Estudio de las preferencias para el vino blanco y el vino tinto utilizando métodos de clasificación binaria

Study of preferences for white and red wine using binary classification methods

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

Luis Felipe Alvarado Pegueros
Datio America, Ciudad de México, México
mcdfelipea@gmail.com
ORCID: 0000-0001-8873-1083

Víctor Flores Rodríguez
IDS Comercial, Ciudad de México, México
nocte.magister@gmail.com
ORCID: 0000-0001-7798-7075

Noé Amir Rodríguez
Universidad UTEL, Estado de México, México
narodriguez@cidesi.edu.mx
ORCID: 0000-0001-5892-0625

doi: https://doi.org/10.36825/RITI.08.16.003

Recibido: Junio 17, 2020
Aceptado: Agosto 01, 2020

Resumen: La aplicación de métodos de la minería de datos nos permite la detección de una serie de patrones que pueden existir en los datos que analizamos pero que no son fáciles de detectar a simple vista. En este caso se aplicaron algunas técnicas para pronosticar las preferencias del sabor del vino a partir de una serie de características físico - químicas de su composición, tanto del vino tinto como del vino blanco, bebidas que han sido del gusto de muchas personas a nivel internacional a través del tiempo. El conjunto de datos que se empleó en este trabajo fue tomado de Vino Verde del Norte de Portugal. Estos datos cuentan con un grupo de variables que permitieron aplicar métodos de clasificación para lograr pronosticar las preferencias del sabor del vino sustentado en el criterio de los clientes. Para lograr este objetivo se emplearon los métodos: análisis discriminante, regresión logística y redes neuronales. Los resultados obtenidos demostraron que para los dos conjuntos de datos los resultados son

muy parecidos cuando aplicamos los tres métodos mencionados. La capacidad discriminante de los modelos permite distinguir claramente la separación de los dos grupos para la clasificación.

Palabras clave: *Minería de Datos, Métodos de Clasificación, Redes Neuronales, Análisis Discriminante, Regresión Logística.*

Abstract: The application of data mining methods allows us to detect a series of patterns that may exist in the data we analyze but are not easy to detect a simple view. In this case, we apply some techniques to predict the frequencies of the taste of wine from a series of physical - chemical characteristics of its composition, both wine and white wine, drinks that have been liked by many people internationally for a long time The data set that was used in this work was taken from Green Wine from the North of Portugal. These data had a group of variables that allowed applying classification methods to predict the flavor specifications of the wine based on the criteria given by the customers. The methods were used to achieve this objective: discriminant analysis, logistic regression and neural networks. The results showed that for the two data sets the results are very similar when the three specific methods are applied. The discriminant capacity of the models makes it possible to clearly distinguish the separation of the two groups for classification.

Keywords: *Data Mining, Classification Methods, Neural Networks, Discriminating Analysis, Logistic Regression.*

1. Introducción

Es conocida la aceptación que tiene la ingesta de vino a nivel internacional, sobre todo, existe una discusión que ha durado muchos años acerca de que si el vino puede ayudar a disminuir la probabilidad de ocurrencias de eventos cardiacos [1], como el infarto agudo al miocardio entre otras afecciones, varios autores escriben al respecto y tienen diferentes opiniones a partir de los resultados que han obtenido en sus investigaciones [2] y [3]. Hasta el momento existe un conceso bastante generalizado de que el vino tinto si ayuda a prevenir este tipo de eventos en los seres humanos, que en muchos casos pueden llegar a ser muy serios, e incluso, causar la muerte.

El presente estudio se basa en algunos métodos de análisis multivariante y de minería de datos para predecir la preferencia del sabor del vino en función de los criterios de los clientes y de algunas de sus características físico - químicas, tanto del vino tinto como del vino blanco.

En este trabajo se demuestra que aplicando los métodos mencionados se obtienen modelos que permiten hacer un buen pronóstico de las preferencias del vino empleando técnicas de clasificación, curvas ROC (*Receiver Operating Characteristics*) y la calibración del modelo. En cuanto a los datos se modifica el criterio de clasificación para el pronóstico, en este caso se crea una variable dicotómica en lugar de trabajar con los valores entre uno y diez, es decir, se crean dos grupos, uno con valor *cero* donde significa que la valoración de los clientes fue de puntuación cinco o menor y valor *uno* cuando la valoración del cliente fue mayor a cinco.

Saber la preferencia del sabor del vino con anticipación puede ayudar a los importadores de vinos, dueños de restaurantes, bares y locales en general donde se expende esta bebida a adquirir aquellos tipos de vinos con características que le sean del agrado a sus clientes, también lo es para los productores, pues sabiendo los gustos de los consumidores pueden planificar mejor su producción.

Se pretende con este trabajo tener una aproximación a la realidad empleando algunas técnicas de la minería de datos y el aprendizaje de máquinas tratando de detectar patrones que permitan hacer pronósticos válidos y ofrecer lo mejor para sus clientes. Es importante indicar que los criterios de preferencias del vino pueden variar entre zonas del planeta en función de sus costumbres alimentarias y culturales.

2. Materiales y métodos

2.1 Los datos

En el presente trabajo se utilizó el lenguaje de programación *R* (versión 3.5.1) y la interfaz de desarrollo *RStudio* (versión 1.1.456). Se emplearon los programas en lenguaje *R* que facilitaron obtener los resultados de los métodos de clasificación mencionados. Se construyeron las matrices de confusión y las curvas ROC para cada uno de los modelos planteados y se aplicó la prueba de Hosmer y Lemeshow [4] para su calibración.

Los datos fueron generados entre los años 2004 y 2009 de la región noroeste de Portugal [5] y tomados de la página de *UCI Machine Learning Repository* [6] los cuales se encuentran disponibles para que el público haga uso de ellos.

Están divididos en datos de vino tinto y vino blanco en archivos independientes, en ambos casos tienen 11 variables físico - químicas y la variable de valoración del usuario, estas son:

a) Acidez fija [*Fixed acidity*] (g (*tartaric acid*)/dm3)
b) Acidez volátil [*Volatile acidity*] (g (*acetic acid*) /dm3)
c) Ácido cítrico [*Citric acid*] (g/dm3)
d) Azúcar residual [*Residual sugar*] (g/dm3)
e) Cloruros [*Chlorides*] (g (*sodium chloride*)/dm3)
f) Dióxido de azufre libre [*Free sulfur dioxide*] (mg/dm3)
g) Dióxido de azufre total [*Total sulfur dioxide*] (mg/dm3)
h) Densidad [*Density*] (g/cm3)
i) Ph
j) Sulfatos [*Sulphates*] (g (*potassium sulphate*)/dm3)
k) Alcohol [*Alcohol*] (vol.%)
l) Calidad [*Quality*] (*value* 0/1)

La variable a pronosticar es *calidad*, esta toma valores entre uno y diez dependiendo de la calificación que el cliente le asigna en función de sus gustos. Para el presente estudio la variable *calidad* toma valores cero si el valor dado por el cliente es menor o igual a cinco y si es mayor entonces se le asigna el valor uno, es decir, se separan los datos en *Recomendado* con el valor *uno* y *No Recomendado* con el valor *cero*, esto facilita mucho el trabajo de predicción cuando se apliquen los métodos de clasificación, los cuales fueron: análisis discriminante, regresión logística y redes neuronales.

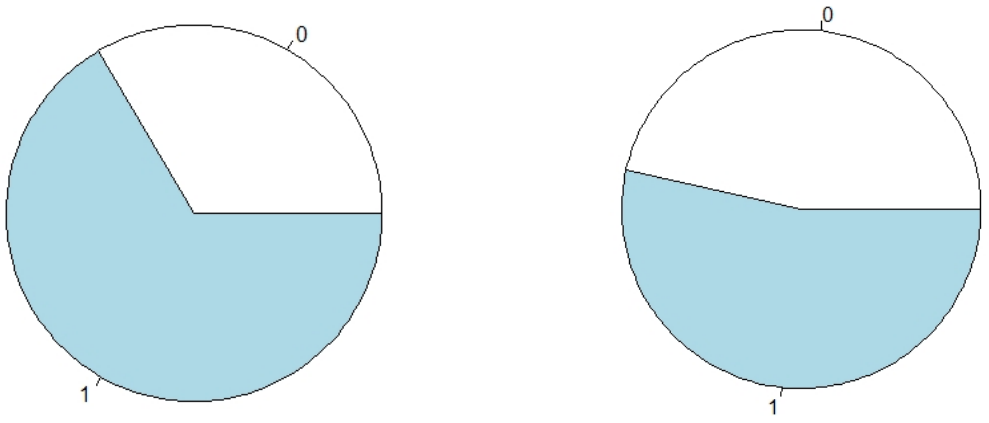
En todos los casos se dividió el conjunto de datos en 75% para entrenamiento del modelo y el 25% para su validación. Los datos del vino blanco cuentan con 4898 registros y del vino tinto con 1599, se puede observar en la tabla 1 como se distribuyen los valores en la variable *calidad.*

Tabla 1. Valores de la variable *calidad* en ambos tipos de vino.

Vino	**Blanco**		**Tinto**	
variable *calidad*	**0**	**1**	**0**	**1**
Total	1640	3258	744	855
%	33.48	66.52	46.53	53.47

Fuente: Elaborada por los autores.

En los Figura 1 se observa el comportamiento de la variable *calidad* en base a sus valores (cero y uno).



(a) (b)

Figura 1. Comportamiento de los valores de la variable calidad para ambos tipos de vino, (a) vino blanco y (b) vino tinto. Fuente: Elaborada por los autores.

2.2 Análisis Discriminante

Este método se emplea para determinar a qué grupo pertenece un individuo en función de las variables clasificadoras o predictoras, en cualquier caso, el individuo siempre podrá pertenecer a un solo grupo de los posibles a ser clasificados, a partir de las variables que describen a ese individuo. En este caso la variable que se trata de predecir es la variable dependiente categórica con dos posibles valores, cero y uno, para cualquier combinación de datos de las variables independientes se dará como resultados uno de estos dos valores. El valor *uno* se corresponde con la clasificación de *Vino Recomendado* y el valor *cero* con la clasificación de *Vino No Recomendado*. Para un estudio más profundo de este tema puede consultar [7].

2.3 Regresión Logística

La regresión logística es muy parecida al análisis discriminante visto anteriormente, este método también se emplea para predecir a una variable dependiente de tipo no métrica. Cuando la variable dependiente toma solo dos posibles valores se estaría hablando de la regresión logística binomial, es decir, es una variable dicotómica y en el caso de regresión logística multinomial la variable dependiente puede tomar un valor dentro de varios posibles.

En este caso sería una regresión logística binomial pues se trata de predecir la variable *calidad* la cual es una variable dicotómica. Para profundizar en este tema se puede consultar [4], [7], [8] y [9].

2.4 Redes Neuronales

Las redes neuronales es un método diferente a los anteriormente mencionados, en este se trata de simular un proceso de aprendizaje aplicando diferentes herramientas de optimización, las cuales permiten, luego de una serie de recorridos a través de la estructura de la red neuronal, ajustar la función que facilita la predicción de la variable dependiente. Los errores encontrados en el proceso se emplean para retroalimentar el proceso de cálculo.

En el presente trabajo se emplean las redes neuronales para predecir nuestra variable dependiente, que siempre tomará valores de *cero* o *uno*. Primeramente, realizamos la normalización de los datos con el objetivo de no introducir errores en los resultados. Para profundizar en este tema se puede consultar [7] y [10].

3. Resultados y Discusión

Se realizó la prueba de significancia estadística para saber si los grupos generados son significativamente distintos respecto a las variables independientes consideradas. En este caso la prueba es estadísticamente significativa, el

valor obtenido de p es 2.2 e-16, como $p < 0.01$, se puede concluir que las variables clasificadoras tienen una capacidad discriminante significativa.

3.1 Regresión Logística

Es importante mencionar, que para los resultados obtenidos en todos los métodos aplicados que veremos a continuación, los llamados *valores positivos (cero)* coinciden con el *Vino No Recomendado* mientras los *valores negativos* (*uno*) con el *Vino Recomendado*.

Al aplicar este método se obtuvo la matriz de confusión que se muestra en la Tabla 2.

Tabla 2. Matriz de confusión al aplicar el método de regresión logística.

		Referencia			
		Vino Blanco		**Vino Tinto**	
		0	**1**	**0**	**1**
Predicción	**0**	249	206	148	79
	1	138	631	33	140

Fuente: Elaborada por los autores.

Tabla 3. Medidas resultantes del modelo de predicción en la regresión logística.

REGRESIÓN LOGÍSTICA		
Medidas	**Vino Blanco**	**Vino Tinto**
Exactitud	0.719	0.72
Sensibilidad	0.6434	0.8177
Especificidad	0.7539	0.6393
VP+	0.5473	0.652
VP-	0.8205	0.8092

Fuente: Elaborada por los autores.

En la Tabla 3 se presentan las estadísticas del modelo de predicción al aplicar este método. Se observa que la exactitud del modelo *(Accuracy*) es del 71.9% para el vino blanco y del 72 % para el tinto, en este caso son prácticamente iguales. La sensibilidad del modelo (*Sensitivity*) 64.34% para el vino blanco y del 81.77% para el tinto, aquí se puede apreciar una diferencia bastante significativa en cuanto a la sensibilidad del modelo para ambos tipos de vino, es decir, la sensibilidad del modelo indica el porciento de positivos que son clasificados como positivos [11]. En cuanto a la especificidad del modelo (*Specificity*) es diferente para cada tipo de vino, pero no tan marcada como la sensibilidad, en este caso fue de 75.39% para el vino blanco y del 63.93% para el tinto, la especificidad indica el porciento de negativos que son clasificados como negativos.

En cuanto a los valores de predicción positivos (*VP+*) los valores fueron de 54.73% para el vino blanco y del 65.2% para el tinto, esta medida indica la probabilidad de que un valor sea positivo si resultó positivo en la predicción, en ambos casos son valores bajos, pero el más significativo es el del vino blanco porque es un valor bastante bajo. Los resultados obtenidos para los valores de predicción negativos (*VP-*) fueron muy parecidos y con un porciento bastante alto, para el caso del vino blanco fue de 82.05% y para el tinto fue de 80.92%, cabe mencionar que esta medida indica la probabilidad de que un valor sea negativo si resultó negativo en la predicción. Se observa que el modelo está prediciendo con mayor confianza cuando se trata del *Vino Recomendado*.

En cuanto a la calibración del modelo aplicando la prueba de Hosmer y Lemeshow [4] muestra que existe una buena calibración para ambos tipos de vino, siendo el valor para el vino blanco de 0.666 y para el vino tinto de 0.178.

3.2 Análisis Discriminante

En las Tablas 4 y 5 se observan los resultados de la matriz de confusión y las estadísticas del modelo de predicción al aplicar el método de análisis discriminante. Para este método la exactitud fue ligeramente mayor al obtenido en la regresión logística, para el vino blanco fue del 74.84% y para el tinto del 76%, en ambos casos son valores significativos indicando una buena precisión en el modelo predictivo.

Los valores de sensibilidad obtenidos fueron del 70.13% para el vino blanco y 68.91% para el tinto y la especificidad fue del 76.42% para el vino blanco y del 82.61% para el tinto, en este caso el valor que se obtuvo para el vino blanco es parecido al de la regresión logística, pero aumentó considerablemente en el caso del vino tinto.

En los valores de predicción positivos es importante señalar que el resultado para el vino blanco es realmente bajo, en este caso fue del 50%, lo que indica que la probabilidad de que un valor sea positivo si resultó positivo en la predicción es muy bajo, esto se podría comparar con el lanzamiento de una moneda al aire. Para el caso del vino tinto es significativo el valor pues es del 78.7%.

Los valores predictivos negativos son bastante altos, lo que indican una alta probabilidad de acertar en el pronóstico cuando los resultados son para *Vinos Recomendados*, para el vino blanco fue del 88.38% y para el tinto del 74.03%.

Tabla 4. Matriz de confusión al aplicar el método de análisis discriminante.

		Referencia			
		Vino Blanco		**Vino Tinto**	
		0	**1**	**0**	**1**
Predicción	**0**	216	216	133	36
	1	92	700	60	171

Fuente: Elaborada por los autores.

Tabla 5. Medidas resultantes en el modelo de predicción del análisis discriminante.

ANÁLISIS DISCRIMINANTE		
Medidas	**Vino Blanco**	**Vino Tinto**
Exactitud	0.7484	0.76
Sensibilidad	0.7013	0.6891
Especificidad	0.7642	0.8261
VP+	0.5	0.787
VP-	0.8838	0.7403

Fuente: Elaborada por los autores.

3.3 Redes Neuronales

Se corrieron en varias ocasiones los conjuntos de datos y se fueron modificando los parámetros de la función *neuralnet{neuralnet}* con el objetivo de encontrar la mejor combinación de estos y emplear el modelo de predicción que alcance valores de exactitud mayores. Los mejores resultados se obtuvieron al emplear el algoritmo *rprob+* con el conjunto de datos del vino blanco y *rprob-* con el del vino tinto. Para los datos del vino blanco se

decidió trabajar con dos capas ocultas, de cinco y dos neuronas respectivamente, mientras que para el conjunto de datos del vino tinto se emplearon dos capas ocultas de tres neuronas cada una.

Tabla 6. Matriz de confusión al aplicar el método de redes neuronales.

		Referencia			
		Vino Blanco		**Vino Tinto**	
		0	**1**	**0**	**1**
Predicción	**0**	277	155	138	31
	1	119	673	51	180

Fuente: Elaborada por los autores.

Tabla 7. Medidas resultantes en el modelo de predicción de redes neuronales.

REDES NEURONALES		
Medidas	**Vino Blanco**	**Vino Tinto**
Exactitud	0.7761	0.795
Sensibilidad	0.6995	0.7302
Especificidad	0.8128	0.8531
VP+	0.6412	0.8166
VP-	0.8497	0.7792

Fuente: Elaborada por los autores.

En las Tablas 6 y 7 se pueden apreciar los resultados de la matriz de confusión para el análisis predictivo del modelo. En cuanto a la exactitud se puede observar que son los mejores que se obtuvieron en los tres métodos que se aplicaron a cada conjunto de datos, para el vino blanco fue de 77.61% y para el tinto del 79.5%.

La sensibilidad del modelo predictivo no fue muy diferente a lo que se había obtenido en los dos métodos visto con anterioridad en este trabajo, para el vino blanco fue del 69.95% y para el tinto del 73.02%. La especificidad fue del 81.28% en el vino blanco y del 85.31% en el tinto. Se puede apreciar entonces que fueron los resultados más altos en cuanto a la especificidad del modelo considerando los dos tipos de vinos a la vez.

En cuanto a los valores predictivos positivos el comportamiento del modelo es muy parecido a lo que se ha obtenido en los otros dos ya analizados, siempre el valor en el vino blanco es menor al del vino tinto, en este caso es del 64.12% para el vino blanco el cual es el mejor para este tipo de vino si lo comparamos con los resultados de la regresión logística y el análisis discriminante. Para el caso del vino tinto el valor fue del 81.66%, que también es el mayor que se obtuvo en todos los métodos considerados en este trabajo. En cuento a los valores de predicción negativos para ambos conjuntos de datos fue alto, del 84.97% para el vino blanco y del 77.92% para el tinto. Estos resultados comparativos se pueden apreciar en las Tablas 8 y 9.

Se recomienda emplear otros métodos de clasificación para comparar los resultados con los obtenidos en este trabajo y ver si son más efectivos en la predicción [12].

Otra vía para determinar la capacidad discriminante de los modelos fue emplear las curvas ROC, estas también indican la capacidad discriminante de los modelos analizados en el trabajo [11] los cuales permiten hacer una distinción entre los dos grupos, en este caso es, si el vino es *no recomendado* o si es *recomendado.* En ambos casos los valores de AUC (*Area Under the Curve*) son mayores al 80% indicando que son modelos con una buena capacidad discriminante. En las Figuras 2 y 3 se pueden apreciar estos resultados.

Tabla 8. Resultados comparativos de las medidas de la matriz de confusión.

	REGRESIÓN LOGÍSTICA		ANÁLISIS DISCRIMINANTE		REDES NEURONALES	
Medidas	**Vino Blanco**	**Vino Tinto**	**Vino Blanco**	**Vino Tinto**	**Vino Blanco**	**Vino Tinto**
Exactitud	0.719	0.72	0.7484	0.76	0.7761	0.795
Sensibilidad	0.6434	0.8177	0.7013	0.6891	0.6995	0.7302
Especificidad	0.7539	0.6393	0.7642	0.8261	0.8128	0.8531
VP+	0.5473	0.652	0.5	0.787	0.6412	0.8166
VP-	0.8205	0.8092	0.8838	0.7403	0.8497	0.7792
Prevalencia	0.3162	0.4525	0.2516	0.4825	0.3235	0.4725

Fuente: Elaborada por los autores.

Tabla 9. Valores de la exactitud (*Accuracy*) de los diferentes métodos de clasificación vistos en el trabajo

Valores de exactitud de los modelos predictivos		
Método	**Vino Blanco**	**Vino Tinto**
REGRESIÓN LOGÍSTICA	0.719	0.72
ANÁLISIS DISCRIMINANTE	0.7484	0.76
REDES NEURONALES	0.7761	0.795

Fuente: Elaborada por los autores.

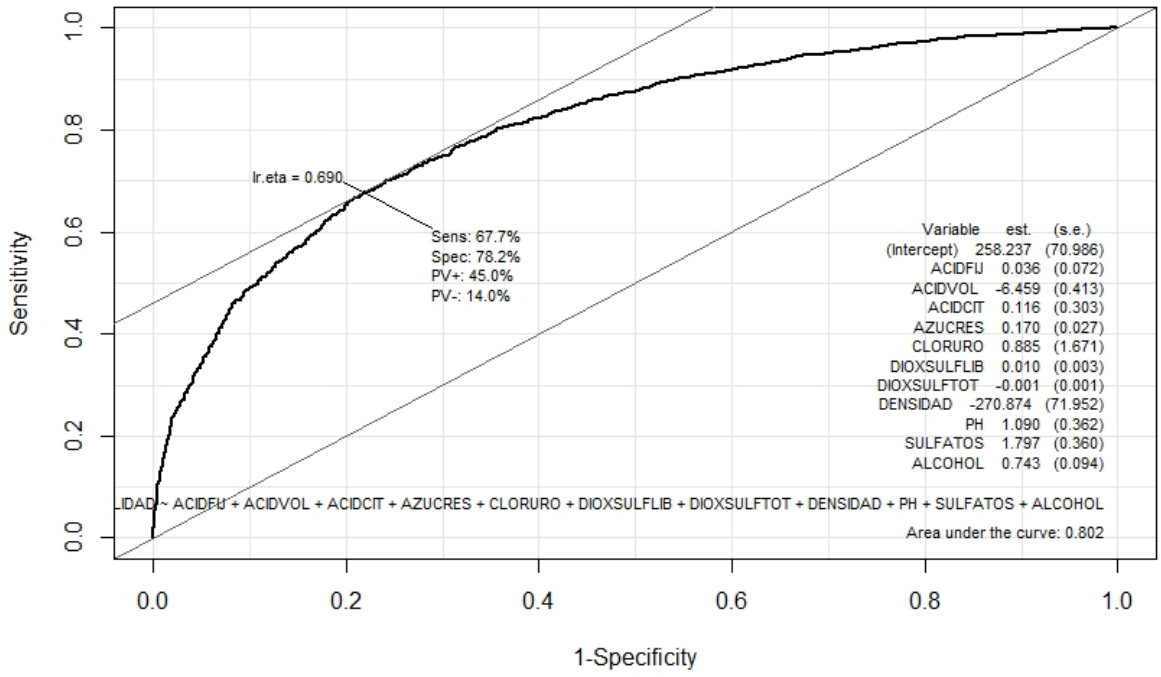


Figura 2. Curva ROC para el modelo de pronóstico con los datos del vino blanco.
Fuente: Elaborada por los autores.

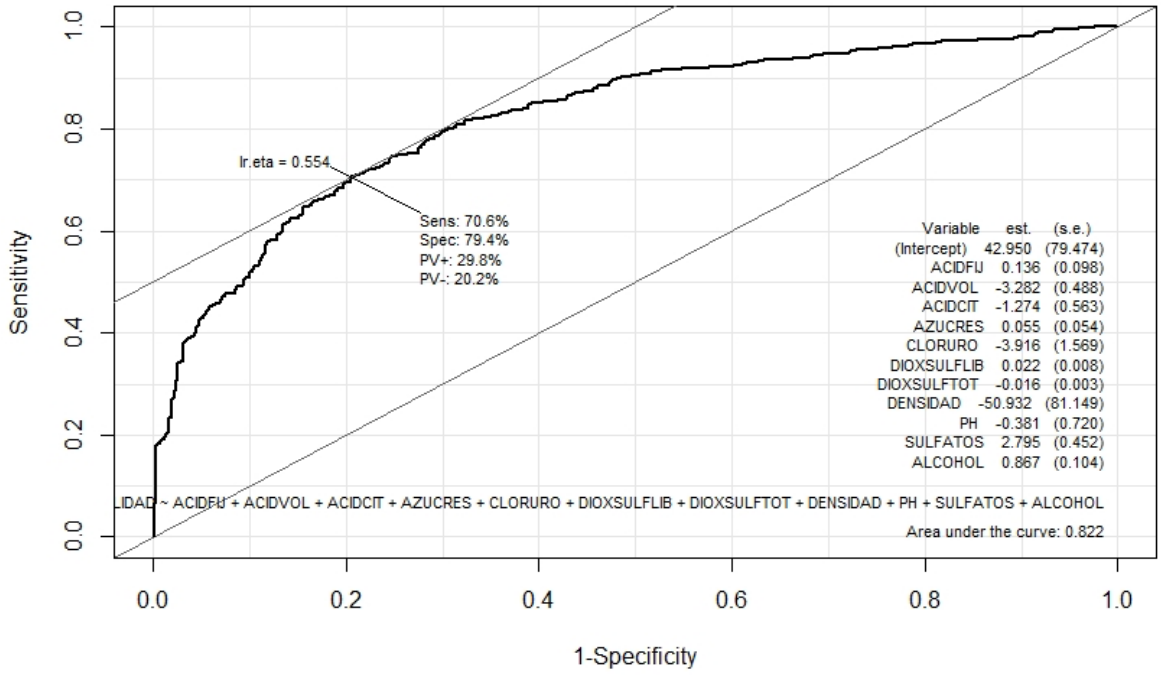


Figura 3. Curva ROC para el modelo de pronóstico con los datos del vino tinto.
Fuente: Elaborada por los autores.

4.- Conclusiones

Después de aplicar los tres métodos de clasificación a los dos conjuntos de datos, tanto del vino blanco como del vino tinto, se puede concluir que el método predictivo con más exactitud fue el de las redes neuronales, para ambos conjuntos de datos y es el que se recomienda para hacer un pronóstico sobre las preferencias del vino basados en estos datos. Las pruebas estadísticas hechas nos confirman que los tres métodos cuentan con capacidad discriminante que permiten hacer una distinción clara entre los dos grupos que se pretenden pronosticar.

La sensibilidad es muy variable entre los métodos predictivos, siendo el mejor pronóstico el del vino tinto cuando se emplea la regresión logística. La especificidad mejor fue la obtenida por el método de redes neuronales para ambos tipos de vinos, los cuales están por encima del 80%. El valor predictivo positivo (clasificados como *Vinos No Recomendados*) en algunos casos fue muy bajo, siendo el más notable el obtenido en el análisis discriminante que fue del 50%, para el vino tinto en general fue un mejor pronóstico, siendo el más alto el de las redes neuronales que estuvo por encima del 80%. Los valores predictivos negativos (clasificados como *Vinos Recomendados*) en general los resultados dieron mejores pronósticos, casi todos por encima del 80%, excepto en el vino tinto, que estuvieron por encima del 74% en el análisis discriminante y en las redes neuronales.

Se puede concluir que los tres métodos de clasificación analizados son válidos para predecir las preferencias de los dos tipos de vinos, tanto el blanco como el tinto, basándonos en los conjuntos de datos utilizados en este trabajo. Se recomienda emplear otros métodos de clasificación para compararlos con los resultados obtenidos en el presente trabajo y ver si son más efectivos en el pronóstico de las preferencias del vino.

5. Referencias

[1] Huerta, I. R. (1998). Revista Española de Cardiología. Recuperado de: https://www.revespcardiol.org/es-vino-corazon-articulo-X0300893298002947?redirect=true

[2] Doll, R., Peto, R., Hall, E., Wheatley, K., Gray, R. (1994). Mortality in relation to consumption of alcohol: 13 years observations on male British doctors. *The BMJ (Clinical research ed.)*, 309 (6959), 911-918. doi: https://doi.org/10.1136/bmj.309.6959.911

[3] Kannel, W. B., Curtis Ellison R. (1996). Alcohol and coronary heart disease: the evidence for a protective effect. *Clinica Chimica Acta,* 246 (1-2), 59-76. doi: https://doi.org/10.1016/0009-8981(96)06227-4

[4] Hosmer, D. W., Lemesbow, S. (1980). A Goodness-of-Fit Tests for the Multiple Logistic Regression Model. *Communications in Statistics-Theory and Methods*, 9 (10), 1043-1069. doi: https://doi.org/10.1080/03610928008827941

[5] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47 (4), 547-553. doi: https://doi.org/10.1016/j.dss.2009.05.016

[6] Cortez, P. (2009). *UCI-Machine leaning repository*. Recuperado de: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

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

[8] Pearson, R. K. (2018). *Exploratory Data Analysis Using R.* Boca Raton, US: CRC Press-Taylor & Francis Group.

[9] Henao Zuluaga, K. J., Correa Morales, J. C. (2018). Regresión Logística Bivariable para Tablas de Contingencia Usando Metodología GSK. *Revista Comunicaciones en Estadística*, 11 (2), 153–170.

[10] Wiley, M. H. (2018). *R Deep Learning Essentials.* UK: Packt Publishing Ltd.

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

[12] del Castillo Collazo, N. (2020). Predicción en el diagnóstico de tumores de cáncer de mama empleando métodos de clasificación. *Revista de Investigación en Tecnología de la Información (RITI)*, 8 (15), 96-104. doi: https://doi.org/10.36825/RITI.08.15.009