UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales

ANÁLISIS INTELIGENTE DE DATOS

Trabajo práctico

Fabiana A. Rossi

Agosto, 2021

INTRODUCCIÓN

El cáncer pancreático es una neoplasia extremadamente mortal, con una incidencia del 30% de los tumores diagnosticados a nivel global durante el año 2020 [1].Una vez diagnosticado, la tasa de supervivencia en los siguientes 5 años es de apenas el 10%. No obstante, si el mismo es detectado en etapas tempranas, la probabilidad de supervivencia aumenta. Desafortunadamente, muchos casos de cáncer pancreático no muestran síntomas detectables sino hasta que el mismo ha generado metástasis [2]. En este sentido, un diagnóstico temprano de los pacientes con cáncer pancreático representa una necesidad clínica insatisfecha y una ventana de oportunidad para el desarrollo de nuevas herramientas que asistan a la identificación temprana de esta afección.

En virtud de lo expuesto en el parágrafo anterior, el objetivo de este trabajo fue la implementación de diversas técnicas de clasificación supervisada y no supervisada sobre un conjunto de datos publicado recientemente, el cuál cuenta con biomarcadores de orina de pacientes con o sin adenocarcinoma pancreático ductal (PDAC, el tipo de cáncer pancreático más recurrente) [2]. Entre los marcadores de la base de datos se encontraban creatinina, LYVE1, REG1B, y TFF1. La proteína creatinina suele utilizarse como un marcador de la función renal, YVLE1 es un receptor de los vasos linfáticos y podría estar relacionado con metástasis, REG1B es una proteína asociada a la regeneración del páncreas, y TFF1 es una proteína que posiblemente esté asociada con la regeneración y reparación del tracto urinario [2].

Los resultados de este trabajo indicaron que el método de regresión logística resulta más adecuado para la clasificación de pacientes sospechados de padecer una neoplasia pancreática, y que el método de *clustering* de *k-means* reproduce más fielmente la distribución de los datos de la base de datos utilizada.

DATOS

La base de datos fue obtenida a partir de la publicación de Debernardi y colaboradores [2]. La misma cuenta con la edad (años), sexo, valores de los biomarcadores urinarios creatinina [mg/ml], LYVE1 [ng/ml], REG1B [ng/ml], y TFF1 [ng/ml] para pacientes en tres categorías de la variable diagnóstico: sanos, con alguna condición benigna, y con PDAC. Los datos fueron generados en 4 centros de investigación. En este trabajo se decidió utilizar los datos de pacientes cuyo diagnóstico era “normal” (sanos) o “maligno” (PDAC), conformando una base de datos de 381 registros, y se utilizó la totalidad de los datos sin tener en cuenta su proveniencia. Si bien esto último podría ser potencialmente contraproducente dado que los modelos supervisados estarían sesgados en el aprendizaje por los datos de un origen con mayor representación, y podría haber efectos de tipo “*batch*” no contemplados en el análisis, se entiende que el análisis de este factor escapa al objetivo del presente trabajo práctico y por ende no será tenido en cuenta.

La base de datos original presenta otros atributos que no fueron utilizados en este análisis, como ser el grado de avance de la enfermedad y otros biomarcadores urinarios. La primera variable se descartó dado que se prevé que dicha información no estaría disponible en el momento de implementación del modelo. Por otro lado, no se utilizaron los otros biomarcadores presentes en la base de datos original, ya que los mismos estaban representados en apenas un 50% de los registros.

METODOLOGÍA

En primer lugar, se realizó un pre-procesamiento de la base de datos y un análisis exploratorio de las variables y su interrelación segregando los registros de acuerdo al valor de la variable diagnóstico. A continuación se llevó a cabo un estudio estadístico para determinar si los datos seguían una distribución normal multivariada, y para testear la homocedasticidad de las matrices de covarianzas entre ambos grupos. Por otro lado, se evaluó si existían diferencias entre los vectores de estimadores de posición de la subpoblación de pacientes con diagnóstico “normal” y aquella con diagnóstico “maligno”.

Para los modelos supervisados, se dividió el set de datos en subconjuntos de entrenamiento o prueba en una proporción 70:30. Se realizó un escalado de todos los datos utilizando los parámetros de media y desvío estándar del subconjunto de entrenamiento, y se aplicaron distintos modelos: análisis de discriminante lineal (LDA), cuadrático (QDA) y cuadrático robusto, discriminante regularizado (RDA), máquinas de soporte vectorial (SVM) y regresión logística. Para el modelo de SVM se utilizaron distintas funciones *kernel* (lineal, radial o sigmoideo).

La exactitud de las predicciones de los modelos (ensayados sobre los datos del subconjunto de prueba) se determinó mediante la construcción de curvas ROC y la medición del área debajo de las mismas (AUC). Las curvas ROC indican gráficamente el compromiso de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. Se eligió la medición del AUC dado que representa un valor que puede utilizarse para comparar la exactitud de las predicciones de distintos modelos, sin que el mismo dependa del umbral elegido.

Por otro lado, se aplicaron métodos no supervisados como *t-SNE*, *k-means* y *clustering* jerárquico. Para los dos primeros modelos, se utilizaron la totalidad de los datos de la base original. En el caso del modelo de *clustering* jerárquico se utilizó una muestra de la base de datos original, de tamaño 100 (balanceada para las clases de la variable “diagnóstico”).

En el caso del modelo de *t-SNE* (que estrictamente no es un método de *clustering*, sino una herramienta que permite visualizar la distribución espacial de un conjunto de datos multidimensional en un espacio de representación de menor

dimensión), se analizaron distintos valores de *perplexity*. Dicho parámetro está relacionado con la cantidad de vecinos cercanos que utiliza el algoritmo para definir la topología de la distribución de los puntos. Se analizó también el efecto de la estandarización.

En el caso de *k-means*, se aplicó el modelo utilizando distinto número de centroides, con o sin estandarización de los datos numéricos y utilizando distinto número de variables.

Por último, en el modelo de *clustering* jerárquico se utilizaron las distancias Euclidea o Manhattan para calcular la matriz de distancias entre los registros, y se eligieron distinto número de *clusters*, cuyas distancias se calcularon utilizando las distancias completa, promedio, simple o de Ward.

RESULTADOS

ANÁLISIS SUPERVISADO

El análisis exploratorio de los datos indicó que las etiquetas de la variables diagnóstico y sexo estaban balanceadas (~50% en cada etiqueta, respectivamente), pero que los registros con diagnóstico "maligno" presentaban una mayoría de pacientes de sexo masculino, mientras que el diagnóstico "normal" estaba preferentemente representado por mujeres. Asimismo, el diagnóstico "maligno" estaba asociado a una mayor edad de los pacientes, y valores mayores en los biomarcadores. Un análisis de la correlación lineal de las variables numéricas con el método de Pearson indicó que las variables creatinina, LYVE1, REG1B y TFF1 muestran una correlación positiva significativa entre todas las combinaciones de a pares posibles, siendo más marcada para TFF1-creatinina y TFF-LYVE1 en los pacientes con diagnóstico "normal" y TFF1-REG1B en los pacientes con diagnóstico "maligno" (Figura 1).

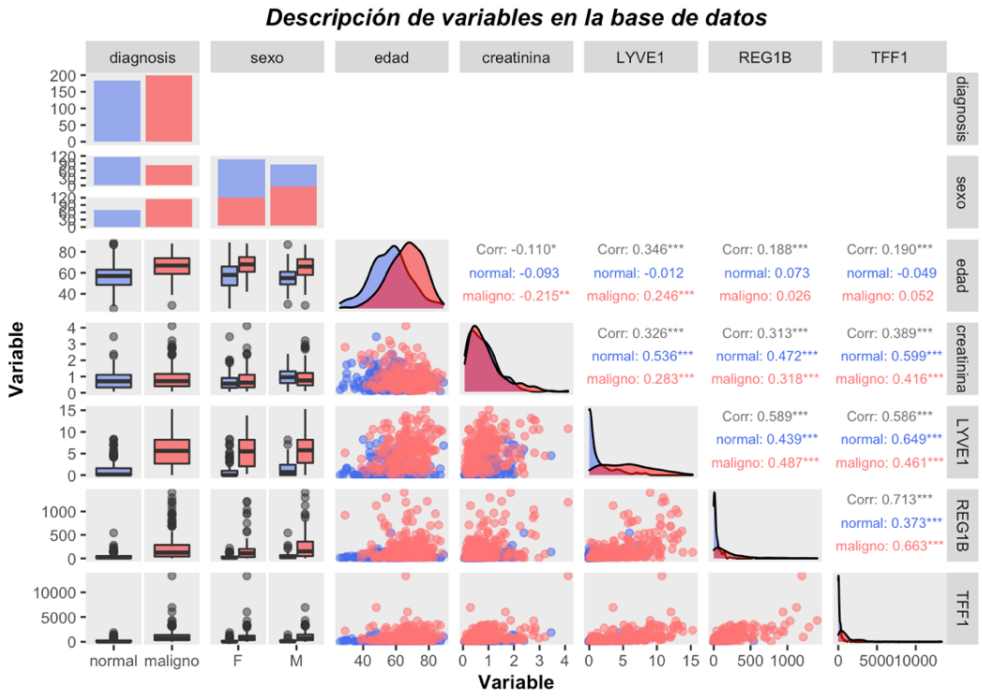

Figura 1: Análisis exploratorio de las variables en la base de datos en función de los niveles de la variable diagnóstico. Para las variables numéricas, se calculó el coeficiente de correlación lineal de Pearson.

El análisis estadístico de los datos indicó que las muestras de los pacientes con diagnóstico "normal" o "maligno" no seguían una distribución univariada normal en ninguna de las variables salvo la edad (según el test Anderson-Darling) (Figura 2 y Tabla 1), y consecuentemente tampoco seguían una distribución multivariada (se testeó utilizando el test Shapiro-Wilk) (Tabla 2).

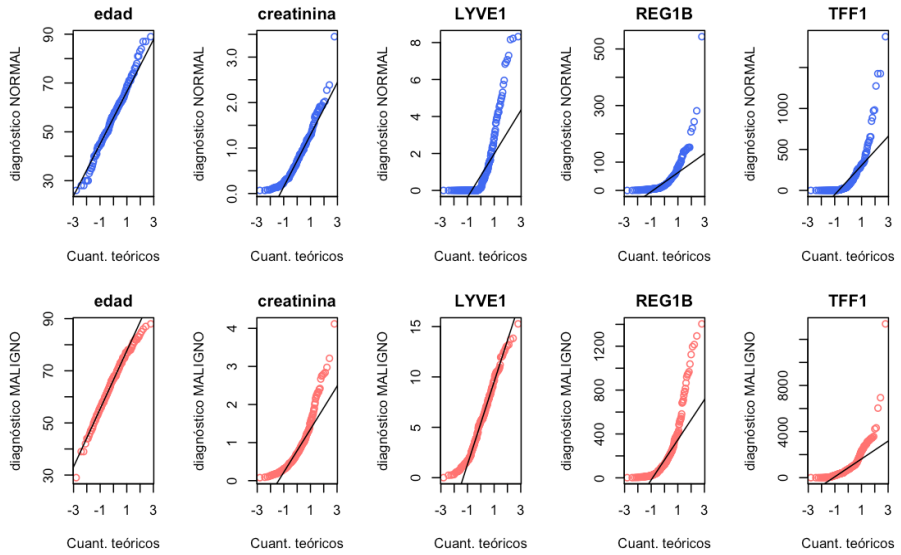

Figura 2: qqplot de cuantiles de distribución de las variables de los datos con diagnóstico "normal" o "maligno" en función de los cuantiles teóricos de una distribución normal. La recta roja indica la forma de la curva si y siguiera una distribución normal.

	Diagnóstico	
Variable	Normal	Maligno
Edad	0,5182	0,0608
Creatinina	0,0000	0,0000
LYVE1	0,0000	0,0002
REG1B	0,0000	0,0000
TFF1	0,0000	0,0000

Tabla 1: p-valor de las pruebas de normalidad univariada mediante el test de Anderson-Darling para las variables de las muestras con diagnóstico normal o maligno.

A continuación se analizó la homocedasticidad de las matrices de varianzas y covarianzas de cada grupo utilizando los test M de Box y Levene (más robusto ante la falta de normalidad). En ambos casos se rechazó la hipótesis nula que estipulaba que las matrices eran iguales (Tabla 2).

Previo al comienzo del procesamiento y aplicación de los métodos de clasificación, se realizaron pruebas que permitieron determinar que la diferencia de los vectores de medias de cada grupo era distinta a cero (Hotelling y npmv, Tabla 2). En otras palabras, se concluyó que existen variables discriminantes entre las consideradas para realizar el análisis, que podrían permitir diferenciar los grupos de pacientes.

		Diagnóstico		
Test	**Evalúa**	Normal	Maligno	**Resultado**
Shapiro-Wilk	Normalidad multivariada	$1{,}22^{-19}$	1.97^{-19}	Rechazo Ho
M de Box	Homocedasticidad	$1{,}81^{-164}$		Rechazo Ho
Levene	Homocedasticidad	$9{,}06^{-13}$		Rechazo Ho
Hotelling T^2	Igualdad en vector de medias	0		Rechazo Ho
npmv	Igualdad en vector de medias	0		Rechazo Ho

Tabla 2: Análisis estadístico delos datos de la base de datos. Se realizaron test de normalidad multivariada, de homocedasticidad de las matrices de varianzas y covarianzas, y de igualdad de vectores de medias paramétrico (Hotelling, asintóticamente normal) y no paramétrico (npmv).

Los modelos de clasificación LDA y QDA tienen como supuesto el cumplimiento de normalidad multivariada. Asimismo, el modelo de LDA supone homocedasticidad de las matrices de varianzas y covarianzas de las muestras analizadas. La Tabla 2 indica que tales supuestos no se cumplen.

Dado que varios autores han evaluado la robustez de los métodos LDA y QDA en ausencia de cumplimiento de los supuestos, y concluyeron que son útiles para la tarea de clasificación en determinadas condiciones [3-5], en este trabajo se decidió evaluar su rendimiento en el conjunto de prueba, aún cuando se sobreentiende que la falta de cumplimiento de los supuestos podría aumentar el error de predicción de los modelos de LDA y QDA. De modo adicional, se evaluaron las predicciones de una alternativa robusta de QDA.

Asimismo se aplicó el modelo RDA. Dicho método es más robusto a la colinealidad de variables, y su implementación involucra resampleo por *bootstrap* en la selección de hiperparámetros (gamma y lambda), por lo cuál es menos sensible a la falta de normalidad.

Por otro lado, las máquinas de soporte vectorial o la regresión logística son métodos más versátiles que tienen menor cantidad de supuestos, y consecuentemente se espera tengan una mejor *performance* en su implementación para clasificar pacientes con diagnóstico "normal" o "maligno".

A fin de evaluar el rendimiento de los modelos anteriormente nombrados, se entrenaron los modelos con distinto número de variables numéricas, y se determinó la exactitud de las predicciones mediante la determinación del AUC en las curvas ROC (Tabla 3 y Figura 3).

	Combinación de variables				
Variables	**1**	**2**	**3**	**4**	**5**
Creatinina	+	+	+	-	-
Edad	+	+	+	+	+
LYVE1	+	+	+	+	+
REG1B	+	+	-	+	+
TFF1	+	-	-	-	+
Modelo	**Valor de AUC**				
LDA	0,860	0,857	0,851	0,853	0,855
QDA	0,929	0,911	0,877	0,886	0,911
RDA	0,858	0,858	0,872	0,854	0,858
LogR	0,889	0,870	0,853	0,865	0,872
SVM_Lineal	0,894	0,872	0,854	0,870	0,891
SVM_Radial	0,915	0,900	0,887	0,866	0,877
SVM_Sigmoideo	0,855	0,822	0,773	0,809	0,841

Tabla 3: Análisis del performance de los distintos métodos de clasificación supervisada mediante el cálculo del área debajo de la curva (AUC) de la curva ROC, utilizando distintas combinaciones de variables. Para cada combinación de variables se indica la presencia y ausencia con un + o -, respectivamente. La escala de colores (determinada para cada modelo) indica mayores valores de AUC en verde, y menores en color rojo.

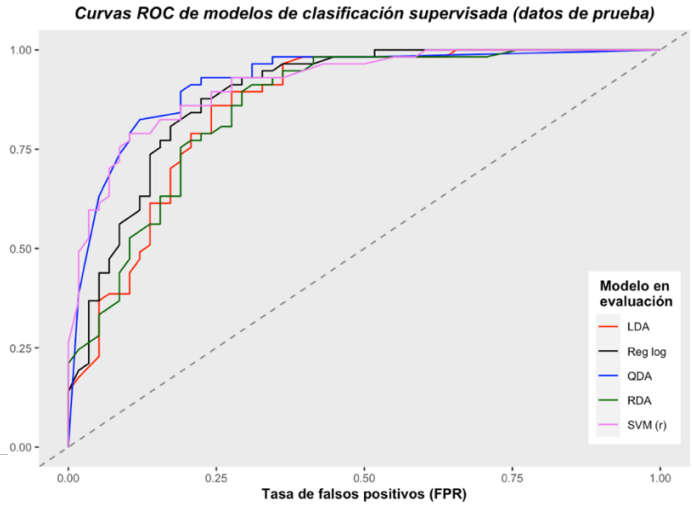

Figura 3: Curvas ROC de los modelos de clasificación supervisada aplicados sobre el subconjunto de datos de prueba y utilizando todas las variables.

Tal como puede observarse en la Tabla 3, el modelo de LDA fue uno de los métodos de peor *performance* según la métrica de AUC. El valor máximo se observó cuando todas las variables numéricas fueron utilizadas como variables discriminantes (combinación 1). Si bien este método resultó efectivo para separar las subpoblaciones estudiadas en esta cohorte, sólo permite realizar separaciones lineales de los datos.

En este sentido, el método de QDA y RDA (ligeramente más complejos, y con mayor varianza) podrían clasificar mejor los casos del subconjunto de prueba.

El análisis de los datos de este trabajo indicó que esto fue cierto para el modelo de QDA, que mostró el máximo valor de AUC en la combinación de variables 1 (al compararlo con el resto de los modelos ensayados). De hecho, una exploración visual de la representación de los datos originales (Figura 4A) y de las predicciones ingenuas del modelo QDA (Figura 4B) en un espacio de dimensiones determinados por las dos primeras componentes de un análisis de componentes principales (APC) indicó que la etiqueta de la variable diagnóstico era coincidente con su predicción, en la mayor parte de los casos.

Cuando se evaluó una alternativa robusta para QDA, mediante el cálculo de la matriz de varianzas y covarianzas con el método MCD, la precisión (*accuracy*) de la predicción *naive* fue del 54.3%, mientras que aquella del modelo QDA fue de 80.8%.

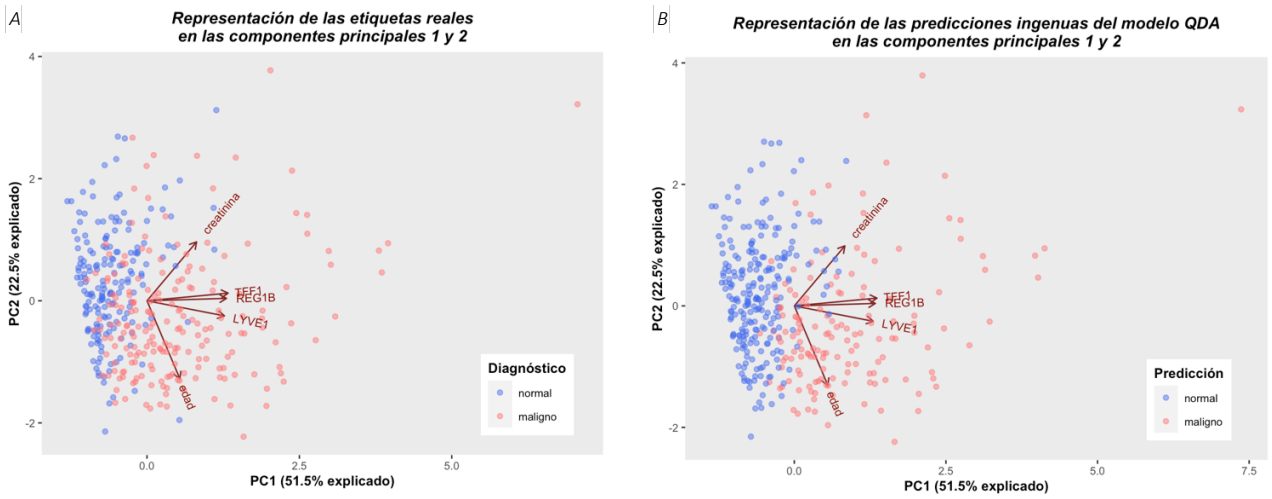

Figura 4: Representación de los valores reales (A) y de las predicciones ingenuas del modelo QDA (B) en las dos primeras componentes principales de un análisis de ACP. Se grafican además las variables originales. Los datos fueron coloreados de acuerdo al valor de la variable diagnóstico (o su predicción).

Por el contrario, el valor de AUC del modelo de RDA fue ligeramente superior al modelo LDA y sólo cuando se consideraron sólo 3 de las 5 variables numéricas de la base de datos (combinación 3, Tabla 3).

A continuación se aplicó el modelo de SVM utilizando distintas funciones *kernel*: lineal, radial y sigmoideo. Tal como se comentó anteriormente, SVM es un modelo que no requiere el cumplimiento de los supuestos de normalidad y homocedasticidad. Nuestros resultados indicaron que la *performance* del modelo SVM fue ligeramente inferior al mejor modelo QDA. Para todas las funciones *kernel* utilizadas, el mayor valor de AUC se obtuvo cuando todas las variables numéricas eran consideradas. Particularmente, la función radial fue la de mejor *performance*, seguida de la función lineal y por último, la sigmoidea (Tabla 3).

Por último, se realizó una clasificación de los datos del subconjunto de prueba mediante el modelo de regresión logística. Al igual que SVM, dicho modelo no requiere la aceptación de los supuestos de normalidad y homocedasticidad entre las matrices de varianza y covarianza de los grupos estudiados, y es particularmente sensible al desbalance de clases en la variable *target*. Dado que la proporción de datos con el valor "normal" o "maligno" era 48:52 en el subconjunto de entrenamiento, se consideró apropiada su aplicación sin realizar procedimientos particulares. Dicho método se implementó utilizando la transformación *logit*.

Si bien los datos de la Tabla 3 indican que el modelo de regresión logística no fue aquel de mejor exactitud en la predicción de pacientes con diagnóstico "maligno" en comparación con los otros modelos ensayados, se decidió ahondar en su exploración por ser un modelo con alto poder explicativo. Desde un punto de vista clínico, esto resulta particularmente relevante ya que permite a los profesionales de la salud aplicar estrategias adecuadas de manera inmediata en virtud del valor de distintas variables en los pacientes.

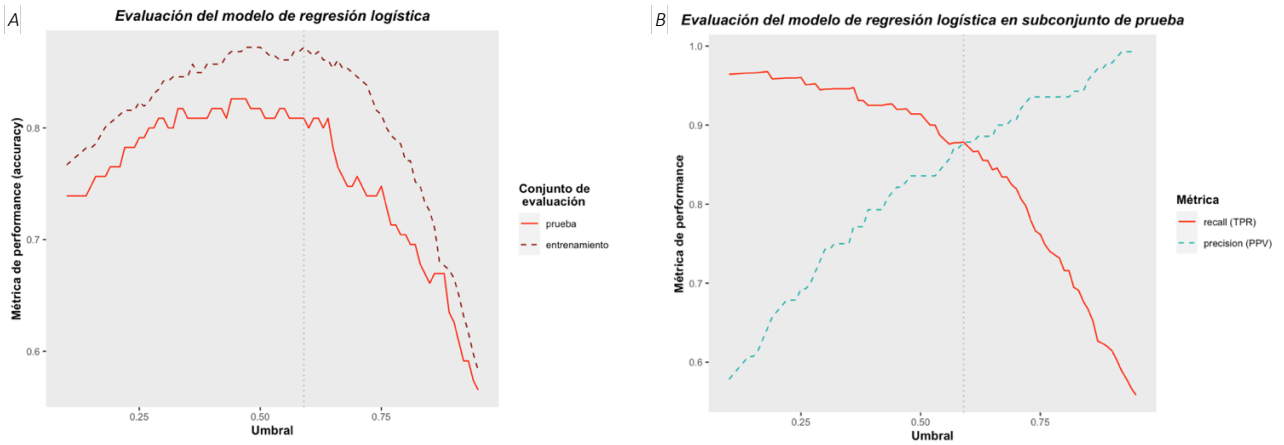

Figura 5: (A) Evaluación del modelo de regresión logística mediante la métrica de accuracy en el subconjunto de prueba (línea continua roja) o "naive" en el subconjunto de entrenamiento (línea punteada bordó) en función del umbral. (B) Evaluación del modelo de regresión logística mediante las métricas de recall (o tasa de verdaderos positivos TPR, sensibilidad- línea continua roja) o precisión (Valor predictivo positivo PPV, especificidad- línea punteada verde) en función del umbral. La línea vertical punteada (en color gris) indica el umbral en donde las curvas de recall y precisión se intersectan.

Si uno utilizara otras métricas de *performance* para evaluar al modelo de regresión logística como ser sensibilidad, especificidad o *accuracy*, debería determinar un umbral que favorezca el cumplimiento de los objetivos del modelo.

Si el objetivo de la aplicación de este modelo fuera la identificación de las personas con diagnóstico maligno, entonces la tasa de falsos negativos debería ser baja y esto requiere alta sensibilidad (*recall*). Dicha métrica, que indica la tasa de verdaderos positivos, es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo. En este caso se consideraría más "costoso" no diagnosticar a alguien potencialmente enfermo respecto de realizar pruebas innecesarias en un paciente sano con un diagnóstico falso positivo. No obstante, un valor de *recall* muy alto viene acompañado por un valor de especificidad muy bajo y un alto número de falsos positivos (Figura 5B). Las implicancias de esto último se interpretan como un elevado número de individuos sanos que son sometidos a pruebas subsiguientes (costo económico), ansiedad, posible estigma, etc. En este escenario, entonces, se decidió optar por un valor umbral intermedio en donde las curvas que distinguen cada métricas se intersectan. Para ese valor umbral, la métrica de *accuracy* (precisión) del modelo fue de un 80% (Figura 5A).

Coef.	Variable	Valor coef.	Odds ratio
β_0	Ordenada al origen	1,222	3,396
β_1	edad	0,823	2,279
β_2	creatinina	-0,728	0,482
β_3	LYVE1	1,538	4,456
β_4	REG1B	0,678	1,989
β_5	TFF1	2,538	12,663

*Tabla 4: Análisis de los coeficientes (coef.) del modelo de regresión logística para cada una de las variables, y del valor de odds ratio (aumento de riesgo), calculado como $e^{coeficiente}$. $logit(p) = \beta_0 + \beta_1$*edad + β_2*creatinina + β_3*LYVE1 + β_4*REG1B + β_5*TFF1, siendo p la probabilidad de tener un diagnóstico maligno.*

Los resultados obtenidos permitieron además analizar el impacto de las distintas variables predictoras sobre la variable diagnóstico y su sentido. Los resultados se indican en la Tabla 4.

Un estudio más detallado de la Tabla 4 indica que TFF1 es la variable que más efecto tiene en la predicción del diagnóstico maligno. En otras palabras, el incremento en una unidad de TFF1 (dejando fijas el resto de las variables) aumenta las chances de

obtener un diagnóstico maligno en 12,66 veces. Las variables LYVE1, edad y REG1B son aquellas que siguen en orden de importancia para predecir un diagnóstico maligno. Por último, la variable creatinina tiene un coeficiente negativo y un valor de odds ratio de 0,482. Esto implica que (dejando el resto de las variables fijas), un aumento en una unidad de la variable creatinina aumenta las chances de obtener un diagnóstico "normal" en 2 veces, respecto de obtener un diagnóstico "maligno" [1/0,482= 2,07].

Dado su poder explicativo y *performance* para este conjunto de datos, se considera que este modelo resulta superador al resto de los modelos estudiados en este trabajo, y el más adecuado para la clasificación de pacientes sospechados de padecer PDAC.

ANÁLISIS NO SUPERVISADO

A continuación, se aplicaron métodos no supervisados como *t-SNE, k-means* y *clustering* jerárquico.

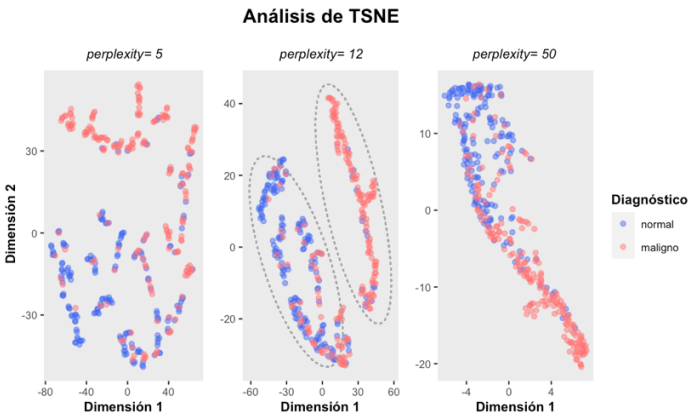

Figura 6: Análisis de TSNE para la totalidad de registros de la base de datos, sin la variable diagnóstico. Cada punto representa un paciente, y los mismos se colorearon de acuerdo al valor de su variable diagnóstico.

Si bien *t-SNE* es una herramienta que permite visualizar la distribución espacial de un conjunto de datos multidimensional en un espacio de representación de menor dimensión y no un método de *clustering*, en algunas oportunidades es útil para analizar si existe agrupamiento intrínseco de los registros. Esto es porque puntos con alta similitud tienen una alta probabilidad de permanecer cercanos en el nuevo espacio de representación. *t-SNE* es un método no paramétrico, no lineal y exploratorio.

En el caso de este trabajo, se aplicó el método de *t-SNE* sobre todos los registros de la base de datos, excluyendo la variable diagnóstico. Se analizaron distintos valores de *perplexity,* dentro del rango recomendado por los autores del modelo [6], y se colorearon los registros (cada punto) de acuerdo a la etiqueta de la variable diagnóstico. Tal como puede observarse en la Figura 6, un valor de *perplexity* de 12 permitió visualizar un posible agrupamiento de los datos en dos categorías, con puntos enriquecidos en las etiquetas de la variable diagnóstico en cada uno. En caso de que la información del valor de la variable *target* no estuviera disponible, el hecho de que se observan 2 agrupamientos podría facilitar la selección de parámetros (k) para otros métodos de *clustering*, como ser *k-means*. Dicho método se repitió con datos estandarizados, pero no se encontraron agrupamientos evidentes con valores de *perplexity* de 5, 10, 15, 20 o 25 (imágenes en archivo html).

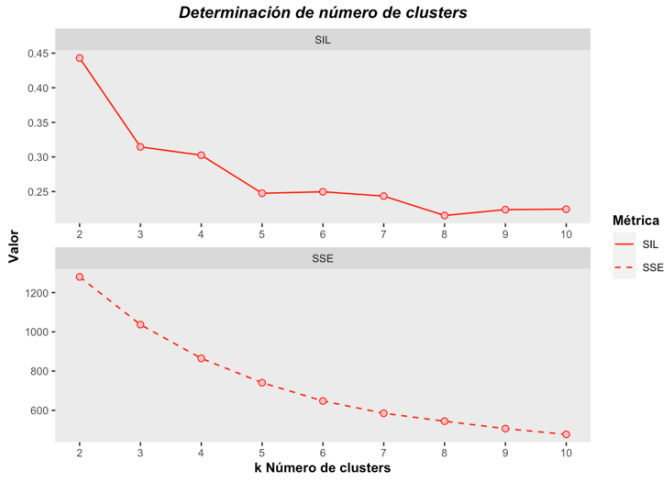

Figura 7: Cálculo de la métrica de Silhouette (SIL) y de la suma de cuadrados del error (SSE).

Con el objetivo de aplicar la técnica de *k-means* se realizaron dos cálculos para determinar el número óptimo de *clusters* de la técnica. El primero de ellos consiste en determinar el *score* de Silhouette, que permite estimar la coherencia dentro del análisis de grupos. Valores cercanos a 1 indican que el objeto está bien emparejado con su propio cúmulo y mal emparejado con los cúmulos vecinos. Por otro lado, y con el objeto de aplicar el método de codo (o "*elbow method*"), se calculó la suma de cuadrados del error (SSE) para distinto número de *clusters*. Dicho método sugiere elegir el k a partir del cuál la disminución del SSE entre valores de k es casi nula, y se produce un quiebre en la curva.

Según el *score* de Silhouette, se debería tomar un valor de k *clusters* = 2 para aplicar la técnica de *k-means*. Según el método del codo, no parece haber un valor de k claro a partir del cuál la pendiente que describe la disminución del SSE se acerca a cero (la disminución es paulatina). Por tal motivo, se decidió aplicar la técnica de *k-means* utilizando distintos parámetros y condiciones experimentales. En primer lugar se realizó un *clustering* con k=2 y k=3 en el conjunto de datos total, utilizando todas las variables numéricas estandarizadas. Los resultados de la asignación de los puntos a cada *cluster* puede observarse en la Figura 8. En dicha figura puede observarse que cuando se utiliza k=2, la cantidad de puntos en cada *cluster* es diferente, siendo el *cluster* 2 de mayor tamaño que el *cluster* 1. Asimismo, el *cluster* 1 está

conformado por registros con diagnóstico maligno, mientras que el *cluster* 2 está conformado por registros mixtos. Estos mismos resultados se observan en la Tabla 5 (experimento 1), en donde se muestra que el *cluster* 1 un 94,32% de los registros con diagnóstico maligno y el *cluster* 2 tiene un 60,75% de los datos de la categoría normal. Complementariamente se analizó el resultado de utilizar un número de *clusters* igual a 3 (k=3). En este caso, la Figura 8 muestra que el *cluster* 1 es aquel que tiene menor cantidad de puntos y que los mismos pertenecen a la categoría "maligno" en la variable diagnóstico. Los *clusters* 2 y 3 presentan mayor cantidad de registros y los mismos son de ambas categorías de la variable diagnóstico. El experimento 6 de la Tabla 5 muestra un análisis más detallado: los *clusters* 1 y 3 captan, mayoritariamente, registros con diagnóstico maligno, y el *cluster* 2 capta una gran proporción de los registros con diagnóstico "normal".

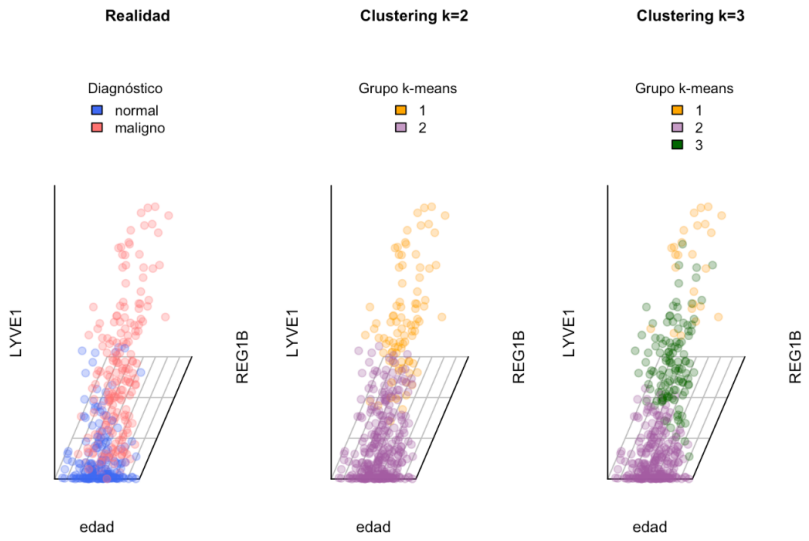

Figura 8: Representación gráfica en 3 dimensiones de los registros con sus variables estandarizadas utilizando las variables LYVE1, REG1B y edad. El color de los puntos indica (A) la categoría de la variable diagnóstico, (B) y (C) el cluster al que pertenecen los puntos cuando se utiliza k=2 o k=3, respectivamente.

CANTIDAD DE CLUSTERS = 2				
	cluster	normal	maligno	% cat. mayoritaria
1	***Variables numéricas escaladas***			
	Cluster1	5	83	94,32
	Cluster2	178	115	60,75
2	***Variables numéricas sin escalado***			
	Cluster1	1	40	97,56
	Cluster2	182	158	53,53
3	***Sin la variable TFF1, con escalado***			
	Cluster1	13	108	89,26
	Cluster2	170	90	65,38
4	***Sin las variables TFF1/creatinina, con escalado***			
	Cluster1	179	102	63,7
	Cluster2	4	96	96
5	***Sin las variables REG1B/ creatinina, con escalado***			
	Cluster1	177	81	68,6
	Cluster2	6	117	95,12
CANTIDAD DE CLUSTERS = 3				
6	***Variables numéricas estandarizadas***			
	Cluster1	0	25	100
	Cluster2	167	70	70,46
	Cluster3	16	103	86,55
7	***Variables numéricas sin escalar***			
	Cluster1	0	24	100
	Cluster2	8	64	88,89
	Cluster3	175	110	61,4

Tabla 5: Análisis del número y proporción de datos de cada categoría de la variable diagnóstico en los clusters obtenidos al utilizar distintas condiciones experimentales.

Al evaluar el efecto de la falta de estandarización de los datos en el *clustering* con k=2 o k=3 (experimentos 2 y 7 de la Tabla 5, respectivamente), se puede observar que se generan *clusters* con registros con diagnóstico maligno más pequeños, y *clusters* mixtos más grandes, al compararlo con la versión estandarizada de dichos experimentos. Es decir que la ausencia de esa transformación empeora el modo en que los datos son agrupados.

Por último se evaluó el efecto de reducir el número de variables utilizadas en el modelo de *clustering*. En los experimentos 3, 4 y 5 de la Tabla 5 se indican los resultados de *k-means* al remover las variables TFF1, TFF1 + creatinina o REG1B + creatinina, respectivamente. Puede observarse que, en comparación con el experimento 1, eliminar alguna de las variables mejora el enriquecimiento de los *clusters* en los registros de cada clase. En particular, el experimento 5 segrega los datos en una proporción ~ 70% con diagnóstico normal en el *cluster* 1 y ~95% con diagnóstico maligno en el *cluster* 2, y aumenta el enriquecimiento de una dada etiqueta en cada grupo (respecto a lo que sucede en el experimento 1). En otras palabras, eliminar las variables REG1B + creatinina (que coincidentemente en la regresión logística eran aquellas variables que tenían coeficientes de menor magnitud) mejora el modo en que los datos son agrupados.

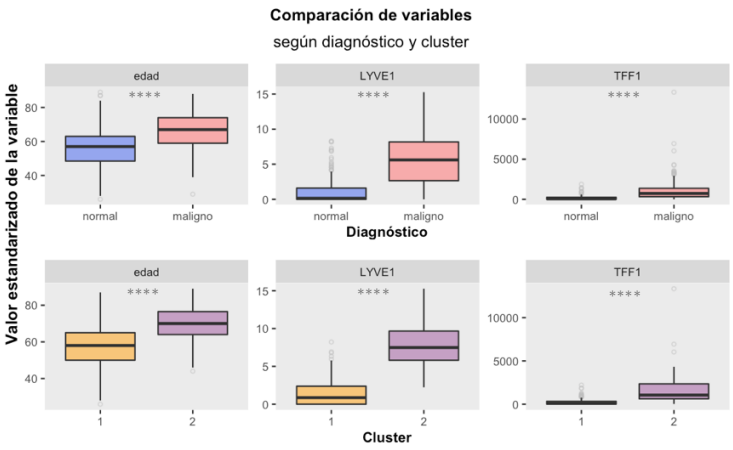

Figura 9: Distribución de las variables utilizadas en el experimento 5 (edad, LYVE1 y TFF1) de acuerdo al valor de la variable diagnóstico o al número de cluster de k-means. (****) indica un p-valor < 0,0001 en un test de comparación de medias Z entre los tratamientos.

Complementariamente se evaluó la distribución de cada una de las variables utilizadas en el experimento 5 (edad, LYVE1 y TFF1) de manera individual, de acuerdo al valor de la variable diagnóstico o al número de *cluster* de *k-means*. Los resultados de la Figura 9 permiten concluir que los

grupos generados por *k-means* siguen una distribución de las variables semejante a aquellas en las cuales se agrupan los datos por el valor de la variable diagnóstico. Dado que el número de registros en cada grupo era mayor a 30, se aproximó la distribución por el teorema central del límite y aplicó un test z de diferencia de medias para dos muestras. Los test individuales indicaron que existían diferencias significativas en las medias de ambos grupos (normal o maligno) para cada una de las variables estudiadas. Este fenómeno también se comprobó para los datos cuando fueron segregados de acuerdo al número de *cluster*. Es decir, la clasificación por *k-means* permitió obtener grupos que se diferenciaban de manera significativa en sus variables edad, LYVE1 y TFF1.

Un análisis integral de todos los experimentos permitió observar que los datos con diagnóstico "normal" se agrupaban en ciertos *cluster* con algunos registros con diagnóstico "maligno", mientras que ciertos registros con PDAC se agrupaban "puros". Este resultado indica que existen registros de pacientes con PDAC que presentan valores extremos en las variables ensayadas, y por ende son fácilmente separables del conjunto de datos, mientras que hay otros registros en la misma categoría que se asemejan más a los pacientes con diagnóstico normal, y por ende se agrupan en un mismo *cluster*.

Por último, se utilizó una muestra de tamaño 100 de la base de datos original, con el objeto de aplicar la técnica de *clustering* jerárquico. Dicha muestra se encontraba balanceada para las clases de la variable diagnóstico en una proporción 1:1, la cual fue omitida en el análisis. Se utilizaron las distancias Euclidea o Manhattan para calcular la matriz de distancias entre los registros con variables estandarizadas, se eligieron distinto número de *clusters*, y las distancias entre los mismos se calcularon utilizando las distancias completa, promedio, simple o de Ward. El coeficiente de correlación cofenético, que indica la proximidad entre las distancias originales y finales de los registros (luego de agrupados), fue calculado para las distintas matrices de distancias (la calculada con distancia Manhattan y aquella calculada con distancia Euclidea) y utilizando los distintos métodos para calcular distancia entre clusters (Tabla 6).

Manhattan			
completo	**promedio**	**simple**	**ward**
0,736	0,872	0,854	0,725
Euclidea			
completo	**promedio**	**simple**	**ward**
0,787	0,870	0,826	0,649

Tabla 6: Coeficiente de correlación cofenético para los distintos métodos utilizados para calcular distancias entre clusters (completo, promedio, simple, Ward), a partir de matrices de distancias de registros calculadas con las distancias de Manhattan o Euclídea.

En la Figura 10 se muestra un dendograma jerárquico de dos *clusters*, para el cuál la matriz de distancias entre registros se construyó utilizando la distancia de Manhattan y la distancia entre *clusters* se calculó con la distancia de Ward. El color de las líneas del dendograma indica distintos *clusters*, y el color de los nodos indica el valor de la variable diagnóstico.

La Tabla 7 muestra un análisis más detallado del número de registros en cada *cluster*, el valor de su variable diagnóstico y el porcentaje de la clase mayoritaria en cada *cluster*.

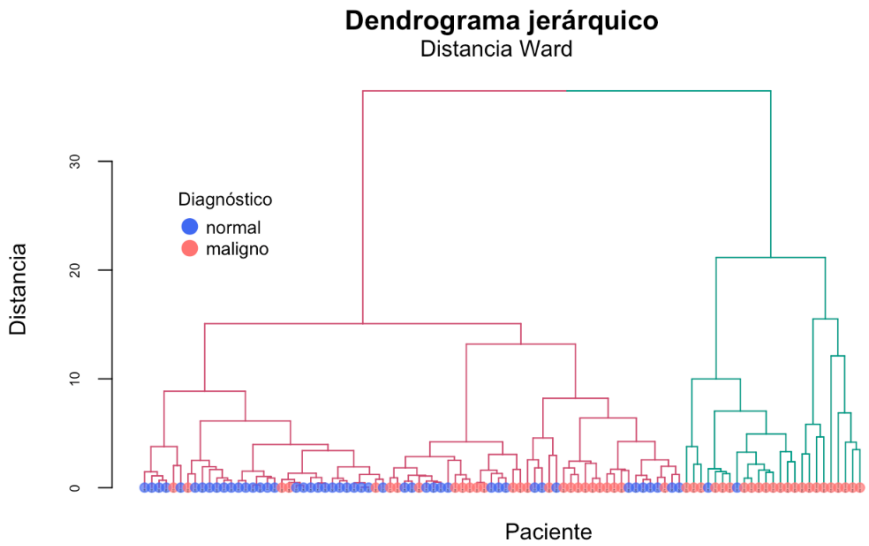

Figura 10: Dendograma jerárquico de una muestra de 100 datos de la base de datos, en el cuál la matriz de distancias entre registros se construyó utilizando la distancia de Manhattan y la distancia entre clusters se calculó con la distancia de Ward. El color de las líneas del dendograma indica distintos clusters, y el color de los nodos indica el valor de la variable diagnóstico.

CANTIDAD DE CLUSTERS = 2			
cluster	**normal**	**maligno**	**% cat. mayoritaria**
1	**ward**		
cluster1	48	27	64
cluster2	2	23	92
2,3	**promedio / simple**		
cluster1	50	49	50,51
cluster2	0	1	100
4	**completo**		
cluster1	50	45	52,63
cluster2	0	5	100
CANTIDAD DE CLUSTERS = 3			
5	**ward**		
cluster1	48	27	64,0
cluster2	0	9	100
cluster3	2	14	87,5
8	**completo**		
cluster1	49	29	62,82
cluster2	1	16	94,12
cluster3	0	5	100

Tabla 7: Análisis del número y proporción de datos de cada categoría de la variable diagnóstico en los clusters obtenidos al utilizar distintas condiciones experimentales.

Al analizar la Figura 10 se puede observar claramente que el número de registros en cada *cluster* está desbalanceado, siendo uno de los *clusters* de menor tamaño y enriquecido en una única etiqueta de la variable diagnóstico de los registros que lo conforman. Este mismo resultado puede observarse en la Tabla 7, en el experimento 1: el *cluster* 1 tiene 48 registros con diagnóstico normal y 27 con diagnóstico maligno, mientras que el *cluster* 2 tiene 23 registros con diagnóstico maligno y

solamente 2 con diagnóstico normal. Para el caso de los experimentos 2 y 3 los resultados son semejantes (distancias promedio y simple): el *cluster* 1 era mayoritario y tenía una proporción aproximadamente 1:1 de las dos etiquetas, y el *cluster* 2 tenía tan sólo 1 registro de la categoría diagnóstico maligno. El experimento 4 presentó una tendencia similar.

Al realizar nuevos dendogramas utilizando un número de *clusters* igual a 3 y con cualquiera de las variantes ensayadas (experimentos 5 y 6), se observa que la distribución de los registros en cada *cluster* es desigual. El *cluster* 1 presenta mayor cantidad de registros, con diagnóstico "normal" o "maligno" en una proporción ~5:3, y el *cluster* 2 y 3 presenta pocos registros, mayoritariamente con diagnóstico "maligno" (Tabla 7).

En virtud de los resultados de la Tabla 6, se esperaba que los resultados del *clustering* jerárquico determinado por el método de distancias promedio fuera el de mayor exactitud al agrupar los datos. No obstante, la Tabla 7 indicó que, el método de calculo de distancias de Ward resultó más adecuado.

Finalmente, se ensayó el efecto de la remoción de distintas variables (TFF1 o REG1B + creatinina) o la no-estandarización de los datos en el agrupamiento y rendimiento del método de *clustering* jerárquico. Los dendogramas obtenidos en ambos casos mostraron un agrupamiento de peor calidad, en la que se formaban *clusters* con muy pocos registros y *clusters* de etiquetas mixtas (imágenes en archivo html). Si bien uno no tendría *a priori* la etiqueta de la variable diagnóstico en un método de clasificación no supervisada, un análisis de diferencias de medias de los *clusters* podría esclarecer si los mismos fueron o no representativos de la realidad.

Respecto de los métodos de clasificación no supervisado, se concluye entonces que el método de *t-SNE* es una buena herramienta de visualización, que asiste en la selección de número de *clusters* a utilizar en otros métodos, tales como los utilizados en este trabajo. El método de *k-means* superó al *clustering* jerárquico en el modo en que agrupó los registros de la base de datos, dado que reprodujo más fielmente la distribución de los datos en cada categoría de la variable diagnóstico.

CONCLUSIONES

En el presente trabajo práctico se aplicaron distintas técnicas de clasificación supervisada y no supervisada sobre un conjunto de datos que contenía información de sexo, edad y varios biomarcadores urinarios, de pacientes con diagnóstico "normal" o que padecían PDAC (diagnóstico maligno).

El objetivo de los métodos de clasificación supervisada era la predicción de la variable diagnóstico en un subconjunto de datos previamente desconocidos por el modelo. A tal fin, en primer lugar, se dividió el conjunto de datos en un subconjunto de entrenamiento y otro de prueba, en una proporción 7:3. Luego se evaluó la *performance* de los modelos de LDA, QDA, RDA, SVM (utilizando distintas funciones kernel) y regresión logística, mediante la construcción de curvas ROC y la determinación del AUC correspondiente. La *performance* de los modelos se ordenó, en manera decreciente, del siguiente modo: QDA > SVM_R > SVM_L, regresión logística > RDA > LDA > SVM_S, siendo (R), (L) y (S) las funciones *kernel* radial, lineal y sigmoidea, respectivamente. Dichos valores se obtuvieron cuando todas las variables numéricas eran consideradas, salvo para el modelo de RDA, en el cuál el máximo valor de AUC se consiguió cuando se descartaron las variables REG1B y TFF1.

Si bien la distribución multivariada de ambas muestras no era normal ni homocedástica, los métodos de LDA y QDA resultaron robustos ante la violación de dichos supuestos. Estos resultados están en concordancia con publicaciones previas, que indican los modelos antes mencionados son robustos ante la ausencia de los supuestos en determinadas condiciones [3-5]. De hecho, el modelo de QDA fue aquél que tuvo mayor valor de AUC, en comparación con el resto de los métodos.

Dado su poder explicativo y *performance* para este conjunto de datos, se consideró que el modelo de regresión logística resulta superador al resto de los modelos estudiados en este trabajo, y el más adecuado para la clasificación de pacientes sospechados de padecer PDAC. En caso de implementación de este modelo, sería interesante evaluarlo cuando el entrenamiento fuera realizado para cada sexo por separado, o cuando el mismo contemplase combinación de variables.

Respecto los métodos de clasificación no supervisado se utilizaron los datos sin la variable diagnóstico. El método de *t-SNE* indicó que los datos presentaban una distribución espacial (en las dimensiones de *t-SNE*) que formaba dos *clusters*. Al colorear cada punto de acuerdo a la etiqueta de la variable diagnóstico, se observó que cada grupo estaba enriquecido en un tipo de diagnóstico. El método de *k-means* superó al *clustering* jerárquico en el modo en que agrupó los registros de la base de datos. Reprodujo más fielmente la distribución de los datos en cada categoría de la variable diagnóstico, y en particular el mejor agrupamiento se obtuvo cuando el número de *clusters* era 2 y las variables REG1B y creatinina no eran incluidas en el análisis. Es notable que dichas variables son aquellas con menores coeficientes (en módulo) en el modelo de regresión logística.

BIBLIOGRAFÍA

1.IARC. (Agencia Internacional para la Investigación en Cáncer), https://gco.iarc.fr/today. Fuente original: GLOBOCAN 2020).

2.Debernardi S., O'Brien H., Algahmdi A.S., Malats N., Stewart G.D., Pljesa-Ercegovac M., Costello E., Greenhalf W., Saad A., Roberts R., Ney A., Pereira S.P., Kocher H.M., Duffy S., Blyuss O., Crnogorac-Jurcevic T. A combination of urinary biomarker panel and PancRISK score for earlier detection of pancreatic cancer: A case-control study. PLoS Med. 2020, 17, e1003489, doi: 10.1371/journal.pmed.1003489.

3.Clarke W.R., Lachenbruch P.A., Broffitt B. How non-normality affects the quadratic discriminant function. Communications in Statistics - Theory and Methods. 2007, 8, 1285-1301, doi: 10.1080/03610927908827830.

4.Lachenbruch P.A., Sneeringer C., Revo L.T. Robustness of the linear and quadratic discriminant function to certain types of non-normality. Communications in Statistics. 2007, 1, 39-56, doi: 10.1080/03610927308827006.

5.Nakanishi H., Sato Y. The performance of the linear and quadratic discriminant functions for three types of non-normal distribution. Communications in Statistics - Theory and Methods. 2007, 14, 1181-1200, doi: 10.1080/03610928508828970.

6.van der Maaten L., Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research. 2008, 2579-2605.

BIBLIOGRAFÍA consultada

- https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/
- https://towardsdatascience.com/interpreting-coefficients-in-linear-and-logistic-regression-6ddf1295f6f1