APRENDIZAJE AUTOMÁTICO DE MÁQUINA
MACC 2021-2
Proyecto

Evaluación del grado de calidad de la leche bovina en función a sus características físico químicas

José Alberto Murcia Navas
joseal.murcia@urosario.edu.co

Profesor:
Edwin Santiago Alférez Baquero
edwin.alferez@urosario.edu.co

1. INTRODUCCIÓN

La leche es el único material producido por la naturaleza para funcionar exclusivamente como fuente de alimento, ya que constituye una fuente nutritiva. La confirmación de esta imagen nutritiva está en el uso extensivo que tiene la leche y sus derivados, como parte de la dieta diaria en los países altamente desarrollados. A consecuencia de esto, estas sociedades gozan casi de una completa carencia de enfermedades nutricionales en la población infantil y adultos jóvenes. En contraste, una elevada proporción de los bebés y niños en los países en desarrollo, donde el suministro de leche es mínimo o nulo, sufren deficiencias nutricionales.

Es por ello, que el desafío para quienes trabajan en el sector lechero no sólo es producir mayor cantidad de leche, sino también, de alta calidad higiénica, y para ello deben contemplarse aspectos fundamentales, como lo son, la higiene microbiológica, química y estética. Tres aspectos, que unidos, pueden contribuir favorablemente a la mejora del sector lechero, con el beneficio consecuente en el desarrollo físico e intelectual de las generaciones venideras

2. DESARROLLO

Para nuestro proyectos se ha hecho uso de un dataset disponible en Kaggle, llamado *Milk Grading*. Este dataset se compone de siete variables independientes tales como pH, temperatura, sabor, olor, porcentaje de grasa, turbidez y color. Generalmente, el grado de calidad de la leche depende de estos parámetros, y juegan un rol vital en análisis predictivos acerca del grado de calidad.

2.1. Análisis descriptivo

En primera instancia realizamos un análisis exploratorio de las variables contempladas en el dataset. De aquí podemos resumir que contamos con temperaturas que van desde 34° hasta 90°

con una media o promedio de 44° en la evaluación de la calidad de la leche. La mayor densidad de datos para la variable temperatura la tenemos en 38º (q1), y le sigue en 45º (q3).

Así mismo, se cuentan con valores ácidos (pH<7.0) y básicos (pH>7.0) para la leche. Y tenemos una mayor densidad de datos para el rango comprendido entre 6.5 y 6.8 (quartiles 1 y 3).

Para las variables sabor, olor, grasa y turbidez contamos con valores 0.0 y 1.0 para hacer referencia a evaluaciones de tipo cualitativo.

2.2. Preprocesamiento de las características

Teniendo en cuenta que la mayoría de algoritmos de aprendizaje automático funcionan mucho mejor si las características están en la misma escala (a excepción de los árboles de decisión), realizamos la normalización y estandarización de nuestros datos.

Por otra parte, el dataset presenta valores continuos para la variable objetivo (Grade), por lo cual se hace necesario realizar un preprocesamiento de éstos datos para poder continuar con el uso de los modelos de regresión logística y KNN. Se llevo a cabo esto mediante el uso de *LabelEncoder* con el fin de hacer uso de los números enteros 0, 1 y 2 para definir la calidad de la leche como mala, regular y buena, respectivamente.

2.3. Técnicas de reducción de la dimensión

Con el objetivo de determinar las variables de mayor importancia para nuestro modelo de clasificación, hicimos uso de las siguientes dos técnicas de reducción de la dimensión:

- feature_selection
- PCA

De aquí pudimos reducir a dos componentes principales (PC) las variables para los modelos de clasificación, las cuales fueron pH y temperatura. El parámetro pH representa por sí solo aproximadamente el 80% de la varianza. Además, podemos ver que los dos primeros componentes principales (pH y temperatura) combinados explican casi el 100% de la varianza en el conjunto de datos.

3. Modelos de Clasificación

Teniendo en cuenta que nuestro proyecto corresponde a un problema de clasificación, hicimos uso de varios modelos y los resultados obtenidos en la validación cruzada para el conjunto de datos original fueron:

Model	Accuracy, cross_val_score
Baggin	0.904
DecisionTreeClassifier	0.847
KNeighborsClassifier	0.991
Perceptron	0.500
RandomForestClassifier	0.996
GradientBoostingClassifier	0.994
XGBClassifier	0.995
XGBClassifier (using Gridsearch)	0.996

A partir de los resultados de lo diferente modelos de clasificación, se puede evidenciar que en general todos clasifican muy bien el conjunto de datos. Sin embargo, los mejores modelos fueron Random Forest y XGBoost con una precisión de 0.996, cercano a los resultados de los modelos Boosting y KNN.

4. Clustering

Para nuestro caso, usamos la técnica K-means y evaluamos los resultados del agrupamiento mediante la métrica de silueta.

En resumen, las siluetas para K-means tienen visiblemente diferentes longitudes y anchuras, lo cual demuestra un agrupamiento relativamente malo. En su lugar, el modelo de clustering jerárgico muestra un resultado aún menos aceptable de acuerdo a lo calculado con la métrica de la silueta.

5. Bibliografía

[1] Raschka Sebastian y Mirjalili, V., "Python Machine Learning". Segunda edición. Marcombo.2019