Predicción en el diagnóstico de tumores de cáncer de mama empleando métodos de clasificación

Prediction in the diagnosis of breast cancer tumors using classification methods

Nelson del Castillo Collazo
IIMAS, UNAM, Ciudad de México, México
nelson.delcastillo@iimas.unam.mx
ORCID: 0000-0002-4187-5511

doi: https://doi.org/10.36825/RITI.08.15.009

Recibido: Marzo 25, 2020
Aceptado: Mayo 06, 2020

Resumen: El presente trabajo consiste en la aplicación de las ciencias de datos con el objetivo de predecir si un tumor de cáncer de mama es benigno o no, para esto se emplean los métodos de clasificación siguientes: redes neuronales, bosques aleatorios y máquina de soporte de vectores. Se utilizó un conjunto de datos del Hospital de la Universidad de Wisconsin relacionados con el cáncer de mama. Se usan las matrices de confusión para conocer las medidas de los modelos de pronósticos y la curva *ROC (Receiver Operating Characteristics)* para determinar la capacidad discriminante de estos, apoyándose en el valor de *AUC (Area Under the Curve)*. Los modelos planteados alcanzan valores de exactitud que indican que se pueden realizar con ellos predicciones muy acertadas, aunque es importante resaltar que el modelo de máquina de soporte de vectores es el que resulta más conveniente utilizar pues su nivel de exactitud en el pronóstico supera el 99%. Se recomienda el empleo de estas técnicas en los hospitales y laboratorios donde se realice la detección de esta enfermedad, pues puede constituir una herramienta de apoyo en el diagnóstico del cáncer de mama.

Palabras clave: *Cáncer de Mama, Bosques Aleatorios, Redes Neuronales, Máquina de Soporte de Vectores, Métodos de Clasificación.*

Abstract: The present work consists of the application of data sciences with the objective of predicting whether a breast cancer tumor is benign or not, for this the following classification method are used: neural networks, random forests and vector support machine. A dataset from the University of Wisconsin Hospital related to breast cancer was used. Confusion matrices are used to know the measurements of the forecast models and the ROC (Receiver Operating Characteristics) curve to determine the discriminant capacity of these, based on the value of AUC (Area Under the Curve). The proposed models reach accuracy values that indicate that very accurate predictions can be made with them, although it is important to highlight that the vector support machine model is the most convenient to use since its level of accuracy in the forecast exceeds 99 %. The use of these techniques is recommended in hospitals and laboratories where the detection of this disease is carried out, since it can be a support tool in the diagnosis of breast cancer.

Keywords: *Breast Cancer, Random Forests, Neural Networks, Vector Support Machine, Classification Methods.*

1. Introducción

El cáncer es una enfermedad que aqueja a parte de la población mundial, si se compara con otras enfermedades el nivel de mortalidad es bastante alto, por ejemplo, se estima que en los Estados Unidos de América para el año 2020 [1] se deben presentar 1,806,950 casos de personas con cáncer y se estima que mueran 606,520, si se visualiza en una estadística diaria serían aproximadamente 4,950 casos a los que se les detectaría la enfermedad y de ellos morirían 1,660.

Ahora, de los datos anteriores se cree que se podrían presentar 279,100 casos de cáncer de mama, incluyendo a ambos sexos, esto es un elemento que indica la importancia de contar con herramientas que permita al personal médico involucrado en la detección de esta enfermedad apoyarse en las nuevas tecnologías.

El propósito de este trabajo es mostrar que las técnicas de inteligencia artificial son de gran utilidad para determinar si un tumor de cáncer de mama es benigno o maligno. El objetivo radica en la posibilidad de implementarlos en los diferentes espacios médicos donde se trabaje la detección de esta enfermedad.

En el presente trabajo se muestran los resultados obtenidos al aplicar los métodos de clasificación redes neuronales, bosque aleatorios y máquina de soporte de vectores para realizar la predicción pues su nivel de exactitud resulta ser por encima del 96% en todos los casos.

2. Materiales y métodos

2.1 Los datos

Los datos fueron tomados de la página de *MLData* [2] relacionados con el cáncer de mama, los cuales corresponden a categorías de núcleos celulares para predecir si un tumor de cáncer de mama es benigno o maligno, estos fueron generados en el año 1992 en el Hospital de la Universidad de *Winsconsin*. Cuentan originalmente con 569 registros, pero fueron eliminados todos aquellos a los que les faltaba información en algunas de sus variables, por lo que se utilizaron en el presente trabajo solo 554.

El conjunto de datos está formado por 10 variables, 9 de ellas independientes y la décima es la que se busca pronosticar, es decir, la variable dependiente donde se identifica como tumor maligno o benigno según sea el caso.

En la Tabla 1 se pueden apreciar las variables que integran este conjunto de datos, incluida la variable C*lase* que es la que se quiere predecir. En la Tabla 2 se puede ver como se distribuyen los datos según los tipos de tumores. Es importante mencionar que los tumores benignos se identifican en la variable *Clase* con el valor 2 y los malignos con el 4, con esto se trató de respetar los valores originales que tienen los datos.

Tabla 1. Nombres de las variables que intervienen en el estudio.

Nombre de la variable	Descripción
Clump_thickness:	Espesor del grupo
Uniformity_of_cell_size:	Uniformidad del tamaño de la celda
Uniformity_of_cell_shape:	Uniformidad de la forma celular
Marginal_adhesion:	Adhesión marginal
Single_epithelial_cell_size:	Tamaño de célula epitelial única
Bare_nuclei:	Núcleos desnudos
Bland_chromatin:	Cromatina blanda
Normal_nucleoli:	Nucleolos normales
Mitosis:	Mitosis
Class:	Clase

Fuente: Elaborada por el autor.

Tabla 2. Distribución de los datos según el tipo de tumor.

Variable	**2**	**4**
Clase	***(tumor benigno)***	***(tumor maligno)***
Total	348	206
%	62.82	37.18

Fuente: Elaborada por el autor.

Para el análisis de los datos y la aplicación de los diferentes métodos de clasificación se empleó el lenguaje R en su versión 3.6.1 y la interfaz de desarrollo *RStudio* en su versión 1.2.5001. Se emplearon códigos en lenguaje R que permitieron obtener los resultados de los métodos que se mencionan en el presente trabajo. Se generaron las matrices de confusión, la curva *ROC* y el valor de *AUC* [3] de los modelos de pronósticos planteados.

2.2 Métodos de clasificación

Se emplearon tres métodos de clasificación para realizar el pronóstico y determinar si los tumores de cáncer de mama eran benignos o malignos, estos fueron: redes neuronales (*neural networks*), bosques aleatorios (*random forest*) y máquina de soporte de vectores (*support vector machines* - SVM). En cada uno de los modelos los datos se dividieron en conjunto de entrenamiento 75% y de validación del modelo 25%.

Las redes neuronales son un método que trata de simular el cerebro humano aplicando algunas técnicas de optimización, con el fin de poder predecir el valor de una variable dependiente en función de un grupo de variables independientes, este proceso se considera de aprendizaje, pues los errores que se encuentran en los resultados sirven para retroalimentar el proceso de cálculo, este proceso se repite una serie de veces hasta converger. En este caso se emplean las redes neuronales para predecir la variable dependiente, la cual siempre tomará valores de cero o uno. Primeramente, se realiza la normalización de los datos con el objetivo de no introducir errores en los resultados por la fuerza que pudieran ejercer valores con un rango muy alto. Para profundizar en este tema se puede consultar [4], [5] y [6].

Los bosques aleatorios son una técnica muy empleada en la actualidad cuando se aplican métodos de aprendizaje de máquinas pues es un algoritmo bastante sencillo en su cálculo. Básicamente se basa en la generación de bosques de árboles aleatorios, es decir, se generan muchos árboles de forma aleatoria y los combina permitiendo obtener una precisión mucho más exacta, por eso el nombre de bosque. Para profundizar en este tema se puede consultar [6] y [7].

La máquina de soporte de vectores (SVM) es un método muy empleado en la actualidad cuando se trabaja con problemas donde se aplican las técnicas de aprendizaje de máquina, a pesar de que se comenzó a utilizar a principios de los noventa. Este método en sus inicios se empleó para pronósticos de variables de tipo binario, pero ya en la actualidad se emplea para clasificación múltiple. Para su mejor comprensión se deben manejar conceptos como: el hiperplano, el clasificador de margen máximo y el vector de soporte. Si desea profundizar en este tema puede consultar [6] donde se explica a detalle este método.

3. Resultados y discusión

Para cada modelo planteado se empleó una semilla diferente en la generación de los números aleatorios al momento de separar los grupos de entrenamiento y validación del modelo por esto, para cada uno de ellos no se cuenta con una matriz de confusión óptima igual.

3.1 Redes Neuronales

En la Tabla 3 se muestra la matriz de confusión óptima para el conjunto de datos de validación del modelo. El método de pronóstico debe ser capaz de reproducir esta matriz, lo cual es muy difícil de lograr. En la medida que el modelo pueda generar una matriz lo más parecida posible a la original entonces la predicción será mejor.

Tabla 3. Matriz de confusión óptima del conjunto de datos de validación del modelo.

		Referencia	
		2	**4**
Predicción	**2**	90	0
	4	0	48

Fuente: Elaborada por el autor.

Tabla 4. Matriz de confusión resultante al aplicar el modelo de red neural al conjunto de datos de validación.

		Referencia	
		2	**4**
Predicción	**2**	87	3
	4	2	46

Fuente: Elaborada por el autor.

Tabla 5. Medidas resultantes de la matriz de confusión.

REDES NEURONALES	
Medidas	**Valores (%)**
Exactitud:	96.38
Sensibilidad:	97.75
Especificidad:	93.88
VP+:	96.67
VP-:	95.83

Fuente: Elaborada por el autor.

Se empleó una estructura o topología de la red neuronal de una capa oculta con cinco nodos, el número de repeticiones para el entrenamiento de la red neuronal fue de diez y los datos fueron normalizados garantizando que todos estuvieran entre los valores cero y uno. En la Tabla 4 se pueden apreciar los valores que se obtuvieron al calcular la matriz de confusión.

Si se comparan las dos matrices mostradas en las Tablas 3 y 4 se puede apreciar que en la predicción de los tumores benignos solo se equivoca en tres casos, es decir, se debían predecir 90 y solo se predicen 87, mientras que de los tumores malignos se debían predecir 48 pero solo clasificaron correctamente 46 de ellos. Esto se puede considerar como un muy buen pronóstico.

En la Tabla 5 se presentan las diferentes medidas que se obtuvieron de calcular la matriz de confusión al aplicar este modelo. La exactitud del modelo es del 96.38%, este es un valor considerado alto, lo que indica que un nuevo conjunto de datos que se desee predecir tendría el mismo porcentaje de probabilidades de que la predicción sea correcta.

En este caso, se asume que los valores positivos son los valores que se identifican con los tumores benignos, es decir, con el valor dos en la variable *Clase*. La sensibilidad es el porciento de valores positivos que son clasificados como positivos [3], en este caso fue del 97.75%. La especificidad es el porciento de negativos que son clasificados como negativos, es decir, el porciento de valores de tumores malignos que fueron clasificados de esa manera; en este caso fue del 93.88%, es el menor valor que se obtuvo dentro de las medidas de este modelo, pero sigue siendo un valor alto si se compara con los resultados de otros tipos de datos [8].

Los valores de predicción positivos (VP+) indican la probabilidad de que un valor sea positivo si resultó positivo en la predicción, el valor es del 96.67%. En el caso de los valores de predicción negativos (VP-) indican la probabilidad de que un valor sea negativo si resultó negativo en la predicción, para esta medida el modelo determinó que es del 95.83%. Tanto para el VP+ como para el VP- los resultados que alcanza son considerados como buenas probabilidades en la predicción.

Se consideran que todos los valores que se obtuvieron en las diferentes medidas de la matriz de confusión son altos, lo que estarían indicando que este modelo de redes neuronales implementado es válido para realizar predicciones que determinen si un tumor es benigno o maligno.

3.2 Bosques aleatorios (Random Forest)

En la Tabla 6 se muestra la matriz de confusión óptima para el conjunto de datos de validación del modelo de bosques aleatorios. En este modelo la clase positiva es la de los tumores benignos (dos) mientras que la clase negativa serían los malignos (cuatro).

Para la generación de este modelo no fue necesario la normalización de los datos. Se calcularon 500 árboles aleatorios. En la Tabla 7 se muestran los resultados de la matriz de confusión que se obtienen del modelo con los datos de validación. Al comparar las dos matrices, en la predicción de los valores de tumores malignos coinciden los de la matriz óptima con los que se predicen por el modelo (51 casos). Esto ocurre pocas veces y está indicando que el modelo es muy certero en cuanto al pronóstico que hace. Sin embargo, para el caso de los tumores benignos el modelo no alcanza a predecir correctamente cuatro de ellos, reconociéndolos como tumores malignos cuando en realidad no lo son.

En la Tabla 8 se presentan las diferentes medidas que se obtuvieron de calcular la matriz de confusión al aplicar el modelo de bosques aleatorios.

Tabla 6. Matriz de confusión óptima del conjunto de datos de validación del modelo de bosques aleatorios.

		Referencia	
		2	**4**
Predicción	**2**	87	0
	4	0	51

Fuente: Elaborada por el autor.

Tabla 7. Matriz de confusión resultante al aplicar el modelo de bosques aleatorios al conjunto de datos de validación.

		Referencia	
		2	**4**
Predicción	**2**	83	4
	4	0	51

Fuente: Elaborada por el autor.

Tabla 8. Medidas resultantes de la matriz de confusión.

BOSQUES ALEATORIOS	
Medidas	**Valores (%)**
Exactitud:	97.1
Sensibilidad:	95.4
Especificidad:	100

VP+:	100
VP-:	92.73

Fuente: Elaborada por el autor.

La exactitud en este modelo es un poco mejor de la que se obtuvo con el de las redes neuronales, en este caso fue del 97.1%. La sensibilidad es del 95.4%, la cual es más baja si se comparan con la obtenida en las redes neuronales. La especificidad es del 100%, esto es un indicativo de que el porciento de negativos que son clasificados como negativos no tuvo ningún error en el pronóstico.

La medida de VP+ es del 100%, lo que indica que si el modelo predice un valor positivo entonces es 100% probable de que lo es. En cuento a VP- se obtiene un 92.73%, indicando la probabilidad de si un valor negativo es pronosticado como negativo. Este es la medida más pequeña que se obtuvo con este modelo. En la Fig. 1 se puede apreciar las variables que tienen más peso en la predicción al emplear este modelo.

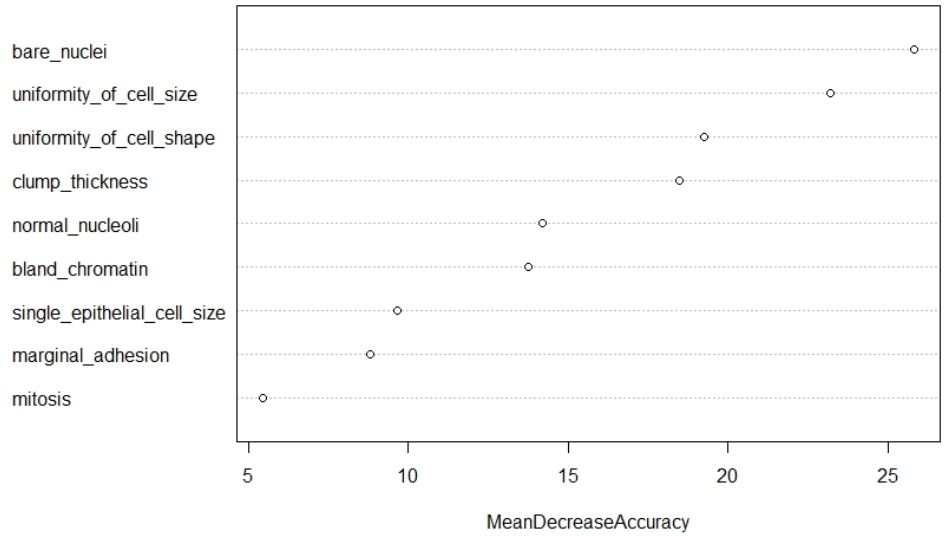

Figura 1. Contribución relativa de las variables predictoras en el pronóstico cuando se emplea el modelo de bosques aleatorios.
Fuente: Elaborada por el autor.

La variable predictora *bare_nuclei* (núcleos desnudos) es la que mayor contribución hace al pronóstico, luego le sigue *uniformity_of_cell_size* (uniformidad del tamaño de la celda). La variable mitosis es la de menor importancia o peso, mientras que el resto se encuentra distribuida en parejas con una mínima diferencia de peso entre ellas con valores que se diferencian en cinco unidades.

3.3 Máquina de soporte de vectores (support vector machines - SVM)

Se corrieron varias veces este método variando el valor del núcleo, tanto en el entrenamiento como en la validación, se utilizaron los núcleos siguientes: radial, lineal, polinomial y sigmoidal. En la Tabla 9 se puede apreciar como varió el número de vectores de soporte y la exactitud del modelo para cada uno de estos núcleos.

Tabla 9. Resumen de los resultados obtenidos para cada núcleo en la función *svm {e1071}*.

Núcleo	No. Vectores de Soporte	Exactitud (%)
radial	79	99.28
lineal	42	99.28
polinomial	73	97.83
sigmoidal	27	97.83

Fuente: Elaborada por el autor.

Se puede apreciar que los núcleos radial y lineal alcanzan los mejores valores de exactitud por lo que se determinó elegir el radial y emplear su matriz de confusión como referencia en este modelo de predicción.

En las Tablas 10 y 11 se pueden observar las matrices de confusión óptima y la que se calcula empleando los datos del conjunto de validación. En la Tabla 12 se presentan las diferentes medidas de la matriz de confusión calculada para el núcleo radial.

La exactitud de este modelo de predicción es del 99.28%. El resto de las medidas tienen valores que confirman la precisión del modelo para realizar pronósticos sobre si un tumor es benigno o no. Los valores obtenidos son mayores, si se comparan con los de los otros dos métodos presentados en este trabajo.

Tabla 10. Matriz de confusión óptima de los datos de validación para el modelo SVM.

		Referencia	
		2	**4**
Predicción	**2**	94	0
	4	0	44

Fuente: Elaborada por el autor.

Tabla 11. Matriz de confusión al aplicar el modelo de SVM al conjunto de datos de validación.

		Referencia	
		2	**4**
Predicción	**2**	93	1
	4	0	44

Fuente: Elaborada por el autor.

Tabla 12. Medidas resultantes de la matriz de confusión.

MÁQ. DE SOP. DE VECT.	
Medidas	**Valores (%)**
Exactitud:	99.28
Sensibilidad:	98.94
Especificidad:	100
VP+:	100
VP-:	97.78

Fuente: Elaborada por el autor.

3.4 Curvas ROC

La curva *ROC* se utilizó en el presente trabajo para determinar la capacidad discriminante de los modelos empleados [6], que permiten hacer una distinción entre los dos tipos de valores que puede tomar la variable dependiente, en este caso si es un tumor benigno (dos) o maligno (cuatro).

En la Figura 2 se puede apreciar la curva *ROC* y el valor de *AUC* el cual es del 99.5%, lo que indica que el modelo asigna mayores valores de probabilidad de ocurrencia a los casos donde el suceso ocurre realmente.

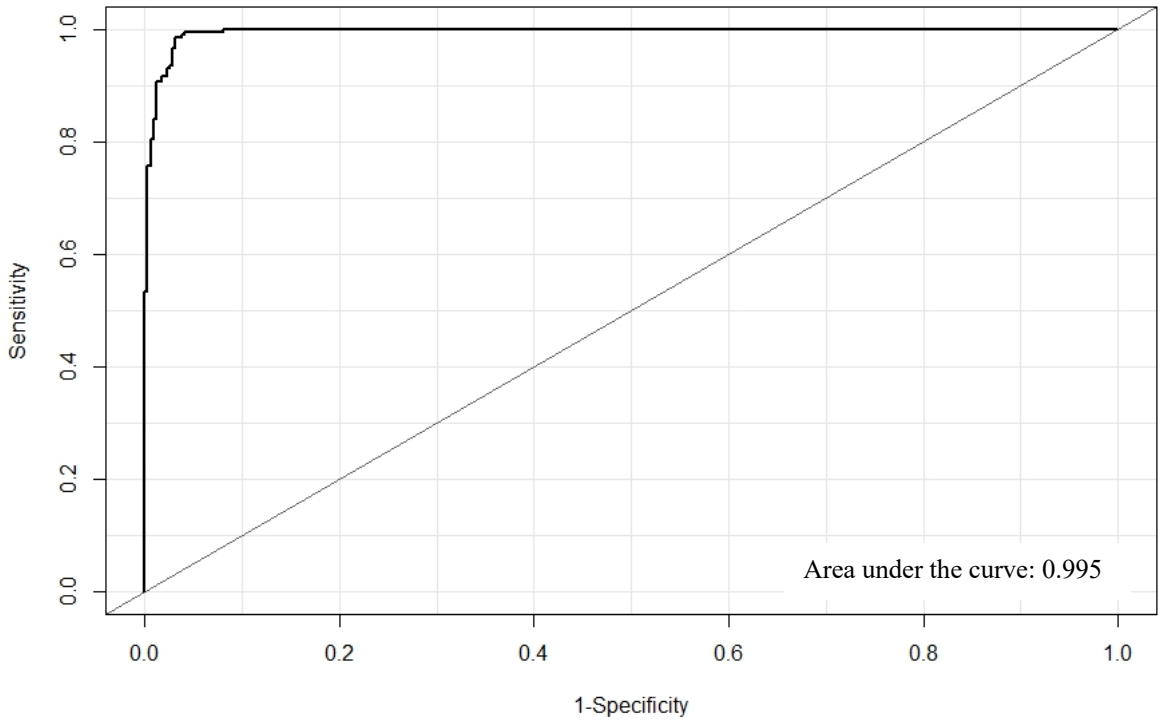

Figura 2. Curva *ROC* y valor de *AUC*.
Fuente: Elaborada por el autor.

4. Conclusiones

Se comprueba una vez más que los métodos de clasificación: redes neuronales, bosques aleatorios y máquinas de soporte de vectores son herramientas muy útiles dentro del aprendizaje de máquinas cuando es de interés realizar pronósticos con variables binomiales.

El modelo de máquina de soporte de vectores fue el que arrojó los mejores resultados en la predicción del tipo de tumor que se predijo, teniendo un solo error en el pronóstico al utilizar el conjunto de datos de validación y dando un valor de exactitud muy alto. Los otros dos modelos presentados, también lograron muy buenas predicciones por lo que se recomienda emplearlos, aunque sus valores de exactitud hayan estado un poco por debajo del modelo de SVM.

Los resultados que se lograron en este trabajo pueden ser aplicados en la detección temprana del cáncer de mama, pues constituye una herramienta que apoya la labor del personal médico encargado de determinar si un tumor es benigno o maligno. Se recomienda su empleo en los hospitales especializados y laboratorios médicos que se dediquen a la detección de este tipo de cáncer.

La curva *ROC* generada muestra una excelente exactitud diagnóstica, el valor de *AUC* indica una alta capacidad discriminante del modelo.

El conjunto de datos con el que se trabajó tiene una relación de poco más de 55 registros por variables, se considera que es un *set* de datos válido para realizar la predicción de su variable dependiente, aunque sería muy recomendable hacer más experimentos con un *set* de datos mucho mayor con el fin de comprobar los resultados obtenidos en cada modelo empleado en el presente trabajo.

Se recomienda el uso del lenguaje R y de la interfaz de desarrollo *RStudio* para el análisis y procesamiento del conjunto de datos. Son herramientas muy valiosas en la implementación de los algoritmos y en el análisis de los resultados.

5. Agradecimientos

Deseo agradecer a la Dra. Rosa María Macías Herrera y al especialista Adrián Durán Chavesti por su invaluable y constante apoyo en la elaboración del presente artículo.

6. Referencias

[1] American Cancer Society. (2019). *Surveillance, Epidemiology, and End Results (SEER). National Cancer Institute*. Recuperado de: https://cancerstatisticscenter.cancer.org/?_ga=2.151790777.1241982100.1584820087-1304861891.1584820087#!/

[2] MLData. (2018). *Breast Cancer: Predict if tumor is benign or malignant*. Recuperado de: https://www.mldata.io/dataset-details/breast_cancer/

[3] Aldás, J., Uriel, E. (2017). *Análisis Multivariante aplicado con R* (2da Ed.). Madrid, España: Ediciones Paraninfo.

[4] Hair, J. F., Anderson, R. E., Tatham, R. L., Black, W. C. (1999). *Análisis Multivariante*. Madrid: Prentice Hall.

[5] Hodnett, M., Wiley, J. F. (2018). *R Deep Learning Essentials* (2da Ed.). UK: Packt Publishing Ltd.

[6] Cirillo, A. (2017). *R Data Mining. Implement data mining techniques through practical use cases and real-world datasets.* Birmingham, Mumbai: Packt Publishing Ltd.

[7] Villalba Bergado, F. (2017). *Aprendizaje supervisado en R*. Recuperado de: https://fervilber.github.io/Aprendizaje-supervisado-en-R/bosques.html

[8] del Castillo Collazo, N. (2020). Incidencias en el pronóstico al aplicar reducción de variables. Un ejemplo práctico. *Revista de Investigación en Tecnología de la Información (RITI)*, 8 (15), 50-69. doi: https://doi.org/10.36825/RITI.08.15.006