FACULTAD DE INFORMÁTICA

PRÁCTICA 1

Aprendizaje Automático I

Borja Souto | Nina López | Carmen Lozano | Carlos Hermida

2º GCID Curso 2021/2022

UNIVERSIDADE DA CORUÑA

1. Introducción

Según el estudio de (Cáncer de mama - Estadísticas, 2022), el cáncer de mama es la quinta causa de muerte en mujeres a nivel mundial, siendo la enfermedad tumoral maligna más común. En 2020, se estima que 684 996 mujeres murieron de cáncer de mama y se diagnosticaron 2 261 419 casos nuevos en todo el mundo. La tasa de supervivencia a 5 años una vez detectado el cáncer es del 90 por ciento, esta tasa de supervivencia incluye a todas las mujeres con cáncer de mama, sin importar la etapa.

El apoyo para concienciar sobre esta enfermedad y los fondos para la investigación han ayudado a avanzar en el diagnóstico y tratamiento. Los índices de supervivencia han aumentado, y el número de muertes asociadas continúa reduciéndose, las muertes por cáncer de mama disminuyeron un 40 por ciento de 1989 a 2017, según la ACS. Esto se debe en su mayor parte a factores como una detección más temprana, un nuevo acercamiento personalizado al tratamiento, y una mejor comprensión de la enfermedad.

Las redes neuronales artificiales (RNA), son una herramienta muy potente para el análisis de conjuntos de datos donde hay relaciones no lineales entre los datos que se estudian y la información a predecir, estas están jugando un papel importante como método de predicción y clasificación en el campo de la medicina.

En relación a lo anterior, las posibles ventajas de resolver el problema tratado con nuestra base de datos sería ayudar a predecir la posibilidad de supervivencia, tras 5 años, de una mujer que se somete a una operación de cáncer de mama, mediante una RNA.

2. Descripción del problema

La base de datos que utilizamos es *Haberman's Survival Data Set,* la cogimos del repositorio de la Universidad de California al que se puede acceder desde el siguiente enlace https://archive.ics.uci.edu/ml/datasets/haberman%27s+survival . Contiene casos de un estudio llevado a cabo entre los años 1958 y 1970 en la Universidad de Chicago Billing´s Hospital sobre la supervivencia de los pacientes que se han sometido a una cirugía por cáncer de mama.

Dispone de 306 instancias, cada una de ellas tiene 4 atributos numéricos (incluyendo la salida) con sus respectivos valores. El primero (Age of patient at time of operation), nos dice la edad del paciente en el momento de la operación. El segundo (Patient's year of operation), el año de operación. El tercero (Number of positive axillary nodes detected) el número de ganglios axilares positivos detectados. La salida del problema será un 1 si el paciente sobrevive 5 años o más, y un 2 si muere antes de 5 años.

En la siguiente tabla (Figura 1) mostramos el mínimo, máximo, media y desviación típica de los atributos de entrada:

	VARIABLE	MÍNIMO	MÁXIMO	MEDIA	DESVIACIÓN TÍPICA
1	Edad en el momento de operación	30	83	52.4575	10.8035
2	Año de operación	58	69	-	3.2494
3	Ganglios axilares detectados	0	52	4.02614	7.18965

Figura 1. Resumen de los datos de los pacientes: mínimo, máximo, media y desviación típica

En el caso de la variable "Año de operación", en la base de datos se guardan los valores con únicamente dos cifras, por lo que los mostramos así en la Figura 1 (por ejemplo, el año 1958 viene representado como 58). Además, para este atributo, no mostramos la media, dado que no lo consideramos información relevante. Lo que sí que consideramos importante es saber el año en el que más operaciones se realizaron, cosa que se puede ver en el histograma de la Figura 2.

Los valores de los atributos "Edad del paciente" y "Número de ganglios axilares detectados" no están acotados. Sin embargo, el atributo "Año de operación" sí lo está, pues el estudio fue realizado entre los años 1958 y 1970.

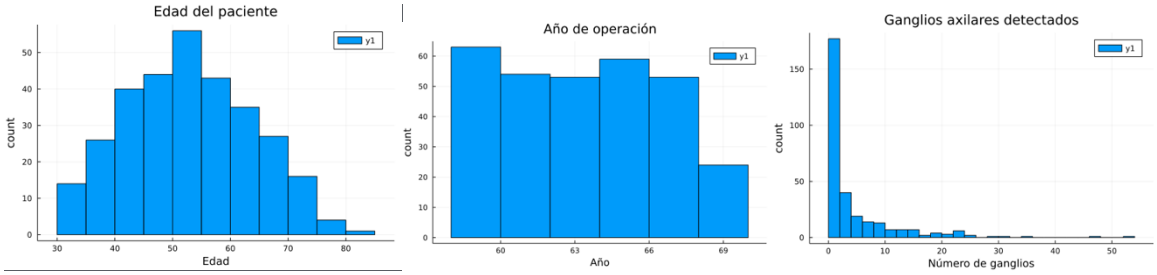

Figura 2. Histogramas de los atributos de entrada

Podemos ver en la Figura 3 que la clase 1 de supervivencia (es decir, vivir en los siguientes 5 años) tiene muchos más datos, que la clase 2 de no supervivencia (morir antes de que pasen 5 años). La primera tiene 225 datos, es decir un 74% del conjunto, mientras que la segunda tiene 81, es decir, un 26%. Vemos por lo tanto que la distribución en clases está sesgada, cosa que puede que influya a la hora de resolver el problema, pero que a nivel real es bueno, ya que sobrevive más gente de la que no sobrevive.

Figura 3. Histograma del atributo de salida (survival status)

Ya son muchos los estudios que usan la inteligencia artificial y, más concretamente las redes de neuronas artificiales para el diagnóstico médico, Baxt (Baxt, 1995) mostró la exactitud predictiva de estos modelos. Algunos autores como Ravdin (Ravdin & Clark, 1992) o Jefferson (Jefferson et al., 1997) modelaron sistemas para la predicción de una recaída tras haber sido operados de cáncer de mama y pulmón respectivamente, otros como Ting (Ting et al., 2019) modelaron sistemas para ayudar a los médicos en el diagnóstico de cáncer de mama categorizando las imágenes médicas entrantes como malignas, benignas o sanas.

Mudunuru & Skrzypek (Mudunuru & Skrzypek, 2020) realizó una investigación con el objetivo de comparar el resultado de las predicciones de supervivencia del cáncer de mama con las redes de neuronas artificiales, árboles de decisión y modelos logísticos. Para comparar estos modelos calcularon la precisión y concluyeron que en los modelos desarrollados con métodos RNA y árboles de decisión apenas se encuentran diferencias significativas, aunque el método RNA proporcionó una mejor especificidad.

3. Desarrollo

3.1. Descripción

El problema que se nos plantea es: dado un paciente determinado del que conocemos la edad del paciente cuando le operaron, el año de la operación y el número de ganglios axilares positivos ¿sufrirá una recaída tras 5 años desde la operación? Para responder a esta pregunta entrenaremos una red neuronal artificial (RNA), en concreto un perceptrón multicapa, que nos ayude a predecir si un paciente podrá sobrevivir o no en un período de 5 años, y además probaremos también a predecirlo usando un árbol de decisión (decision tree, DT), una máquina de soporte vectorial (SVM) y el método de los k vecinos más cercanos (kNN). Todo esto lo hacemos mediante el método de validación cruzada, en el que se hacen varias iteraciones *(folds*), y cada una tiene diferentes conjuntos de entrenamiento y de test. De cada *fold* se obtiene su precisión y su puntación F1 (análisis estadístico de precisión mediante un test) y finalmente se hace la media de todos estos. Nosotros empleamos 10 iteraciones para llevar a cabo nuestro estudio.

La validación cruzada lo que garantiza es que, si en algún caso se toma un conjunto de entrenamiento que no cumple las características que debería cumplir uno ideal, esto no va a tomarse por resultado final, sino que solo como parte de él, de manera que la media de los resultados va a rondar siempre los mismos valores, da igual la cantidad de veces que se ejecute.

Consideramos todas las variables de la base de datos relevantes para nuestro problema. La edad, porque generalmente, cuanto más joven sea una persona, más propensa es a recuperarse con facilidad de una operación, y suelen tener mejor salud que las personas de más avanzada edad. El año de operación lo consideramos importante porque a medida que pasa el tiempo, las técnicas de medicina avanzan y, por tanto, los métodos de operación, o materiales usados serán más seguros en años más recientes, y por tanto estos estarán relacionados con el éxito de las operaciones. Y, por último, el número de ganglios es también relevante, porque es algo relacionado directamente con el cáncer de mama; cuantos más ganglios axilares se tengan, peor.

Los valores de los atributos de nuestra base de datos requieren de normalización. La forma más adecuada de realizar esto, para el atributo “Año de operación”, sería mediante máximo y mínimo, pues los valores están acotados. Sin embargo, para los atributos “Número de ganglios axilares detectados” y “Edad del paciente”, la forma más correcta sería mediante media y desviación típica, pues los valores son resultado de una medición y algún valor puede salirse y ser muy alto o muy bajo (lo que crearía una normalización errónea en caso de hacerlo con el método del mínimo y máximo).

3.2. Resultados

Una vez cargada la base de datos y normalizados los datos, definimos los parámetros que vamos a utilizar, como la topología, la tasa de aprendizaje, el número máximo de ciclos por entrenamiento, el porcentaje de patrones que se usarán para validación, el número máximo de ciclos en los que, si no mejora el *loss* en el conjunto de validación, se para el entrenamiento, y el número de veces que se entrena la RNA para cada *fold*, al no ser determinístico el entrenamiento. Además, definimos los parámetros característicos de cada tipo de modelo (SVM, DT, kNN), que se le pasan en formato de diccionario. A continuación, entrenamos la red neuronal artificial, las SVM, los árboles de decisión, y los KNN.

Cabe mencionar que, como la implementación de las RNA, es una hecha por nosotros, para comprobar que entrenada bien y no sobreentrenaba la red, obteníamos los *loss* y dejabamos de entrenar la red cuando los valores de estos no mejoraban. En la Figura 5 podemos ver los valores de los *loss* para los conjuntos de entrenamiento y test en cada iteración

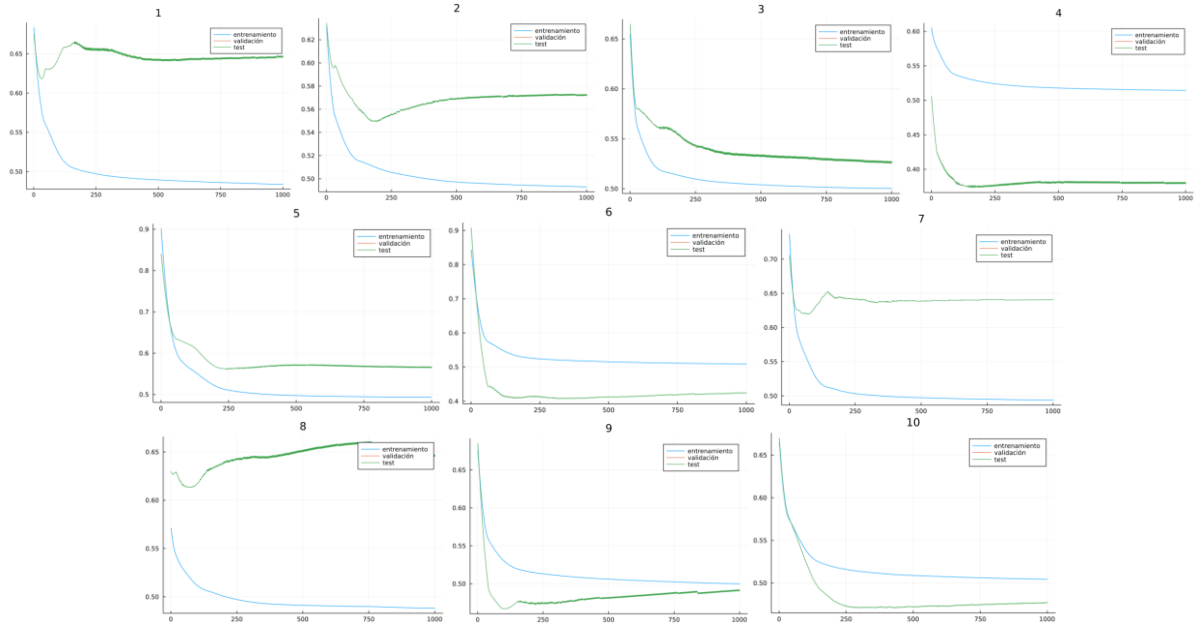

Figura 5. Gráficas correspondientes a cada iteración donde se muestran en verde los loss *del conjunto de test y en azul los del conjunto de entrenamiento.*

Comenzamos con las RNA, probado diferentes topologías hasta obtener una que nos de buenos resultados, probando varias veces cada topología. Algunas de las mejores topologías que probamos son [4,3], [2,2] y [2], siendo esta última la mejor de las tres. Los resultados que obtuvimos con esta se pueden ver en la Figura 6.

FOLD	ACCURACY (EN %)	F1 SCORE (EN %)
1/10	64.90	77.78
2/10	74.00	81.71
3/10	70.65	82.19
4/10	85.81	91.68
5/10	77.29	85.03
6/10	85.10	91.89
7/10	69.80	79.09
8/10	69.40	78.67
9/10	76.27	84.89
10/10	80.27	87.35

MEDIA ACCURACY	MEDIA F1 SCORE	DESVIACIÓN TÍPICA
75.34	-	6.93
-	84.03	5.09

Figura 6. Tablas de los resultados de la RNA con mejor topología

Probamos a continuación con la SVM, de nuevo con varios valores para los parámetros. Tras probar con distintos tipos de *kernel*, escogimos el *kernel* de función de base radial (rbf) que es el *kernel* predeterminado usado dentro del algoritmo de clasificación, este permite construir círculos o hiperesferas. También le pasamos los parámetros grado del *kernel*, gamma y C, siendo este último una forma de controlar el sobreajuste. Finalmente obtuvimos los mejores resultados para kernelDegree = 4, kernelGamma = 3 y C = 2 (aunque también obtuvimos buenos resultados para los valores 4, 4, 3 respectivamente), que son los que se muestran en la Figura 7.

FOLD	ACCURACY (EN %)	F1 SCORE (EN %)
1/10	29.03	38.89
2/10	80.65	87.5
3/10	77.42	86.27
4/10	64.52	78.43
5/10	80.65	88.46
6/10	67.74	80.00
7/10	80.00	87.50
8/10	86.67	92.00
9/10	66.67	79.17
10/10	70.00	80.85

MEDIA ACCURACY	MEDIA F1 SCORE	DESVIACIÓN TÍPICA
70.33	-	16.27
-	79.91	15.13

Figura 7. Tablas de los resultados de la SVM con mejores parámetros

Tal y como hicimos para los anteriores modelos, probamos también con diferentes profundidades para el modelo de árbol de decisión, siendo en este caso claro que el mejor de los resultados salía con una profundidad de 4. Esto lo mostramos en la figura 8.

FOLD	ACCURACY (EN %)	F1 SCORE (EN %)
1/10	80.65	88.46
2/10	70.97	80.85
3/10	80.65	86.96
4/10	61.29	74.99
5/10	70.97	82.35
6/10	83.87	90.91
7/10	66.67	78.26
8/10	80.00	86.96
9/10	63.33	70.27
10/10	40.00	47.06

MEDIA ACCURACY	MEDIA F1 SCORE	DESVIACIÓN TÍPICA
69.84	-	13.13
-	78.71	12.85

Figura 8. Tablas de los resultados del árbol de decisión con profundidad ideal

Por último, probamos varias veces con diferentes valores para el parámetro “numNeighbors” que hace referencia al número de datos más cercanos que se van a tener en cuenta a la hora de clasificar una nueva observación. Es decir, un dato se clasifica en la clase a la que pertenezcan la mayoría de sus vecinos. Es por esto que se suele coger un número de vecinos impar para evitar los empates. Probamos, por tanto, con 3 y 5 siendo 3 el número con el que obteníamos mejores resultados, que mostramos en la Figura 9.

FOLD	ACCURACY (EN %)	F1 SCORE (EN %)
1/10	77.42	86.79
2/10	77.42	85.11
3/10	74.19	83.33
4/10	70.97	79.07
5/10	80.65	87.5
6/10	58.06	73.47
7/10	73.33	83.33
8/10	30.0	27.57
9/10	66.66	78.26
10/10	73.33	81.82

MEDIA ACCURACY	MEDIA F1 SCORE	DESVIACIÓN TÍPICA
68.20	-	14.85
-	76.63	17.74

*Figura 9. Tablas de los resultados de kNN con un valor de “*numNeighbors” ideal

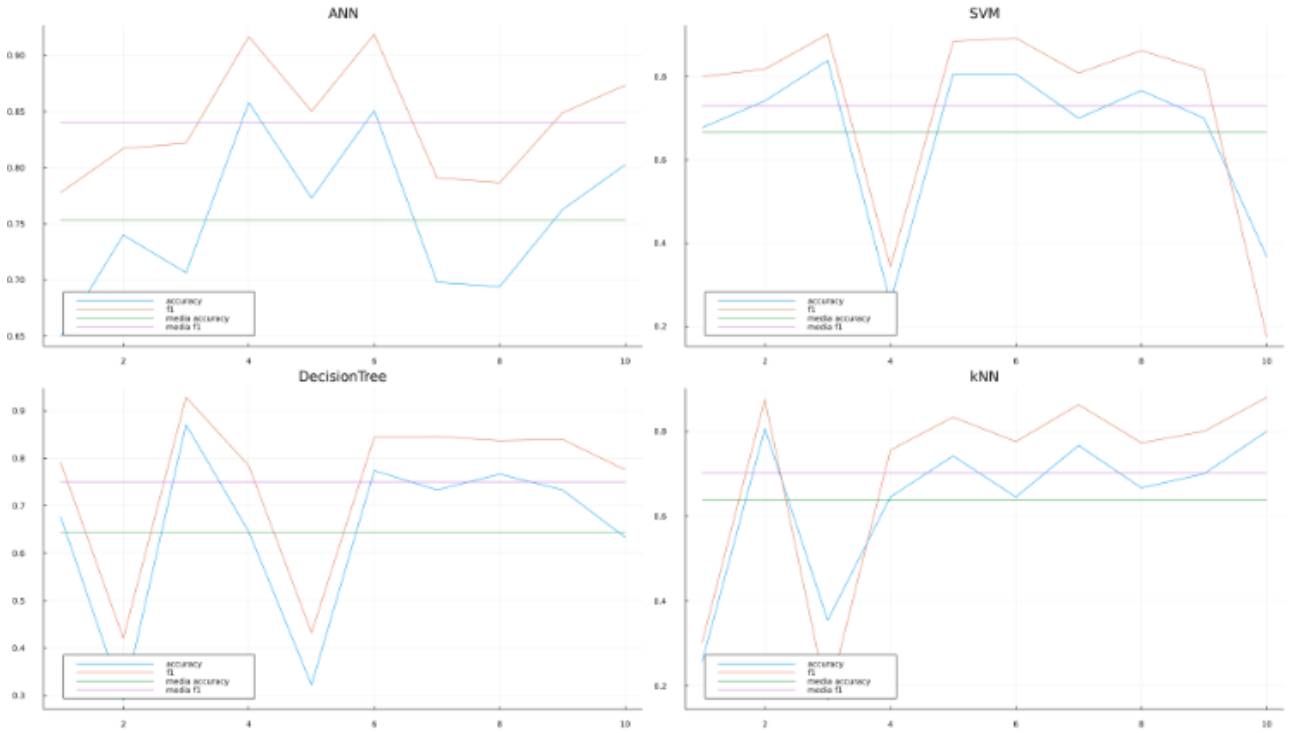

Figura 10. Gráficas de los resultados obtenidos para cada tipo de modelo

3.3 Discusión

Al finalizar este procedimiento de pruebas concluimos que los resultados obtenidos para la probabilidad de sobrevivir tras cinco años desde la operación no son aceptables, con una perspectiva desde el campo de la medicina, ya que todos los modelos nos dan aproximadamente un 70% de probabilidad de acierto. Sin embargo, desde el punto de vista teórico del funcionamiento de estos modelos, un 70% de precisión a la hora de predecir no es un resultado tan malo, pero aún así, la prioridad es al aplicarlo a nuestro problema real.

En base a los resultados obtenidos en el apartado anterior vemos que la RNA con topología [2] (1 capa oculta de 2 neuronas) es la que mejor precisión (*accuracy*) tiene, siendo esta de un 75.66%, y un F1-score de un 84.26%, con una desviación típica baja en ambos casos. Sin embargo, la SVM, DT y kNN dan resultados parecidos entre ellos, ligeramente menores a los de la RNA, y además, con unas desviaciones típicas bastante mayores. Por lo tanto, concluimos que el mejor de los modelos es la RNA.

En todos los modelos hemos obtenido algún *fold* con unos valores de *accuracy* y *F1-score* muy bajos respecto a los demás. Esto puede ser debido a que al hacer c*ross-validation*, en el momento de crear el conjunto de entrenamiento, se escogieron instancias que juntas no cumplían las características ideales de dicho conjunto, las cuáles son: ser significativo y ser representativo.

4. Conclusiones

Lo que hicimos por lo tanto hasta ahora, fue analizar y normalizar las diferentes variables de la base de datos de *Haberman's Survival Data Set* en función de sus características. Una vez hecho esto probamos 4 modelos diferentes que nos ayudarían a predecir la supervivencia tras 5 años, que serían máquinas de soporte vectorial (SVM), redes neuronales artificiales (RNA), árboles de decisión y el método de los k vecinos más cercanos (kNN).

Para cada uno de estos modelos y con el método de validación cruzada probamos diferentes valores para sus correspondientes parámetros y nos quedamos con el mejor de cada modelo. Finalmente, de entre estos nos quedamos con la predicción de la RNA porque su nivel de precisión era más alto y menos variable en comparación con las otras.

Reflexionando sobre lo anterior, llegamos a la conclusión de que, aunque los resultados de la RNA son los mejores, no son en realidad buenos, pues al tratarse de una base de datos relacionada con la medicina, poder decir tan solo con un 75% de seguridad/precisión quién va a sobrevivir los 5 años siguientes a su operación, no es algo aceptable, ya que es un asunto en el que está implicada la vida.

Si presentara una precisión de 95% o superior, sería algo que se podría aceptar, aunque probablemente se llevarían a cabo estudios sobre la persona en concreto al tratarse de un tema delicado.

5. Trabajo futuro

El siguiente paso dentro de este estudio podría ser tener un mayor número mayor de pacientes, así como considerar más atributos clínico-patológicos como el tamaño del tumor o los receptores de estrógenos que podrían ser relevantes para poder entrenar mejor nuestra red y por consiguiente obtener mejores resultados.

Para obtener nuestros resultados, consideramos relevantes todos los atributos, puesto que teníamos pocos, pero en un futuro estudio en el que contamos con más cantidad de atributos, y quizá más relevantes que los actuales, podría aplicarse un análisis de componentes principales (PCA).

REFERENCIAS BIBLIOGRÁFICAS

Baxt, W. G. (1995). Application of artificial neural networks to clinical medicine. The Lancet, 346(8983). https://doi.org/10.1016/S0140-6736(95)91804-3

Cáncer de mama - Estadísticas. (2022, 24 marzo). Cancer.Net. https://www.cancer.net/es/tipos-de-c%C3%A1ncer/c%C3%A1ncer-de-mama/estadisticas

Jefferson, M. F., Pendleton, N., Lucas, S. B., & Horan, M. A. (1997). Comparison of genetic algorithm neural network with logistic regression for predicting outcome after surgery for patients with nonsmall cell lung carcinoma. Cancer, 79(7). https://doi.org/10.1002/(SICI)1097-0142(19970401)79:7<1338::AID-CNCR10>3.0.CO;2-0

Mudunuru, V. R., & Skrzypek, L. A. (2020). A comparison of artificial neural network and decision trees with logistic regression as classification models for breast cancer survival. International Journal of Mathematical, Engineering and Management Sciences, 5(6). https://doi.org/10.33889/IJMEMS.2020.5.6.089

Ravdin, P. M., & Clark, G. M. (1992). A practical application of neural network analysis for predicting outcome of individual breast cancer patients. Breast Cancer Research and Treatment, 22(3). https://doi.org/10.1007/BF01840841

Ting, F. F., Tan, Y. J., & Sim, K. S. (2019). Convolutional neural network improvement for breast cancer classification. Expert Systems with Applications, 120. https://doi.org/10.1016/j.eswa.2018.11.008