Control de calidad del vino tinto “Vino verde” mediante modelos de aprendizaje automático

Da Silva, J., Díaz, B., Fariña, M.

Resumen:

Proponemos un enfoque de modelos basados en aprendizaje automático para predecir las preferencias de sabor del vino en los humanos, que se basa en información fácilmente disponible en las pruebas analíticas recolectadas en el paso a la certificación. Un gran conjunto de datos es considerado, con muestras de vino tinto “Vinho verde” (de Portugal). Se realizó una limpieza del conjunto de datos y dos clasificaciones con técnicas de aprendizaje automático mediante máquinas de soporte vectorial (SVM) y árboles de decisión, bajo procedimientos computacionales eficientes. Dicho modelo es útil para apoyar las evaluaciones de cata de vinos del enólogo y mejorar la producción de vino. Además, técnicas similares pueden ayudar en el marketing objetivo al modelar gustos de los consumidores de nichos de mercado.

Palabras clave: *aprendizaje automático, clasificación, vectores de máquina soporte, árboles de decisión.*

1. Introducción

Vino verde es la denominación de origen que designa aquellos vinos producto de uvas originarias de la región de Vinho verde en Portugal, cuya vendimia y producción se lleva a cabo exclusivamente en la mencionada región (*Vinho Verde*, s. f.). Por vino tinto se entiende aquel producto de la vinificación de mostos de uvas tintas o aquellas cuyo jugo es tinto (A.L. Waterhouse, 2004).

Los métodos de aprendizaje automático engloban numerosas técnicas a la vez que algoritmos para extraer información, estimar dependencia o estructura desconocida de un sistema a partir de los datos otorgados, mediante un número limitado de observaciones del tipo entrada/salida (Pajares, 2011). El aprendizaje automático supervisado, divide al conjunto de pruebas en dos subconjuntos; uno de mayor tamaño (entrenamiento) y el otro de prueba, donde se comprueba la validez del modelo sobre los datos restantes (Tuya et al., 2007).

Mediante el aprendizaje supervisado se pueden realizar tareas de regresión para encontrar una función que modele los datos con el menor error así como de clasificación, que generaliza una estructura conocida al aplicarla a los nuevos datos (Jiménez, 2018).

Entre los modelos de clasificación se encuentran las máquinas de vectores de soporte (SVM por sus siglas en inglés) realizan tareas de clasificación encontrando el hiperplano que maximiza el margen entre clases, siendo los vectores que definen el hiperplano los de soporte (Bobadilla, 2021).

Otro modelo empleado en SVM es el de árboles de decisión, el cual es un modelo jerárquico de decisiones y sus correspondientes consecuencias, mediante particiones secuenciales de un conjunto de datos que maximice las diferencias de la variable dependiente (Antonio et al., 2019; Rokach, 2008).

El objetivo principal de la presente investigación es el de emplear modelos de aprendizaje supervisado para clasificar la calidad del vino tinto *Vinho verde* a partir de sus caracteres fisicoquímicos principales, con carácter descriptivo y predictivo, y realizar así las inferencias correspondientes a los métodos evaluados.

2. Materiales y Métodos

Se empleó el conjunto de datos que describe la calidad del vino tinto, bajo la denominación de origen *Vinho verde*, en función a la acidez fija, volátil, contenido de ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre densidad, pH, sulfatos y alcohol de éste (Cortez et al., 2009).

En primera instancia, se realizó la limpieza y tratamiento de los datos con la correspondiente eliminación de *outliers* y, posteriormente se emplearon modelos de aprendizaje automático supervisado clasificación mediante el software estadístico *RStudio* en su versión 2022.7.1.554.

Adicionalmente se modificó el conjunto de datos original mediante un condicional detallado en la Tabla 1 con relación al puntaje (0 al 10) de calidad originalmente otorgado.

Puntaje	Calidad
Menor o igual a 5	Mal vino
Mayor a 5	Buen vino

Tabla 1. Parámetro modificado de calidad en base a condición.

Es importante destacar que las 11 variables independientes en estudio se encuentran en distintas escalas (Tabla 2) y por tanto fue necesario realizar una estandarización previa.

#	Parámetro	Unidad
1	Acidez fija	$g_{\text{ácido tartárico}}/dm^3$
2	Acidez volátil	$g_{\text{ácido acético}}/dm^3$
3	Ácido cítrico	g/dm^3
4	Azúcar residual	g/dm^3
5	Cloruros	g_{NaCl}/dm^3
6	Dióxido de azufre libre	mg/dm^3
7	Dióxido de azufre total	mg/dm^3
8	Densidad	g/cm^3
9	pH	-
10	Sulfatos	$g_{\text{sulfato de potasio}}/dm^3$
11	Alcohol	vol %

Tabla 2. Parámetros fisicoquímicos del vino tinto.

Los modelos de aprendizaje supervisado automático empleados para realizar la clasificación del vino en base a la calidad fueron los de vectores de máquina soporte y árboles de decisión detallados a continuación.

2.1. Vectores de máquina soporte

Se evaluó en primer lugar un modelo de kernel **lineal**, implementando un método de muestreo de validación cruzada de 10 capas, para garantizar que los resultados obtenidos sean independientes de la partición realizada para el entrenamiento (70% del total de datos y la prueba (30% restante) del modelo. El costo de penalización aplicada por violar el margen durante el proceso de ajuste se determinó mediante la evaluación de un rango de valores comprendidos entre 0.001 y 5, seleccionando el que otorgue el mejor ajuste.

El segundo modelo evaluado fue el de kernel tipo **radial**, habiendo evaluado un rango de costo por penalización entre 0.001 y 10 y gamma comprendido entre 0.5 y 4 respectivamente.

Se generó la matriz de confusión para los modelos estudiados, detallando en ella los verdaderos positivos y negativos obtenidos así como falsos positivos y negativos respectivamente.

A partir de ésta se obtuvieron la **exactitud**, que indica cuán cerca está el resultado con respecto al valor verdadero, la **sensibilidad** y la **especificidad**, indicando ambas la capacidad del estimador para discriminar los casos positivos de aquellos negativos, actuando como métricas de calidad.

Como método alternativo a las métricas expuestas anteriormente se utilizaron las curvas **ROC** (*Receiver Operating Characteristic*), que consisten en una representación gráfica del rendimiento del clasificador en función a la proporción de verdaderos positivos y de falsos positivos y cuya área es sinónimo de la calidad del clasificador.

2.2. Árboles de decisión

Se evaluó el modelo de Árboles de Decisión del tipo Clasificación, utilizando las 11 variables independientes disponibles de la base de datos.

Se utilizaron las librerías “rpart” para la generación del árbol de decisiones y “rpart.plot” para poder representar de manera gráfica los resultados obtenidos.

Para la partición de datos en el conjunto de entrenamiento y prueba se utilizaron una semilla con tamaño de 1000. La muestra para el conjunto de datos de entrenamiento fue el 70% de los datos. La muestra de prueba utilizada fueron los 30% restantes de los datos.

Una vez hecha la partición de datos para entrenamiento y prueba se procedió a utilizar los mismos con el modelo. Obtenidos los resultados del primer modelo se procedió a realizar la “poda” del mismo adoptando como parámetro **CP** un valor de 0,028 para este procedimiento.

Se generó una matriz de confusión para el modelo en cuestión, detallando en ella los verdaderos positivos y negativos obtenidos así como falsos positivos y negativos respectivamente.

Del mismo modo que con SVM, se utilizaron las métricas de **exactitud**, **sensibilidad** y **especificidad** a la vez que la **curva ROC** con su área correspondiente como métricas para evaluar el desempeño del clasificador obtenido con respecto a los valores verdaderos así como falsos obtenidos en el proceso.

3. Resultados

3.1. Vectores de máquina soporte

El modelo de clasificación **lineal** empleó un costo de penalización de 0.1 y requiere de **621 vectores soporte**, divididos en dos clases, de 310 y 311 vectores respectivamente. En la Tabla 3 se observa que el presente modelo cuenta con 156 verdaderos positivos y 182 verdaderos negativos contra 55 falsos positivos y 66 falsos negativos, lo que se traduce en que el modelo de aprendizaje automático escogido posee una exactitud de 0.74, sensibilidad de 0.74 y especificidad de 0.73 aproximadamente, detallados en la Tabla 4.

	Valores predichos	
Valores reales	156	55
	66	182

Tabla 3. Matriz de confusión para modelo lineal de vector de máquina soporte.

Exactitud	Sensibilidad	Especificidad
0.7363834	0.7393365	0.733871

Tabla 4. Métricas para modelo lineal mediante vector de máquina soporte.

La proporción verdaderos positivos contra falsos positivos se ilustró mediante la curva ROC, la cual se presenta en la Figura 1 donde el área bajo la misma fue de 0.7366037.

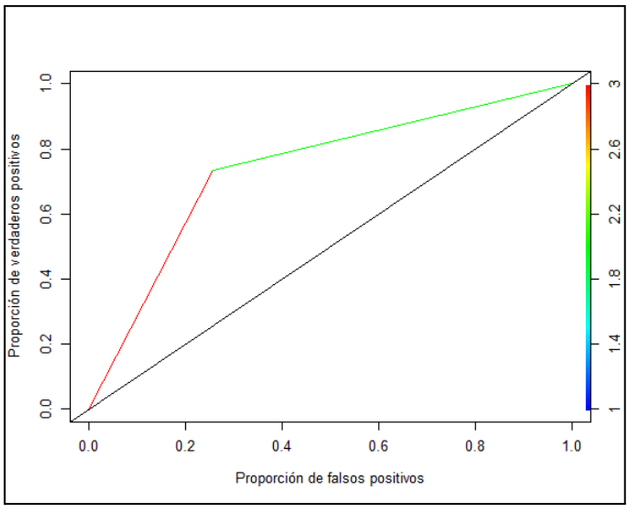

Figura 1. Curva ROC modelo lineal SVM

Al aumentar la dimensión de los datos mediante el kernel **radial** con los mejores parámetros que han sido un costo de penalización de 1 y gamma igual a 0.5, se obtuvieron dos clases, con un total de **906 vectores soporte** divididos en 472 y 434 respectivamente.
El clasificador otorgó 163 verdaderos positivos y 195 verdaderos negativos contra 48 falsos positivos y 53 falsos negativos (Tabla 5), traducidos en una exactitud de 0.78, sensibilidad de 0.77 y especificidad de 0.79 aproximadamente (Tabla 6).

	Valores predichos	
Valores reales	163	48
	53	195

Tabla 5. Matriz de confusión para modelo radial mediante vectores de máquina soporte.

Exactitud	Sensibilidad	Especificidad
0.7777778	0.7677725	0.7862903

Tabla 6. Métricas para modelo radial mediante vector de máquina soporte.

Al comparar los verdaderos positivos contra falsos positivos se obtuvo la curva ROC cuya área es de 0.7794011.

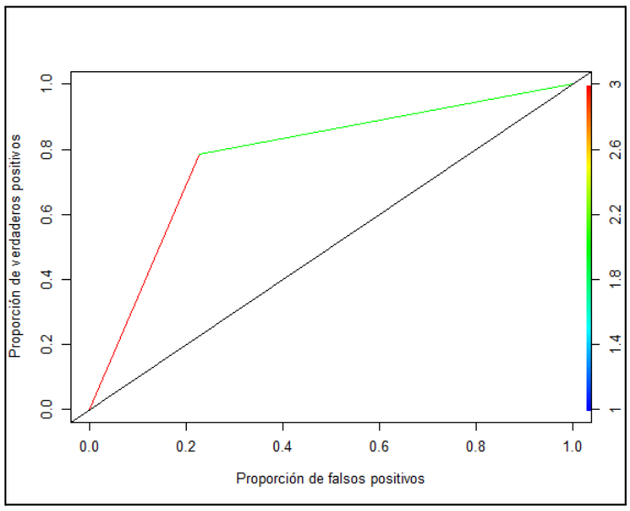

Figura 2. Curva ROC modelo radial SVM

En lo que refiere a vectores de máquina soporte, ambos tipos de kernel han demostrado una buena capacidad para predecir la clasificación de la calidad del vino, siendo el de tipo radial el cual demostró mejor aptitud debido a su mayor exactitud, sensibilidad y especificidad junto un mayor área bajo la curva (5% mayor).

3.2.Árboles de decisión

Para correcto uso del modelo se insertó una columna de categoría que pueden tener las variables son 1 cuando el vino es considerado de buena calidad y 0 cuando el vino es considerado de mala calidad.
Para el entrenamiento del modelo se emplearon los valores discriminados por tipo de conjunto que figuran en la siguiente tabla:

Conjunto	Porcentaje(%)	Cantidad
Entrenamiento	70	910
Prueba	30	449

Tabla 7. Conjunto de entrenamiento para el modelo de Árbol de decisiones.

Con el conjunto de datos de entrenamiento el primer modelo fue preparado. El primer modelo arroja el árbol de decisiones de la Figura 3.
Con este primer modelo se pudieron observar los siguientes puntos:

- Nivel de profundidad del árbol: 6
- La variables predictoras de mayor peso para predecir la calidad en los vinos rojos son, ordenados de mayor a menor:
  - Alcohol - “trans.alcohol”
  - Sulfatos - “trans.sulphates”
  - Total de Dióxido de Azufre - “trans.total.sulfur.dioxide”

- Ácidos volátiles - “trans.volatile.acidity”
- Acidez fija - “trans.fixed.acidity”

Este modelo sugiere que la cantidad de alcohol contenida en una muestra de análisis fisicoquímica es la variable que más incidencia tiene para determinar si este será de buena calidad o no.

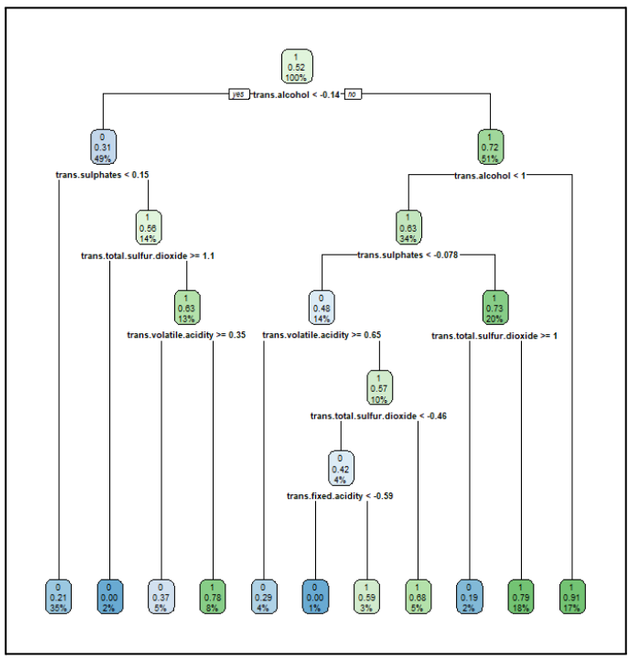

Figura 3. Árbol de decisiones - Primer Modelo

El primer modelo debe ser tratado para evitar el *Overfitting*. Para poder realizar esto se estableció un criterio de poda para el árbol de decisiones. El mismo se estableció utilizando el gráfico representado en la Figura 4.

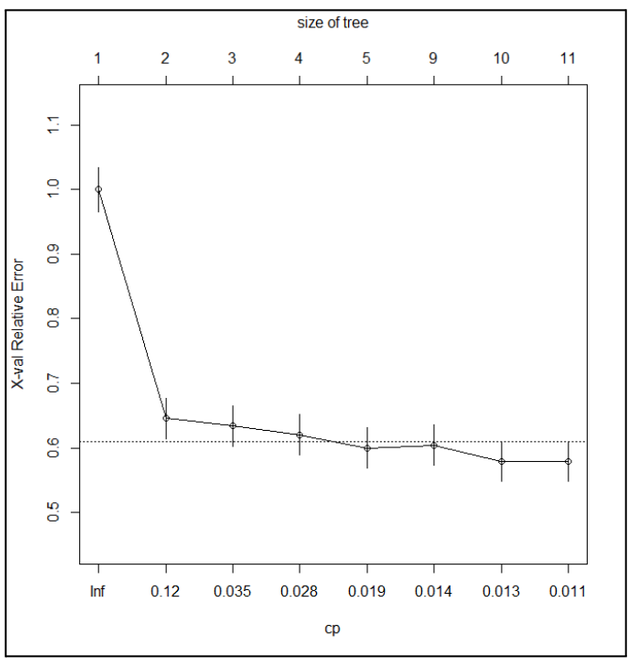

Figura 4. Error relativo en profundidad del árbol.

Se seleccionó como índice CP el valor 0.028 de acuerdo a lo observado en la Figura 4. Este valor se utilizó para poder realizar la poda del árbol y evitar el *Overfitting*.

Como resultado de la poda podemos observar en el nuevo modelo generado en la Figura 5 que el principal predictor para predecir la calidad del vino rojo es que la cantidad de alcohol contenida debe ser mayor a -0,14. Seguido de la cantidad de sulfatos presentes en la muestra que para que el resultado sea un buen vino debe ser mayor a 0,15 y por último la cantidad total de Dióxido de Azufre que debe ser menor a 1,1. Cabe mencionar que se debe tener en cuenta que todos los valores anteriormente mencionados están escalados.

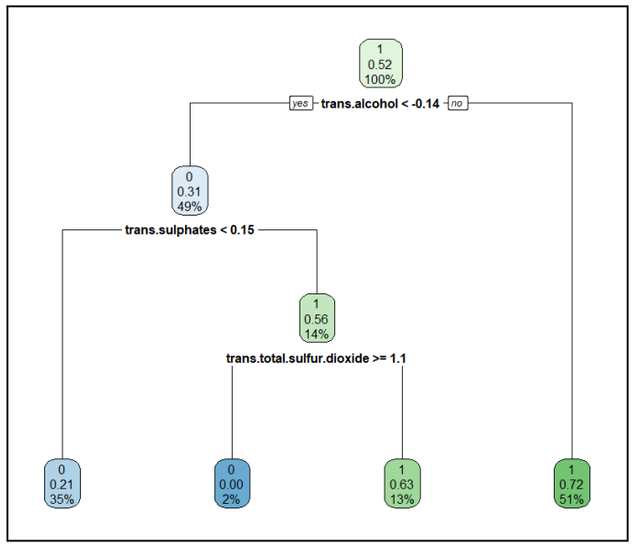

Figura 5. Árbol de decisiones - Segundo Modelo

El segundo modelo que se encuentra representado de manera gráfica en la Figura 5 arrojó los resultados observados en la Tabla 8 donde se puede apreciar que cuenta la detección de 112 Verdaderos Negativos y 166 Verdaderos Positivos contra 61 Falsos Positivos y 50 Falsos Negativos.

	Valores Predichos	
Valores reales	112	62
	50	166

Tabla 8. Matriz de confusión para modelo de Árbol de Decisiones representado en la Figura 5.

Esto se ve representado en la Tabla 9 donde se puede apreciar una exactitud de 0,71, una sensibilidad de 0,64 y una especificidad de 0,76.

Exactitud	**Sensibilidad**	**Especificidad**
0,71	0,64	0,76

Tabla 9. Métricas para modelo de Clasificación por Árbol de decisiones.

La proporción verdaderos positivos contra falsos positivos se ilustró mediante la curva ROC, la cual se presenta en la Figura 6 donde el área bajo la misma fue de 0.7060983.

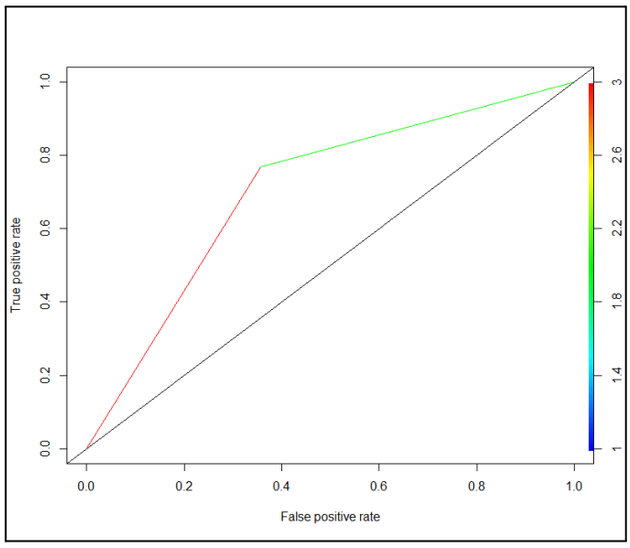

Figura 6. Curva ROC modelo de clasificación para Árbol de Decisiones.

4. Conclusión

Ambos métodos de aprendizaje automático supervisado estudiados bajo el alcance de la presente investigación han demostrado una buena capacidad para clasificación del vino tinto "Vinho verde" en base a las preferencias sensoriales del consumidor, siendo el de vectores de máquinas soporte (SVM) con kernel radial el que mayor exactitud, sensibilidad y especificidad ofrece.

Sin embargo, es importante destacar que el método de árboles de decisión presenta una ventaja adicional con respecto al SVM, puesto que visualiza el modo en que los atributos estudiados particionan la población con respecto a la relevancia de éstos sobre la variable dependiente, omitiendo aquellos que tengan menor relevancia. Gracias a ésto se podrían reducir los parámetros previo al entrenamiento de otra técnica de aprendizaje si se deseara entrenar otro modelo a partir de dicha reducción.

Por lo ya expuesto se considera que los dos métodos analizados bien podrían complementarse entre sí dependiendo del objetivo, sea clasificar el vino a partir de muestreos fisicoquímicos o bien estudiar a fondo la influencia de los parámetros sobre la calidad sensorial en base a la importancia que hayan demostrado ejercer al momento de clasificar la calidad sensorial del consumidor.

5. Referencias

Antonio, V. V. J., Julio, G. A., Francisco, P. R., & Mauricio, B. P. (2019). *Métodos de Data Science aplicados a la Economía y a la Dirección y Administración de Empresas*. UNED. https://books.google.com.py/books?id=rCi6DwAAQBAJ

Bobadilla, J. (2021). *Machine Learning y Deep Learning: Usando Python, Scikit y Keras*. Ediciones de la U. https://books.google.com.py/books?id=iAAyEAAAQBAJ

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. https://doi.org/10.1016/j.dss.2009.05.016

Jiménez, J. L. Á. (2018). *UF2215—Herramientas de los sistemas gestores de bases de datos. Pasarelas y medios de conexión*. Editorial Elearning, S.L. https://books.google.com.py/books?id=9V5WDwAAQBAJ

Pajares, G. (2011). *Aprendizaje automático* (1a ed). Ra-Ma : Ediciones de la U.

Red Wine Color. Revealing the Mysteries», A.L. Waterhouse y J.A. Kennedy, Eds., ACS Symposium Series 886. American Chemical Society: Washington, DC, 2004.

Rokach, L. (2008). *Data Mining with Decision Trees: Theory and Applications*. World Scientific. https://books.google.com.py/books?id=GlKIIR78OxkC

Tuya, J., Ramos Román, I., & Dolado Cosín, J. (2007). *Técnicas cuantitativas para la gestión en la ingeniería del software*. Netbiblo.

Vinho Verde. (s. f.). Recuperado 16 de febrero de 2023, de https://www.vinhoverde.pt/en/denomination-of-controlled-origin