Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales
Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Trabajo de Especialización

Detección de Anomalías y Predicción de Mortalidad en Terapia Intensiva

Francisco Valentini

Noviembre de 2019

Supervisor:
Marcelo Soria

Resumen

El presente trabajo hace un análisis no supervisado y supervisado de pacientes de la Unidad de Terapia Intensiva del Beth Israel Deaconess Medical Center de Boston ingresados entre 2001 y 2012, usando la información recopilada durante el primer día de internación. En primer lugar, se desarrolla un indicador de casos atípicos implementando las técnicas de Kurtosis plus Specific Directions y de Redes Neuronales Autoencoders. En segundo lugar, se genera un predictor de mortalidad durante la primera semana ajustando un Modelo Aditivo Generalizado (GAM) con regularización. Por último, se evalúa la capacidad predictiva del indicador de atipicidad y del predictor GAM frente a tres scores de severidad estándar en unidades de terapia intensiva.

Índice

1. Introducción

El presente trabajo se propone hacer un análisis no supervisado y supervisado de pacientes de la Unidad de Terapia Intensiva del Beth Israel Deaconess Medical Center de Boston ingresados entre 2001 y 2012.

En lo que refiere al aprendizaje no supervisado, generamos un indicador de atipicidad o anomalía de los pacientes en términos de las mediciones fisiológicas y características físicas registradas durante las primeras 24 horas después de su ingreso. En particular, ajustamos dos indicadores usando las técnicas de *Kurtosis plus Specific Directions* (Peña y Prieto 2007) y de Autoencoder (Thompson etal. 2002). Estos indicadores podrían usarse en la práctica como insumo para tratar pacientes así como también para el desarrollo de investigaciones.

En cuanto al aprendizaje supervisado, generamos un predictor de mortalidad dentro de los 7 días posteriores a la internación usando como insumo los atributos de las primeras 24 horas usados para la tarea no supervisada. En particular, ajustamos un Modelo Aditivo Generalizado con penalización de tipo *lasso* (Chouldechova y Hastie 2015). La elección del modelo se debe a que permite captar no linealidaes y alcanzar mejor precisión en la clasificación que modelos más simples, a la vez que admite interpretar el impacto de los principales factores asociados a la mortalidad por separado, dada la aditividad del modelo. Asimismo, la regularización realiza selección automática de variables relevantes a la vez que trata la potencial multicolinealidad que trae aparejada el uso de atributos correlacionados.

El predictor GAM es validado mediante *model stacking* (ver sección 2.2.3.2) en una regresión logística junto al indicador de anomalía ajustado con el Autoencoder, y junto a tres predictores estándar de severidad que también usan información de las primeras 24 horas: SOFA (Vincent etal. 1996), SAPSII (Le Gall, Lemeshow, y Saulnier 1993) y OASIS (Johnson, Kramer, y Clifford 2013). El objetivo de esta etapa es evaluar la significatividad del aporte de ambos indicadores –el predictor de mortalidad y el indicador de atipicidad– a la predicción de mortalidad en la primera semana.

2. Materiales y métodos

2.1. Fuentes de datos

El estudio se hizo con MIMIC-III, una base de libre acceso con datos anonimizados de pacientes internados en la Unidad de Terapia Intensiva (UTI) del Beth Israel Deaconess Medical Center de Boston entre 2001 y 2012 (Johnson etal. 2016).

MIMIC-III incluye datos demográficos, mediciones horarias de signos vitales, resultados de pruebas de laboratorio, procedimientos, medicamentos, notas de cuidadores, informes de imágenes y eventos relevantes tanto dentro como fuera del hospital –por ejemplo, la mortalidad. Se puede acceder a una descripción del esquema completo de tablas en el repositorio oficial del MIT Laboratory for Computational Physiology. Las operaciones de extracción y preprocesamiento de los datos se describen en la sección 3.

2.2. Algoritmos

2.2.1. Imputación de datos faltantes

Para imputar valores numéricos faltantes durante la etapa de preprocesamiento (ver sección 3.3) se usó el procedimiento de *median polish* de Tukey (ver Hoaglin, Mosteller, y Tukey (1983)).

La técnica parte del supuesto de que el valor x_{ij} que toma cada registro i para cada covariable j se modela como un modelo lineal y aditivo $x_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, tal que μ representa el "valor común" o efecto general del set de datos, α_i es el efecto por fila, β_j es el efecto por columna y ε_{ij} son fluctuaciones aleatorias.

El procedimiento de *median polish* consiste en estimar μ, α_i y β_j iterativamente, sustrayendo las medianas por fila y columna de los datos hasta que la mediana de los residuos se acerque a cero, dado un máximo de iteraciones prefijado. En lugar de un modelo de ANOVA para la estimación del modelo aditivo, elegimos *median polish* por su robustez a valores atípicos.

Considerando que el objetivo final no es inferir los parámetros del modelo, sino imputar valores de variables con distintas unidades de medidas, resumimos la técnica de imputación implementada de la siguiente manera:

1. Se estandarizan las columnas sustrayendo las medianas y diviendo por los desvíos absolutos medianos (MAD) de cada una, omitiendo los valores faltantes (si algún MAD equivale a cero, se sumaron pequeños desvíos en relación a la escala de la variable para computarlo).
2. Se estiman el efecto general y los efectos por fila y columna usando *median polish* omitiendo los valores faltantes.
3. Se imputan los valores faltantes de los datos normalizados sumando correspondientemente los efectos estimados en el paso anterior.
4. Se reescalan las variables a su escala original multiplcando por los MADs y sumando las medianas calculadas en el primer paso.

Esta técnica fue aplicada una vez descartadas las variables con una alta proporción de valores faltantes (ver sección 3.3).

2.2.2. Detección de anomalías

Para identificar casos anómalos nos propusimos generar indicadores de *outlyingness* que indiquen el grado de atipicidad o anomalía de cada observación. Los indicadores usan los predictores del set completo de internaciones ya preprocesado, sin incluir la variable respuesta de mortalidad del análisis supervisado. No es de interés en esta aplicación identificar puntos de corte en los indicadores que separen lo anómalo de lo no-anómalo.

Para la tarea de detección de anomalías, los algoritmos descritos en esta sección se ajustaron inicialmente sobre todo el set de datos y posteriormente se compararon visualmente. Sin

embargo, como también era de interés analizar si el indicador ajustado con un autoencoder contribuye a la predicción de mortalidad, lo entrenamos en segunda instancia siguiendo un criterio acorde a una tarea supervisada, que se describe en la sección 2.2.3.2.

2.2.2.1. Kurtosis plus Specific Directions (KSD)

La detección de observaciones atípicas en datos multivariados puede lograrse hallando una proyección univariada de los datos que indique el grado de atipicidad o *outlyingness* de cada uno: si una observación dada tiene un valor alto en esta proyección, entonces se clasifica como outlier multivariado.

De esta manera, sean X la matriz de datos y x una observación particular tal que $x \in R^p$, siendo p la cantidad de atributos, el problema consiste en hallar una dirección a tal que $||a|| = 1$ y tal que:

$$t(x) = \max_a \left| \frac{x'a - \hat{\mu}(a'X)}{\hat{\sigma}(a'X)} \right|$$

donde $\hat{\mu}$ y $\hat{\sigma}$ son estimadores robustos de posición y dispersión, respectivamente (Maronna etal. 2019).

El procedimiento de *Kurtosis plus Specific Directions* (KSD) propuesto por Peña y Prieto (2007) intenta hallar a usando las dos siguientes observaciones:

1. Una pequeña proporción de valores atípicos tiende a incrementar la curtosis de una distribución –las colas se vuelven más pesadas– mientras que una proporción alta disminuye la curtosis –la distribución se vuelve más bimodal.
2. Se puede aumentar la probabilidad de obtener direcciones buenas –que detecten valores atípicos– mediante un mecanismo de muestreo estratificado.

Por lo tanto, el método KSD busca iterativamente proyecciones que maximicen y minimicen la curtosis, así como también direcciones obtenidas al azar. Sobre estas direcciones se computa $t(x)$, que se usa en el presente trabajo como indicador de anomalía o atipicidad de una estadía en la UTI. Si bien el método descrito en Peña y Prieto (2007) posee un paso adicional para definir qué observaciones efectivamente son outliers –el cual recibe el nombre de "checking"–, en esta aplicación usamos la cantidad $t(x)$ directamente como indicador del grado de atipicidad y obviamos este paso.

2.2.2.2. Autoencoder

Otro método posible para detectar anomalías consiste en ajustar un modelo que busque reconstruir las propias observaciones, de modo que aquellos casos con un error de reconstrucción alto –desvíos altos con respecto al modelo aprendido– se puedan catalogar como *outliers*.

Para este fin usamos la arquitectura de redes neuronales conocida como *autoencoder* (AE) (ver Thompson etal. (2002) y Japkowicz, Myers, y Gluck (1995)). Un AE consiste en una red

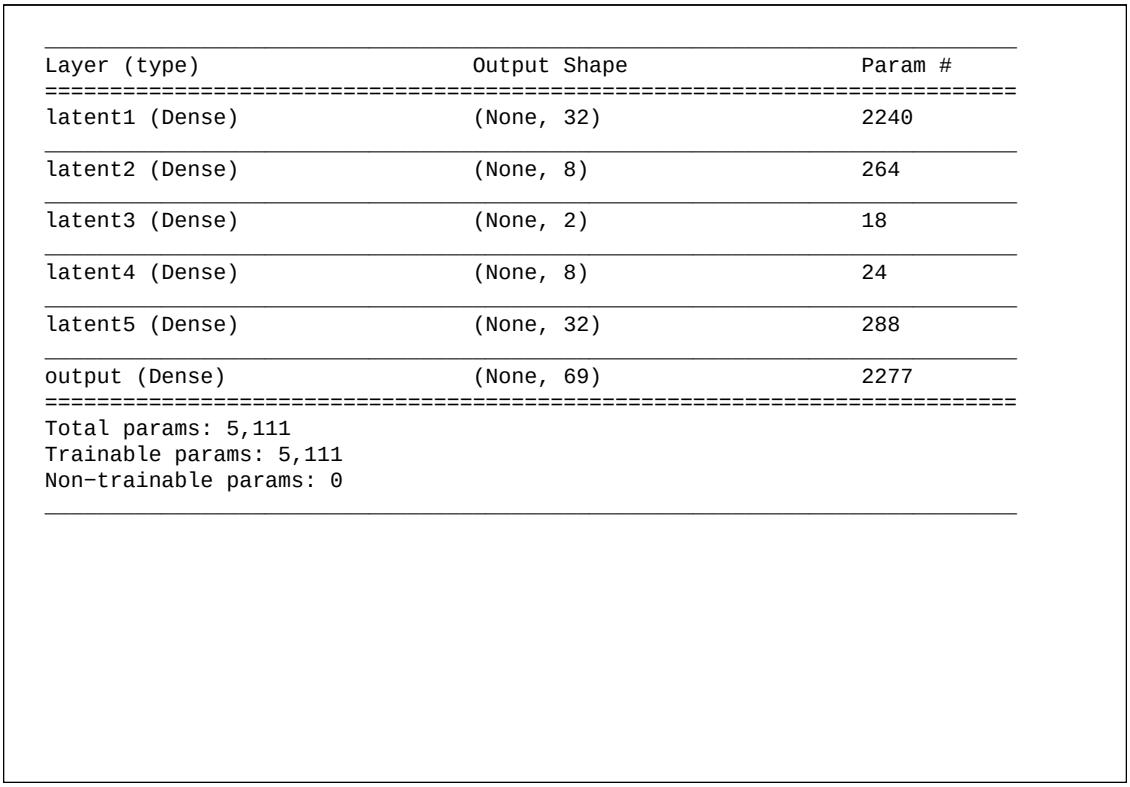

Figura 1: Estructura del autoencoder implementado

neuronal con tantas neuronas como features en la capa de entrada y tantas neuronas como features en la capa de salida, y cuyos pesos se ajustan para intentar reproducir el input.

En particular, los AE llamados *undercomplete* reducen la cantidad de neuronas en las capas intermedias en relación a las capas de entrada/salida, a fin de generar una representación en baja dimensión de los datos. Las salidas de las capas intermedias funcionan como un espacio latente o *embedding* que capturan las dimensiones más importantes de la distribución multivariada. Al reconstruir los datos desde esta representación latente se obtiene una reconstrucción de cada registro en la dimensión original, libre de ruido y anomalías. Entonces, el error de reconstrucción de una observación dada, que es el error entre el punto original y su reconstrucción, se usa como un indicador de *outlyingness*.

Formalmente, partiendo de una observación dada $x \in R^p$, el AE codifica x con múltiples representaciones intermedias $x^{int} \in R^k$ tal que k representa la cantidad de neuronas de una capa intermedia dada y tal que $k < p$. La representación de menor dimensión es luego decodificada como $\hat{x} \in R^p$ y se computa el score de *outlyingness* como $t(x) = Ave(x - \hat{x})^2$.

La estructura del AE implementado –sin considerar la capa de input– se describe en la Figura 1. Usamos una red *feed-forward* y *fully-connected* con cinco capas intermedias (*latent1* hasta *latent5*) tal que los datos llegan a representarse en dos dimensiones en la capa *latent3*. Todas las capas intermedias usan funciones de activación ReLU mientras que la capa de salida usa activaciones lineales. Los pesos se entrenaron mediante el algoritmo de optimización Adam.

2.2.3. Predicción de mortalidad

El objetivo de aprendizaje supervisado es la predicción de mortalidad durante los 7 días siguientes a la internación ($mor7$) –el evento de mortalidad se codifica como $mor7 = 1$ y la supervivencia como $mor7 = 0$. Para este fin se ajusta un Modelo Aditivo Generalizado regularizado (GAM) usando como covariables mediciones realizadas durante las primeras 24 horas de internación (ver Sección 3.2). El predictor GAM es validado mediante *model stacking* en una regresión logística junto a otros predictores estándar de mortalidad y al indicador de anomalía AE, a fin de determinar la significatividad de su aporte a la predicción de $mor7$.

Para realizar ambos ejercicios los registros fueron separados al azar entre un set de entrenamiento (85 % de los datos) y un set de test (15 % de los datos).

2.2.3.1. Modelo Aditivo Generalizado (GAM)

Los Modelos Aditivos Generalizados (GAMs) permiten modelar relaciones no lineales entre cada covariable x_j y la respuesta Y ajustando funciones suaves y no lineales $f_j(x_j)$ para cada covariable por separado. En esta aplicación, en la medida en que el target es binario, ajustamos un GAM de regresión logística según la expresión:

$$log\left(\frac{p(Y=1|X)}{1-p(Y=1|X)}\right) = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i$$

según la cual el *logodds* de $mor7$ es una función no lineal y aditiva de las covariables x_j. Transformando las *logodds* se puede obtener el resultado del modelo como una probabilidad de mortalidad estimada.

Para ajustar el modelo usamos la metodología presentada por Chouldechova y Hastie (2015). La misma propone un enfoque de estimación penalizada que permite ajustar cada f_j como cero, lineal o no lineal según lo sugerido por los datos. Gracias a este mecanismo de *feature selection* se pueden filtrar variables irrelevantes o redundantes automáticamente, ajustar funciones lineales cuando son una buena aproximación o capturar relaciones no lineales fuertes cuando están presentes. El parámetro que regula el grado de penalización es λ; para $\lambda = 0$ se conservan todas las covariables en el modelo, mientras que para $\lambda \to \infty$ el efecto de todos los predictores tiende a cero. Todas las covariables se estandarizan con las medias y desvíos estimados con los datos de entrenamiento antes de realizar el ajuste.

Debido al doble objetivo que perseguimos en el aprendizaje supervisado –entender el impacto de los principales factores asociados a la mortalidad (1) y optimizar la precisión en la predicción de mortalidad (2)– ajustamos el GAM penalizado de las dos maneras siguientes:

(1) Se usa un parámetro de regularización lo suficientemente alto como para identificar y visualizar el efecto de a lo sumo las 30 covariables más importantes, usando todos los datos de entrenamiento. Esta decisión tiene una justificación puramente de inferencia, dado que el λ que optimiza la capacidad preditiva en sets de validación selecciona demasiadas variables con coeficiente no nulo, dificultando la interpretación.

(2) Para optimizar la capacidad predictiva se usa el valor del parámetro λ que optimiza el *accuracy* a la hora de predecir nuevos casos. El valor de λ se halla mediante 10-fold Cross-Validation usando los datos de entrenamiento. La performance final del predictor (*GAMpred*) se evalúa en el set de test.

2.2.3.2. Stacking

Esta etapa consiste en ajustar una regresión logística usando como covariables (1) el indicador de atipicidad *AE*, (2) el predictor *GAMpred* y (3) tres scores estándar de mortalidad que usan información de las primeras 24 horas de internación; y usando como variable respuesta *mor7*.

El indicador *AE* se ajusta con los datos de entrenamiento mientras que *GAMpred* se ajusta como se indicó en la Sección 2.2.3.1. En base a estos modelos ajustados se extraen las predicciones sobre el set de test que se usan como covariables en el *stacked model*. La regresión logística se ajusta con los datos de test remuestrados en 500 sets de bootstrap. A partir de la evaluación de la significatividad media de los coeficientes de (1) y (2) en las regresiones de bootstrap se puede evaluar la contribución a la predicción de mortalidad.

Cabe destacar que para evaluar correctamente la capacidad predictiva de los dos modelos (*AE* y *GAMpred*) sería necesario incluir en la validación todos las etapas del proceso de modelado además de la estimación de la probabilidad: en particular, todas las decisiones de preprocesamiento de los datos descritas en la Sección 3.

2.3. Software usado

Para almacenar y preprocesar la base MIMIC-III usamos el motor de bases de datos PostgreSQL. Para todos los análisis restantes usamos el lenguaje R (R Core Team 2016), haciendo uso en particular de las librerías *tidyverse* (Wickham 2017), *gamsel* (Chouldechova, Hastie, y Spinu 2018) y *keras* (Chollet, Allaire, y others 2017).

3. Preprocesamiento

3.1. Extracción de datos

De MIMIC-III conservamos atributos personales y mediciones fisiólogicas de cada paciente registrados en las primeras 24 horas de internación en la UTI médica –no se consideran pacientes de otros tipos de unidades (por ejemplo, neonatal o quirúrgica). Para cada paciente consideramos únicamente el primer ingreso al hospital y a la UTI, a la vez que eliminamos pacientes con transferencias entre unidades del hospital. La variable respuesta fue construida observando el evento de mortalidad en una ventana de 7 días posteriores a las primeras 24 horas de internación.

Con la información disponible calculamos tres indicadores estándar de severidad que también usan la información de las primeras 24 horas: el *Sequential Organ Failure Assessment* SOFA

Cuadro 1: Mediciones durante las primeras 24 horas de internación

Variable	Definición
admission_age	Edad (en años)
aniongap	Brecha aniónica (en milliequivalentes/litro)
bicarbonate	Bicarbonato (en milliequivalentes/litro)
bilirubin	Bilirrubina (en miligramos por decilitro)
bun	Nitrógeno ureico en sangre (en miligramos por decilitro)
chloride	Cloruro en sangre (en miliequivalente por litro)
creatinine	Creatinina en sangre (en miligramos por decilitro)
diasbp	Presión arterial diastólica (en milímetros de mercurio)
glucose_lab	Glucosa en sangre según laboratorio (en miligramos por decilitro)
glucose_vitals	Glucosa en sangre según signos vitales (en miligramos por decilitro)
heartrate	Frecuencia cardíaca (en latidos por minuto)
hematocrit	Hematocrito (en %)
hemoglobin	Hemoglobina (en gramos por decilitro)
inr	Ratio Internacional Normalizado de tiempo de protrombina
lactate	Lactato (en milimoles por litro)
meanbp	Presión arterial media (en milímetros de mercurio)
platelet	Plaquetas (en miles por milímetro cúbico)
potassium	Potasio (en miliequivalentes por litro)
pt	Tiempo de protrombina (en segundos)
ptt	Tiempo de tromboplastina parcial activado (en segundos)
resprate	Frecuencia respiratoria (en respiraciones por minuto)
sodium	Sodio en sangre (en milimoles por litro)
spo2	Saturación de oxígeno capilar periférica (en %)
sysbp	presión arterial sistólica (en milímetros de mercurio)
tempc	Temperatura corporal (en grados Celsius)
urineoutput	Gasto urinario (en mililitros)
wbc	Conteo de glóbulos blancos (en células por litro)
weight	Peso (en kilogramos)

(Vincent etal. 1996), el *Simplified Acute Physiology Score II* SAPS II (Le Gall, Lemeshow, y Saulnier 1993) y el *Oxford Acute Severity of Illness Score* OASIS (Johnson, Kramer, y Clifford 2013).

Para generar las vistas y tablas necesarias para extraer los datos usamos el código disponible en Pollard etal. (2017).

3.2. Generación de atributos

Como resultado de la extracción de atributos obtuvimos un set de covariables medidos una o más veces durante las primeras 24 horas, que se muestran en el Cuadro 1. Aquellas covariables con más de una medición durante el primer día de internación se resumieron usando estadísticos de mínimo, máximo, media y coeficiente de tendencia con respecto al tiempo.

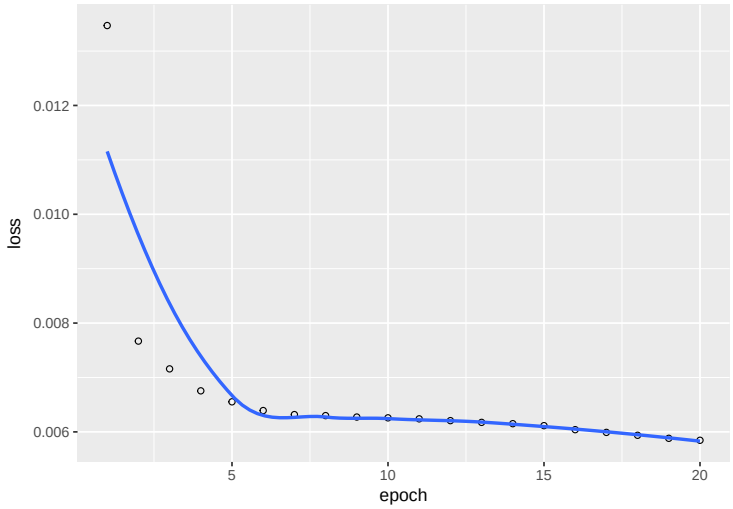

Figura 2: Entrenamiento de autoencoder para detección de anomalías

3.3. Tratamiento de datos faltantes

Eliminamos las variables con un porcentaje de faltantes mayor al 50 %, y para completar los valores faltantes restantes aplicamos la técnica de *median polish*.

3.4. Tratamiento de valores atípicos

Durante el preprocesamiento eliminamos aquellos registros con mediciones afectadas seguramente por errores de carga –por ejemplo, pacientes con más de 120 años, con nitrógeno ureico en sangre negativo o con producción de orina superior a los 100 litros.

Como resultado de las tareas de preprocesamiento descritas en la Sección 3 generamos un set de datos listo para entrenar los algoritmos –previo al remuestreo– conformado por 69 covariables y 11297 registros.

4. Modelado y resultados

4.1. Detección de anomalías

En la presente sección se presentan los resultados de aplicar los procedimientos de KSD y AE tal como se describe en la sección 2.2.2.

En la Figura 2 se presenta el detalle de la evolución del error cuadrático medio para cada *epoch* del entrenamiento del autoencoder. El error se estabiliza aproximadamente a partir de la recorrida nº 15 por el set de datos.

Al comparar los indicadores de *outlyingness* AE y KSD en escala logarítmica se observa que ambos tienen una distribución asimétrica –lo cual es de esperar porque se tratan de

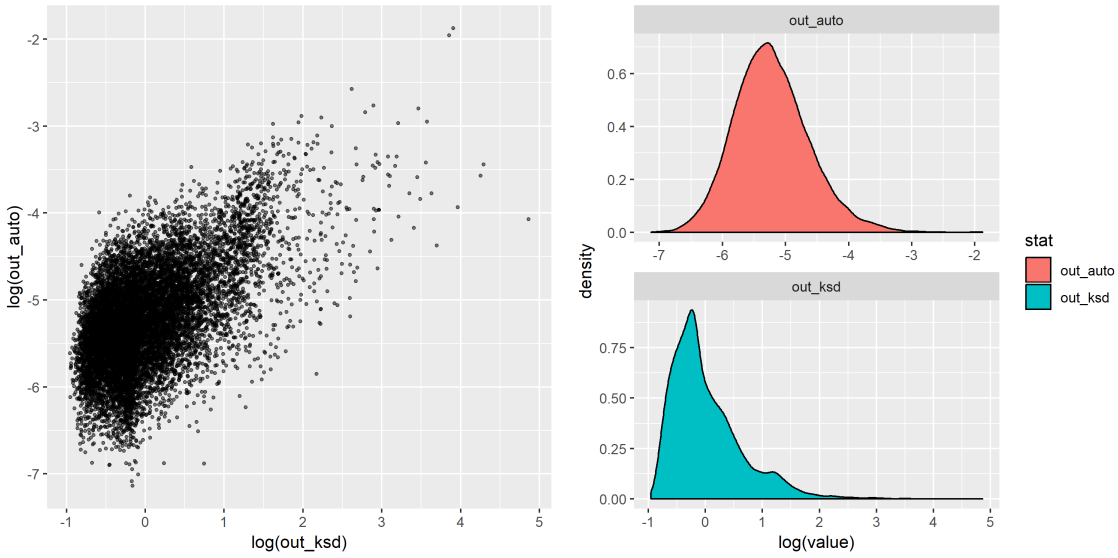

Figura 3: Outlyingness KSD vs. autoencoder (escala logarítmica)

indicadores de atipicidad; la asimetría es más marcada en el caso de KSD, lo cual sugiere que se pueden catalogar más pacientes como atípicos usando esta técnica (ver Figura 3). Por otra parte, en el diagrama de dispersión se observa que los indicadores presentan un grado de asociación no despreciable, arrojando un coeficiente de correlación de Spearman –en la escala original– de 0.51.

4.2. Predicción de mortalidad

4.2.1. Modelo Aditivo Generalizado (GAM)

4.2.1.1. Factores de riesgo de mortalidad

En la presente sección presentamos los resultados de entrenar un GAM con regularización persiguiendo un objetivo de inferencia –presentado como objetivo (1) en la sección 2.2.3.1. Todas las covariables consideradas para entrenar el GAM según las transformaciones descritas en la sección 3.2 se indican en ell Cuadro 2.

Para conservar no más de 30 de las covariables más relevantes en términos de capacidad predictiva escogimos un valor de $\lambda = 1.578$ –el Cuadro 3 indica la cantidad de predictores identificados por el GAM con un efecto no nulo sobre la probabilidad de mortalidad para cada posible valor de λ en el *path* de penalización.

En la Figura 4 se representa visualmente el efecto *ceteris-paribus* estimado de cada covariable sobre el *log-odds* de la mortalidad –en verde se indican los efectos linales y en rojo, los no lineales. Las covariables de los ejes horizontales se presentan estandarizadas según las medias y varianzas del dataset de entrenamiento.

Se identifican únicamente cuatro efectos no lineales, que corresponden al potasio mínimo (v23), la frecuencia cardíaca (v37), la tendencia en el tiempo de la presión arterial media

Cuadro 2: Predictores incluidos en GAM

variable	name	variable	name	variable	name
v1	admission_age	v23	potassium_min	v45	diasbp_mean
v2	urineoutput	v24	potassium_max	v46	diasbp_trend
v3	aniongap_min	v25	ptt_min	v47	meanbp_min
v4	aniongap_max	v26	ptt_max	v48	meanbp_max
v5	bicarbonate_min	v27	inr_min	v49	meanbp_mean
v6	bicarbonate_max	v28	inr_max	v50	meanbp_trend
v7	bilirubin_min	v29	pt_min	v51	resprate_min
v8	bilirubin_max	v30	pt_max	v52	resprate_max
v9	creatinine_min	v31	sodium_min	v53	resprate_mean
v10	creatinine_max	v32	sodium_max	v54	tempc_min
v11	chloride_min	v33	bun_min	v55	tempc_max
v12	chloride_max	v34	bun_max	v56	tempc_mean
v13	glucose_min_labs	v35	wbc_min	v57	tempc_trend
v14	glucose_max_labs	v36	wbc_max	v58	spo2_min
v15	hematocrit_min	v37	heartrate_min	v59	spo2_max
v16	hematocrit_max	v38	heartrate_max	v60	spo2_mean
v17	hemoglobin_min	v39	heartrate_mean	v61	spo2_trend
v18	hemoglobin_max	v40	sysbp_min	v62	glucose_min_vitals
v19	lactate_min	v41	sysbp_max	v63	glucose_max_vitals
v20	lactate_max	v42	sysbp_mean	v64	glucose_mean
v21	platelet_min	v43	diasbp_min	v65	glucose_trend
v22	platelet_max	v44	diasbp_max	v66	weight

Cuadro 3: Cantidad de predictores en el regularization path del GAM

NonZero	Lin	NonLin	%Dev	Lambda	NonZero	Lin	NonLin	%Dev	Lambda
0	0	0	0.00000	20.040	19	18	1	0.2376	3.262
2	2	0	0.01024	19.440	21	19	2	0.2550	2.484
3	3	0	0.06117	14.810	22	19	3	0.2587	2.338
5	5	0	0.09510	11.620	25	22	3	0.2623	2.201
6	6	0	0.11110	10.620	26	23	3	0.2673	2.010
8	8	0	0.14400	8.333	27	23	4	0.2704	1.892
9	9	0	0.16470	7.163	28	24	4	0.2766	1.676
10	10	0	0.18600	5.974	29	25	4	0.2798	1.578
12	12	0	0.19240	5.623	31	26	5	0.2813	1.531
13	13	0	0.21140	4.550	32	26	6	0.2829	1.485
14	14	0	0.21400	4.414	33	25	8	0.2932	1.202
16	16	0	0.22750	3.681	34	24	10	0.2970	1.097
18	17	1	0.23000	3.572	35	24	11	0.2995	1.033
					36	25	11	0.3008	1.002

Cuadro 4: Área debajo de la curva ROC: GAM y scores de severidad estándar

GAM	OASIS	SAPSII	SOFA
87.27	84.69	83.58	78.7

(v50) y la temperatura corporal media (v56). Las cuatro funciones estimadas tienen el mismo patrón: para valores relativemente altos y bajos de estos atributos la probabilidad de mortalidad estimada es más alta, mientras que baja para valores intermedios.

En cuanto a la significatividad empírica del efecto, se destacan por su efecto negativo sobre la probabilidad de muerte las covariables de gasto urinario (v2), presión arterial sistólica media (v40) y saturación de oxígeno capilar periférica –media (v60) y con su coeficiente de tendencia (v61). Por su parte, el lactato mínimo (v19), el lactato máximo (v20), el tiempo de protrombina mínimo (v27), el mínimo conteo de glóbulos blancos (v35) y el crecimiento de la temperatura corporal en el tiempo (v57) presentan un efecto positivo significativo sobre la mortalidad –a medida que toman valores más altos entre los distintos pacientes crece la probabilidad estimada de muerte en la primera semana.

4.2.1.2. Predicción

Para ajustar un GAM que optimice la capacidad predictiva –según el objetivo (2) planteado en la sección 2.2.3.1– se escogió un valor del parámetro de penalización λ tal que optimice el error de clasificación estimado por 10-fold Cross Validation. En particular, elegimos el modelo con mayor penalización (λ más alto) tal que el error medio se encuentre a un desvío estándar del error medio mínimo –este valor se destaca en la barra vertical derecha de la Figura 5. Como se observa en el eje horizontal superior, dicho valor implica un modelo que conserva 40 covariables con efectos no nulos sobre la probabilidad de muerte estimada.

Para comparar la performance del modelo ajustado con los tres scores de severidad replicados –OASIS, SOFA y SAPSII– graficamos la curva de ROC (Figura 6) y computamos el área debajo de la curva, *AUC* (Cuadro 4), a la vez que comparamos la distribución de la probabilidad estimada entre sobrevivientes y fallecidos durante la primera semana (Figura 7). Los tres resultados se obtuvieron calculando los indicadores a partir de los datos de test.

El predictor *GAMpred* desarrollado presenta una *AUC* superior a los scores estándar. Sin embargo cabe tener en cuenta las siguientes consideraciones a la hora de hacer la comparación:

(1) Si bien la *AUC* es mayor, la curva no se ubica consistentemente por encima de la curva del resto de los predictores –en particular, *GAMpred* se ve superado por OASIS y SAPSII en el primer tramo de la curva.

(2) Los cuatro predictores tienen problemas de sensibilidad para cualquier punto de corte que se pueda plantear si fuera necesario clasificar: mientras que los casos negativos (no-mortalidad) se ajustan con una probabilidad baja exitosamente, los casos de mortalidad

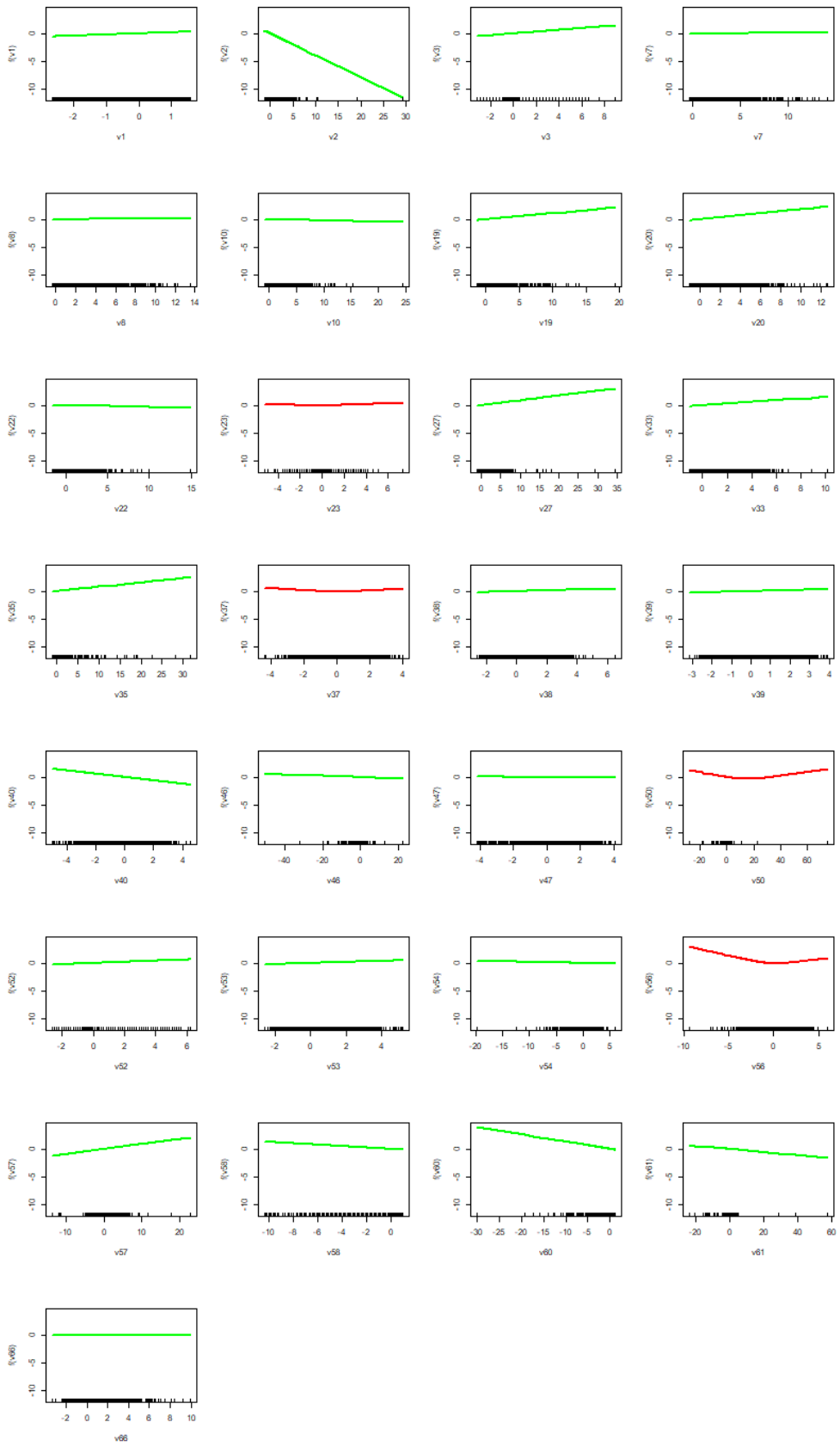

Figura 4: Variables explicativas del GAM (lambda=1.578)

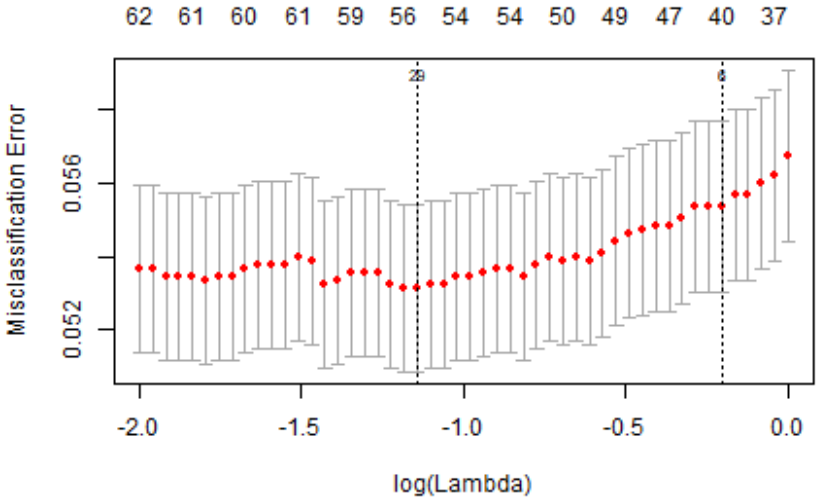

Figura 5: Error de clasificación de 10-fold CV según penalización del GAM

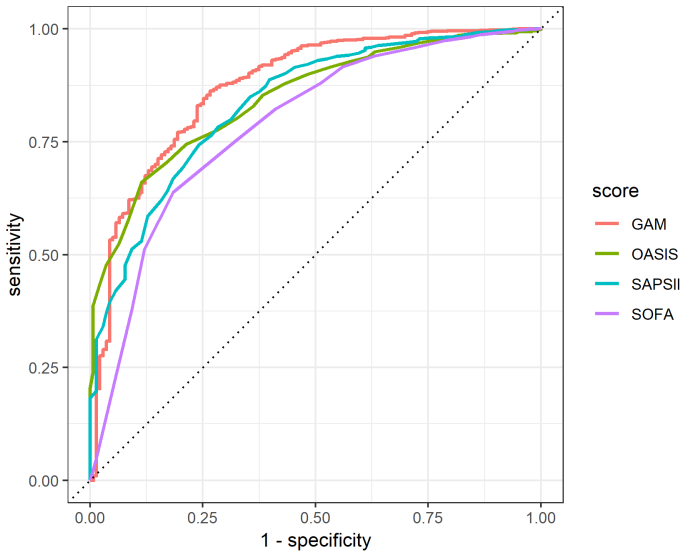

Figura 6: Curva de ROC: GAM y scores de severidad estándar

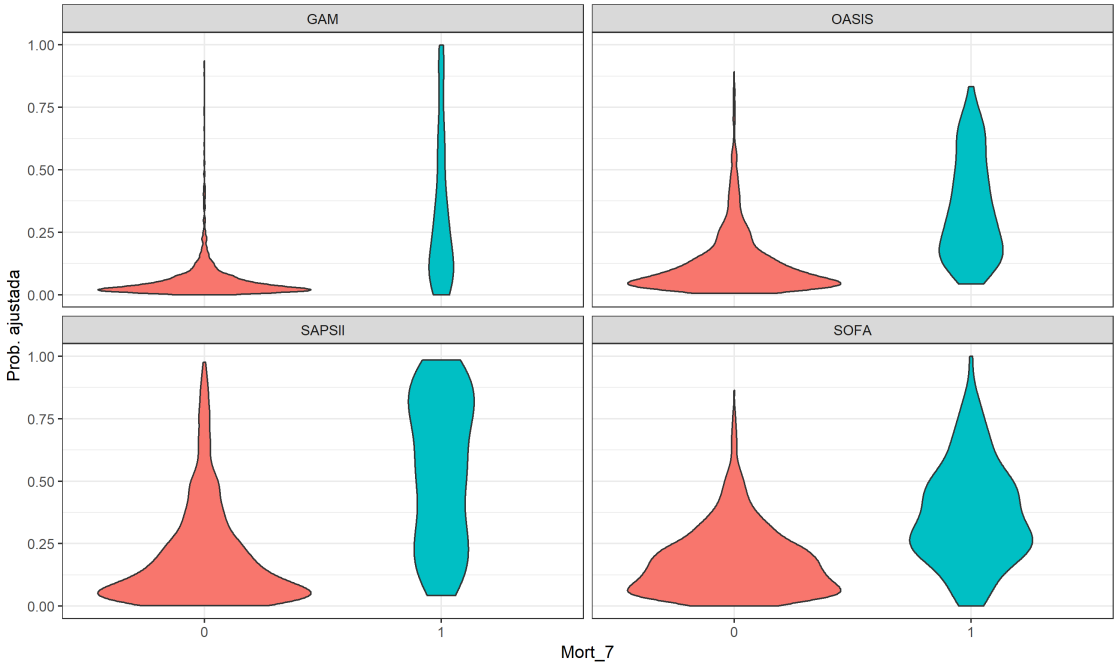

Figura 7: Distribución de positivos y negativos: GAM y scores de severidad estándar

efectiva no logran estimarse en buena medida con una probabilidad alta, como se observa en la Figura 7.

(3) Se cuenta con una sola estimación de AUC en test para cada indicador. La bondad de la comparación sería superior si se realizara más de una partición de remuestro en el set de datos, lo que permitiría evaluar el poder predictivo con la media de múltiples estimaciones de AUC para cada predictor.

(4) Para que la comparación sea válida sería necesario:

- Predecir mortalidad con *GAMpred* en pacientes de otros hospitales en otros momentos del tiempo, ya que los tres scores de severidad no fueron ajustados con datos del Beth Israel Deaconess Medical Center, mientras que *GAMpred*, sí.
- Incluir en la validación las etapas de preprocesamiento descritas en la sección 3 –por ejemplo, la imputación de faltantes con *median polish* debería realizarse con los datos de test únicamente.

4.2.2. Stacking

En la Figura 8 presentamos el correlograma de Pearson de las variables incluidas en la regresión logística *stacked* ajustada con los registros de test: el indicador de atipicidad AE en escala logarítmica (outl_auto_log), el predictor *GAMpred* (pred_gam) y los tres scores estándar de severidad (oasis, sapsii y sofa).

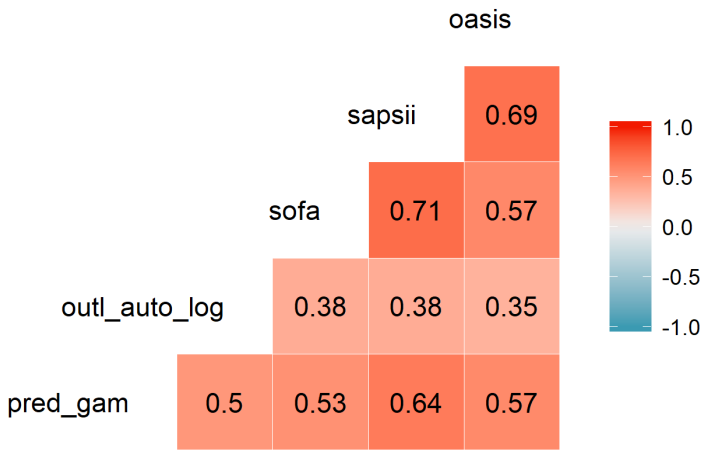

Figura 8: Correlograma de los predictores en la regresión logística

Cuadro 5: Resultados de regresión logística (bootstrap de 500 repeticiones)

term	Coef. (media)	Des.Est. Coef.	p-valor (media)
(Intercept)	-0.59421	2.32179	0.49123
oasis	2.26464	0.85540	0.04766
outl_auto_log	1.04342	0.66906	0.23998
pred_gam	4.81714	0.97749	0.00006
sapsii	1.20604	0.73865	0.19490
sofa	0.01756	0.04101	0.45180

Ningún par de regresores presenta una correlación bivariada de Pearson que se pueda considerar como muy alta, lo cual indica que todos captan factores de mortalidad relativamente distintos y que podemos asumir la ausencia de colinealidad –al menos en un sentido bivariado– en la regresión logística.

Con el fin de aumentar el grado de confianza en la inferencia que hacemos de la regresión estimada, ajustamos el modelo en 500 sets de bootstrap usando como base el set de test; computamos la media y el desvío de los coeficientes estimados, así como también el promedio de los p-valores.

Como se indica en el Cuadro 5, *predGAM* es consistentemente el predictor con más importancia para la predicción de mortalidad, seguido por OASIS. Esto es razonable porque, si bien *predGAM* no fue ajustado con los datos de test, fue ajustado con datos del mismo hospital –de esta manera, es más que razonable que capte mejor las características del proceso generador de datos que los indicadores estándar.

Por su parte, el indicador de atipicidad AE no presenta un aporte significativo una vez que se consideran los predictores propios de mortalidad (*predGAM* y los tres scores estándar).

Referencias

Chollet, François, JJ Allaire, y others. 2017. «R Interface to Keras». https://github.com/rstudio/keras; GitHub.

Chouldechova, Alexandra, y Trevor Hastie. 2015. «Generalized Additive Model Selection». http://arxiv.org/abs/1506.03850.

Chouldechova, Alexandra, Trevor Hastie, y Vitalie Spinu. 2018. *gamsel: Fit Regularization Path for Generalized Additive Models*. https://CRAN.R-project.org/package=gamsel.

Hoaglin, David C., Frederick Mosteller, y John Wilder Tukey. 1983. *Understanding robust and exploratory data analysis*. New York; Chichester: John Wiley & Sons.

Japkowicz, Nathalie, Catherine Myers, y Mark Gluck. 1995. «A Novelty Detection Approach to Classification», IJCAI'95,, 518-23. http://dl.acm.org/citation.cfm?id=1625855.1625923.

Johnson, Alistair E. W., Andrew A Kramer, y Gari D Clifford. 2013. «A new severity of illness scale using a subset of acute physiology and chronic health evaluation data elements shows comparable predictive accuracy». *Critical care medicine* 41 (7): 1711-8.

Johnson, Alistair E. W., Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, y Roger G Mark. 2016. «MIMIC-III, a freely accessible critical care database». *Scientific data* 3: 160035.

Le Gall, Jean-Roger, Stanley Lemeshow, y Fabienne Saulnier. 1993. «A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study». *JAMA* 270 (24): 2957-63. https://doi.org/10.1001/jama.1993.03510240069035.

Maronna, R. A., R. D. Martin, V. J. Yohai, y M. Salibián-Barrera. 2019. *Robust Statistics: Theory and Methods (with R)*. Wiley Series in Probability and Statistics. Wiley. https://books.google.com.ar/books?id=K5RxDwAAQBAJ.

Peña, Daniel, y Francisco J Prieto. 2007. «Combining Random and Specific Directions for Outlier Detection and Robust Estimation in High-Dimensional Multivariate Data». *Journal of Computational and Graphical Statistics* 16 (1): 228-54. https://doi.org/10.1198/106186007X181236.

Pollard, Tom, Alistair Johnson, Jim Blundell, erinhong, Paris Nicolas, Eric Carlson, Mike Wu, etal. 2017. «MIT-LCP/mimic-code: MIMIC-III v1.4». https://doi.org/10.5281/zenodo.821872.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Thompson, Benjamin, R. J. II, Jai Choi, Mohamed El-Sharkawi, Ming-Yuh Huang, y Carl Bunje. 2002. «Implicit learning in autoencoder novelty assessment». *Proceedings of the International Joint Conference on Neural Networks* 3 (febrero): 2878-83. https://doi.org/10.1109/IJCNN.2002.1007605.

Vincent, J. -L., R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, y L. G. Thijs. 1996. «The SOFA (Sepsis-related Organ Failure Assess-

ment) score to describe organ dysfunction/failure». *Intensive Care Medicine* 22 (7): 707-10. https://doi.org/10.1007/BF01709751.

Wickham, Hadley. 2017. *tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.