98.52% NTPROBNP_men1.04
##   14.97% TNFRI_men0.499
##    7.39% VAP1_men0.0836
```

El árbol consta de 4 decisiones de profundidad. En este árbol se observa como se clasifican los pacientes según las diferentes combinaciones de las

variables. Seguidamente, se obtiene la matriz de confusión de los datos de entrenamiento y, al final, tenemos las características usadas para la clasificación. En la matriz de confusión, se observa que la tasa de error del árbol en los datos de entrenamiento es del 8.3% (45 mal clasificados). No obstante, los árboles de decisión tienden a sobreajustar el modelo en los datos de entrenamiento. De ahí la importancia de evaluarlos en un conjunto de prueba, tal como se muestra a continuación:

```
## Confusion Matrix and Statistics
##
##          Bio_test_labels
## Bio_pred mimic ictus
##    mimic     9    34
##    ictus   105   618
```

El rendimiento del árbol de clasificación en estos datos no es muy bueno, con una precisión global de solo el 81.85% y un índice de concordancia de 0.04.

Se intenta mejorar la precisión del árbol de decisión, con el aumento en el número de pruebas que trata de dividir los pacientes de otra forma para obtener diferentes algoritmos. Con el parámetro 'trials' se indica el número de árboles de decisión separados para usar en el equipo 'boosted'. El algoritmo para de añadir árboles si reconoce que los ensayos adicionales no parecen mejorar la precisión. Se utilizan 10 'trials', un número que parece bastante estandarizado, ya que pueden llegar a reducir las tasas de error en los datos de test sobre el 25%.

Con 10 iteraciones el modelo clasifica los individuos con una tasa de error de sólo el 3.3% (18 pacientes mal clasificados). El comportamiento del modelo en el conjunto de datos de prueba es el siguiente:

```
## Confusion Matrix and Statistics
##
##                  Bio_test_labels
## Bio_boost_pred10 mimic ictus
##             mimic     9    34
##             ictus   105   618
```

Se consigue mejorar el rendimiento del árbol de clasificación, incrementando la precisión hasta el 81.85%, aunque con un índice de concordancia bajo de 0.04. No obstante, el error de clasificación sigue siendo alto, del 18.15%.

Técnica 6: Random Forest

Se entrena el modelo con un Random Forest de 1000 árboles, obteniendo el siguiente resultado:

```
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 13.49%
## Confusion matrix:
##       mimic ictus    class.error
## mimic     6    72 0.923076923077
## ictus     1   462 0.002159827214
```

El Random Forest incluye 3 variables en cada árbol. La tasa de error es del 13.49%. Este error corresponde al 'out-of-bag error (OOB) rate', el cual es un estimador insesgado del error en un conjunto de datos de test. No obstante, se comprueba como rinde realmente el algoritmo en los datos de test:

```
## Confusion Matrix and Statistics
##
##              Bio_test_labels
## Biorf_predict mimic ictus
##         mimic     6    13
##         ictus   108   639
```

El rendimiento del algoritmo Random Forest no es muy bueno, con una precisión del 84.2% y un índice de concordancia de 0.05, por lo que el error de clasificación es del 15.8%, insuficiente para nuestras expectativas creadas.

Para mejorar el modelo, se realiza un 10-fold crossvalidation, se repite 10 veces y se escoge el mejor modelo. Además se prueban los modelos cambiando el parámetro de número de variables a entrar en los modelos para acabar seleccionando el mejor, obteniendo el siguiente resultado:

```
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 487, 486, 487, 486, 488, 487, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy      Kappa
##   2     0.8608388921  0.08430997024
##   4     0.8597205387  0.13346628944
##   6     0.8560023505  0.14313914286
##   8     0.8545344641  0.14993859655
```

El kappa más alto en un random forest ha sido de 0.15, produciéndose con un mtry=8 (número de variables explicativas introducidas). La precisión del modelo en los datos de entrenamiento es del 85.5%. El rendimiento en la muestra de validación es el siguiente:

```
## Confusion Matrix and Statistics
##
##                Bio_test_labels
## Biomrf_predict mimic ictus
##          mimic    13    23
##          ictus   101   629
```

El rendimiento del algoritmo Random Forest con un 10-fold cross-validation no mejora sustancialmente el anterior, alcanzando una precisión similar, del 83.81% y un índice de concordancia mayor, de 0.11, por lo que el error de clasificación sigue siendo alto.

Resumen de los resultados obtenidos con biomarcadores

En la siguiente tabla se muestra un resumen de los resultados que se han obtenido de cada algoritmo con sus variantes para clasificar el diagnóstico de los pacientes en los datos de test:

	Precisión	Kappa	Error
k-NN (k=23)	85.12%	0	14.88%
k-NN (k=5)	84.6%	0.06	15.4%
Naive-Bayes (mejor corte)	79.63%	0.11	20.37%
Naive-Bayes (Laplace=1)	79.24%	0.1	20.76%
ANN (1 nodo oculto)	79.37%	0.16	20.63%
ANN (2 nodos ocultos)	79.63%	0.17	20.37%
ANN (3-fold-cv)	82.77%	0.14	17.23%
SVM (lineal)	85.12%	0	14.88%
SVM (rbf)	85.12%	0	14.88%
SVM (3-fold-cv poly)	83.81%	0.09	16.19%
Árbol de decisión	81.85%	0.04	18.15%
A.decisión (boosted)	81.85%	0.04	18.15%
Random forest (1000 a.)	84.2%	0.05	15.8%
R.forest (10-fold-cv)	83.81%	0.11	16.19%

Tabla 3. Resumen del rendimiento obtenido por los diferentes algoritmos (biomarcadores en clasificación ictus/mimic)

No hay ningún algoritmo que rinda bien, por lo que parece que no hay ningún modelo basado sólo en biomarcadores que tenga una buena validación. Los algoritmos que rinden peor son el de k-NN y el SVM. Estos algoritmos casi no son capaces de detectar ningún mimic. La mayor precisión que tienen corresponde al % de ictus que hay en la muestra, por lo que no aportan nada al estudio. El algoritmo de Naive-Bayes muestra un mayor error, pero ya es capaz de detectar algunos mimics, aunque no conseguimos mejorar la predicción con *Laplace=1*. Se podrían categorizar los datos de otra forma, pero es difícil pensar que obtendremos una mejor clasificación, ya que se han utilizado los puntos de corte óptimos de cada biomarcador. En una red neuronal artificial, tampoco conseguimos alcanzar una buena precisión, pero también es capaz de detectar más mimics. Además, son los algoritmos donde encontramos valores de kappa más altos. Un mejor balance de precisión y kappa se alcanza con una red neuronal entrenada con 3-fold crossvalidation. Los árboles de decisión también muestran un pobre rendimiento. El Random Forest entrenado con un 10-fold cross-validation muestra un mejor balance de precisión y concordancia.

En resumen, los mejores modelos obtenidos para clasificar a los pacientes a partir de los biomarcadores son el Random Forest entrenado con un 10-fold cross-validation, el cual nos da un error de clasificación de sólo el 16.19%, obteniendo un índice de concordancia de k=0.11 y la red neuronal artificial de 2 capas ocultas con un 3-fold cross-validation con un error del 17.23% y un kappa=0.14. En general, la precisión de estos modelos es baja, con pobres índices de concordancia, por lo que parece que los biomarcadores por sí solos poco pueden aportar a la clasificación de ictus vs mimics.

3.1.5 Técnicas clasificación en variables clínicas

Técnica 1: k-Nearest Neighbour

Para aplicar el algoritmo kNN, el cual se basa en distancias, necesitamos que todas las variables clínicas estén en la misma escala. Así, se ha creado una nueva base con todas las variables numéricas normalizadas para que tomen

valores entre 0 y 1, tanto variables numéricas, como variables ordinales y nominales, en cuyo caso se transforman en variables *dummy*.

Se entrena el algoritmo con una k=23, obteniendo el siguiente rendimiento en la muestra de validación:

```
## Confusion Matrix and Statistics
##
##                  stroke_test_labels
## stroke_test_pred mimic ictus
##            mimic    11     6
##            ictus   103   646
```

La precisión global del modelo es del 85.77%, el índice de concordancia kappa del 0.13 y el test de McNemar es significativo, ya que la predicción tiende a dar más diagnósticos de ictus. Hay muy buena sensibilidad pero muy poca especificidad para la predicción de ictus

Se prueba con diferentes valores de k para intentar mejorar el rendimiento del modelo. En la siguiente tabla se muestra un resumen de los resultados obtenidos para cada valor de k. Se puede observar que hay otros valores de k que mejoran ligeramente la predicción de nuestro algoritmo inicial. El algoritmo con k=26 obtiene una mayor precisión. Se alcanza un kappa máximo con k=5, aunque la precisión es peor que la del modelo anterior.

Valores k	% clasificados correctamente	Índice Kappa
1	79.11%	0.14
3	83.03%	0.16
5	84.07%	0.17
11	84.73%	0.15
16	85.12%	0.12
21	85.25%	0.07
26	86.42%	0.16

Tabla 4. Rendimiento de los algoritmos k-NN según el valor de k (variables clínicas en clasificación ictus/mimic)

Técnica 2: Naive Bayes

Para entrenar el algoritmo de Naive-Bayes se categorizan las variables numéricas clínicas igual que con los biomarcadores, buscando un punto de corte óptimo para cada variable.

El algoritmo de Naive-Bayes muestra el siguiente rendimiento en la muestra de validación:

```
## Confusion Matrix and Statistics
##
##           Actual
## Predicción mimic ictus
##      mimic    36    53
##      ictus    78   599
```

La precisión global de solo el 82.9% y el índice Kappa de 0.26. La sensibilidad es alta y la especificidad ha mejorado aunque sigue habiendo un número elevado de falsos negativos, que son ictus clasificados como mimics.

Se intenta mejorar la predicción entrenando el algoritmo con *Laplace=1* y se obtiene el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##           Actual
## Predicción mimic ictus
##      mimic    37    55
##      ictus    77   597
```

Se mejora un poco el índice Kappa (0.26), y la precisión global del modelo (82.77%) es similar que con *Laplace=0*. Se ha probado con otros valores de Laplace, pero apenas se mejora la precisión del modelo.

Técnica 3: Artificial Neural Network

Las redes neuronales trabajan mejor con los datos agrupados alrededor de 0, por lo que se usan los datos normalizados que se han utilizado para entrenar los algoritmos k-NN. Se entrena el algoritmo con **un nodo oculto** y se obtiene una red neuronal con el siguiente rendimiento en la muestra de validación:

```
## Confusion Matrix and Statistics
##
## prediction mimic ictus
##      mimic    21    14
##      ictus    93   638
```

La precisión global del modelo es del 86.03% y el índice de concordancia kappa es 0.23.

El entrenamiento de un modelo neuronal artificial con **dos nodos ocultos** rinde de la siguiente manera:

```
## Confusion Matrix and Statistics
##
##
## prediction2 mimic ictus
##       mimic    38    74
##       ictus    76   578
```

La precisión global del modelo es del 80.42% y el índice de concordancia kappa de 0.22. No hay mejora. Además, se ha probado con más nodos ocultos, empeorando el rendimiento del modelo.

Se entrena el modelo con **dos nodos ocultos** mediante un **3-fold crossvalidation** y se obtiene el siguiente rendimiento en la muestra de validación:

```
## Confusion Matrix and Statistics
##
##             stroke_test_labels
## prediction3f mimic ictus
##        mimic    25    23
##        ictus    89   629
```

La precisión global del nuevo modelo es del 85.38%, superior al modelo anterior, pero sin mejorar al de 1 nodo oculto.

Técnica 4: Support Vector Machine (SVM)

Con un algoritmo SVM entrenado con una **función lineal** se obtiene el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##                   stroke_test_labels
## lineal_predictions mimic ictus
##              mimic     0     0
##              ictus   114   652
```

La precisión global del modelo SVM utilizando la función lineal es del 85.12% y el índice de concordancia es nulo, ya que todos los pacientes los clasifica como ictus.

Si se utiliza la **función RBF** para entrenar el modelo, se obtiene el siguiente resultado:

```
## Confusion Matrix and Statistics
##
##                stroke_test_labels
## rbf_predictions mimic ictus
##           mimic    12     5
##           ictus   102   647
```

El rendimiento del modelo con la función RBF mejora, ya que se aumenta la precisión y la concordancia.

Se intenta mejorar el algoritmo entrenando primero un **modelo lineal con 3-fold crossvalidation** y se evalúa el rendimiento, pero se obtiene exactamente el mismo rendimiento que cuando no se realiza validación cruzada. Si se entrena el modelo con la **función RBF con 3-fold crossvalidation** también se obtiene el mismo resultado que con el algoritmo sin validación cruzada, o sea, no hay mejora.

Se prueba por último un **SVM con un Kernel polinomial con 3-fold crossvalidation** y se evalúa el rendimiento:

```
## Confusion Matrix and Statistics
##
##                     stroke_test_labels
## prediction3fsvmPoly mimic ictus
##               mimic    16     9
##               ictus    98   643
```

Con el kernel polynomial se mejora levemente los anteriores algoritmos, alcanzando una precisión global del 86% y un kappa de 0.19.

Técnica 5: Arbol de Decisión

El entrenamiento de los datos mediante un árbol de decisión da el siguiente resultado:

```
## Decision tree:
##
## NIHSS_basal > 1: ictus (465.4/42.7)
## NIHSS_basal <= 1:
## :...HTA = No: mimic (29.3/7.1)
##     HTA = Sí: ictus (46.3/13.1)
##
##
## Evaluation on training data (541 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       3   63(11.6%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##      22    56    (a): class mimic
##       7   456    (b): class ictus
##
##
##  Attribute usage:
##
##   99.26% NIHSS_basal
##   14.60% HTA
```

El árbol consta de 3 decisiones de profundidad y está formado sólo por las variables NIHSS y la HTA. Con esta combinación, se observa que la tasa de error del árbol en los datos de entrenamiento es del 11.6% (63 mal clasificados). El rendimiento del árbol en los datos de prueba es el siguiente:

```
## Confusion Matrix and Statistics
##
##            stroke_test_labels
## stroke_pred mimic ictus
##       mimic    17    23
##       ictus    97   629
```

Se obtiene una precisión global del 84.33% y un índice de concordancia de 0.16.

Se intenta mejorar el rendimiento aumentando la precisión de los árboles de decisión. Con 10 iteraciones, el modelo clasifica los individuos con una tasa de error del 10.4%. El comportamiento del modelo en el conjunto de datos de prueba es el siguiente:

```
## Confusion Matrix and Statistics
##
##                     stroke_test_labels
## stroke_boost_pred10 mimic ictus
##               mimic    17    14
##               ictus    97   638
```

Se ha conseguido mejorar el rendimiento del árbol de clasificación, incrementando la precisión hasta el 85.51% y el índice Kappa a 0.18.

Técnica 6: Random Forest

Entrenamiento del modelo:

```
##
## Call:
##  randomForest(formula = stroke_train_labels ~ ., data =
stroke_train2,      ntree = 1000)
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 11.65%
## Confusion matrix:
##       mimic ictus   class.error
## mimic    21    57 0.73076923077
## ictus     6   457 0.01295896328
```

El Random Forest ha incluido 1000 árboles y ha probado 4 variables en cada uno. La tasa de error es del 11.65%. El rendimiento del modelo en los datos de prueba se muestra a continuación:

```
## Confusion Matrix and Statistics
##
##              stroke_test_labels
## Biorf_predict mimic ictus
##         mimic    16    14
##         ictus    98   638
```

El rendimiento del algoritmo Random Forest no es demasiado malo, con una precisión del 85.38% y un índice de concordancia de 0.17.

Para intentar mejorar el modelo, se realiza un 10-fold crossvalidation, obteniendo el siguiente resultado:

```
## Random Forest
##
## 541 samples
##  17 predictor
##   2 classes: 'mimic', 'ictus'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 487, 486, 487, 486, 488, 487, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy      Kappa
##   2     0.8841651738  0.2926101349
##   4     0.8856603774  0.3407876903
##   6     0.8823231688  0.3482374638
##   8     0.8799293565  0.3512248558
##
## Kappa was used to select the optimal model using the largest
value.
## The final value used for the model was mtry = 8.
```

El kappa más alto en el algoritmo ha sido de k=0.351 y se ha producido con un mtry=8. La precisión del modelo es del 88%. El rendimiento del modelo en la muestra de validación es el siguiente:

```
## Confusion Matrix and Statistics
##
##               stroke_test_labels
## Biomrf_predict mimic ictus
##          mimic    21    19
##          ictus    93   633
```

El algoritmo Random Forest con un 10-fold crossvalidation mejora ligeramente el anterior, alcanzando una precisión del 85.38% y un índice de concordancia de 0.21.

Resumen de los resultados obtenidos con variables clínicas

En la siguiente tabla se muestra un resumen de los resultados que se han obtenido de cada algoritmo con sus variantes para clasificar el diagnóstico de los pacientes en los datos de prueba:

	Precisión	kappa	Error
k-NN (k=23)	85.77%	0.13	14.23%
k-NN (k=5)	84.07%	0.17	15.93%
k-NN (k=26)	86.42%	0.16	13.58%
Naive-Bayes (mejor corte)	82.9%	0.26	17.1%
Naive-Bayes (Laplace=1)	82.77%	0.26	17.23%
ANN (1 nodo oculto)	86.03%	0.23	13.97%
ANN (2 nodos ocultos)	80.42%	0.22	19.58%
ANN (3-fold-cv)	85.38%	0.24	14.62%
SVM (lineal)	85.12%	0	14.88%
SVM (rbf)	86.03%	0.15	13.97%
SVM (3-fold-cv poly)	86.03%	0.19	13.97%
Árbol de decisión	84.33%	0.16	15.67%
A.decisión (boosted)	85.51%	0.18	14.49%
Random forest (1000 a.)	85.38%	0.17	14.62%
R.forest (10-fold-cv)	85.38%	0.21	14.62%

Tabla 5. Resumen del rendimiento obtenido por los diferentes algoritmos (variables clínicas en clasificación ictus/mimic)

Los algoritmos que parecen que rinden mejor son las redes neuronales artificiales, específicamente aquella con 1 nodo oculto o la creada con validación cruzada. El peor algoritmo es el SVM entrenado con una función lineal, el cual clasifica a todos los pacientes como ictus. En general, la precisión de los modelos es relativamente baja, con pobres índices de concordancia, por lo que la clínica por sí sola no nos permite obtener una buena clasificación de los pacientes.

3.1.6 Técnicas clasificación en Clínica + Biomarcadores

Técnica 1: k-Nearest Neighbour

Se entrena el algoritmo con una k=23 tomando todos los biomarcadores y variables clínicas como variables explicativas y se obtiene el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##                  stroke_test_labels
## stroke_test_pred mimic ictus
##            mimic     4     2
##            ictus   110   650
```

La precisión global del modelo es del 85.38% y el índice de concordancia kappa es 0.05. Hay muy buena sensibilidad (99.6%) pero muy poca especificidad (3.5%). Se intenta mejorar el modelo probando con diferentes valores de k. En la siguiente tabla se muestra un resumen de los resultados obtenidos para cada valor de k. Se puede observar que hay otros valores de k que mejoran ligeramente la predicción del algoritmo inicial. Con el algoritmo con k=26 se obtiene una mayor precisión. Se alcanza un kappa máximo con k=11.

Valores k	% clasificados correctamente	Índice Kappa
1	78.72%	0.08
3	81.59%	0.03
5	83.29%	0.07
11	84.99%	0.09
16	85.12%	0.08
21	85.12%	0.02
26	85.64%	0.07
31	85.12%	0.01

Tabla 6. Rendimiento de los algoritmos k-NN según el valor de k (variables clínicas+biomarcadores en clasificación ictus/mimic)

Técnica 2: Naive Bayes

Se procede a realizar el algoritmo de Naive-Bayes con todas las variables categorizadas, y se comprueba cuál es su rendimiento en los datos de prueba:

```
## Confusion Matrix and Statistics
##
##           Actual
## Predicción mimic ictus
##      mimic    47    89
##      ictus    67   563
```

Parece que este algoritmo no rinde demasiado bien, con una precisión global de solo el 79.63%. La sensibilidad es alta y la especificidad ha mejorado,

pero hay un número elevado de falsos negativos, que son ictus clasificados como mimics.

Se prueba la predicción del modelo con *laplace=1* en los mismos datos categorizados:

```
## Confusion Matrix and Statistics
##
##           Actual
## Predicción mimic ictus
##      mimic    46    86
##      ictus    68   566
```

Mejora un poco el índice Kappa (0.255), pero la precisión global del modelo (79.9%) es igual que con *Laplace=0*. Se ha probado con otros valores de Laplace, pero apenas se mejora la precisión del modelo.

Técnica 3: Artificial Neural Network

Se entrena primero el algoritmo con **un nodo oculto** y se obtiene la siguiente red neuronal:

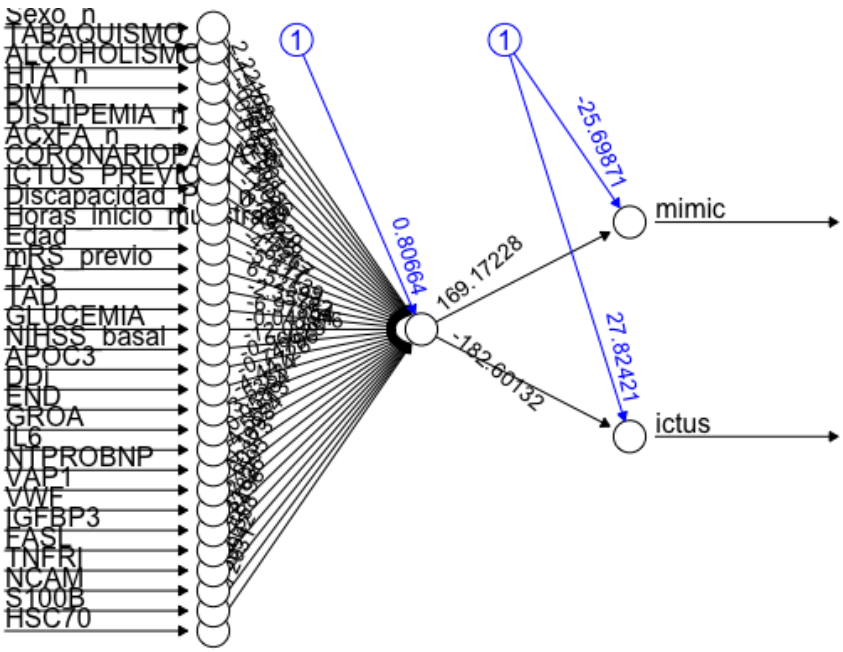


Figura 5. Red neuronal con un nodo oculto (variables clínicas+biomarcadores en clasificación ictus/mimic)

En este modelo se han realizado un total de 3823 pasos con un error de estimación igual a 34. Rendimiento de la red:

```
## Confusion Matrix and Statistics
##
## prediction mimic ictus
##      mimic    40    45
##      ictus    74   607
```

La precisión global del modelo es del 84.46% y el índice de concordancia kappa es 0.31.

El entrenamiento de un modelo neuronal artificial con dos nodos ocultos da como resultado la siguiente red:

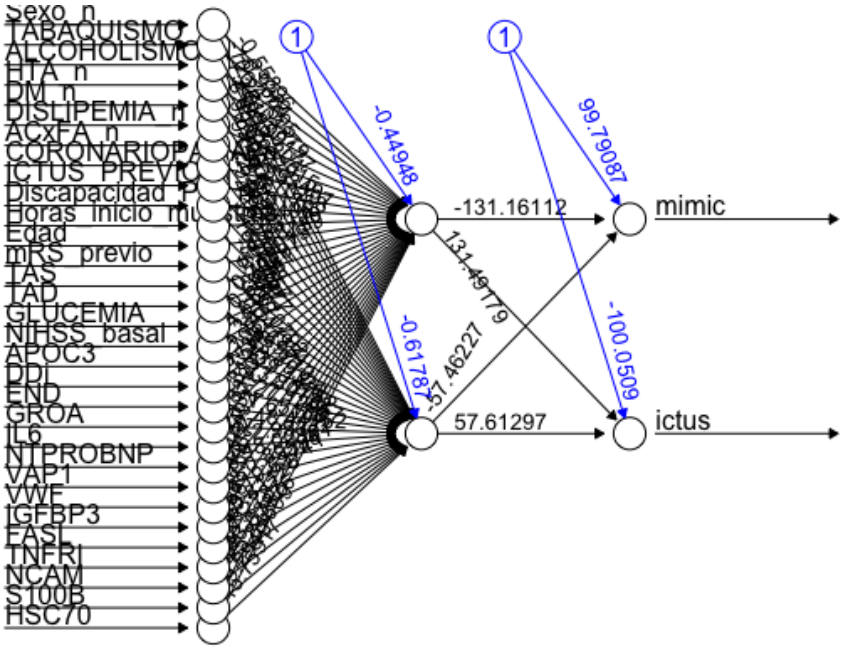


Figura 6. Red neuronal con dos nodos ocultos (variables clínicas + biomarcadores en clasificación ictus/mimic)

En este modelo se han realizado un total de 10402 pasos con un error de estimación igual a 25.02, menor que cuando utilizamos sólo un nodo oculto. El rendimiento de este algoritmo se muestra a continuación:

```
## Confusion Matrix and Statistics
##
## prediction2 mimic ictus
##       mimic    41    61
##       ictus    73   591
```

La precisión global del modelo es del 82.51% y el índice de concordancia Kappa del 0.28. No hay mejora. Además, con más nodos ocultos se empeora el rendimiento del modelo.

Se intenta mejorar el modelo con **dos nodos ocultos** mediante un **3-fold crossvalidation** y se obtiene el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##             stroke_test_labels
## prediction3f mimic ictus
##        mimic    34    39
##        ictus    80   613
```

La precisión global del nuevo modelo es del 84.46%, algo superior al modelo anterior, aunque sin mejorar al de 1 nodo oculto.

Técnica 4: Support Vector Machine (SVM)

El rendimiento de un algoritmo SVM entrenado con la **función lineal** es el siguiente:

```
## Confusion Matrix and Statistics
##
##                   stroke_test_labels
## lineal_predictions mimic ictus
##              mimic    27    23
##              ictus    87   629
```

La precisión global del modelo SVM utilizando la función lineal es del 85.64% y el índice de concordancia kappa es de 0.262.

El resultado del modelo SVM entrenado con la **función RBF** es el siguiente:

```
## Confusion Matrix and Statistics
##
##                stroke_test_labels
## rbf_predictions mimic ictus
##           mimic    10     4
##           ictus   104   648
```

Se aumenta levemente la precisión (85.9%) pero se empeora el kappa (0.127).

Se intenta mejorar el modelo SVM realizado con la función lineal mediante un 3-fold crossvalidation**,** pero se obtiene exactamente el mismo resultado que cuando no se realiza validación cruzada. Lo mismo ocurre cuando se realiza validación cruzada al modelo SVM con la función RBF.

Si se realiza un modelo SVM con **kernel polinomial y 3-fold crossvalidation,** el algoritmo rinde de la siguiente manera:

```
## Confusion Matrix and Statistics
##
##                     stroke_test_labels
## prediction3fsvmPoly mimic ictus
##               mimic     9     4
##               ictus   105   648
```

Así, con una precisión del 85.7% y un kappa de 0.114, se observa que el kernel polinomial tampoco mejora los anteriores algoritmos.

Técnica 5: Arbol de Decisión

Entrenamiento de los datos:

```
## Evaluation on training data (541 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##      30   19( 3.5%)   <<
##
##     (a)   (b)    <-classified as
##    ----  ----
##      61    17    (a): class mimic
##       2   461    (b): class ictus
##
##  Attribute usage:
##
##   99.26% NIHSS_basal
##   51.20% ICTUS_PREVIO
##   41.22% ACxFA
##   31.42% APOC3
##   28.47% HTA
##   27.36% VWF
##   22.37% Edad
##   12.01% GROA
##   10.35% END
##    9.61% IL6
##    8.87% Discapacidad_Pre
```

```
##    8.87% DDi
##    5.55% TAS
##    5.18% CORONARIOPATIA
##    4.44% DM
##    2.40% VAP1
##    1.85% mRS_previo
##    1.85% NCAM
##    1.66% Sexo
##    1.66% TAD
```

El árbol consta de 30 decisiones de profundidad. Se puede observar que, según la construcción del árbol, el NIHSS es la variable que más influye en la clasificación de los pacientes. Hay algunos biomarcadores que influyen también de forma importante, como el APOC3 y el VWF. En la matriz de confusión, se observa que la tasa de error del árbol en los datos de entrenamiento es del 3.5% (19 mal clasificados). A continuación se muestra el rendimiento del árbol en la muestra de validación:

```
## Confusion Matrix and Statistics
##
##             stroke_test_labels
## stroke_pred mimic ictus
##       mimic    30    62
##       ictus    84   590
```

El rendimiento del árbol de clasificación no es muy bueno, con una precisión global de solo el 80.94% y un índice de concordancia de 0.18.

En el intento por mejorar el rendimiento del modelo aumentando la precisión del árbol de decisión, se observa que con 10 iteraciones el modelo clasifica los individuos con una tasa de error nula, del 0%, o sea, no hay ningún paciente mal clasificado (ver anexos). El comportamiento del modelo en el conjunto de datos de prueba es el siguiente:

```
## Confusion Matrix and Statistics
##
##                      stroke_test_labels
## stroke_boost_pred10 mimic ictus
##                mimic    20    21
##                ictus    94   631
```

Se ha conseguido mejorar el rendimiento del árbol de clasificación, incrementando la precisión hasta el 84.99% y el índice de concordancia a 0.19. No obstante, el error de clasificación sigue siendo algo alto.

Técnica 6: Random Forest

Entrenamiento del algoritmo:

```
## Call:
##  randomForest(formula = stroke_train_labels ~ ., data =
stroke_train2,      ntree = 1000)
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 11.83%
## Confusion matrix:
##       mimic ictus    class.error
## mimic    15    63 0.807692307692
## ictus     1   462 0.002159827214
```

El Random Forest ha incluido 1000 árboles y ha probado 5 variables en cada uno. La tasa de error es del 11.83%. Rendimiento del modelo:

```
## Confusion Matrix and Statistics
##
##              stroke_test_labels
## Biorf_predict mimic ictus
##         mimic    16     8
##         ictus    98   644
```

Así, el Random Forest tiene una precisión de el 86.16% y un índice de concordancia de 0.19.

Para mejorar el modelo, se realiza un 10-fold crossvalidation, se repite 10 veces y se escoge el mejor modelo, obteniendo el siguiente resultado:

```
## Random Forest
##
## 541 samples
##  31 predictor
##   2 classes: 'mimic', 'ictus'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 487, 486, 487, 486, 488, 487, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy      Kappa
##   2     0.8717558605  0.1666872289
##   4     0.8791641573  0.2491604084
##   6     0.8812051966  0.2828012925
##   8     0.8808552824  0.2986181428
##
```

```
## Kappa was used to select the optimal model using the largest
value.
## The final value used for the model was mtry = 8.
```

El kappa más alto en un random forest ha sido de k=0.299 y se ha procucido con un mtry=8. La precisión del modelo es del 88.1%. El rendimiento en la muestra de validación es el siguiente:

```
##                stroke_test_labels
## Biomrf_predict mimic ictus
##          mimic    21     9
##          ictus    93   643
```

El rendimiento del algoritmo Random Forest con un 10-fold crossvalidation mejora ligeramente el anterior, alcanzando una precisión del 86.68% y un índice de concordancia de 0.24, por lo que el error de clasificación ya no es tan alto, del 13.32%.

Resumen de resultados obtenidos en clínica + biomarcadores

En la siguiente tabla se muestra un resumen de los resultados que se han obtenido de cada algoritmo con sus variantes para clasificar el diagnóstico de los pacientes en los datos de test:

	Precisión	**kappa**	**Error**
k-NN (k=23)	85.38%	0.05	14.62%
k-NN (k=26)	85.64%	0.07	14.36%
Naive-Bayes (mejor corte)	79.63%	0.26	20.37%
Naive-Bayes (Laplace=1)	79.9%	0.25	20.1%
ANN (1 nodo oculto)	84.46%	0.31	15.54%
ANN (2 nodos ocultos)	82.51%	0.28	17.49%
ANN (3-fold-cv)	84.46%	0.28	15.54%
SVM (lineal)	85.64%	0.26	14.36%
SVM (rbf)	85.9%	0.13	14.1%
SVM (3-fold-cv poly)	85.77%	0.11	14.23%
Árbol de decisión	80.94%	0.18	19.06%
A.decisión (boosted)	84.99%	0.19	15.01%
Random forest (1000 a.)	86.16%	0.19	13.84%
R.forest (10-fold-cv)	86.68%	0.24	13.32%

Tabla 7. Resumen del rendimiento obtenido por los diferentes algoritmos (variables clínicas+biomarcadores en clasificación ictus/mimic)

El algoritmo que mejor rinde es un Random Forest entrenado con 10-fold CV, con una precisión del 86.7% y un índice Kappa de 0.24. El modelo con mayor concordancia es el obtenido con una red neuronal con 1 nodo oculto, el cual tiene un índice Kappa de 0.31. Los modelos con peor concordancia son los entrenados con k-NN, y los de peor precisión los entrenados con Naive-Bayes. No hay ningún algoritmo con el que lleguemos a superar el 90% de precisión en la validación.

En general, la precisión de los modelos es baja, con pobres índices de concordancia, por lo que parece que aún añadiendo los biomarcadores a la clínica no se obtiene una buena clasificación de los pacientes.

3.1.7 Comparación rendimiento de técnicas en los diferentes grupos de variables (Biomarcadores, Clínica y Clínica+Biomarcadores)

En la siguiente tabla se muestra el rendimiento de los modelos construidos con biomarcadores, variables clínicas y con biomarcadores añadidos a la clínica:

	Biomarcadores		**Clínica**		**Clínica + Biomarcadores**	
	Precisión	**kappa**	**Precisión**	**kappa**	**Precisión**	**kappa**
k-NN (k=23)	85.12%	0	85.77%	0.13	85.38%	0.05
k-NN (k=5)	84.6%	0.06	84.07%	0.17	83.29%	0.07
k-NN (k=26)	85.12%	0	86.42%	0.16	85.64%	0.07
Naive-Bayes (mejor corte)	79.63%	0.11	82.9%	0.26	79.63%	0.26
Naive-Bayes (Laplace=1)	79.24%	0.1	82.77%	0.26	79.9%	0.25
ANN (1 nodo oculto)	79.37%	0.16	86.03%	0.23	84.46%	0.31
ANN (2 nodos ocultos)	79.63%	0.17	80.42%	0.22	82.51%	0.28
ANN (3-fold-cv)	82.77%	0.14	85.38%	0.24	84.46%	0.28
SVM (lineal)	85.12%	0	85.12%	0	85.64%	0.26
SVM (rbf)	85.12%	0	86.03%	0.15	85.9%	0.13
SVM (3-fold-cv poly)	83.81%	0.09	86.03%	0.19	85.77%	0.11
Árbol de decisión	81.85%	0.04	84.33%	0.16	80.94%	0.18
A.decisión (boosted)	81.85%	0.04	85.51%	0.18	84.99%	0.19
Random forest (1000 a.)	84.2%	0.05	85.38%	0.17	86.16%	0.19
R.forest (10-fold-cv)	83.81%	0.11	85.38%	0.21	86.68%	0.24

Tabla 8. Rendimiento obtenido por los algoritmos en los diferentes grupos de variables para clasificar ictus/mimic

Se observa que el rendimiento de los algoritmos entrenados sólo con las variables clínicas es mejor que el rendimiento de los biomarcadores. Cuando se añaden los biomarcadores a la clínica, no se consigue mejorar el rendimiento de todos los algoritmos. Sólo se mejora el rendimiento en la red neuronal artificial entrenada con 2 nodos ocultos, el SVM entrenado con la función lineal y los Random Forests construido con 1000 árboles, sin y con un 10-fold crossvalidation.

3.1.8 Comparación con estudio publicado en Stroke

En el análisis del estudio con los mismos datos publicados en Stroke, el rendimiento del modelo obtenido mediante regresión logística dio como resultado una precisión del 85.9% y una concordancia de k=0.085 en la clasificación de ictus/mimic.

El mejor modelo obtenido mediante las técnicas de Machine Learning utilizadas en este estudio ha tenido un rendimiento algo superior al modelo de regresión logística, con una precisión del 86.7% y una k=0.24, por lo que se podría decir que los Random Forests mejoran ligeramente la predicción realizada para ictus/mimic.

3.2 Clasificación Tipo Ictus: Isquémico-Hemorrágico

3.2.1 Análisis descriptivo

Del total de 1115 pacientes con ictus, 627 son hombres y 488 mujeres. La edad media es de 72 años y los factores de riesgo más frecuentes son la hipertensión arterial (n=831), la dislipemia (n=545), la fibrilación auricular (n=351) y la diabetes mellitus (n=282). 186 son fumadores, 188 tienen ictus previo y 163 tienen discapacidad previa. Las medias de tensión arterial sistólica y diastólica son 158,9 y 84,6 mmHg respectivamente, el nivel medio de glucemia es de 135,9 mg/dl y la puntuación mediana de la escala NIHSS es de 8 (0-42).

La variable respuesta, de la cual nos interesa predecir su clasificación, es el tipo de ictus, del cual obtenemos que 941 son isquémicos y 174 hemorrágicos, que representan un 84.4% y un 15.6% de la muestra respectivamente.

3.2.2 Relación variables clínicas con grupos de clasificación

Los ictus isquémicos son mucho mayores que los hemorrágicos, con una diferencia en la media de edad de 4.5 años (72.7 vs 68.2) y un p-valor que es estadísticamente muy significativo ($p<0.001$, t de Student).

Hay un mayor porcentaje de hombres entre los ictus hemorrágicos (69%) que entre los isquémicos (53,9%). El test de la ji-cuadrado nos revela que la mayor proporción de hombres mostrada en los ictus hemorrágicos con respecto a los isquémicos es estadísticamente significativa ($p=0.0002$).

Entre los factores de riesgo, hay un mayor porcentaje de hipertensión arterial en los ictus hemorrágicos ($p=0.032$). Por contra, en los ictus isquémicos se observan más pacientes con fibrilación auricular ($p<0.001$) y coronariopatía ($p=0.003$).

No hay diferencias significativas en la discapacidad previa de los pacientes entre los 2 grupos ($p=0.3$, U de Mann-Whitney).

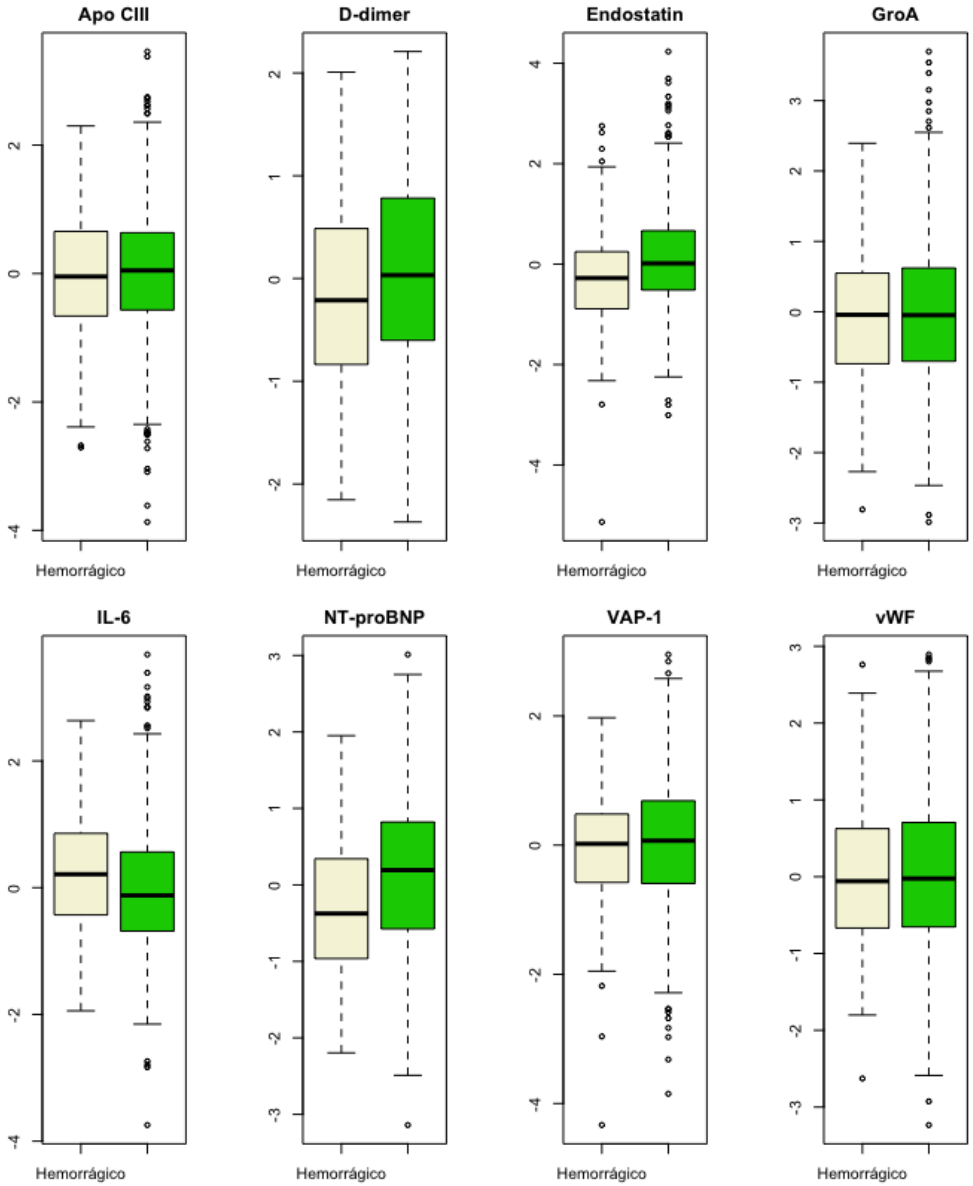
Las diferencias en la presión arterial sistólica son muy amplias, con una media que está incrementada en unos 18 mmHg en los ictus hemorrágicos

con respecto a los isquémicos (p<0.001, t de Student). Los ictus hemorrágicos también muestran valores significativamente superiores en la presión arterial diastólica (p<0.001, t-test de Welch) y en la glucemia (p<0.001, t-test de Welch).

Los pacientes con ictus hemorrágico muestran un mayor grado de severidad, con puntuaciones mucho más altas en la escala NIHSS (p<0.001, U de Mann-Whitney).

3.2.3 Distribución biomarcadores por grupo de clasificación

A continuación se muestra la distribución de los pacientes según su clasificación de ictus isquémico (verde) o hemorrágico (marfil):



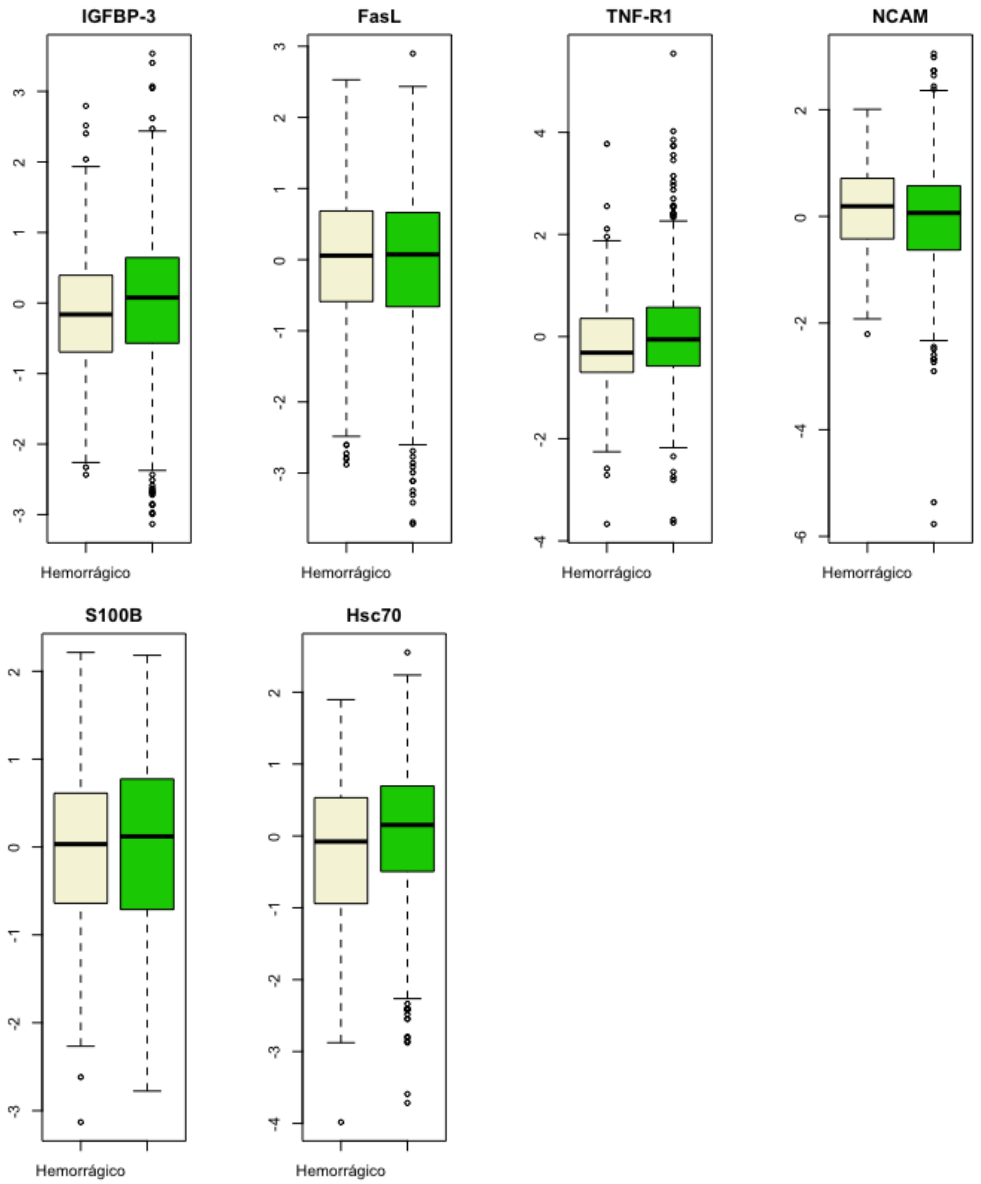


Figura 7. Distribución de los biomarcadores según grupo de clasificación (tipo de ictus)

Los marcadores que estuvieron significativamente más elevados en los pacientes con ictus isquémico fueron: D-Dimer (p=0.004, t de Student), Endostatin (p<0.001, t de Student), NT-proBNP (p<0.001, t de Student), TNFR-1 (p=0.002, t de Student) y HSC70 (p=0.005, t de Student). Por contra son los hemorrágicos los que tienen valores más altos de IL-6 (p=0.002).

3.2.4 Técnicas clasificación en biomarcadores

Se dividen las muestras en entrenamiento y prueba. Se analizan 463 casos en el grupo de entrenamiento y 652 en el grupo de prueba.

Técnica 1: k-Nearest Neighbour

Como el grupo de entrenamiento incluye 463 casos, se utiliza la raíz cuadrada de éstos para establecer el valor de k, que en este caso es 22. Tras entrenar el algoritmo se obtiene el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##              Bio_test_labels
## Bio_test_pred Hemorrágico Isquémico
##   Hemorrágico           0         0
##   Isquémico           100       552
```

La precisión global del modelo es del 84.66% y el índice de concordancia kappa es 0. El algoritmo está clasificando a todos los pacientes como ictus isquémico, por lo que no es bueno para diagnosticar a estos pacientes.

Se intenta mejorar el modelo modificando el valor k en el algoritmo. En la siguiente tabla se muestra un resumen de los resultados obtenidos probando con diferentes valores de k. Se puede observar que para valores de k pequeños se obtienen índices Kappa mayores. No obstante, los algoritmos tienden a clasificar los ictus hemorrágicos como isquémicos. Se alcanza un índice máximo de Kappa de 0.17 con un valor de k igual 3.

Valores k	% clasificados correctamente	Índice Kappa
1	78.22%	0.15
3	83.13%	0.17
5	84.51%	0.13
11	84.66%	0.03
16	84.82%	0.03
21	84.66%	0
26	84.66%	0

Tabla 9. Rendimiento de los algoritmos k-NN según el valor de k (biomarcadores en clasificación isquémico/hemorrágico)

Técnica 2: Naive Bayes

Como se ha comentado anteriormente, el algoritmo de Naive-Bayes sólo trabaja con variables categóricas, por lo que se categorizan todos los biomarcadores por el punto de corte óptimo. Seguidamente se entrena el algoritmo, obteniendo el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##              Actual
## Predicción   Hemorrágico Isquémico
##   Hemorrágico          10        19
##   Isquémico            90       533
```

Este algoritmo no rinde demasiado bien, con una precisión global de solo el 83.28%. La sensibilidad es alta, aunque hay una baja especificidad y un número elevado de falsos negativos, que son ictus isquémicos clasificados como hemorrágicos.

Si se intenta mejorar el modelo, entrenando el algoritmo con un valor de *Laplace=1*, se obtiene el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##              Actual
## Predicción   Hemorrágico Isquémico
##   Hemorrágico           9        19
##   Isquémico            91       533
```

No se mejora la predicción del modelo. La precisión global (83.13%) es peor que con *Laplace=0*, por lo que este estimador no aportaría más beneficio al algoritmo. La prueba con otros valores de Laplace, ha dado una mejor predicción cuando éste es igual a 8:

```
## Confusion Matrix and Statistics
##
##              Actual
## Predicción   Hemorrágico Isquémico
##   Hemorrágico           7         4
##   Isquémico            93       548
```

Así, entrenando el algoritmo con un *Laplace=8* se obtiene una precisión en la clasificación del tipo de ictus del 85,12% y un índice Kappa del 0.098.

Técnica 3: Artificial Neural Network

Entrenando el algoritmo con **un nodo oculto** se obtiene la siguiente red neuronal:

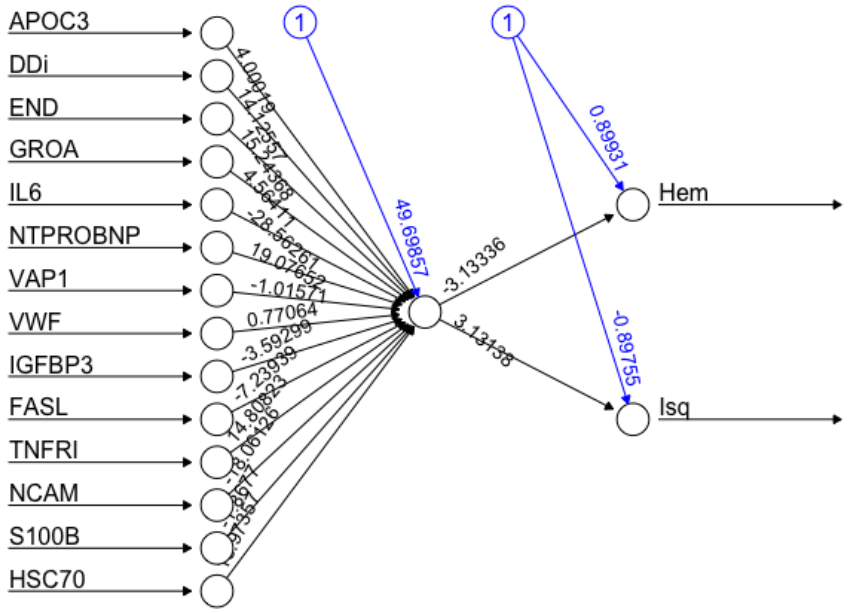


Figura 8. Red neuronal con un nodo oculto (biomarcadores en clasificación isquémico/hemorrágico)

En este modelo se han realizado un total de 4524 pasos con un error de estimación igual a 47.83. Rendimiento del modelo:

```
## Confusion Matrix and Statistics
##
##
## prediction    Hemorrágico Isquémico
##   Hemorrágico          35         55
##   Isquémico            65        497
```

La precisión global del modelo es del 81.6% y el índice de concordancia kappa es 0.26. No hay muy buena precisión y hay un alto número de falsos negativos.

Si se realiza el modelo neuronal con **dos nodos ocultos**, se obtiene la siguiente red:

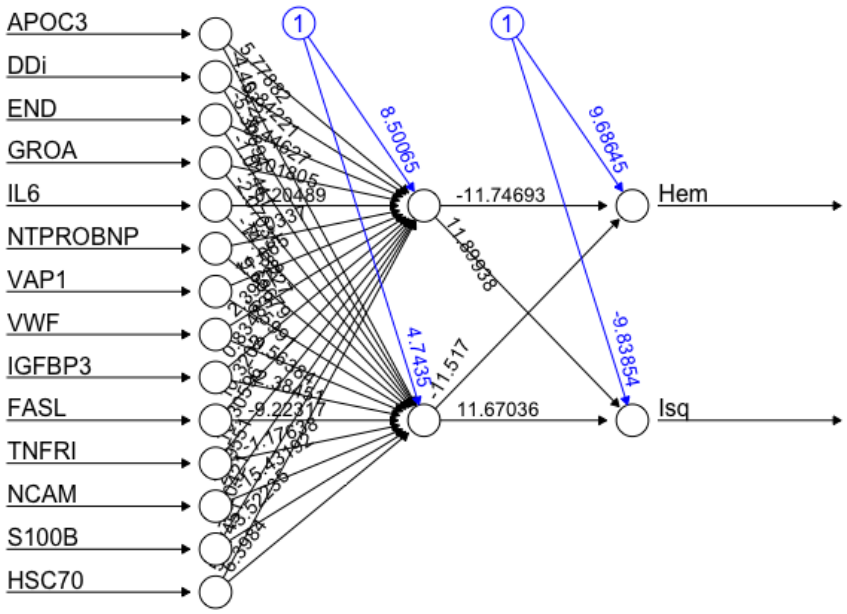


Figura 9. Red neuronal con dos nodos ocultos (biomarcadores en clasificación isquémico/hemorrágico)

En este modelo se han realizado un total de 1540 pasos con un error de estimación igual a 39.81, menor que cuando se utiliza sólo un nodo oculto.

Rendimiento del modelo neuronal con dos nodos ocultos:

```
## Confusion Matrix and Statistics
##
## prediction2   Hemorrágico Isquémico
##   Hemorrágico          18        47
##   Isquémico            82       505
```

La precisión global del modelo es del 80.21% y el índice de concordancia Kappa es 0.11. No hay mejora, siendo mejor el algoritmo con un nodo oculto. Se ha probado con más nodos ocultos pero no mejoran el rendimiento del modelo.

Para intentar mejorar el algoritmo, se entrena el modelo con **dos nodos ocultos** mediante un **3-fold crossvalidation** y se obtiene que la precisión global del nuevo modelo en la muestra de validación es del 84.66%, clasificando todos los ictus isquémicos como hemorrágicos. Con un índice Kappa=0, este modelo no rinde mejor que los anteriores modelos sin validación cruzada.

Técnica 4: Support Vector Machine (SVM)

Se obtiene el algoritmo SVM entrenando los datos con una **función lineal,** obteniendo el siguiente rendimiento en la muestra de validación:

```
## Confusion Matrix and Statistics
##
##                   Bio_test_labels
## lineal_predictions Hemorrágico Isquémico
##        Hemorrágico           0         0
##        Isquémico           100       552
```

La precisión global del modelo SVM utilizando la función lineal es del 84.66%. Todos los pacientes los clasifica como ictus isquémico.

Si se utiliza la **función RBF**, se obtiene exactamente el mismo rendimiento que con el algoritmo anterior.

Se intentan mejorar los modelos anteriores utilizando validación cruzada, pero se obtiene el mismo resultado. Lo mismo ocurre cuando se entrena un modelo con un **kernel polinomial y 3-fold crossvalidation.**

Técnica 5: Arbol de Decisión

El entrenamiento de los datos mediante un árbol de decisión da el siguiente resultado:

```
## Decision tree:
##  Isquémico (463/74)
##
## Evaluation on training data (463 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       1   74(16.0%)   <<
##
##     (a)   (b)    <-classified as
##    ----  ----
##             74    (a): class Hemorrágico
##            389    (b): class Isquémico
```

No se realiza ninguna partición significativa en el árbol. El rendimiento, por tanto, es nulo.

Técnica 6: Random Forest

Entrenamiento del modelo:

```
## Call:
##  randomForest(formula = Bio_train_labels ~ ., data = Bio_train2,
ntree = 1000)
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 17.06%
## Confusion matrix:
##             Hemorrágico Isquémico   class.error
## Hemorrágico           0        74 1.00000000000
## Isquémico             5       384 0.01285347044
```

El Random Forest ha incluido 1000 árboles, probando 3 variables en cada uno. La tasa de error es del 17.06%. A continuación se muestra cuál es su rendimiento en la muestra de prueba:

```
## Confusion Matrix and Statistics
##
##              Bio_test_labels
## Biorf_predict Hemorrágico Isquémico
##   Hemorrágico           4         5
##   Isquémico            96       547
```

El algoritmo de Random Forest tiene una precisión del 84.51% y un índice de concordancia de 0.05, por lo que el error de clasificación es del 15.49%.
Para mejorar el modelo, se realiza un 10-fold crossvalidation, se repite 10 veces y se selecciona el mejor algoritmo:

```
## Random Forest
##
## 463 samples
##  14 predictor
##   2 classes: 'Hemorrágico', 'Isquémico'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 416, 416, 417, 416, 417, 417, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy      Kappa
##   2     0.8359441875  -0.007778396484
##   4     0.8326879433   0.006505544868
##   6     0.8320452256   0.017611159356
##   8     0.8314067222   0.035607868831
```

```
##
## Kappa was used to select the optimal model using the largest
value.
## The final value used for the model was mtry = 8.
```

El Kappa más alto en un Random Forest ha sido de 0.035 y se ha producido con un mtry=8. La precisión del modelo es del 83.1%. El rendimiento del modelo en el grupo de prueba es el siguiente:

```
## Confusion Matrix and Statistics
##
##                Bio_test_labels
## Biomrf_predict Hemorrágico Isquémico
##    Hemorrágico           9         12
##    Isquémico            91        540
```

El Random Forest con un 10-fold cross-validation no mejora sustancialmente el anterior, alcanzando una precisión similar, del 84.2% y un índice de concordancia mayor, de 0.1. El error de clasificación es del 15.8%.

Resumen de los resultados obtenidos con biomarcarcadores

En la siguiente tabla se muestra un resumen de los resultados que se han obtenido de cada algoritmo con sus variantes para clasificar el diagnóstico de los pacientes en los datos de test:

	Precisión	**kappa**	**Error**
k-NN (k=22)	84.66%	0	15.34%
k-NN (k=5)	84.51%	0.13	15.49%
Naive-Bayes (mejor corte)	83.28%	0.09	16.72%
Naive-Bayes (Laplace=8)	85.12%	0.1	14.88%
ANN (1 nodo oculto)	81.6%	0.26	18.4%
ANN (2 nodos ocultos)	80.21%	0.11	19.79%
ANN (3-fold-cv)	84.66%	0	15.34%
SVM (lineal)	84.66%	0	15.34%
SVM (rbf)	84.66%	0	15.34%
SVM (3-fold-cv poly)	84.66%	0	15.34%
Árbol de decisión	84.66%	0	15.34%
A.decisión (boosted)	84.66%	0	15.34%
Random forest (1000 a.)	84.51%	0.05	15.49%
R.forest (10-fold-cv)	84.2%	0.1	15.8%

Tabla 10. Resumen del rendimiento obtenido por los diferentes algoritmos (biomarcadores en clasificación isquémico/hemorrágico)

No hay ningún algoritmo que rinda bien, por lo que parece que no hay ningún modelo basado sólo en biomarcadores que tenga una buena validación. Los algoritmos que rinden peor son el SVM y los árboles de decisión, los cuales no son capaces de detectar ningún ictus hemorrágico. El algoritmo de Naive-Bayes con Laplace=8 es el que muestra un menor error, capaz de dectar algunos ictus hemorrágicos. En una red neuronal artificial, tampoco se consigue alcanzar una buena precisión, pero también es capaz de detectar más ictus hemorrágicos Además, son los algoritmos donde se obtiene un valor de kappa más alto, aunque no se mejora con una red neuronal entrenada con validación cruzada. Los Random Forest también muestran un pobre rendimiento.

En general, la precisión de los modelos es relativamente baja, con un pobre índice de concordancia, por lo que parece que los biomarcadores por sí solos poco pueden aportar a esta clasificación.

3.2.5 Técnicas clasificación en variables clínicas

Técnica 1: k-Nearest Neighbour

Se normalizan los datos clínicos y se entrena el algoritmo con una k=22, obteniendo el siguiente rendimiento en los datos de prueba:

```
## Confusion Matrix and Statistics
##
##                  stroke_test_labels
## stroke_test_pred Hemorrágico Isquémico
##      Hemorrágico           0         0
##      Isquémico           100       552
```

La precisión global del modelo es del 84.66%, el índice de concordancia kappa es 0. No se detecta ningún ictus hemorrágico con este algoritmo.

En la siguiente tabla se muestra un resumen de los resultados obtenidos para diferentes valores de k. Se puede observar que hay otros valores de k que mejoran ligeramente la predicción del algoritmo inicial. Obtenemos mayor índice de Kappa con una k=1, aunque la precisión es bastante baja.

Valores k	% clasificados correctamente	Índice Kappa
1	75.31%	0.1
3	78.99%	0.07
5	81.6%	0.03
11	84.66%	0.03
16	84.51%	0
21	84.51%	0
26	84.66%	0

Tabla 11. Rendimiento de los algoritmos k-NN según el valor de k (variables clínicas en clasificación isquémico/hemorrágico)

Técnica 2: Naive Bayes

Se categorizan las variables clínicas, se entrena el algoritmo y se obtiene el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##              Actual
## Predicción   Hemorrágico Isquémico
##   Hemorrágico          21        22
##   Isquémico            79       530
```

La precisión global es del 84.51% y el índice Kappa es 0.22. La sensibilidad es alta y la especificidad ha mejorado aunque hay un alto número de falsos positivos, que son ictus hemorrágicos clasificados como isquémicos.

El rendimiento del modelo con *laplace=1* es el siguiente:

```
## Confusion Matrix and Statistics
##
##              Actual
## Predicción   Hemorrágico Isquémico
##   Hemorrágico          21        23
##   Isquémico            79       529
```

No mejora el algoritmo construido con *Laplace=0.* Se ha probado con otros valores de Laplace, pero no mejoran la precisión del modelo.

Técnica 3: Artificial Neural Network

Con el entrenamiento de un modelo neuronal artificial con **un nodo oculto**, se obtiene el siguiente rendimiento en la muestra de validación:

```
## Confusion Matrix and Statistics
##
##
## prediction    Hemorrágico Isquémico
##   Hemorrágico          40        52
##   Isquémico            60       500
```

La precisión global del modelo es del 82.82% y el índice de Kappa es 0.32.

Si se entrena el modelo con **dos nodos ocultos**, el rendimiento es el siguiente:

```
## Confusion Matrix and Statistics
##
##
## prediction2   Hemorrágico Isquémico
##   Hemorrágico          30        34
##   Isquémico            70       518
```

La precisión global del modelo es algo mayor, del 84.05%, aunque el índice de concordancia kappa es menor, del 0.28. El entrenamiento de la red con más nodos ocultos empeora su rendimiento.

Se intenta mejorar el rendimiento con un modelo neuronal artificial de **dos nodos ocultos con 3-fold crossvalidation,** obteniendo el siguiente resultado:

```
## Confusion Matrix and Statistics
##
##              stroke_test_labels
## prediction3f  Hemorrágico Isquémico
##   Hemorrágico          22         7
##   Isquémico            78       545
```

La precisión global del nuevo modelo es del 86.96%, superior a los modelos anteriores. El índice kappa es igual a 0.29.

Técnica 4: Support Vector Machine (SVM)

Con un algoritmo SVM entrenado con una **función lineal** se obtiene el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##                   stroke_test_labels
## lineal_predictions Hemorrágico Isquémico
##        Hemorrágico           0         0
##        Isquémico           100       552
```

La precisión global del modelo SVM utilizando la función lineal es del 84.66%. No hay buena predicción, ya que todos los ictus hemorrágicos los clasifica como isquémicos.

Si se utiliza la **función RBF** para entrenar el modelo, se obtiene el siguiente resultado:

```
## Confusion Matrix and Statistics
##
##                stroke_test_labels
## rbf_predictions Hemorrágico Isquémico
##     Hemorrágico           1         0
##     Isquémico            99       552
```

Hay sólo una mejoría muy leve en la precisión. Se intentan mejorar los modelos anteriores utilizando validación cruzada, pero no se obtiene un buen resultado. Lo mismo ocurre cuando se entrena un modelo con un **kernel polinomial y 3-fold crossvalidation**, con una baja precisión y una concordancia nula.

Técnica 5: Arbol de Decisión

Entrenamiento del modelo:

```
## Evaluation on training data (463 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##      13   50(10.8%)   <<
##
##     (a)   (b)    <-classified as
```

```
##    ----  ----
##      29    45    (a): class Hemorrágico
##       5   384    (b): class Isquémico
##
##  Attribute usage:
##
##  100.00% ALCOHOLISMO
##   93.09% ACxFA
##   66.74% NIHSS_basal
##   27.86% TAS
##   19.87% GLUCEMIA
##   18.14% Sexo
##    9.29% DM
##    7.34% ICTUS_PREVIO
##    5.83% Discapacidad_Pre
##    4.54% Horas_inicio_muestras
```

El árbol consta de 13 decisiones de profundidad. La tasa de error del árbol en los datos de entrenamiento es del 10.8% (50 mal clasificados). Hay varios condicionantes en el árbol, con mayor contribución el NIHSS basal y varios factores de riesgo como el alcoholismo y la fibrilación auricular. El rendimiento del árbol en los datos de validación es el siguiente:

```
## Confusion Matrix and Statistics
##
##              stroke_test_labels
## stroke_pred   Hemorrágico Isquémico
##   Hemorrágico          26        28
##   Isquémico            74       524
```

La precisión global del árbol es de solo el 84.36%, con un índice de concordancia de 0.26.

Para mejorar el algoritmo, se entrena el árbol con 10 iteraciones, y el modelo resultante clasifica los individuos con una tasa de error del 8.2%. El comportamiento del modelo en la muestra de validación es el siguiente:

```
## Confusion Matrix and Statistics
##
##                     stroke_test_labels
## stroke_boost_pred10 Hemorrágico Isquémico
##         Hemorrágico          17        19
##         Isquémico            83       533
```

No se ha conseguido mejorar el rendimiento del árbol de clasificación, manteniendo la precisión del 84.36% y empeorando el índice de concordancia a 0.18.

Técnica 6: Random Forest

Entrenamiento del modelo con un Random Forest de 1000 árboles:

```
## Call:
##  randomForest(formula = stroke_train_labels ~ ., data =
stroke_train2,      ntree = 1000)
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 15.98%
## Confusion matrix:
##             Hemorrágico Isquémico   class.error
## Hemorrágico           7        67 0.90540540541
## Isquémico             7       382 0.01799485861
```

El random forest de 1000 árboles ha probado 4 variables en cada uno. La tasa de error es del 15.98%. El modelo rinde como sigue:

```
## Confusion Matrix and Statistics
##
##              stroke_test_labels
## Biorf_predict Hemorrágico Isquémico
##   Hemorrágico           9         3
##   Isquémico            91       549
```

Con el algoritmo Random Forest se obtiene una precisión del 85.58% y un índice de concordancia de 0.13.

Para intentar mejorar el modelo, se realiza un 10-fold crossvalidation, obteniendo el siguiente resultado:

```
## Random Forest
##
## 463 samples
##  17 predictor
##   2 classes: 'Hemorrágico', 'Isquémico'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 416, 416, 417, 416, 417, 417, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy      Kappa
##   2     0.8443442286  0.06396845480
##   3     0.8436779731  0.08922155419
##   5     0.8456349060  0.12868374919
##   6     0.8439001953  0.12783999657
##   8     0.8451997122  0.14446051941
```

```
##
## Kappa was used to select the optimal model using the largest
value.
## The final value used for the model was mtry = 8.
```

El kappa más alto en un random forest ha sido de k=0.144 y se ha producido con un mtry=8. La precisión del modelo es del 84.5%. El rendimiento del modelo en el grupo de prueba es el siguiente:

```
## Confusion Matrix and Statistics
##
##                stroke_test_labels
## Biomrf_predict Hemorrágico Isquémico
##    Hemorrágico          13         5
##    Isquémico            87       547
```

El rendimiento del algoritmo Random Forest con un 10-fold cross-validation mejora ligeramente el anterior, alcanzando una precisión del 85.89% y un índice de concordancia de 0.18.

Resumen de los resultados obtenidos con variables clínicas

En la siguiente tabla se muestra un resumen de los resultados que se han obtenido de cada algoritmo con sus variantes para clasificar el diagnóstico de los pacientes en los datos de prueba:

	Precisión	kappa	Error
k-NN (k=22)	84.66%	0	15.34%
k-NN (k=5)	81.6%	0.03	18.4%
Naive-Bayes (mejor corte)	84.51%	0.22	15.49%
Naive-Bayes (Laplace=1)	84.36%	0.22	15.64%
ANN (1 nodo oculto)	82.82%	0.32	17.18%
ANN (2 nodos ocultos)	84.05%	0.28	15.95%
ANN (3-fold-cv)	86.96%	0.29	13.04%
SVM (lineal)	84.66%	0	15.34%
SVM (rbf)	84.82%	0.02	15.18%
SVM (3-fold-cv poly)	84.66%	0	15.34%
Árbol de decisión	84.36%	0.26	15.64%
A.decisión (boosted)	84.36%	0.18	15.64%
Random forest (1000 a.)	85.58%	0.13	14.42%
R.forest (10-fold-cv)	85.89%	0.18	14.11%

Tabla 12. Resumen del rendimiento obtenido por los diferentes algoritmos (variables clínicas en clasificación isquémico/hemorrágico)

Los algoritmos que rinden peor son el de k-NN y el SVM. Estos algoritmos casi no son capaces de detectar ningún ictus hemorrágico. El algoritmo de Naive-Bayes ya es capaz de detectar algunos hemorrágicos, aunque no se consigue mejorar la predicción con Laplace=1. Las redes neuronales son los algoritmos donde se encuentran valores de Kappa más altos. Un mejor balance de precisión y Kappa se alcanza con una red neuronal entrenada con validación cruzada. Los árboles de decisión y los Random Forest no llegan a superar los resultados de las redes neuronales.
En general, no hay ningún algoritmo que rinda muy bien, por lo que parece que un modelo basado sólo en características clínicas no tiene una buena validación.

3.2.6 Técnicas clasificación en Clínica + Biomarcadores

Técnica 1: k-Nearest Neighbour

Con el entrenamiento de un algoritmo k-NN con variables clínicas y biomarcadores se obtiene el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##                  stroke_test_labels
## stroke_test_pred Hemorrágico Isquémico
##      Hemorrágico           0         2
##      Isquémico           100       550
```

La precisión global del modelo es del 84.36% y el índice kappa es -0.01.

En la siguiente tabla se muestra un resumen de los resultados obtenidos para diferentes valores de k. Se puede observar que hay otros valores de k que mejoran ligeramente la predicción de nuestro algoritmo inicial. Se alcanza un kappa máximo con k=5, aunque el rendimiento es bastante pobre:

Valores k	% clasificados correctamente	Índice Kappa
1	74.39%	0.06
3	80.83%	0.08
5	83.74%	0.1
11	84.66%	0.04
16	84.51%	0.01
21	84.51%	0
26	84.66%	0

Tabla 13. Rendimiento de los algoritmos k-NN según el valor de k (variables clínicas+biomarcadores en clasificación isquémico/hemorrágico)

Técnica 2: Naive Bayes

El algoritmo de Naive-Bayes entrenado con variables clínicas y biomarcadores categorizados rinde de la siguiente manera:

```
## Confusion Matrix and Statistics
##
##              Actual
## Predicción   Hemorrágico Isquémico
##   Hemorrágico          43        47
##   Isquémico            57       505
```

La precisión global del algoritmo es del 84.05%, con un índice de kappa de 0.36. Se prueba la predicción del modelo con *laplace=1* y se obtiene el siguiente rendimiento:

```
## Confusion Matrix and Statistics
##
##              Actual
## Predicción   Hemorrágico Isquémico
##   Hemorrágico          41        47
##   Isquémico            59       505
```

No hay una mejora ni del índice Kappa ni de la precisión global del modelo (83.74%).

Se ha probado con otros valores de Laplace, pero no se mejora la precisión ni la concordancia del modelo. Así que el mejor algoritmo entrenado con Naive Bayes es el primero que se ha probado, con un *Laplace=0*.

Técnica 3: Artificial Neural Network

A continuación se muestra la red neuronal entrenada con **un nodo oculto:**

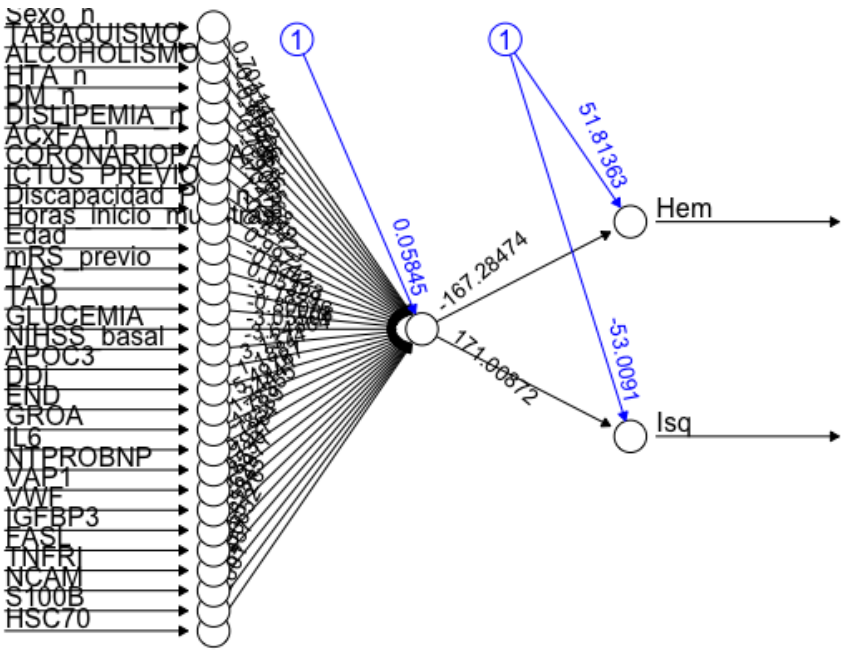


Figura 10. Red neuronal con un nodo oculto (variables clínicas + biomarcadores en clasificación isquémico/hemorrágico)

En este modelo se han realizado un total de 4705 pasos con un error de estimación igual a 31.01. Rendimiento de la red neuronal:

```
## Confusion Matrix and Statistics
##
## prediction    Hemorrágico Isquémico
##   Hemorrágico          43         43
##   Isquémico            57        509
```

La precisión global del modelo es del 84.66% y el índice de concordancia Kappa es 0.37.

El entrenamiento de un modelo neuronal artificial con **dos nodos ocultos** da como resultado la siguiente red:

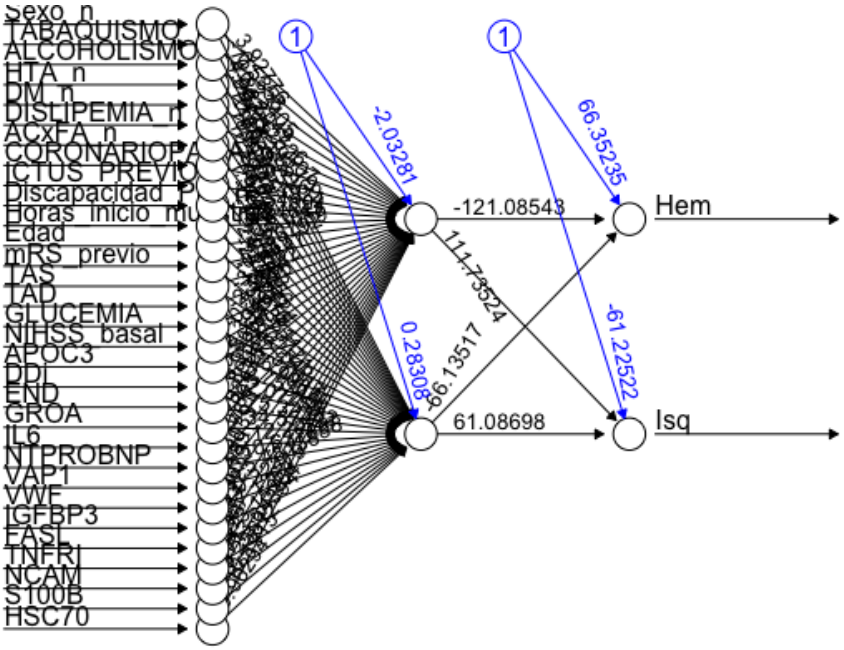


Figura 11. Red neuronal con dos nodos ocultos (variables clínicas + biomarcadores en clasificación isquémico/hemorrágico) (pág.70)

En este modelo se han realizado un total de 43830 pasos con un error de estimación igual a 24.98, menor que cuando se ha utilizado sólo un nodo oculto. Rendimiento del modelo con dos nodos ocultos:

```
## Confusion Matrix and Statistics
##
## prediction2   Hemorrágico Isquémico
##   Hemorrágico          36         43
##   Isquémico            64        509
```

La precisión global del modelo es del 83.59% y el índice de concordancia Kappa es 0.31. No hay mejora. Además, con más nodos ocultos se empeora el rendimiento del modelo.

El rendimiento de una red neuronal entrenada con **dos nodos ocultos y 3-fold crossvalidation** es el siguiente:

```
## Confusion Matrix and Statistics
##
##              stroke_test_labels
## prediction3f  Hemorrágico Isquémico
##   Hemorrágico          39         25
##   Isquémico            61        527
```

Así, la precisión global del nuevo modelo es del 86.81%, superior al modelo anterior, y mejorando al de 1 nodo oculto.

Técnica 4: Support Vector Machine (SVM)

Rendimiento del modelo SVM entrenado con una **función lineal**:

```
## Confusion Matrix and Statistics
##
##                    stroke_test_labels
## lineal_predictions Hemorrágico Isquémico
##        Hemorrágico          31         19
##        Isquémico            69        533
```

La precisión global del modelo SVM utilizando la función lineal es del 86.5%. En este caso, sí hay una precisión aceptable y un índice de concordancia moderado (Kappa=0.346). Un modelo SVM entrenado con la **función RBF** rinde mucho peor que el anterior con una precisión del 84.8% y un Kappa=0.016.

Se intenta mejorar el modelo SVM realizado con la función lineal mediante un 3-fold crossvalidation**,** pero se obtiene exactamente el mismo resultado que cuando no se realiza validación cruzada. La validación cruzada en el modelo SVM con la función RBF tampoco mejora el rendimiento. También se realiza un modelo SVM con **kernel polinomial y 3-fold crossvalidation**, pero el rendimiento es malo, siendo mejor el algoritmo entrenado con la función lineal.

Técnica 5: Arbol de Decisión

El árbol de decisión entrenado con variables clínicas y biomarcadores es el siguiente:

```
## Evaluation on training data (463 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##      17   37( 8.0%)   <<
```

```
##     (a)   (b)    <-classified as
##    ----  ----
##      41    33    (a): class Hemorrágico
##       4   385    (b): class Isquémico
##
##  Attribute usage:
##
##  100.00% ALCOHOLISMO
##   91.79% END
##   89.42% ACxFA
##   68.68% NIHSS_basal
##   25.27% TAD
##   16.41% TABAQUISMO
##   11.66% NCAM
##    9.94% Sexo
##    6.48% GLUCEMIA
##    5.40% DISLIPEMIA
##    3.02% TNFRI
##    2.38% mRS_previo
##    1.94% S100B
```

Este árbol consta de 17 decisiones de profundidad y tiene una tasa de error del 3.5% (19 mal clasificados). En la muestra de validación se comporta de la siguiente manera:

```
## Confusion Matrix and Statistics
##
##              stroke_test_labels
## stroke_pred   Hemorrágico Isquémico
##   Hemorrágico          24        31
##   Isquémico            76       521
```

El rendimiento del árbol de clasificación en estos datos es moderado, con una precisión global de solo el 83.59% y un índice de concordancia de 0.23.

En el intento por mejorar el rendimiento del modelo aumentando la precisión del árbol de decisión, se observa que con 10 iteraciones el modelo clasifica los individuos con una tasa de error del 1,1% (ver anexos). El comportamiento del modelo en la muestra de validación es el siguiente:

```
## Confusion Matrix and Statistics
##
##                     stroke_test_labels
## stroke_boost_pred10 Hemorrágico Isquémico
##         Hemorrágico          22        20
##         Isquémico            78       532
```

Se ha conseguido mejorar el rendimiento del árbol de clasificación, incrementando la precisión hasta el 84.97% y el índice Kappa a 0.24.

Técnica 6: Random Forest

El entrenamiento de un Random Forest da como resultado el siguiente modelo:

```
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 15.77%
## Confusion matrix:
##             Hemorrágico Isquémico    class.error
## Hemorrágico           2        72 0.972972972973
## Isquémico             1       388 0.002570694087
```

El Random Forest incluye 1000 árboles y prueba 5 variables en cada uno. La tasa de error es del 15.77%. En los datos de validación se comporta de la siguiente manera:

```
## Confusion Matrix and Statistics
##
##              stroke_test_labels
## Biorf_predict Hemorrágico Isquémico
##   Hemorrágico           5         1
##   Isquémico            95       551
```

El rendimiento del algoritmo no es demasiado bueno, con una precisión del 85.28%, un índice de concordancia bajo (kappa=0.08) y un error de clasificación del 14.72%.

Para intentar mejorar el modelo, se realiza un 10-fold crossvalidation, se repite 10 veces y se escoge el mejor modelo, obteniendo el siguiente:

```
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 416, 416, 417, 416, 417, 417, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy      Kappa
##   2     0.8406944187  0.004406779661
##   4     0.8415452770  0.014158320794
##   6     0.8404631514  0.026422711143
##   8     0.8393713640  0.033784189286
##
## Kappa was used to select the optimal model using the largest
value.
## The final value used for the model was mtry = 8.
```

El kappa más alto en un random forest ha sido de k=0.0337 y se ha producido con un mtry=8. La precisión del modelo es del 83.9% . El rendimiento en la muestra de validación es el siguiente

```
## Confusion Matrix and Statistics
##
##                stroke_test_labels
## Biomrf_predict Hemorrágico Isquémico
##    Hemorrágico           7         4
##    Isquémico            93       548
```

El rendimiento del algoritmo Random Forest con un 10-fold cross-validation mejora ligeramente el anterior, alcanzando una precisión del 85.12% y un índice de concordancia de 0.1, por lo que el error de clasificación ya no es tan alto, del 14.88%.

Resumen de resultados obtenidos con clínica + biomarcadores

En la siguiente tabla se muestra un resumen de los resultados que se han obtenido de cada algoritmo con sus variantes para clasificar el tipo de ictus de los pacientes en los datos de test:

	Precisión	kappa	Error
k-NN (k=22)	84.36%	-0.01	15.64%
k-NN (k=5)	83.74%	0.1	16.26%
Naive-Bayes (mejor corte)	84.05%	0.36	15.95%
Naive-Bayes (Laplace=1)	83.74%	0.34	16.26%
ANN (1 nodo oculto)	84.66%	0.37	15.34%
ANN (2 nodos ocultos)	83.59%	0.31	16.41%
ANN (3-fold-cv)	86.81%	0.4	13.19%
SVM (lineal)	86.5%	0.35	13.5%
SVM (rbf)	84.82%	0.02	15.18%
SVM (3-fold-cv poly)	85.89%	0.15	14.11%
Árbol de decisión	83.59%	0.23	16.41%
A.decisión (boosted)	84.97%	0.24	15.03%
Random forest (1000 a.)	85.28%	0.08	14.72%
R.forest (10-fold-cv)	85.12%	0.1	14.88%

Tabla 14. Resumen del rendimiento obtenido por los diferentes algoritmos (variables clínicas+biomarcadores en clasificación isquémico/hemorrágico)

El algoritmo que rinde mejor es el construido por la red neuronal artificial con un validación cruzada, el cual alcanza una precisión del 86.8% y un índice Kappa de 0.4. El SVM entrenado con la función lineal también rinde bastante bien, con valores que se aproximan a los obtenidos por la red neuronal. Los algoritmos que rinden peor son los k-NN y los árboles de decisión. El algoritmo de Naive-Bayes muestra un mayor error, pero concordancias moderadas, aunque no conseguimos mejorar la predicción con *Laplace=1*.

3.2.7 Comparación rendimiento de técnicas en los diferentes grupos de variables (Biomarcadores, variables clínicas y clínica+ biomarcadores)

En la siguiente tabla se muestra el rendimiento de los modelos construidos con los biomarcadores, las variables clínicas y con los biomarcadores añadidos a la clínica:

	Biomarcadores		**Clínica**		**Clínica + Biomarcadores**	
	Precisión	**kappa**	**Precisión**	**kappa**	**Precisión**	**kappa**
k-NN (k=22)	84.66%	0	84.66%	0	84.36%	-0.01
k-NN (k=5)	84.51%	0.13	81.6%	0.03	83.74%	0.1
Naive-Bayes (mejor corte)	83.28%	0.09	84.51%	0.22	84.05%	0.36
Naive-Bayes (Laplace=8 y 1)	85.12%	0.1	84.36%	0.22	83.74%	0.34
ANN (1 nodo oculto)	81.6%	0.26	82.82%	0.32	84.66%	0.37
ANN (2 nodos ocultos)	80.21%	0.11	84.05%	0.28	83.59%	0.31
ANN (3-fold-cv)	84.66%	0	86.96%	0.29	86.81%	0.4
SVM (lineal)	84.66%	0	84.66%	0	86.5%	0.35
SVM (rbf)	84.66%	0	84.82%	0.02	84.82%	0.02
SVM (3-fold-cv poly)	84.66%	0	84.66%	0	85.89%	0.15
Árbol de decisión	84.66%	0	84.36%	0.26	83.59%	0.23
A.decisión (boosted)	84.66%	0	84.36%	0.18	84.97%	0.24
Random forest (1000 a.)	84.51%	0.05	85.58%	0.13	85.28%	0.08
R.forest (10-fold-cv)	84.2%	0.1	85.89%	0.18	85.12%	0.1

Tabla 15. Rendimiento obtenido por los algoritmos en los diferentes grupos de variables para clasificar isquémico/hemorrágico (pág.75)

En general, el rendimiento de los algoritmos entrenados sólo con las variables clínicas es mejor que el rendimiento de los biomarcadores. Cuando se añaden los biomarcadores a la clínica, no se consigue mejorar el rendimiento de todos los algoritmos. Se mejora el rendimiento de las redes neuronales artificiales, algoritmos SVM y el árbol de decisión entrenado por Boosted. El mejor algoritmo obtenido, con un Kappa=0.4 y un error de clasificación del 13.19% es el entrenado mediante una red neuronal artificial con validación cruzada.

3.2.8 Comparación con estudio publicado en Stroke

En el análisis del estudio con los mismos datos publicados en Stroke, el rendimiento del modelo obtenido mediante regresión logística dio como resultado una precisión del 86.9% y una concordancia de k=0.230 en la clasificación de ictus isquémico/hemorrágico.
El mejor modelo obtenido mediante las técnicas de Machine Learning utilizadas en este estudio ha tenido un rendimiento similar al modelo de regresión logística, con una precisión del 86.8% y una k=0.40. No se supera la precisión pero sí se obtiene una mayor concordancia entre la predicción y la clasificación real del tipo de ictus.

4. Conclusiones

El mejor algoritmo para clasificar ictus/mimic se ha obtenido mediante un Random Forest entrenado con validación cruzada, consiguiendo una precisión del 86.7% en la muestra de validación.

El mejor algoritmo para diferenciar el ictus isquémico del hemorrágico se ha obtenido mediante una red neuronal artificial entrenada con 2 nodos ocultos y validación cruzada, consiguiendo una precisión del 86.8% en la muestra de validación.

La precisión obtenida por el algoritmo entrenado mediante Random Forest es algo superior a la obtenida por un modelo de regresión logística en la clasificación de ictus/mimic. En cambio, el rendimiento obtenido por la red neuronal artificial en la clasificación del tipo de ictus no supera el obtenido mediante regresión logística, con precisiones similares.

Así, aunque se ha conseguido mejorar algo la clasificación ictus/mimic de los pacientes, no se ha alcanzado la capacidad diagnóstica del 90% que se esperaba obtener a priori, por lo que el éxito del algoritmo es relativo, produciendo una mejora que parece poco relevante. En la clasificación del tipo de ictus, ni se alcanza el 90% de precisión esperada, ni se mejora el rendimiento obtenido por el modelo de regresión logística.

Se puede considerar que se han logrado algunos objetivos del estudio, en cuanto a que se han logrado conseguir los mejores algoritmos para clasificar a los pacientes, obteniendo además un mejor rendimiento en la clasificación de ictus/mimic. Sin embargo, no se han obtenido los resultados esperados, ya que se esperaban obtener mayores capacidades diagnósticas con la aplicación de las técnicas de Machine Learning.

Pueden haber varias razones por las cuales no se ha alcanzado una mayor precisión en la predicción de las clases. Una de ellas podría ser que la asociación con los biomarcadores no es lo suficientemente clara como para clasificarnos bien los grupos. Es cierto también que algunos de los biomarcadores podrían estar tan asociados con algunas características clínicas, como la gravedad inicial (escala NIHSS) que hace que no aporten mucho más a la clasificación de los pacientes. No obstante, se esperaba que

otras relaciones que fueran más allá de lo lineal u ordenado de algunos de los algoritmos entrenados, como las redes neuronales, nos fueran a enlazar los datos de otra forma, comportando una mejora significativa en los rendimientos de los modelos. Sí es cierto que, si se realizan más pruebas, intentando variar los valores de los parámetros o funciones, podríamos llegar a encontrar algoritmos más óptimos, pero se han intentado realizar los modelos con parámetros y funciones que se han considerado adecuados, por lo que no es esperable una mejora sustancial en el rendimiento de nuevos algoritmos.

Aún habiendo utilizado técnicas de Machine Learning que son perfectamente válidas para aplicar a los datos de nuestro estudio, hay algunas de ellas que podrían ser más apropiadas u óptimas para el tipo de variables que tenemos. Se podrían descartar algunas técnicas que, según la la experiencia adquirida en este análisis, serían poco útiles para la clasificación de este tipo de datos y que nos pueden permitir ahorrarnos tiempo en futuros análisis. Así, el primer algoritmo que se podría descartar sería el de Naïve Bayes ya que, primero, es un algoritmo que sólo trabaja con variables categóricas, por lo que es dificil saber como categorizar los biomarcadores para que, en combinación con otros, den una respuesta óptima. El segundo motivo de descarte sería por la suposición que hace de que todas las variables son independientes, cuando normalmente en un análisis de biomarcadores es difícil que esto ocurra. En nuestro caso, además, pese a intentar obtener buenos puntos de corte, ninguno de los algoritmos de Naïve Bayes mejoran las precisiones obtenidas por las variables clínicas al añadir los biomarcadores. Otro algoritmo podría descartarse es el k-NN, más bien porque tenemos uno más potente, que es el Support Vector Machine, que podría decirse que es una mejora del anterior, ya que agrupa a los individuos según su cercanía acorde a los datos creando hiperplanos a partir de diferentes funciones, que ayudan a clasificarlos mejor, teniendo además la posibilidad de ajustar el márgen de error. Además, son algoritmos que no tienden a sobreajustarse. En nuestro caso, se observa que las precisiones obtenidas tanto en la clasificación ictus/mimic como en el tipo de ictus son mejores en los algoritmos SVM que en los k-NN. También podríamos descartar los árboles de decisión, ya que, aunque son muy fáciles de

interpretar, son muy propensos al sobreajuste del modelo. Además, se tiene a disposición de los Random Forests, los cuales entrenan varios conjuntos de árboles combinados, con los que se obtiene una mejor precisión y un menor sobreajuste en el modelo. Las redes neuronales también podrían se considerar un buen método para este tipo de análisis, ya que no hacen suposiciones sobre relaciones entre los datos y pueden modelar patrones más complejos que otras técnicas.

Así, pese a que son algoritmos computacionalmente intensivos y en los cuales existe la necesidad de afinarlos con varios parámetros y/o funciones, los métodos de Machine Learning que funcionarían mejor en este tipo de datos serían las redes neuronales, los algoritmos SVM y los Random Forest, pudiendo descartar el resto de técnicas en futuros análisis.

Otro factor a tener en cuenta en la decisión de escoger una técnica u otra son las variaciones producidas en el rendimiento del algoritmo si se producen cambios en la semilla de aleatorización. La experiencia con estos datos nos revelan que en las redes neuronales se producen algunas variaciones (no muy importantes) en el rendimiento del modelo con el cambio de semilla, sobretodo cuando se realiza validación cruzada. El Random Forest , en cambio, muestra más consistencia en este sentido, ya que los cambios que se producen son mínimos, por lo que, podría ser la técnica de Machine Learning más recomendable en este tipo de estudios.

En el caso concreto de nuestros biomarcadores, no obstante, los modelos de regresión logística seleccionando las características más asociadas de forma individual con cada clasificación de pacientes pueden tener un buen rendimiento, sin tender tanto al sobreajuste como los nuevos algoritmos entrenados. Parece que en estos datos no existen esas relaciones ocultas que normalmente no se detectan mediante los análisis de regresión tradicionales. Además, en los algoritmos creados en este estudio, pese a tener un mayor número de características incluidas en los modelos, no se consigue mejorar de forma relevante el resultado.

Parece que ha habido otros estudios en los que tampoco han hallado un beneficio de los algoritmos de Machine Learning sobre los modelos de regresión logística [23, 24]. En un estudio publicado recientemente [23], se intentó demostrar que las técnicas de Machine Learning mejorarían los

resultados obtenidos mediante modelos de regresión logística para predecir el pronóstico del tratamiento endovascular en el ictus isquémico agudo. Sin embargo, como en nuestro estudio, ha habido poco éxito en el intento.

Como ha pasado en otros estudios [18,19] es de esperar que los algoritmos de Machine Learning mejoren los modelos clásicos de regresión ya que son más eficientes a la hora de procesar relaciones no lineales y interacciones complejas entre variables. Además, el hecho de que analicemos todas las variables de forma simultánea y no una selección de ellas, como en los modelos de regresión logística, podría hacer pensar que casi con seguridad los modelos deberían rendir bastante mejor, pero se ha demostrado que no en todos los estudios es así.

Una de las lecciones aprendidas de este trabajo es que los algoritmos de clasificación no hacen milagros. Para obtener buenas predicciones, los biomarcadores tienen que estar asociados de una forma más clara con la variable respuesta y lo más independientemente posible de las variables clínicas. Los algoritmos creados mediante Machine Learning, pueden detectar otro tipo de relaciones entre variables que nos incrementen su valor predictivo con respecto a la variable respuesta. No obstante, si ese tipo de relaciones no detectadas por los métodos tradicionales no existen, bastaría con utilizar estos últimos para entrenar los modelos, que además, son más fácilmente interpretables.

Como futuras líneas de trabajo queda buscar otros biomarcadores que, en combinación con algunos de los que se han analizado, puedan permitir clasificar a los pacientes de una forma más precisa. No obstante, puede ser una misión difícil, ya que los biomarcadores estudiados han sido los más potentes relacionados con la patología tras lo publicado hasta la fecha. También se podrían estudiar otro tipo de marcadores, como mutaciones relacionadas con procesos vasculares como el ictus e interaccionarlas con los marcadores ya estudiados.

Así que, de momento, hasta que no se obtengan biomarcadores más precisos, seguirán siendo necesarias e imprescindibles las exploraciones complementarias (TC y/o RM cerebral) para acabar diagnosticando correctamente a los pacientes con ictus.

5. Glosario

Términos más relevantes utilizados en la memoria:

- Biomarcadores: Indicadores utilizados para medir procesos y respuestas biológicas. En este estudio, representan una medida objetiva de características moleculares propuestas para mejorar el diagnóstico del ictus. Los biomarcadores utilizados se detallan en la metodología.
- Ictus: Trastorno brusco de la circulación cerebral que altera la función de una determinada región del cerebro.
- Kappa: Índice que muestra la concordancia entre las clases que predicen los modelos y la clasificación real.
- Machine Learning: Conjunto de técnicas con las que se desarrollan algoritmos para identificar patrones que permiten transformar los datos en un determinado objeto mediante aprendizaje automatizado. En el caso de este estudio, se utilizan para predecir unas clases concretas en función de las características de la muestra.
- Mimics: Patología con síntomas muy parecidos que simulan o se confunden con un ictus.
- Precisión del modelo: Porcentaje de pacientes clasificados correctamente por el algoritmo.

6. Bibliografía

1. Miller JR. The importance of early diagnosis of multiple sclerosis. J Manag Care Pharm 2004: 3 (Suppl B): S4-S11.
2. Astigarraga I, González-Granado LI, Allende LM, Alsina L. Síndromes hemofagocíticos: la importancia del diagnóstico y tratamiento precoces. An Pediatr (Barc) 2018; 89 (2): 124.e1-124.e8.
3. Lees KR, Bluhmki E, von Kummer R, Brott TG, Toni D, Grotta JC, et al; ECASS, ATLANTIS, NINDS and EPITHET rt-PA Study Group. Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ECASS, ATLANTIS, NINDS, and EPITHET trials. Lancet 2010; 375: 1695-1703.
4. Wardlaw JM, Murray V, Berge E, del Zoppo GJ. Thrombolysis for acute ischaemic stroke. Cochrane data-base of systematic reviews 2014: CD000213.
5. Hacke W, Donnan G, Fieschi C, Kaste M, von Kummer R, Broderick JP, et al; ECASS, ATLANTIS, NINDS, and EPITHET trials. Lancet 2010; 375: 1695-1703.
6. Ye D, Zhang T, Lou G, Xu W, Dong F, Chen G, Liu Y. Plasma miR-17, miR-20a, miR-20b and miR-122 as potential biomarkers for diagnosis of NAFLD in type 2 diabetes mellitus patients. Life Sci 2018; 208: 201-207.
7. Undén J, Strandberg K, Malm J, Campbell E, Rosengren L, Stenflo J, Norrving B, Romner B, Lindgren A, Andsberg G. Explorative investigation of brain damage and coagulation system activation in clinical stroke differentiation. J Neurol 2009; 256: 72-77.
8. Llombart V, García-Berrocoso T, Bustamante A, Giralt D, Rodríguez-Luna D, Muchada M, Penalba A, Boada C, Hernández-Guillamon M, Montaner J. Plasmatic retinol-binding protein 4 and glial fibrillary acidic protein as biomarkers to differentiate ischemic stroke and intracerebral hemorrhage. J Neurochem 2016; 136: 416-424.
9. Bustamante A, López-Cancio E, Pich S, Penalba A, Giralt D, García-Berrocoso T, et al. Blood biomarkers for the early diagnosis of stroke. The Stroke-Chip Study. Stroke 2017; 48: 2419-2425.

10. Wendt M, Ebinger M, Kunz A, Rozanski M, Waldschidt C, Weber JE, Winter B, Koch PM, Nolte CH, Hertel S, Ziera T, Audebert HJ; STEMO Consortium. Coeptin levels in patients with acute ischemic stroke and stroke mimics. Stroke 2015; 46: 2426-2431.
11. Montaner J, Mendioroz M, Ribó M, Delgado P, Quintana M, Penalba A, et al. A panel of biomarkers including caspase-3 and D-dimer may differentiate acute stroke from stroke-mimicking conditions in the emergency department. J Intern Med 2011; 270: 166-174.
12. Subhashini P, Jaya Krishna S, Usha Rani G, Sushma Chander N, Maheshwar Reddy G, Naushad SM. Application of machine learning algorithms for the differential diagnosis of peroxisomal disorders. J Biochem 2018; (in press).
13. Cardoso I, Almeida E, Allende-Cid H, Frery AC, Rangayyan RM, Azevedo-Marques PM, Ramos HS. Analysis of machine learning algorithms for diagnosis of diffuse lung diseases. Methods Inf Med 2018; (in press).
14. Meyer A, Zverinski D, Pfahringer B, Kempbert J, Kuehne T, Sündermann SH, Stamm C, Hofmann T, Falk V, Eickhoff C. Machine learning for real-time prediction of complications in critical care: a retrospective study. Lancet Respir Med 2018; (in press).
15. Ferroni P, Zanzotto FM, Scarpato N, Riondino S, Nanni U, Roselli M, Guadagni F. Risk assessment for venus thromboembolism in chemotherapy-treated ambulatory cancer patients: a machine learning approach. Med Decis Making 2017; 37: 234-242.
16. Konerman MA, Zhang Y, Zhu J, Higgins PD, Lok AS, Waljee AK. Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data. Hepatology 2015; 61: 1832-1841.
17. Mani S, Chen Y, Li X, Arlinghaus L, Chakravarthy AB, Abramson V, Bhave SR, Levy MA, Xu H, Yankeelov TE. Machine Learning for predicting the response of breast cancer to neoadjuvant chemotherapy. J Am Med Inform Assoc 2013; 20: 688-695.
18. Signal AG, Mukherjee A, Elmunzer BJ, Higgins PS, Lok As, Zhu J, Marrero JA, Waljee AK. Machine learning algorithsms outperform

conventional regression models in predicting development of hepatocellular carcinoma. Am J Gastroenterol 2013; 108: 1723-1730.

19. Kop R, Hoogendoorn M, Teije AT, Büchner FL, Slottje P, Moons LM, Numans ME. Predictive modeling of colorectal cancer using a dedicated pre-processing pipeline on routine electronic medical records. Comput Biol Med 2016; 76: 30-38.
20. http://ictus.sen.es/?page_id=90. Acceso: 19/10/2018.
21. Lantz, Brett. Machine Learning with R, Second Edition. Packt Publishing, Birmingham, 2015.
22. Ruopp MD, Perkins NJ, Whitcomb BW, Shisterman EF. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection Bimo J 2008; 50: 419-430.
23. Van Os HJA, Ramos LA, Hilbert A, van Leeuwen M, van Walderveen MAA, et al; MR CLEAN Registry Investigators. Predicting outcome of endovascular treatment for acute ischemic stroke: Potential value of Machine Learning Algorithms. Front Neurol 2018; 9: 784.
24. Van der Ploeg T, Nieboer D, Steyerberg EW. Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injurt. J Clin Epidemiol 2016; 78:83-89.

7. Anexos

En los siguientes anexos, se encuentran los códigos R y outputs de todos los análisis realizados para obtener los resultados incluidos en la memoria:

- *"Anexo 1. Resultados Ictus vs. Mimic y sólo biomarcadores.pdf"*: Resultados del análisis descriptivo, comparaciones ictus vs mimic y técnicas de Machine Learning empleadas sólo en biomarcadores (R Markdown).
- *"Anexo 2. Resultados Ictus vs. Mimic – Solo Clinica.pdf"*: Resultados del análisis de técnicas de Machine Learning en la comparación de ictus vs mimic empleadas sólo en las variables clínicas (R Markdown).
- *"Anexo 3. Resultados Ictus vs. Mimic – Clinica y Biomarcadores.pdf"*: Resultados del análisis de técnicas de Machine Learning en la comparación de ictus vs mimicc tras añadir los biomarcadores a las variables clínicas (R Markdown).
- *"Anexo 4. Resultados Tipo de ictus y sólo biomarcadores.pdf"*: Resultados del análisis descriptivo de los ictus, comparaciones entre ictus isquémicos y hemorrágicos y técnicas de Machine Learning empleadas sólo en biomarcadores (R Markdown).
- *"Anexo 5. Resultados Tipo de ictus - Sólo Clínica.pdf"*: Resultados del análisis de técnicas de Machine Learning en la clasificación entre ictus isquémico y hemorrágico empleadas sólo en las variables clínicas (R Markdown).
-*"Anexo 6. Resultados Tipo de ictus - Clínica y Biomarcadores.pdf"*: Resultados del análisis de técnicas de Machine Learning en la clasificación entre ictus isquémico y hemorrágico tras añadir los biomarcadores a las variables clínicas (R Markdown).