Predicción de la calidad del vino usando Aprendizaje Automático

Por Victor Sanchez, *Universidad Autonoma de Nuevo Leon, Nuevo Leon, Mexico*

23 de marzo de 2023

ABSTRACT

La amplia gama de herramientas pertenecientes al área de Aprendizaje Automático han permitido explorar y analizar distintos sectores de la industria, entre ellos, el sector alimentario. La predicción en la calidad de los productos es una de las facetas que se pueden desempeñar a tráves de algoritmos de clasificación y regresión.

Keywords: Aprendizaje Automático, Calidad, Vino.

1. Introducción

Uno de los factores mas relevantes a la hora de analizar un producto es su calidad. La calidad de un producto es un determinante crítico de la satisfacción del consumidor (Khan and Ahmed, 2012), misma que puede conducir a un escenario de mejores ventas y con ello, una mayor utilidad o rendimiento para las empresas. El mercado del vino es un excelente ejemplo de esta interacción, negocio donde se han diseñado sistemas de trazabilidad para proteger la calidad y el origen del vino, con métodos químicos combinados con quimiometría, los cuales pueden describir los vinos mediante la aplicación de marcadores químicos específicos (Fikselova et al., 2022).

Con el paso de los años, los sistemas de trazabilidad se han ido modificando en busqueda de una mejora puntual en la medición de la calidad de estos productos, lo cual ha resultado beneficioso para el mercado, el cuál ha ido diversificando sus técnicas y herramientas a fin de comparar resultados. Con la llegada del aprendizaje automatico se dió paso a un nuevo enfoque para resolver problemas (Telikani et al., 2022), expandiendo así la amplia gama de alternativas de solución sobre un tema en particular. En este estudio se presenta un método alternativo para predecir la calidad de los vinos, mediante técnicas de Aprendizaje Automático, para ello, se contemplarán diferentes componentes químicos que posiblemente definan la calidad de un vino, tales como: la concentración de ácidos, cantidad de azúcar, sal, sulfato y dióxido de azufre, densidad de agua, porcentaje de alcohol, entre otros factores más.

2. Algoritmos

Hablar de aprendizaje, es hablar de un concepto con dominio muy amplio. En consecuencia, el campo del aprendizaje automático se ha ramificado en varios subcampos que se ocupan de diferentes tipos de tareas de aprendizaje (Shai and Shai, 2014). Una de estas clasificaciones se suele abordar en base a la interacción que existe entre el .ªlumnoz su entorno, ya sea de manera supervisada o no supervisada; como los algoritmos que a continuación se describen.

2.1. DBSCAN

DBSCAN por sus siglas en inglés, *Density-based spatial clustering of applications with noise*, es un algoritmo de agrupamiento de aprendizaje no supervisado, que como su nombre lo indica forma clusters con una alta densidad de puntos, para ello, define la densidad de los clusters en función de los siguientes dos parámetros:

(i) Radio de la vecindad, donde la vecindad esta dada por:

$$N_\epsilon(p) : q|d(p,q) < \epsilon; p, q \in S$$

(ii) Cota mínima de puntos presentes en la vecindad

$$pmin*$$

Bajo estos dos parametros, el algoritmo categoriza las observaciones en 3 clases:

(i) **Puntos centrales**:

$$p \in S : |N_\epsilon(p)| \geqslant pmin*$$

(ii) **Puntos directamente alcanzables por densidad**:

$$p \in S : p \in N_\epsilon(q) \wedge |N_\epsilon(q)| \geqslant pmin*$$

(iii) **Puntos alcanzables por densidad**: Si existe una cadena de puntos de tamaño n tal que el punto ubicado después de la posición i-ésima es directamente alcanzable por densidad desde el punto de la posición i-ésima, donde i es menor que n.

El algoritmo trabaja de la siguiente manera: partiendo de un punto arbitrario y analizando la categoría a la que pertenece, de ser un punto central se creará un cluster, de lo contrario, se pasará al siguiente punto, repitiendo el proceso anterior; aunque es a partir de esta segunda iteración que las clasificaciones de directamente alcanzables y alcanzables por densidad pueden ser aplicables. El algoritmo termina una vez se hayan procesado todos los puntos.

Los puntos que no son incluidos en los clusters son considerados outliers. El tiempo de complejidad de DBSCAN es $O(n^2)$, donde n es el número de puntos (Zeinab et al., 2021).

2.2. *PCA*

PCA por sus siglas en inglés, *Principal Component Analysis*, es una técnica de aprendizaje no supervisado, que reduce la dimensionalidad de los datos mientras conserva en la medida de lo posible la covarianza de los datos.

Para lograrlo, parte de la dimensionalidad de los datos, supongamos p, y busca iterativamente p vectores unitarios y ortogonales entre sí tales que maximicen la varianza del producto del respectivo vector con el conjunto de datos. Es decir:

$$\begin{aligned} \text{máx}\, var(c_k x) : \forall c_k \in [1, p] \\ c_k' c_k = 1 : \forall c_k \in [1, p] \\ c_q' c_k = 0 : \forall k \neq q \end{aligned}$$

Como el modelo anterior, PCA cuenta támbien con un criterio de selección, en este caso la dimensionalidad a elegir. Siendo uno de los primeros enfoques, la revisión de los eigenvalores y su contribución en la varianza de los datos, hasta obtener un porcentaje deseado de varianza explicada (Cattell, 1966).

2.3. *Regresión*

La regresión es una técnica supervisada que permite estimar el valor de una variable a partir de otras, mediante la definición de una función de estimación. Cuando la dimensión de los datos es mayor a 2 y además se supone que existe una relación lineal entre la variable independiente y las variables dependientes, se dice que la regresión es **lineal múltiple**. Cuya formula está dada por la siguiente ecuación:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$$

Donde Y es la variable que se desea predecir, X_k son las variables predictoras, y ϵ representa el error aleatorio del modelo.

Otra forma de abordar la regresión es a tráves del algoritmo ***Random Forest*** el cuál genera árboles de decisión de forma aleatoria con el objetivo de combinarlos y obtener una predicción más precisa y estable, siendo su principal clave la aleatoriedad. Un bosque aleatorio se compone de L árboles de decisión, cada uno de ellos creados a partir de una muestra aleatoria de los datos de tamaño n, en donde cada árbol se va expandiendo hasta alcanzar una cantidad mínima de nodos mediante un proceso recursivo en el que se seleccionan p variables de P disponibles y se forman nuevos nodos mediante un criterio de ramificación. En una regresión por bosque aleatorio la forma más habitual de obtener una predicción es tomar el promedio de los resultados generados de todos los árboles pertenecientes al bosque aleatorio. Para un punto particular:

$$y_L(x) = \frac{1}{L} \sum_{l=1}^{L} A_l(x)$$

Donde $y_L(x)$ representa la predicción en x bajo un bosque aleatorio de L árboles, y $A_l(x)$ es el resultado obtenido en el l-ésimo árbol aleatorio para x.

2.4. *Clasificación*

Por su parte, la clasificación es un método que estima categorías o clases para un conjunto de variables, una técnica supervisada que es útil cuando lo que se desea es etiquetar datos. Mediante ***Random Forest*** es posible también clasificar datos, combinando los resultados mediante un esquema de votación, siendo los más usuales, votaciones duras y suaves.

En una votación dura se elige aquella clase o etiqueta con mayor frecuencia dentro de todos los clasificadores individuales, en nuestro caso árboles.

$$\hat{y(x)} = \arg\text{máx}\left(F_{C_1}(x), ..., F_{C_W}(x)\right)$$

Mientras que en una votación suave para cada clasificador se estima la probabilidad de cada clase, se promedia entre modelos y se toma la clase con mayor probabilidad.

$$\hat{y(x)} = \arg\text{máx}\, \frac{1}{W} \sum_{w=1}^{L} f_{C_w}(x)$$

Donde W es la cantidad de clases que existen, mientras que F y f representan la distribución de frecuencia y probabilidad de las clases respectivamente.

3. Muestra

Los datos recolectados pertenecen a dos muestras separadas de vinos provenientes del norte de Portugal, la separación de las muestras se ha realizado en base al color de la bebida, esto es sumamente relevante, teniendo en cuenta que ya anteriormente, se ha intentado correlacionar las propiedades cromáticas de los vinos con la composición, el origen, la crianza y la evaluación sensorial (Heredia and Guzman, 1993). En total se tiene una muestra de 4,898 y 1,599 vinos blancos y tintos respectivamente, cada uno de ellos con 12 variables de medición, incluyendo su calidad.

Fig. 1. Clústers *DBSCAN* 2 dimensiones

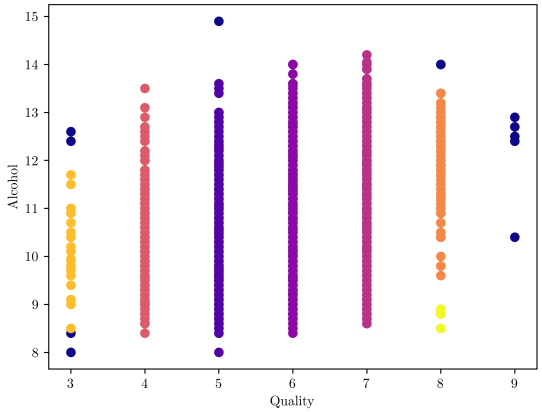
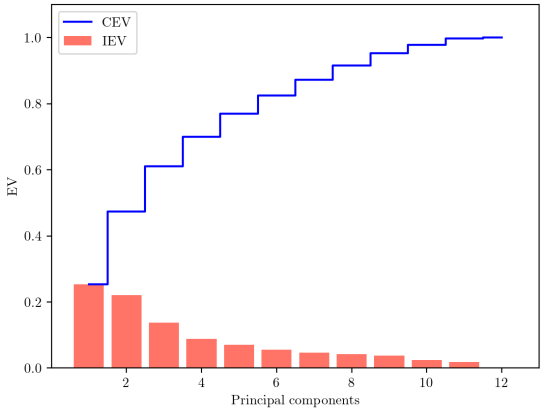


Fig. 2. Ánalisis de varianza *PCA*



4. Resultados

Los algoritmos que previamente se presentaron se han aplicado para la muestra descrita de vinos, con el objetivo de analizar posteriormente sus resultados y definir un mejor criterio para la predición de la calidad del vino.

4.1. DBSCAN

Derivado del desempeño obtenido en el proceso de selección de características, se ha probado el algoritmo DBSCAN para las variables: alcohol y calidad, cuya correlación fue la más alta entre el grupo total de variables.

Los resultados permiten resaltar las ventajas del algoritmo (ver Fig. 1), no es necesario definir el número de clústers a generar, así como la forma de los mismos. Adicionalmente, por definición, es capaz de identificar outliers del conjunto de datos, mismos que pueden ser eliminados para la posterior aplicación de otros modelos.

Sin embargo, al ser una técnica basada en densidad, la densidad de los datos es la que define el desempeño de la misma, de manera que la selección de parámetros juega un papel importante en los resultados obtenidos. Sharma and Sharma (2017) consideraron utilizar la información de los K-vecinos más cercanos en el DBSCAN con la finalidad de lograr una técnica de agrupamiento sin parámetros.

Siguiendo esa misma motivación, es posible reducir el número de escenarios al aplicar la técnica a la dimensionalidad completa de los datos.

	Épsilon				
	2.0	**2.5**	**3.0**	**3.5**	**4.0**
Clústers	4	2	2	1	1
Outliers	245	90	40	19	6

Teniendo como resultado un mayor número de clústers a medida que el radio épsilon seleccionado tienda a ser más pequeño, así como un mayor número de valores atípicos.

4.2. PCA - DBSCAN

Derivado del fin práctico que tiene el Análisis de Componentes Principales, se llevó acabo su aplicación para posteriormente analizar los clústers generados mediante el algoritmo DBSCAN. Bajo una elección de 6 componentes, se han generado únicamente 2 clústers y 138 posibles outliers. La elección de los 6 componentes proviene del porcentaje de variación explicada obtenida a tráves de los mismos (ver Fig. 2).

El épsilon determinado para el DBSCAN de 6 componentes ha sido de 1.5, radio menor a comparación de los seleccionados en las pruebas del DBSCAN de dimensionalidad completa.

4.3. Regresión

Dando seguimiento a la identificación de outliers por DBSCAN, se probaron 3 regresiones distintas: dos regresiones lineales múltiples (MLR), considerando en la segunda de ellas la eliminación de outliers por DBSCAN, y una regresión por Random Forest (RFR). Destacandosé una limitación de estimación en puntuaciones extremas sobre los vinos para los tres modelos, siendo menos restrictiva en la regresión por bosque aleatorio. En la siguiente tabla se presentan los resultados obtenidos por los modelos en una muestra de cinco elementos.

	MLR	**MLR***	**RFR**
4	5.40	5.75	5.43
5	5.99	6.16	5.81
6	5.36	6.12	5.86
7	6.04	6.17	6.49
8	5.95	5.37	7.12

Fig. 3. Importancia de las características

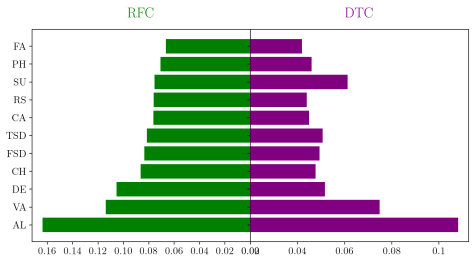
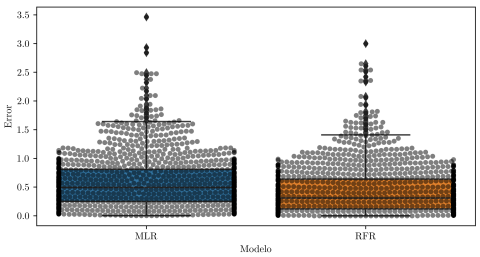


Fig. 4. Errores absolutos de los modelos



4.4. ***Clasificación***

Las puntuaciones en productos si bien suelen dar un mayor detalle en la comparativa de artículos, es la simplicidad de análisis una de las características que buscan los consumidores al momento de comparar dos o más productos. Una manera de simplificar la calidad del vino es a tráves de dos categorías: bueno y malo. En nuestro caso, se considera que un vino es de buena calidad cuando su nota es superior a cinco.

Para la estimación de la categoría de los vinos, se han propuesto dos técnicas de clasificación: Bosque Aleatorio y Árboles de decisión. De manera simultánea se observó como en ambas técnicas la característica más relevante es el alcohol, mientras que el resto de variables mantiene un nivel de importancia casi similar (ver Fig. 3). Importancia de características es una técnica que asigna puntuaciones a las características de los datos en función de su utilidad en los modelos.

5. Desempeño

Los resultados obtenidos por los modelos aplicados pueden ser evaluados a tráves de distintas métricas de desempeño. En el caso de las regresiones realizadas, se han analizado 4 indicadores, los cuales brindan información sobre el error de los modelos.

	MLR	**MLR***	**RFR**
MAE	0.582	0.581	0.434
MSE	0.548	0.574	0.370
RMSE	0.740	0.758	0.608
R2	0.294	0.299	0.524

Respecto a los indicadores MAE y MSE, ambos analizan la distancia de los datos pronosticados contra los observados, sin embargo el MSE penaliza la aparición de outliers en los datos. Por su parte, el indicador R cuadrada describe el porcentaje de variación de la variable dependiente que es explicada por las variables predictoras. De forma aislada no deberían considerarse como un indicador ideal u óptimo para analizar el desempeño de las regresiones realizadas. Se observa que el modelo por bosque aleatorio ha sido más preciso en sus estimaciones a comparación de las regresiones lineales aplicadas, sin ningún beneficio existente sobre el uso del DBSCAN.

Referente a los modelos de clasificación se han considerado las siguientes métricas de desempeño:

(i) Exactitud: Porcentaje de elementos clasificados correctamente. Se recomienza analizar bajo cuidado cuando las clases son desbalanceadas.

(ii) Precisión: Porcentaje de observaciones clasificadas como positivas que son efectivamente positivas.

(iii) Exhaustividad: Porcentaje de observaciones efectivamente positivas, que fueron clasificadas como positivas.

(iv) Valor F: Estadístico que combina las métricas de Precisión y Exhaustividad.

(v) ABC: Área bajo la curva del funcionamiento del receptor. Un valor de uno indica un modelo que perfectamente clasifica tanto elementos como positivos y negativos.

En líneas generales, el algoritmo de bosque aleatorio obtiene un mejor desempeño frente al modelo de árboles de decisión. Se puede observar una área bajo la curva de 0.9, lo que representa una probabilidad de 90 % de que el modelo distinga entre clase negativa y clase positiva.

	RFC	**DTC**
Exactitud	0.831	0.780
Precisión	0.852	0.824
Exhaustividad	0.890	0.834
Valor F	0.871	0.829
ABC	0.902	0.758

6. Experimentación

Como un alcance de la comparación de modelos, se ha utilizado un diseño factorial de un solo factor sobre los modelos de regresión estudiados, particularmente, de los errores absolutos obtenidos por cada modelo. Se destaca una mayor dispersión en el modelo de regresión lineal múltiple (ver Fig. 4), existiendo un desplazamiento de entre 0.1 % y 0.2 % con respecto a los errores subyacentes del modelo de bosque aleatorio.

6.1. ***Revisión Estadística***

Con base al diseño factorial considerado, se han validado las características de cada nivel con el fin de utilizar una prueba

estadística correspondiente que nos permita comparar los modelos. Se ha revisado si la distribución de errores sigue una distribución normal para cada uno de los modelos, mediante una prueba de Shapiro-Wilk, la cuál considera una muestra aleatoria de n elementos y tiene como estadístico

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x_i})^2}$$

Así como una prueba de correlación entre modelos, mediante el estadístico

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

En donde r representa la correlación de Pearson entre ambos conjuntos. Obteniendo como resultado que los errores de los modelos no siguen una distribución normal y existe una dependencia lineal entre ellos. Derivado de las características de los errores, se ha utilizado una prueba de rangos con signo de Wilcoxon, la cual analiza si la mediana de las diferencias de cada par de datos es cero, es decir, provienen de la misma distribución o no. Para ello considera el siguiente estadístico:

$$W^+ = \sum_{z_i > 0} R_i$$

Donde z_i representa la diferencia de errores y R_i el rango asignado a la variable z_i. Al aplicar la prueba se falla rechazando la hipótesis nula, no existiendo suficiente evidencia estadística de que los modelos sigan distintas distribuciones, y con ello, errores muy similares, aunque como se destacó anteriormente los errores en la regresión lineal son ligeramente más altos a comparación del modelo por bosque aleatorio.

Shapiro-Wilk		
	W	**p-valor**
MLR	0.89	0.00*
RFR	0.83	0.00*

Correlación		
	t	**p-valor**
Modelos	0.66	1.91

Wilcoxon		
	W+	**p-valor**
Modelos	227,261	2.70

7. Conclusiones

Hablar de predicción de datos en base a información historica, es hablar también de la calidad de dicha información. Durante los primeros compases del ánalisis realizado, se trabajo mediante una técnica que busca agrupar datos e identificar valores atípicos, siendo esto último relevante posterior al objetivo de predicción que se persiguió. Sin embargo, el algoritmo DBSCAN no es del todo fino y su gran ventaja es también parte de su debilidad, la elección de su radio y la gran dimensionalidad de este conjunto ha generado una variabilidad en los resultados obtenidos, aún así, es destacable que la identificación de outliers se haya mantenido por debajo del $3\,\%$ del total de los datos mediante las distintas instancias tomadas. Posteriormente, al utilizar regresiones, el supuesto de linealidad no fue adecauado, nuevamente el DBSCAN aunque mejora ligeramente un par de indicadores a comparación de la regresión original, no logra superar los resultados obtenidos por *Random Forest Regressor*, el cuál demuestra su capacidad de manejar grandes dimensionalidades, siendo su única desventaja el ánalisis de componentes individuales dentro del modelo. El alcance de árboles de decisión se extendió a modelos de clasificación, en los cuales se destacó la relevancia del porcentaje de alcohol como indicador de la calidad de un vino. Finalmente, el buen desempeño obtenido por el modelo de bosque aleatorio aplicado tanto en regresión como clasificación, lo coloca como un excelente candidato en la predicción de vinos, respaldado además por el diseño de experimentos realizado.

References

Khan, L M. and van Ahmed, R. 2012. A Comparative Study of Consumer Perception of Product Quality: Chinese versus Non-Chinese Products. *Journal of PJETS* **139**, 118–143.

Fikselova, M., Snirc, M., Rybnikar, S., Mezey, J., Jakabova, S., Zelenakova, L., Vlcko, T. and Balaska, M. 2022. The Wine Quality Description of Different Origin Evaluated By Modern Chemometric Approach. *Journal of Microbiology, Biotechnology and Food Sciences* **12**, 1–5.

Telikani, A., Tahmassebi, A., Banzhaf, W. and Gandomi, A. 2022. Evolutionary Machine Learning: A Survey. *ACM Computing Surveys* **54**, 1–35.

Shai, S S. and Shai, B D. 2014. Understanding Machine Learning: From Theory to Algorithms. *Cambridge University Press* **1**, 22.

Zeinab, F., Alireza, B. and Midia, R. 2021. Determining the Parameters of DBSCAN Automatically Using the Multi-Objective Genetic Algorithm. *Journal Of Information Science And Engineering* **37**, 157–183.

Nahid, G., Hamid, S. and Nooshin, H. 2021. K-DBSCAN: An improved DBSCAN algorithm for big data. *The Journal of Supercomputing*.

Jolliffe, I T. 1986. Principal Component Analysis. *Springer*

Cattell, R B. 1966. The scree test for the number of factors. Multivariate behavioral research. *University of Illinois* **1**, 245–276.

Bingham, N H. and Fry, J M. 2017. Regression: Linear Models in Statistics. *Springer*.

Zhang, C. and Ma, Y. 2014. Ensemble Machine Learning: Methods and Applications. *Springer*.

Yan, J., Han, S. and Chen, Y. 2023. Prediction of Traffic Accident Severity Based on Random Forest. *Journal of Advanced Transportation*.

Zakaria, E. M., Niroop, S., Prakash, R. and Shrirang, A. 2022. Random Forest Regressor-Based Approach for Detecting Fault Location and Duration in Power Systems. *Sensors*.

Heredia, F J. and Guzman, M. 1993. The Color Of Wine: A Historical Perspective. *University of Sevilla*.

Sharma, A. and Sharma, A. 2017. KNN-DBSCAN: Using k-nearest neighbor information for parameter-free density based clustering. *Institute of Electrical and Electronics Engineers*.

Torres, J., Rodríguez, L., Méndez, L. and Pérez, I. 2020. Diseño de experimentos para optimizar resistencia e índices de capacidad de un fusible. *Cultura Científica y Tecnológica* **17**, 1–9.