# Incidencia en el pronóstico al aplicar reducción de variables. Un ejemplo práctico

## Incidence in the forecast by applying variable reduction. A practical example

**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.08.15.006

Recibido: Enero 09, 2020
Aceptado: Marzo 06, 2020

**Resumen:** En el presente trabajo se muestra cómo incide en el pronóstico reducir el número de variables con las que se realiza la predicción, para esto se aplicó el análisis factorial. Las variables se agruparon en tres factores. Se emplearon los métodos de clasificación siguientes: análisis discriminante, regresión logística y redes neuronales. Se trabajó con tres grupos de datos, el primero incluye a las variables originales, el segundo las variables que pertenecen a los factores uno, dos y tres y el tercero está compuesto solo por las del factor uno. Se emplearon las matrices de confusión y las curvas *ROC* para conocer la exactitud de los modelos de pronóstico. Se muestran los resultados obtenidos para cada grupo, donde se aprecia que la reducción de variables es muy conveniente para llegar a excelentes resultados en la predicción usando menos recursos; un ejemplo de esto es el caso de la regresión logística donde la diferencia en la exactitud del modelo, entre los dos primeros grupos es menor al tres por ciento.

***Palabras clave:*** *Reducción de Variables, Análisis Factorial, Métodos de Clasificación, Redes Neuronales.*

Abstract: This paper shows how the forecast affects reducing the number of variables with which the prediction is made, for this the factor analysis was applied. The variables were grouped into three factors. The following classification methods were used: discriminant analysis, logistic regression and neural networks. We worked with three groups of data, the first includes the original variables, the second the variables belonging to factors one, two and three and the third is composed only of those of factor one. Confusion matrices and *ROC* curves were used to determine the accuracy of the forecast models. The results obtained for each group are shown, where it is appreciated that the reduction of variables is very convenient to reach excellent prediction results using fewer resources; An example of this is the case of logistic regression where the difference in the accuracy of the model between the first two groups is less than three percent.

***Keywords:*** *Variable Reduction, Factorial Analysis, Classification Methods, Neural Networks.*

## 1. Introducción

En el presente trabajo se muestra cómo influye la reducción de variables en el pronóstico de las preferencias del vino, tanto del vino tinto como del blanco. Los datos fueron tomados de Internet [1]. Se aplica la técnica del análisis factorial para obtener un conjunto de variables menor y conocer cuál es el comportamiento de la predicción; se dividen en tres conjuntos de datos para ambos tipos de vino a partir de los resultados obtenidos en el análisis factorial.

Se hace una comparación de los resultados al aplicar los métodos de clasificación: análisis discriminante, regresión logística y redes neuronales. Se demuestra que aplicando estos métodos se obtienen modelos que permiten hacer un buen pronóstico de las preferencias del vino empleando técnicas de clasificación. En cuanto a la variable dependiente se hace una mejor clasificación de los grupos para el pronóstico según la preferencia de los clientes, en este caso se crea una variable dicotómica en lugar de trabajar con los valores entre cero y diez.

Se pretende en este trabajo tener una aproximación a la realidad empleando algunas técnicas de minería de datos y aprendizaje de máquinas, con el objetivo de determinar la incidencia que tienen en la predicción la reducción de variables. Se llegan a resultados muy interesantes, para los datos que se emplean. Se demuestra que es mejor trabajar con todas las variables en casi todos los grupos de estudio, pero en algunos casos los resultados que se obtienen cuando se hace la reducción de las variables son muy significativos en el pronóstico. Es importante mencionar que los criterios de las preferencias del vino pueden variar mucho en función de los individuos que den su opinión al realizar las pruebas de cata [2] y [3].

## 2. Materiales y métodos

### *2.1 Los datos*

Para la realización de este trabajo se empleó el lenguaje de programación R (versión 3.5.2) y la interface de desarrollo RStudio (versión 1.1.453), la cual cuenta con herramientas que facilitan que el trabajo con dicho lenguaje sea muy cómodo y fácil. Se utilizaron los códigos en lenguaje R que permitieron obtener los resultados de todos los métodos que se mencionan en este trabajo. Se empleó el análisis factorial para reducir el número de variables en los datos y se generaron las matrices de confusión [4] y las curvas ROC para cada uno de los modelos de pronósticos planteados.

Los datos para ambos tipos de vinos fueron generados entre los años 2004 y 2009, son originarios de la región noroeste de Portugal [5] y tomados de la página de *UCI Machine Learning Repository* [1], los cuales se encuentran disponibles para que el público en general pueda hacer 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 (*calidad*), las cuales se pueden observar en la Tabla 1. La variable a pronosticar es *CALIDAD*, esta variable toma valores entre cero y diez dependiendo de la calificación que el cliente le asigna en función de sus gustos.

Para el estudio se modificaron los valores de la variable *CALIDAD*, se le asignó cero si el valor dado por el cliente es menor o igual a cinco y si es mayor entonces se le asigna uno [6]. Dicho esto de otro modo, se separaron los valores de esta variable en *Recomendado* con uno y *No Recomendado* con cero, indicando que es una variable binomial lo que facilita el trabajo cuando se apliquen los métodos de clasificación.

En este trabajo se empleó el análisis factorial para reducir el número de variables del conjunto de datos inicial [5], los métodos de clasificación empleados fueron: análisis discriminante, regresión logística y redes neuronales. En la aplicación de los modelos de pronósticos los datos se dividieron en: conjunto de entrenamiento 75% y conjunto de validación del modelo 25%.

**Tabla 1.** Datos que conforman ambos tipos de vino y el nombre de la variable que le corresponde en este trabajo.

| | |
|---|---|
| ACIDFIJ: | Acidez fija (g(ácido tartárico)/dm3) |
| ACIDVOL: | Acidez volátil (g(ácido acético)/dm3) |
| ACIDCIT: | Ácido cítrico (g/dm3) |
| AZUCRES: | Azúcar residual (g/dm3) |
| CLORURO: | Cloruro (g(sodio cloruro)/dm3) |
| DIOXSULFLIB: | Dióxido de azufre libre (mg/dm3) |
| DIOXSULFTOT: | Dióxido de azufre total (mg/dm3) |
| DENSIDAD: | Densidad (g/cm3) |
| PH: | pH |
| SULFATOS: | Sulfatos (g(potasio Sulfato)/dm3) |
| ALCOHOL: | Alcohol (vol. %) |
| CALIDAD: | Calidad (valores 0/1) |

Fuente: Elaborada por el autor.

**Tabla 2.** Total de registros con valores cero y uno en el vino Blanco.

| **Vino** | **Blanco** | |
|---|---|---|
| **variable** ***calidad*** | **0** **(*No Recomendado*)** | **1** **(*Recomendado*)** |
| Total | 1640 | 3258 |
| % | 33.48 | 66.52 |

Fuente: Elaborada por el autor.

**Tabla 3.** Total de registros con valores cero y uno en el vino Tinto.

| **Vino** | **Tinto** | |
|---|---|---|
| **variable** ***calidad*** | **0** **(*No Recomendado*)** | **1** **(*Recomendado*)** |
| Total | 744 | 855 |
| % | 46.53 | 53.47 |

Fuente: Elaborada por el autor.

Los datos del vino blanco cuentan con 4898 registros y del vino tinto con 1599, se puede observar en las Tablas 2 y 3 cómo se distribuyen los valores cero y uno en la variable *CALIDAD* dependiendo del tipo de vino.

*2.2 Análisis factorial y métodos de clasificación*

Se aplicó la técnica de análisis factorial para reducir la cantidad de variables en los dos conjuntos de datos con el objetivo de conocer con cuántos factores se podría trabajar y cuáles serían las variables más importantes.

Como se mencionó antes, en el presente trabajo se emplean tres métodos de clasificación para hacer el pronóstico de la preferencia del vino en función de los criterios de los clientes, estos métodos son: análisis discriminante, regresión logística y redes neuronales. A continuación se menciona, de forma resumida, en qué consiste cada uno y para qué se utiliza.

El análisis discriminante se emplea para determinar a qué grupo pertenece un individuo en función de las variables predictoras, en cualquier caso, el individuo siempre podrá pertenecer a un solo grupo de los posibles a ser clasificados, a partir de los datos que describen a ese individuo. En este caso la variable *CALIDAD* es la que se intenta predecir, la cual podrá tener dos posibles valores, cero o uno. 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 [5] y [7].

La regresión logística es muy parecida al análisis discriminante, este método también puede emplearse para predecir a una variable dependiente de tipo no métrica. Cuando la variable dependiente toma solo dos posibles valores se habla de la regresión logística binomial, es decir, es una variable dicotómica. Para profundizar en este tema se puede consultar [5], [7] y [8].

Las redes neuronales es un método muy diferente al resto del análisis multivariante visto aquí, en este caso, de alguna manera se 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. Para profundizar en este tema se puede consultar [5], [9], [10] y [11].

## 3. Resultados y discusión

### *3.1 Reducción de variables*

3.1.1 Conjunto de datos correspondientes al vino tinto

Como ya se ha mencionado, se empleó el análisis factorial para la reducción de variables con el objetivo de conocer cómo se comportaría el pronóstico de las preferencias en ambos tipos de vino.

Primeramente, se calcula el contraste de esfericidad de Bartlett para valorar la significación de la matriz de correlación, en este caso se puede desechar la hipótesis de que es una matriz identidad pues el valor que se obtiene es prácticamente cero.

Luego se calcula la prueba de la medida de adecuación o suficiencia de muestreo (*MSA*) con el objetivo de conocer si existe una adecuación de los datos a un modelo de análisis factorial. Se cuenta con una clasificación dada por Kaiser [12] y Kaiser y Rice [13] para saber si el valor dado es adecuado como indicador. En la clasificación que dan indican que un valor de *MSA* menor a 0.5 es inaceptable para aplicar el análisis factorial.

En el caso del conjunto de datos del vino tinto el valor de *MSA* que se obtiene es de 0.46 lo que estaría indicando que no se puede aplicar el análisis factorial, esto se puede observar en la Tabla 4, ahora, Hair [5] plantea que se deben ir eliminando variables del análisis partiendo de la variable que obtenga un valor individual de *MSA* menor, en este caso empezamos con la variable de azúcar residual (AZUCRES**)** que presentó un valor de 0.21.

Se aplicó nuevamente el test y el valor de MSA general subió a 0.52, que ya sería un valor aceptable para poder continuar con el análisis, pero se recomienda también que los valores individuales de MSA sean de 0.5 o mayores. Para lograr esto se fueron eliminando las variables siguientes: alcohol, cloruro, dióxido de azufre total y dióxido de azufre libre, alcanzando el valor general de MSA de 0.68. Los valores individuales fueron todos mayores a 0.5 como se puede apreciar en la Tabla 5.

Luego, como se cumplieron los requisitos mínimos para aplicar el análisis factorial se buscó determinar el número de factores, para esto se emplean los autovalores mayores a 1. Se puede apreciar la Tabla 6 que se recomienda emplear solo dos factores.

Cuando se aplicó el análisis factorial con solo dos factores se obtuvieron comunalidades mayores a uno lo que indica que existe alguna anomalía en el cálculo [10]. Se continuó probando con diferentes números de factores logrando el mejor resultado cuando se utilizaron tres. En el análisis factorial se usó una rotación varimax y se empleó el método de ejes principales. En la Tabla 7 se pueden observar las cargas factoriales para cada uno de los factores, las variables que pertenecen a cada uno de ellos y las comunalidades (h2).

**Tabla 4.** Valores individuales de MSA por variable al comenzar el estudio.

| Variable | MSA |
|---|---|
| ACIDFIJ | 0.45 |
| ACIDVOL | 0.57 |
| ACIDCIT | 0.7 |
| **AZUCRES** | **0.21** |
| CLORURO | 0.48 |
| DIOXSULFLIB | 0.49 |
| DIOXSULFTOT | 0.47 |
| DENSIDAD | 0.38 |
| PH | 0.45 |
| SULFATOS | 0.54 |
| ALCOHOL | 0.3 |

Fuente: Elaborada por el autor.

**Tabla 5.** Valores individuales de MSA al finalizar el test.

| Variable | MSA |
|---|---|
| ACIDFIJ | 0.64 |
| ACIDVOL | 0.61 |
| ACIDCIT | 0.75 |
| DENSIDAD | 0.6 |
| PH | 0.74 |
| SULFATOS | 0.77 |

Fuente: Elaborada por el autor.

**Tabla 6.** Se muestran los autovalores después de aplicar el método de análisis paralelo.

| Factores | Autovalores adaptados |
|---|---|
| 1 | 2.853851 |
| 2 | 1.158152 |
| 3 | 0.810867 |
| 4 | 0.589088 |
| 5 | 0.340345 |
| 6 | 0.247694 |

Fuente: Elaborada por el autor.

**Tabla 7.** Matriz de cargas factoriales.

| Variable | PA1 | PA2 | PA3 | h2 |
|---|---|---|---|---|
| ACIDFIJ | 0.83 | 0.24 | 0.44 | 0.94 |
| PH | -0.68 | -0.24 | -0.14 | 0.54 |
| ACIDVOL | -0.14 | -0.73 | 0.11 | 0.56 |
| ACIDCIT | 0.51 | 0.69 | 0.19 | 0.77 |
| SULFATOS | 0.1 | 0.36 | 0.1 | 0.15 |
| DENSIDAD | 0.31 | 0.05 | 0.9 | 0.91 |

Fuente: Elaborada por el autor.

**Tabla 8.** Valores individuales de MSA por variable al comenzar el estudio.

| Variable | MSA |
|---|---|
| ACIDFIJ | 0.17 |
| ACIDVOL | 0.23 |
| ACIDCIT | 0.72 |
| AZUCRES | 0.33 |
| CLORURO | 0.66 |
| DIOXSULFLIB | 0.59 |
| DIOXSULFTOT | 0.71 |
| DENSIDAD | 0.4 |
| **PH** | **0.16** |
| SULFATOS | 0.19 |
| ALCOHOL | 0.36 |

Fuente: Elaborada por el autor.

En este caso se puede observar que la variable sulfatos (SULFATOS) no tiene una carga factorial grande y su comunalidad es muy baja lo que indica que puede eliminarse del estudio. El análisis factorial nos indica que las variables acidez fija (ACIDFIJ) y el pH (PH) pertenecen al factor uno, las variables acidez volátil (ACIDVOL) y el ácido cítrico (ACIDCIT) pertenecen al dos y finalmente la variable densidad (DENSIDAD) al tres. Para este resultado se obtuvo una varianza acumulada del 64%. Luego de aplicar este método solo quedaron cinco variables; estas fueron las que se emplearon para realizar el pronóstico de la preferencia del vino tinto, de un total inicial de 11.

3.1.2 Conjunto de datos correspondientes al vino blanco

Para el conjunto de datos del vino blanco se realizó el mismo procedimiento que para el vino tinto, se empezó por investigar si se cumplían los requisitos mínimos para aplicar el análisis factorial.

Se calculó el contraste de esfericidad de Bartlett para valorar la significación de la matriz de correlación, en este caso se puede desechar también la hipótesis de que es una matriz identidad pues el valor que se obtiene es prácticamente cero. Se calculó el MSA y se obtuvo un valor de 0.37 lo que indica que no se puede aplicar aún el análisis factorial por lo que se comenzó a eliminar variables una a una a partir de la de menor valor de MSA, esto puede apreciarse en la Tabla 8, el orden en que se eliminaron las variables fue: pH, sulfatos, acidez fija, acidez volátil y azúcar residual.

Luego de eliminar las variables mencionadas se obtuvo un valor de *MSA* general de 0.65 que entra en la clasificación de aceptable para aplicar el análisis factorial, en la Tabla 9 se pueden apreciar los valores individuales de *MSA* y todos están por encima de 0.5. Como se cumplen los requisitos mínimos para aplicar el análisis factorial [5], se busca determinar el número de factores, para esto se seleccionan la cantidad de autovalores mayores a uno, en la Tabla 10 se puede apreciar que después de aplicar el método paralelo [10] se indica emplear solo dos factores.

Cuando se aplicó el análisis factorial con solo dos factores también se obtuvieron comunalidades mayores a 1 lo que indica que existe también alguna anomalía en el cálculo [10]. Se continuó probando con diferentes números de factores logrando el mejor resultado cuando se utilizaron tres. Se empleó una rotación varimax y el método de ejes principales. En la Tabla 11 se pueden observar las cargas factoriales por cada factor, las variables que pertenecen a cada uno de ellos y las comunalidades (h2).

**Tabla 9.** Valores individuales de MSA al finalizar el test.

| Variables | MSA |
|---|---|
| ACIDCIT | 0.64 |
| CLORURO | 0.74 |
| DIOXSULFLIB | 0.63 |
| DIOXSULFTOT | 0.69 |
| DENSIDAD | 0.64 |
| ALCOHOL | 0.63 |

Fuente: Elaborada por el autor.

**Tabla 10.** Se muestran los autovalores después de aplicar el método de análisis paralelo.

| Componente | Autovalores adaptados |
|---|---|
| 1 | 2.604413 |
| 2 | 1.037266 |
| 3 | 0.967571 |
| 4 | 0.777945 |
| 5 | 0.366698 |
| 6 | 0.246104 |

Fuente: Elaborada por el autor.

**Tabla 11.** Matriz de cargas factoriales.

| Variables | PA1 | PA2 | PA3 | h2 |
|---|---|---|---|---|
| DENSIDAD | 0.89 | 0.27 | 0.15 | 0.888 |
| ALCOHOL | -0.76 | -0.19 | -0.32 | 0.72 |
| DIOXSULFTOT | 0.33 | 0.79 | 0.15 | 0.759 |
| DIOXSULFLIB | 0.11 | 0.72 | 0.07 | 0.531 |
| CLORURO | 0.16 | 0.05 | 0.7 | 0.517 |
| ACIDCIT | 0.08 | 0.1 | 0.12 | 0.031 |

Fuente: Elaborada por el autor.

En este caso se puede observar que la variable ácido cítrico (ACIDCIT) no tiene una carga factorial alta y su comunalidad es muy baja (0.031) por lo que se podría eliminar del estudio. El análisis factorial indicó que las variables densidad (DENSIDAD) y alcohol (ALCOHOL) pertenecen al factor uno, las variables dióxido de azufre total (DIOXSULFTOT) y el dióxido de azufre libre (DIOXSULFLIB) pertenecen al dos y finalmente, el cloruro (CLORURO) al tres. Se obtuvo una varianza acumulada del 57%. Luego de aplicar este método solo quedaron cinco variables; estas fueron las que se emplearon para realizar el pronóstico de la preferencia del vino blanco, de un total inicial de 11.

*3.2 Incidencia en el pronóstico*

Se debe mencionar que en los métodos de pronósticos empleados en este trabajo: análisis discriminante, regresión logística y redes neuronales, cuando se habla de valores positivos (cero) se refieren a la clasificación de *Vino No Recomendado* mientras que valores negativos (uno) se refiere a la clasificación de *Vino Recomendado*.

Se crearon tres grupos diferentes para realizar los pronósticos, el primero contiene todas las variables originales del conjunto de datos, es decir, los datos originales, el segundo está formado por las variables que están incluidas en los factores uno, dos y tres resultantes del análisis factorial y el tercero solo por las variables que están incluidas en el factor uno.

Para los tres grupos se utilizó la misma semilla en la generación de los números aleatorios que crearon el conjunto de datos de entrenamiento y validación. Esto permite hacer una buena comparación en cuanto a los resultados que se obtienen en el pronóstico, basados en los grupos de datos mencionados y empleando los tres métodos de clasificación.

3.2.1 Regresión Logística

Primeramente, se muestra la matriz de confusión óptima para cada tipo de vino, en las Tablas 12 y 13 se puede apreciar en la correspondiente a la del vino tinto y blanco respectivamente. Es válido aclarar que los pronósticos son para cada fila de la matriz, por ejemplo, en la Tabla 12, que se corresponde con el vino tinto, el método de pronóstico que se emplee debería predecir para los valores positivos (cero) 181 elementos y ningún negativo (uno), en cambio, debería predecir 219 elementos negativos (uno) y ninguno positivo.

Estas matrices óptimas se basan en los valores que forman los conjuntos de validación de cada tipo de vino y lo que se pretende es que cuando se aplique cualquier método de clasificación permita obtener resultados lo más parecidos a los que se muestran en ambas tablas.

**Tabla 12**. Matriz de confusión óptima para el vino tinto.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 181 | 0 |
| | **1** | 0 | 219 |

Fuente: Elaborada por el autor.

**Tabla 13**. Matriz de confusión óptima para el vino blanco.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 387 | 0 |
| | **1** | 0 | 837 |

Fuente: Elaborada por el autor.

*Vino tinto*

Las matrices de confusión que se obtienen cuando se aplica el método de regresión logística al conjunto de datos del vino tinto se puede apreciar en las Tablas 14, 15 y 16 respectivamente según al grupo que corresponde.

**Tabla 14**. Matriz de confusión del grupo #1.

| | | Referencia | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 148 | 79 |
| | **1** | 33 | 140 |

Fuente: Elaborada por el autor.

**Tabla 15**. Matriz de confusión del grupo #2.

| | | Referencia | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 147 | 104 |
| | **1** | 34 | 115 |

Fuente: Elaborada por el autor.

**Tabla 16**. Matriz de confusión del grupo #3.

| | | Referencia | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 169 | 199 |
| | **1** | 12 | 20 |

Fuente: Elaborada por el autor.

En la Tabla 17 se puede apreciar diferentes medidas de los modelos que se obtienen en cada grupo. El que tiene mayor exactitud es la del grupo #1, este resultado era el que se esperaba que se obtuviera pues cuenta con la totalidad de las variables que intervienen en el estudio, pero es importante observar cómo el valor de exactitud del modelo de pronóstico del grupo #2 es bastante bueno si se toma como referencia el valor de 72% que se obtuvo en el grupo #1. En el caso del grupo #3 es muy bajo, está por debajo del 50% y es evidente que no debería tomarse en cuenta para realizar un pronóstico en el que pueda confiarse.

Se muestran otras medidas importantes del modelo de pronóstico para los tres grupos que se analizan. Se puede observar que la sensibilidad es muy parecida para los grupos #1 y #2. En el caso de la especificidad hay más de 10% de diferencias entre estos dos grupos. En el caso de los valores de pronóstico negativos (VP-) estos son bastante representativos para ambos conjuntos. En el caso de los valores de pronóstico positivos (VP+) tienen una diferencia de alrededor del 8%. Los resultados del modelo que se obtuvo con el grupo #3 no son representativos ni útiles pues la exactitud es muy baja (47.25%), a pesar de que el valor de la sensibilidad es muy alto con un 93.37%, no se debe tener en cuenta.

**Tabla 17.** Medidas resultantes de la regresión logística aplicada al vino tinto.

| **REGRESION LOGISTICA. VINO TINTO** | | | |
|---|---|---|---|
| MEDIDAS | GRUPO #1 | GRUPO #2 | GRUPO #3 |
| Exactitud | 0.72 | 0.655 | 0.4725 |
| Sensibilidad | 0.8177 | 0.8122 | 0.9337 |
| Especificidad | 0.6393 | 0.5251 | 0.0913 |
| VP+ | 0.652 | 0.5857 | 0.45924 |
| VP- | 0.8092 | 0.7718 | 0.625 |

Fuente: Elaborada por el autor.

*Vino blanco*

En el caso del vino blanco se puede apreciar los resultados de la matriz de confusión en las Tablas 18, 19 y 20. Las matrices de confusión de los grupos #1 y #2 son muy parecidas. Para el caso del grupo #3 no es buena pero no está alejada de los resultados que se obtuvieron para los dos primeros conjuntos de datos.

En la Tabla 21 se pueden apreciar los resultados que se obtuvieron en las diferentes medidas para el vino blanco. En este caso la exactitud del modelo es muy parecida para cualquier grupo, el mejor modelo sigue siendo el del grupo #1 pero los valores de los otros dos son muy cercanos, en este caso podríamos tenerlo en cuenta para hacer una predicción si asumimos que el valor del grupo #1 es del 71.9% mientras que el #2 es de 69.04% y del #3 es 67.65%, la diferencia entre el mayor y el menor es solo del 4.25% y entre el grupo #2 y #3 es de 1.39%

**Tabla 18.** Matriz de confusión del grupo #1.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 249 | 206 |
| | **1** | 138 | 631 |

Fuente: Elaborada por el autor.

**Tabla 19.** Matriz de confusión del grupo #2.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 260 | 252 |
| | **1** | 127 | 585 |

Fuente: Elaborada por el autor.

**Tabla 20.** Matriz de confusión del grupo #3.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 258 | 267 |
| | **1** | 129 | 570 |

Fuente: Elaborada por el autor.

**Tabla 21.** Medidas resultantes de la regresión logística aplicada al vino blanco.

| **REGRESION LOGISTICA. VINO BLANCO** | | | |
|---|---|---|---|
| MEDIDAS | GRUPO #1 | GRUPO #2 | GRUPO #3 |
| Exactitud | 0.719 | 0.6904 | 0.6765 |
| Sensibilidad | 0.6434 | 0.6718 | 0.6667 |
| Especificidad | 0.7539 | 0.6989 | 0.6810 |
| VP+ | 0.5473 | 0.5078 | 0.4914 |
| VP- | 0.8205 | 0.8216 | 0.8155 |

Fuente: Elaborada por el autor.

La sensibilidad es prácticamente igual para los tres grupos. En cuanto a la especificidad los valores difieren un poco, pero la diferencia entre el mayor y el menor es de 7.29%. Para los valores de predicción negativos (VP-) todos están por encima del 80% mientras los valores de predicción negativos están alrededor del 50%.

### 3.2.2 Análisis Discriminante

Para aplicar el análisis discriminante es necesario realizar la prueba de significancia estadística [10], para todos los grupos, tanto para el vino tinto como para el vino blanco, los valores que resultaron fueron muy próximos a cero, lo que indica que las variables clasificadoras cuentan con una capacidad discriminante significativa.

*Vino tinto*

En las Tablas 22, 23 y 24 se muestran los resultados de la matriz de confusión para los tres grupos del vino tinto. Para los grupos #1 y #2 los resultados son parecidos, el error mayor está cuando se predicen los valores positivos, en

el grupo #2 difiere por 24 elementos con respecto al grupo #1. Para el grupo #3 los valores difieren considerablemente con el resto de los grupos.

**Tabla 22.** Matriz de confusión del grupo #1.

| | | Referencia | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 133 | 36 |
| | **1** | 60 | 171 |

Fuente: Elaborada por el autor.

**Tabla 23.** Matriz de confusión del grupo #2.

| | | Referencia | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 109 | 60 |
| | **1** | 63 | 168 |

Fuente: Elaborada por el autor.

**Tabla 24.** Matriz de confusión del grupo #3.

| | | Referencia | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 72 | 97 |
| | **1** | 102 | 129 |

Fuente: Elaborada por el autor.

En la Tabla 25 se muestran los resultados del análisis discriminante para los datos del vino tinto. Si se comparan con los del modelo de regresión logística se aprecia que la exactitud para cada grupo aumenta en casi un 4%, esto es importante resaltarlo porque es un indicador de que con el método de análisis discriminante la predicción aumenta.

Tal y como ocurrió con la regresión logística los valores de los grupos #1 y #2 son muy parecidos, tanto en la sensibilidad como en los valores de predicción negativos (VP-). En el caso de la especificidad y de los valores de pronósticos positivos (VP+) difieren alrededor del 9% y 14% respectivamente. En el caso del grupo #3 la exactitud aumentó, no de forma significativa, pues sigue siendo bajo si lo comparamos con los grupos #1 y #2, el resto de las medidas son también bajos y no se deben tener en cuenta para hacer una predicción real.

**Tabla 25.** Medidas resultantes del análisis discriminante aplicado al vino tinto.

| **ANÁLISIS DISCRIMINANTE. VINO TINTO** | | | |
|---|---|---|---|
| MEDIDAS | GRUPO #1 | GRUPO #2 | GRUPO #3 |
| Exactitud | 0.76 | 0.6925 | 0.5025 |
| Sensibilidad | 0.6891 | 0.6337 | 0.4138 |
| Especificidad | 0.8261 | 0.7368 | 0.5708 |
| VP+ | 0.787 | 0.645 | 0.426 |
| VP- | 0.7403 | 0.7273 | 0.5584 |

Fuente: Elaborada por el autor.

*Vino blanco*

En las Tablas 26, 27 y 28 se observan los resultados de la matriz de confusión para los tres grupos del vino blanco. Se aprecian diferencias en las predicciones, sobre todo en los resultados del grupo #3.

**Tabla 26.** Matriz de confusión del grupo #1.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 216 | 216 |
| | **1** | 92 | 700 |

Fuente: Elaborada por el autor.

**Tabla 27.** Matriz de confusión del grupo #2.

| | | **Referencia** | |
|---|---|---|---|
| | | 0 | 1 |
| **Predicción** | 0 | 169 | 263 |
| | 1 | 109 | 683 |

Fuente: Elaborada por el autor.

**Tabla 28.** Matriz de confusión del grupo #3.

| | | **Referenica** | |
|---|---|---|---|
| | | 0 | 1 |
| **Predicción** | 0 | 159 | 273 |
| | 1 | 112 | 680 |

Fuente: Elaborada por el autor.

En la Tabla 29 se aprecian los resultados del análisis discriminante para los datos del vino blanco. Si se comparan con los de la regresión logística los valores de exactitud del modelo en los grupos #2 y #3 se mantienen muy parecidos, casi iguales, pero en el caso del grupo #1 es mayor en 3.65%, es decir, se hace una mejor predicción para el vino blanco si se emplea el modelo de análisis discriminante.

En el caso de los valores de predicción negativos (VP-) en los grupos #1 y #2 son muy parecidos y están por arriba del 86% en ambos casos. La sensibilidad tiene una diferencia del 10% entre los grupos #1 y #2 y la especificidad es muy parecida con una diferencia de poco más del 4%. En cuanto a los valores de predicción positivos (VP+) en ambos son muy bajos, fueron del 50% para el grupo #1 y del 39.12% para el #2.

La exactitud del grupo #3 fue bastante buena cuando se aplicó este modelo, la sensibilidad y los VP+ son considerablemente bajos.

**Tabla 29.** Medidas resultantes del análisis discriminante aplicado al vino blanco.

| **ANÁLISIS DISCRIMINANTE. VINO BLANCO** | | | |
|---|---|---|---|
| MEDIDAS | GRUPO #1 | GRUPO #2 | GRUPO #3 |
| Exactitud | 0.7484 | 0.6961 | 0.6855 |
| Sensibilidad | 0.7013 | 0.6079 | 0.5867 |
| Especificidad | 0.7642 | 0.722 | 0.7135 |
| VP+ | 0.5 | 0.3912 | 0.3681 |
| VP- | 0.8838 | 0.8624 | 0.8586 |

Fuente: Elaborada por el autor.

### 3.2.3 Redes Neuronales

*Vino tinto*

Para el conjunto de datos del grupo #1 se empleó una estructura con dos capas ocultas de tres neuronas cada una, para el #2 se empleó una capa oculta con dos neuronas y para el #3 una capa oculta con una neurona.

En las Tablas 30, 31 y 32 se muestran las matrices de confusión que resultaron de aplicar el modelo de redes neuronales para los tres conjuntos de variables. Se aprecia que el grupo #1 tiene mejores valores de predicción con respecto a las de los grupos #2 y #3. En las tres los resultados están lejos de la matriz de confusión óptima para el vino tinto mostrada en la Tabla 12, aunque se puede apreciar que los resultados del grupo #1 son los mejores que se han obtenido si lo comparamos con los modelos vistos anteriormente.

**Tabla 30.** Matriz de confusión del grupo #1.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 138 | 31 |
| | **1** | 51 | 180 |

Fuente: Elaborada por el autor.

**Tabla 31.** Matriz de confusión del grupo #2.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 111 | 58 |
| | **1** | 68 | 163 |

Fuente: Elaborada por el autor.

**Tabla 32.** Matriz de confusión del grupo #3.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 76 | 93 |
| | **1** | 104 | 127 |

Fuente: Elaborada por el autor.

En la Tabla 33 se muestran los resultados al aplicar el método de redes neuronales a los tres grupos de variables. La exactitud del modelo para el grupo #1 fue la mayor que se obtuvo, teniendo una diferencia con la exactitud del modelo de regresión logística de 7.5% y con el modelo del análisis discriminante de 3.5%. En cuanto a los resultados de los grupos #2 y #3 son muy parecidos a los que se obtuvieron en los tres métodos de predicción aplicados. Esto se puede ver en la Tabla 34.

**Tabla 33.** Medidas resultantes del método de redes neuronales aplicado al vino tinto.

| **REDES NEURONALES. VINO TINTO** | | | |
|---|---|---|---|
| MEDIDAS | GRUPO #1 | GRUPO #2 | GRUPO #3 |
| Exactitud | 0.795 | 0.685 | 0.5075 |
| Sensibilidad | 0.7302 | 0.6201 | 0.4222 |
| Especificidad | 0.8531 | 0.7376 | 0.5773 |
| VP+ | 0.8166 | 0.6568 | 0.4497 |
| VP- | 0.7792 | 0.7056 | 0.5498 |

Fuente: Elaborada por el autor.

En la Tabla 34 la diferencia entre la exactitud del modelo de redes neuronales en los grupos #1 y #2 es de 11%. En general los resultados del segundo grupo son muy parecidos. Los del #3 no son adecuados para tener en cuenta su predicción en cualquiera de los métodos empleados.

**Tabla 34.** Resumen de los valores de exactitud de los diferentes métodos aplicados al vino tinto.

| EXACTITUD DEL MODELO. VINO TINTO | | | |
|---|---|---|---|
| MÉTODOS | GRUPO #1 | GRUPO #2 | GRUPO #3 |
| REG-LOG | 0.72 | 0.655 | 0.4725 |
| ANÁL-DISC | 0.76 | 0.6925 | 0.5025 |
| RED-NEUR | 0.795 | 0.685 | 0.5075 |

Fuente: Elaborada por el autor.

*Vino blanco*

En las Tablas 35, 36 y 37 se muestran las matrices de confusión al emplear las redes neuronales en los tres conjuntos de datos del vino blanco. Los resultados indican que la matriz del grupo #1 es la que más se acerca a la matriz de confusión óptima del vino blanco que se puede apreciar en la Tabla 13.

En la Tabla 38 se muestran los resultados al aplicar el método de redes neuronales a los tres grupos de datos. Los valores de exactitud para los tres son altos, de hecho, el valor en el #2 fue de 71.24% que está muy próximo a la exactitud del grupo #1 al aplicar el método de regresión logística. Esto se puede ver en la Tabla 39.

**Tabla 35.** Matriz de confusión del grupo #1.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 277 | 155 |
| | **1** | 119 | 673 |

Fuente: Elaborada por el autor.

**Tabla 36.** Matriz de confusión del grupo #2.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 220 | 212 |
| | **1** | 140 | 652 |

Fuente: Elaborada por el autor.

**Tabla 37.** Matriz de confusión del grupo #3.

| | | **Referencia** | |
|---|---|---|---|
| | | **0** | **1** |
| **Predicción** | **0** | 172 | 260 |
| | **1** | 126 | 666 |

Fuente: Elaborada por el autor.

**Tabla 38.** Medidas resultantes del método de redes neuronales aplicado al vino blanco.

| REDES NEURONALES. VINO BLANCO | | | |
|---|---|---|---|
| MEDIDAS | GRUPO #1 | GRUPO #2 | GRUPO #3 |
| Exactitud | 0.7761 | 0.7124 | 0.6846 |
| Sensibilidad | 0.6995 | 0.6111 | 0.5772 |
| Especificidad | 0.8128 | 0.7546 | 0.7192 |
| VP+ | 0.6412 | 0.5093 | 0.3981 |
| VP- | 0.8497 | 0.8232 | 0.8409 |

Fuente: Elaborada por el autor.

En la Tabla 39 los valores de exactitud de los grupos #2 y #3 son muy parecidos, lo que indica que los tres conjuntos de datos cuando se aplica cualquier método de pronóstico los resultados son muy buenos. Se debe señalar que en el caso del grupo #1 los mejores resultados se obtienen cuando se emplean las redes neuronales.

**Tabla 39.** Resumen de los valores de exactitud de los diferentes métodos aplicados al vino blanco.

| EXACTITUD DEL MODELO. VINO BLANCO | | | |
|---|---|---|---|
| MÉTODOS | GRUPO #1 | GRUPO #2 | GRUPO #3 |
| REG-LOG | 0.719 | 0.6904 | 0.6765 |
| ANÁL-DISC | 0.7484 | 0.6961 | 0.6855 |
| RED-NEUR | 0.7761 | 0.7124 | 0.6846 |

Fuente: Elaborada por el autor.

3.2.4 Curvas ROC

Las curvas *ROC (Receiver Operating Characteristics)* se emplearon en este trabajo para determinar la capacidad discriminante de los modelos empleados [10], que permiten hacer una distinción entre los dos tipos de valores que puede tomar la variable dependiente.

Los resultados que se obtuvieron indican que para el grupo #1, tanto para el vino tinto como para el vino blanco, los modelos tienen una buena capacidad discriminante, pues como se aprecia en las Figuras 1 y 2, los valores de *AUC (Area Under the Curve)* son mayores al 80%.

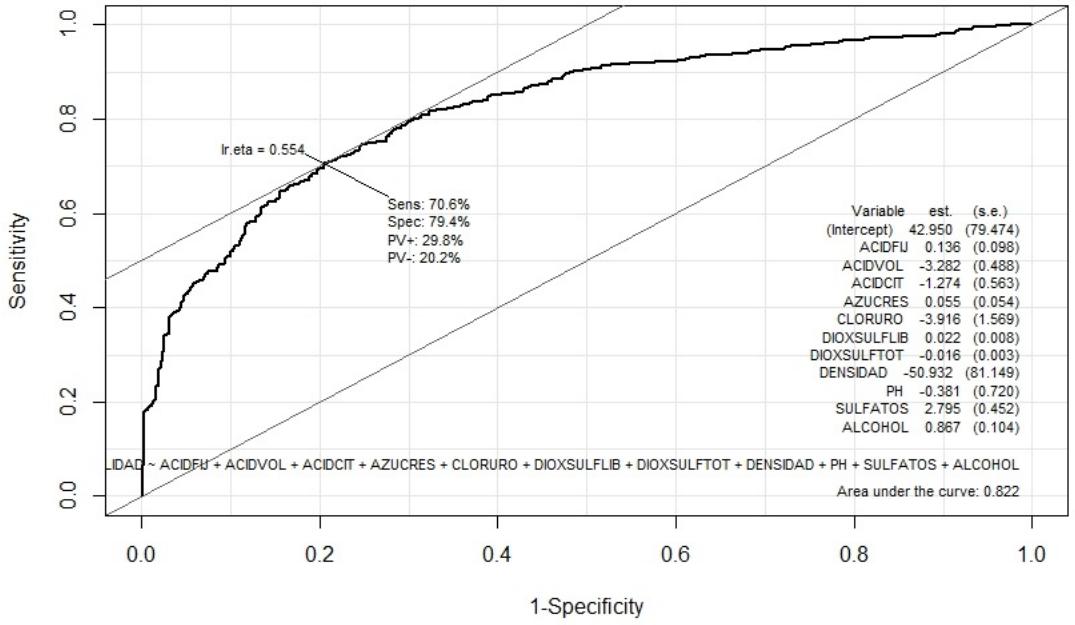


**Figura 1.** Curva *ROC* correspondiente al vino tinto del grupo #1.
Fuente: Elaboración propia.

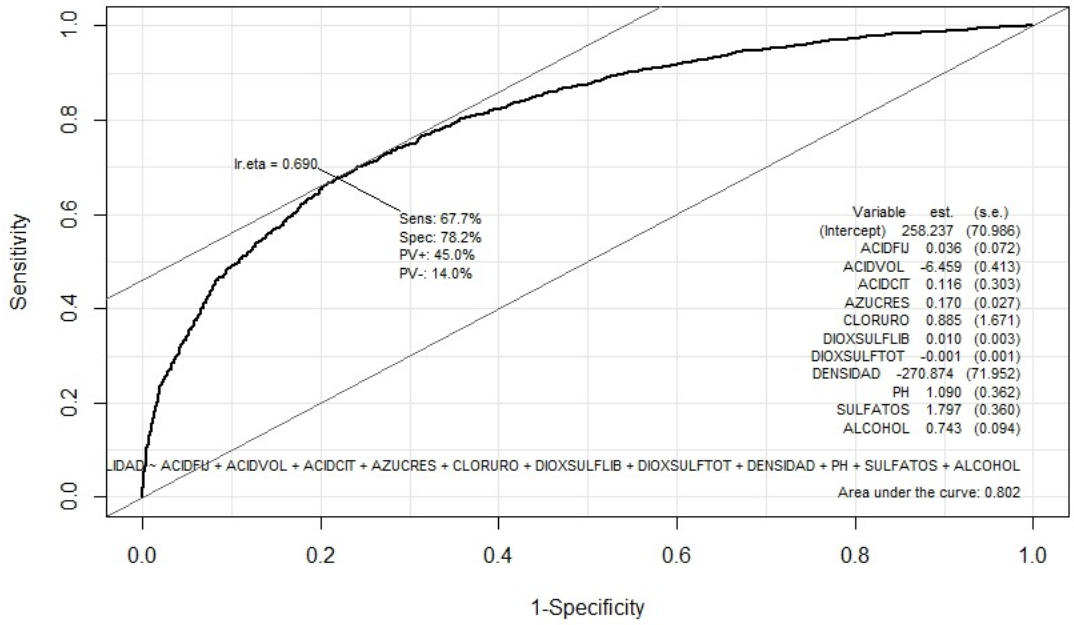


**Figura 2.** Curva *ROC* correspondiente al vino blanco del grupo #1.
Fuente: Elaboración propia.

En el grupo #2 el valor de *AUC*, para ambos tipos de vinos, está por arriba del 75% de probabilidades de que el modelo pueda distinguir entre los vinos *No Recomendados* y los *Recomendados*. Esto resultados se aprecian en las Figuras 3 y 4.

En cuanto al grupo #3 el valor de *AUC* para el vino tinto es de 57.3%, lo cual indica que el modelo tiene muy baja capacidad discriminante entre las dos clases que se analizan. El valor de *AUC* para el vino blanco es del 73.9% lo cual indica que el modelo tiene una buena capacidad discriminante. Estas curvas se muestran en las Figuras 5 y 6.

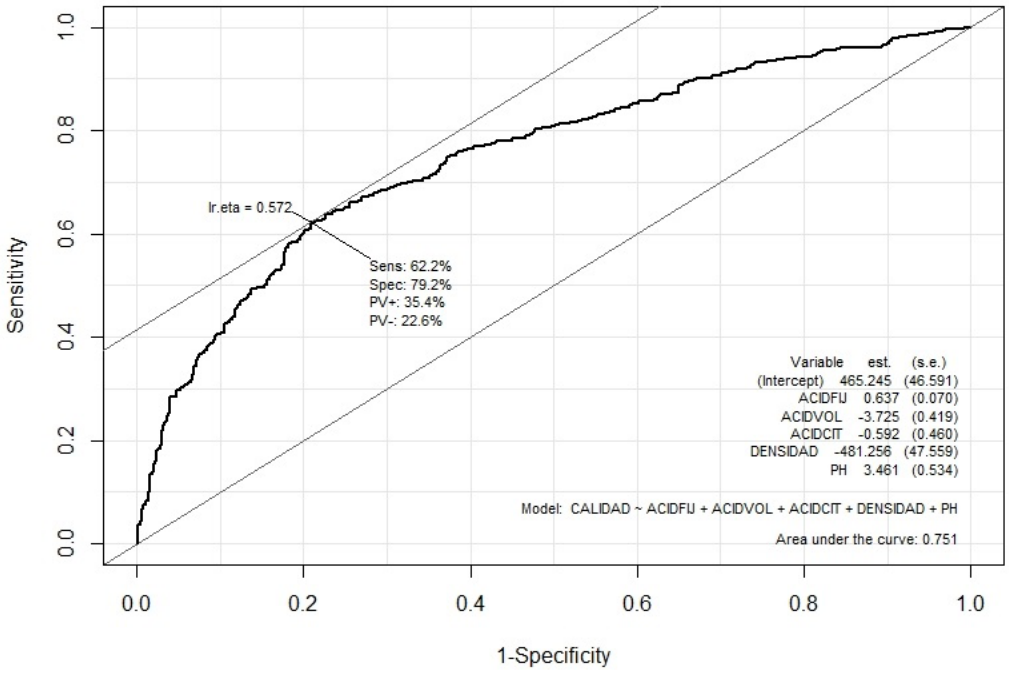


**Figura 3.** Curva *ROC* correspondiente al vino tinto del grupo #2.
Fuente: Elaboración propia.

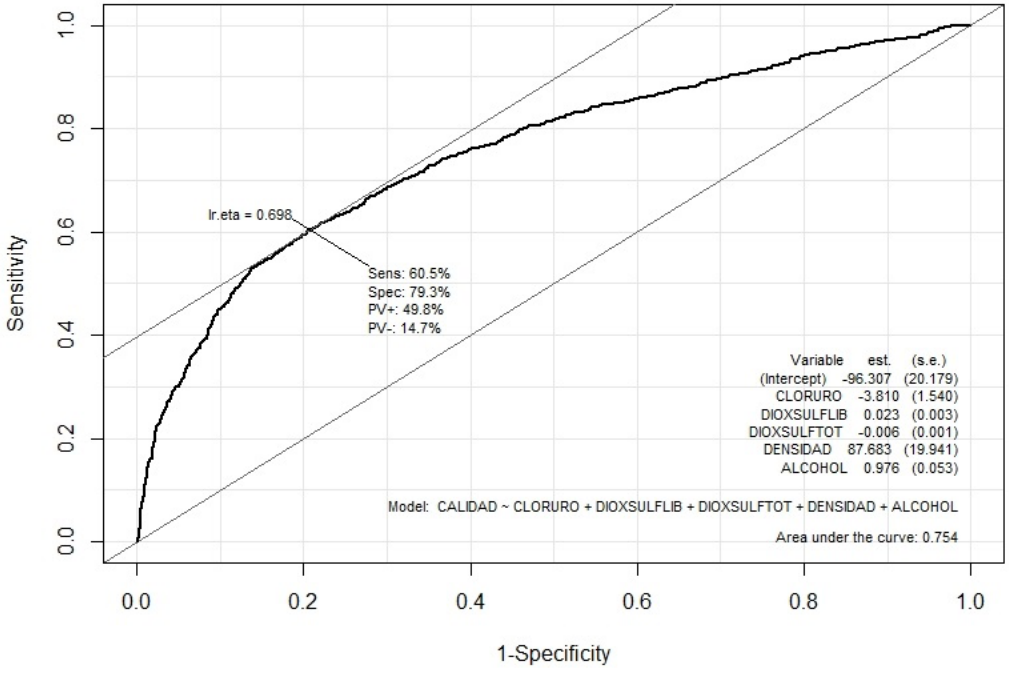


**Figura 4.** Curva *ROC* correspondiente al vino blanco del grupo #2.
Fuente: Elaboración propia

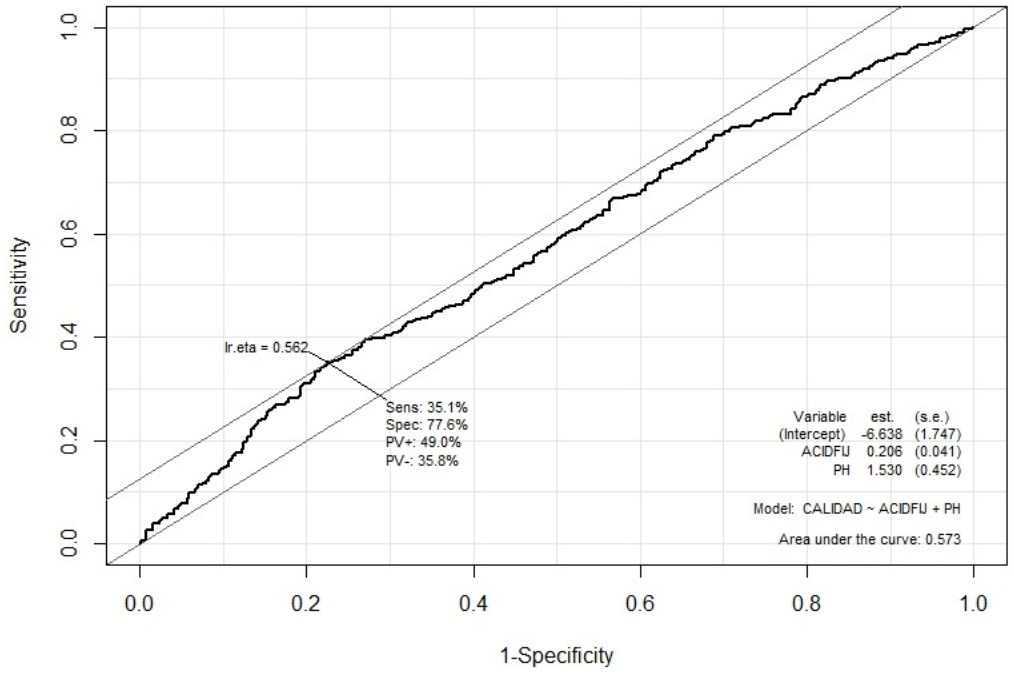


**Figura 5.** Curva *ROC* correspondiente al vino tinto del grupo #3.
Fuente: Elaboración propia.

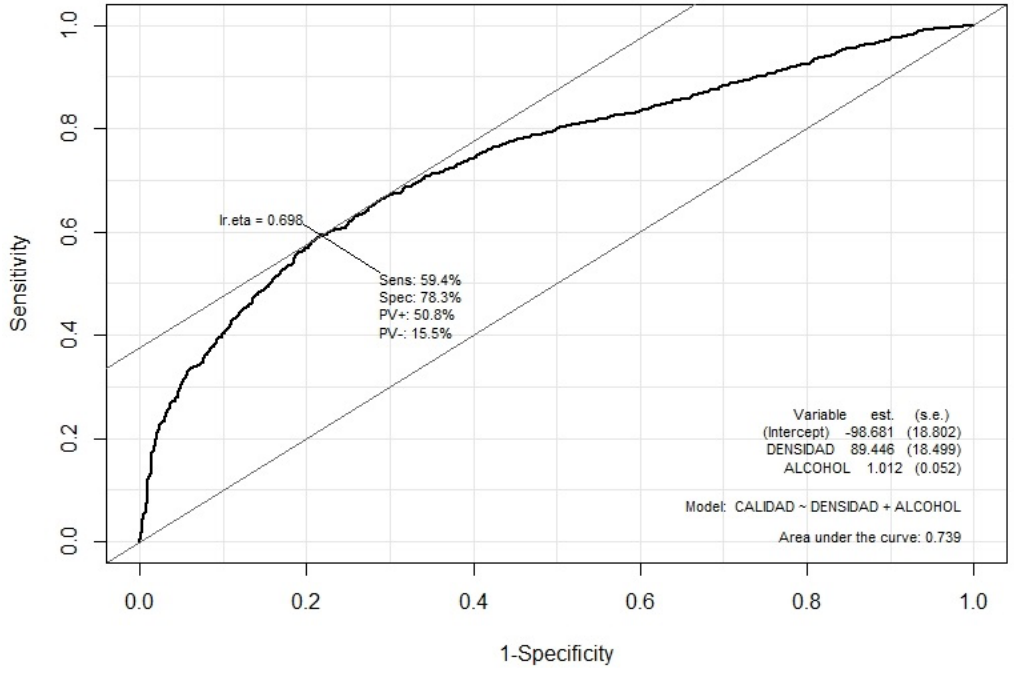


**Figura 6.** Curva *ROC* correspondiente al vino blanco del grupo #3.
Fuente: Elaboración propia.

## 4. Conclusiones

El empleo del lenguaje R resultó ser de muy fácil uso en la programación de los diferentes métodos de clasificación, así como para la implementación del análisis factorial. El RStudio constituyó una herramienta muy útil en la ejecución de los diferentes ejemplos.

Al aplicar el análisis factorial para ambos conjuntos de datos iniciales se obtienen tres factores, en los cuales se agrupan cinco variables, las cuales fueron diferentes para los dos tipos de vinos. En ambos casos se eliminó una variable por tener un valor de comunalidad muy bajo.

Las curvas *ROC* y los valores de *AUC* que se obtuvieron indican que los modelos para los grupos #1 y #2 tienen capacidad discriminante, lo que significa que pueden hacer una buena distinción entre los vinos *Recomendados y No Recomendados*. Para el caso del grupo #3 solo el modelo de pronóstico para el vino blanco resultó tener una buena capacidad discriminante, aunque menor que los obtenidos en los otros dos grupos. Para el modelo del vino tinto del grupo #3 su valor de *AUC* fue muy bajo lo que indica una mala capacidad discriminante y no se debe tener en cuenta para realizar algún tipo de pronóstico.

Al aplicar la matriz de confusión se obtuvo que los valores de exactitud del vino tinto fueron bastante parecidos entre los grupos #1 y #2, esto es un resultado muy interesante pues el grupo #1 está compuesto por las variables originales, las cuales son once, mientras que el #2 tiene solo cinco variables; la diferencia entre los valores de exactitud entre ellos está alrededor del diez por ciento. Los valores del grupo #3 resultaron sobre el cincuenta por ciento para los tres métodos de clasificación. Para el vino blanco fueron mucho mejores, pues los valores de la exactitud de los modelos fueron muy parecidos, teniendo una diferencia muy baja entre ellos. Los mejores modelos siempre son cuando utilizamos las once variables, tanto para el vino tinto como para el blanco, ahora, para este conjunto de datos es recomendable emplear el grupo #2 ya que sus resultados son bastante buenos si se comparan con los que se obtuvieron en el grupo #1.

## 5. Agradecimientos

Deseo agradecer al profesor Dr. Noé Amir Rodríguez Olivares, a la Dra. Rosa María Macias Herrera y al especialista Adrián Durán Chavesti por su invaluable y constante apoyo en la elaboración del presente artículo.

## 6. Referencias

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

[2] Rayo Llerena, I., Marín Huerta, E. (1998). Vino y Corazón. *Revista Española de Cardiología*, 51 (6), 435- 449. Recuperado de: https://www.revespcardiol.org/es-vino-corazon-articulo-X0300893298002947?redirect=true

[3] 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. *BMJ*, 309 (6959), 911-918. doi: https://doi.org/10.1136/bmj.309.6959.911

[4] Moreno Padilla, R. D. (2019). La llegada de la inteligencia artificial a la educación. *Revista de Investigación en Tecnologia de la Información (RITI)*, 7 (14), 260-270. doi: https://doi.org/10.36825/RITI.07.14.022

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

[6] Valencia Ramírez, J. P. (2019). Contratos inteligentes. *Revista de Investigación en Tecnologia de la Información (RITI)*, 7 (14), 1-10. doi: https://doi.org/10.36825/RITI.07.14.001

[7] 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

[8] Pearson, R. K. (2018). *Exploratory Data Analysis Using R*. Boca Raton, US: CRC Press-Taylor & Francis Group.
[9] 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
[10] Aldás, J., Uriel, E. (2017). *Análisis Multivariante aplicado con R* (2da Ed.). Madrid, España: Ediciones Paraninfo .
[11] Hodnett, M., Wiley, J. F. (2018). *R Deep Learning Essentials* (2da Ed.). UK: Packt Publishing Ltd.
[12] Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31-36. doi: https://doi.org/10.1007/BF02291575
[13] Kaiser, H. F., Rice, J. (1974). Litte Jiffy, Mark IV. *Educational and Psychological Measurement*, 34, 111-117. doi: https://doi.org/10.1177/001316447403400115