Predicción de la calidad del vino usando Machine Learning

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

4 de marzo de 2023

ABSTRACT

La amplia gama de herramientas pertenecientes al área de Machine Learning 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: Machine Learning, 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 *Machine Learning*, 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 clustering 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 (*DBSCAN*), 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*

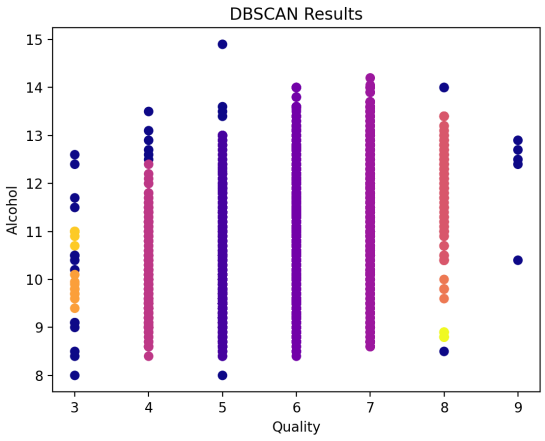
La regresión es una técnica supervisada que permite estimar el valor de una variable a partir de otras mediante la definición de una función de estimación. Cuando la dimensión de los datos es mayor a 2 se dice que la regresión es **múltiple**, si además se supone que existe una relación lineal entre la variable independiente y las variables dependientes, la regresión es de orden lineal. 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, y X_k son las variables predictoras, mientras que 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, la clave está en su aleatoriedad.

Fig. 1. Clústers *DBSCAN* 2 dimensiones



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.

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.

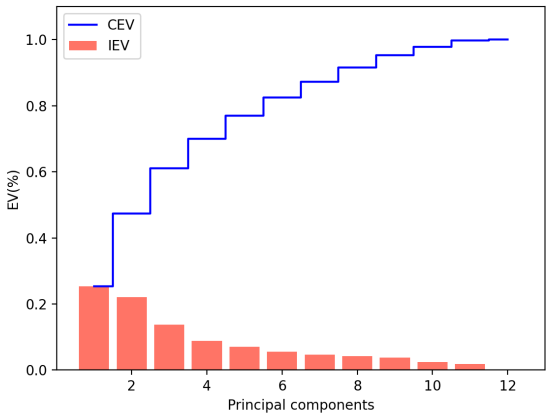
4.1. *DBSCAN*

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. 2), 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)

Fig. 2. Ánalisis de Varianza *PCA*



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.

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). Las instancias fueron medidas a tráves de distintos indicadores de precisión de las regresiones.

	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

Se han tomado en cuenta estos 4 indicadores derivado de la información que aportan del modelo. En el caso del 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.

5. 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 las regresiones, el camino de suponer linealidad en los datos no fue el correcto, 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 poder manejar grandes dimensionalidades, aunque el control sobre el mismo es corto, siendo este un punto a favor o en contra dependiendo del entorno.

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*.

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*.