Universidad Sergio Arboleda

Taller 5-Método Clasificación Vinos

DIOGENES BARRETO ALVAREZ

Métodos de Análisis de Datos

Docente
Luz Estela Gomez

December 13, 2020

Ejercicio Se tiene una una data de vinos los cuales tiene una clasificación de calidad de 1-8 (Variable dependiente) . Vamos a proceder a analizar esta data y clasificar los vinos de acuerdo a su clasificación de calidad aplicando el algoritmo de clasificación K-Nearest Neighbors y probamos el modelo finalmente. La información realizamos el proceso de minería d edatos:

-Problema
-Identificación de Variable.
-Orden y limpieza.
-Filtro.
-Modelo de Clasificación.
-Análisis.
-Implementación.

En el análisis de las variables predicativas notamos que ninguna tiene un comportamiento homogéneo y estas posiblemente afectan las respuesta de la variable dependiente, en este caso calidad(Quality).La decisión es usar todas las variables en el análisis y no descartar ninguna. La siguen te figura muestra el análisis del perfil para dos variables tomadas de manera aleatoria, con posible distribución normal.Estas Variables son numéricas, posiblemente continuas. Sin embargo se anexa el reporte completo de estas como soporte.

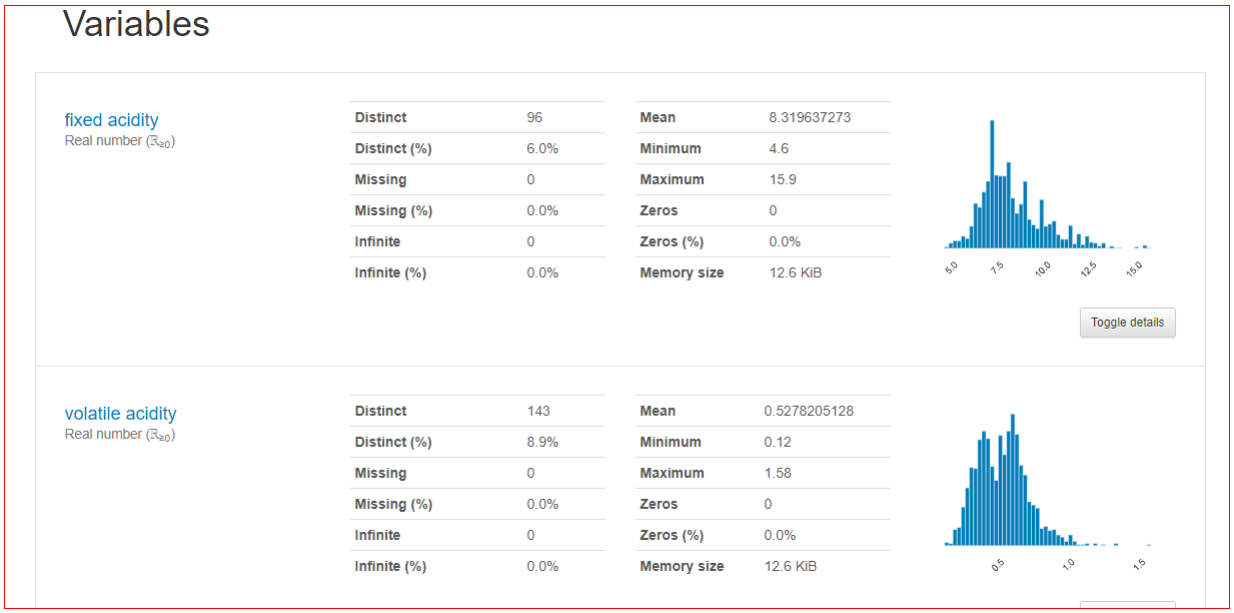


Figure 1: Comportamiento Variable Predicativas

En el análisis de las variables, se muestra de no hay datos faltantes, todas tiene información.

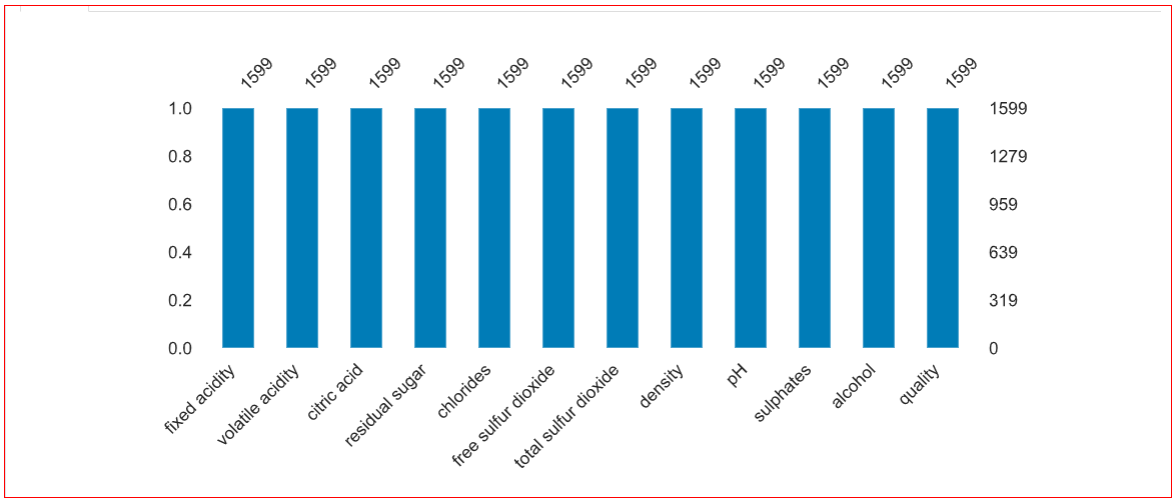


Figure 2: datos completos

En las revisión de la data , se identifico que no existen variables predicativas con valores nulos. esto importante para la identificación en las variables que requieren ser considerados o eliminación de esa posible fila.

```
In [218]: df.isnull().any()
Out[218]:
fixed acidity           False
volatile acidity        False
citric acid             False
residual sugar          False
chlorides               False
free sulfur dioxide     False
total sulfur dioxide    False
density                 False
pH                      False
sulphates               False
alcohol                 False
quality                 False
dtype: bool
```

Figure 3: Validación de Variables sin valores nulos

se hizo las revisión de la estadística descriptiva de las variables. Se pudo observar que algunas de estas presentan valores atípicos alejados de la media. La tabla siguientes muestra el análisis y validamos esta con el gráfico correspondiente a la variable "total sulfur dioxide".

Índice	fixed acidity	olatile acidit	citric acid	esidual suga	chlorides	e sulfur diox	al sulfur diox	density	pH	sulphates	alcohol	quality
count	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599	1599
mean	8.31964	0.527821	0.270976	2.53881	0.0874665	15.8749	46.4678	0.996747	3.31111	0.658149	10.423	5.63602
std	1.7411	0.17906	0.194801	1.40993	0.0470653	10.4602	32.8953	0.00188733	0.154386	0.169507	1.06567	0.807569
min	4.6	0.12	0	0.9	0.012	1	6	0.99007	2.74	0.33	8.4	3
25%	7.1	0.39	0.09	1.9	0.07	7	22	0.9956	3.21	0.55	9.5	5
50%	7.9	0.52	0.26	2.2	0.079	14	38	0.99675	3.31	0.62	10.2	6
75%	9.2	0.64	0.42	2.6	0.09	21	62	0.997835	3.4	0.73	11.1	6
max	15.9	1.58	1	15.5	0.611	72	289	1.00369	4.01	2	14.9	8

Figure 4: Estadística descriptivas de las Variables

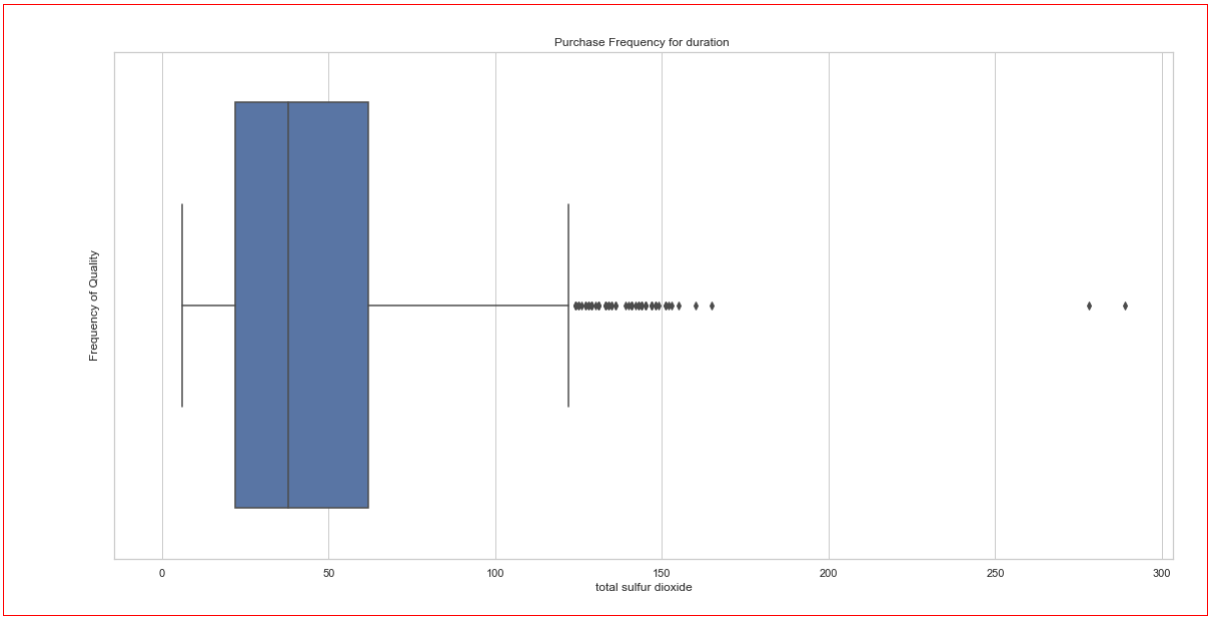


Figure 5: Variable total sulfur dioxide

También realizamos el mapa de calentamiento o matriz de correlación de las variables descriptivas. Se notan parámetros de correlación bajos para algunas Variables.

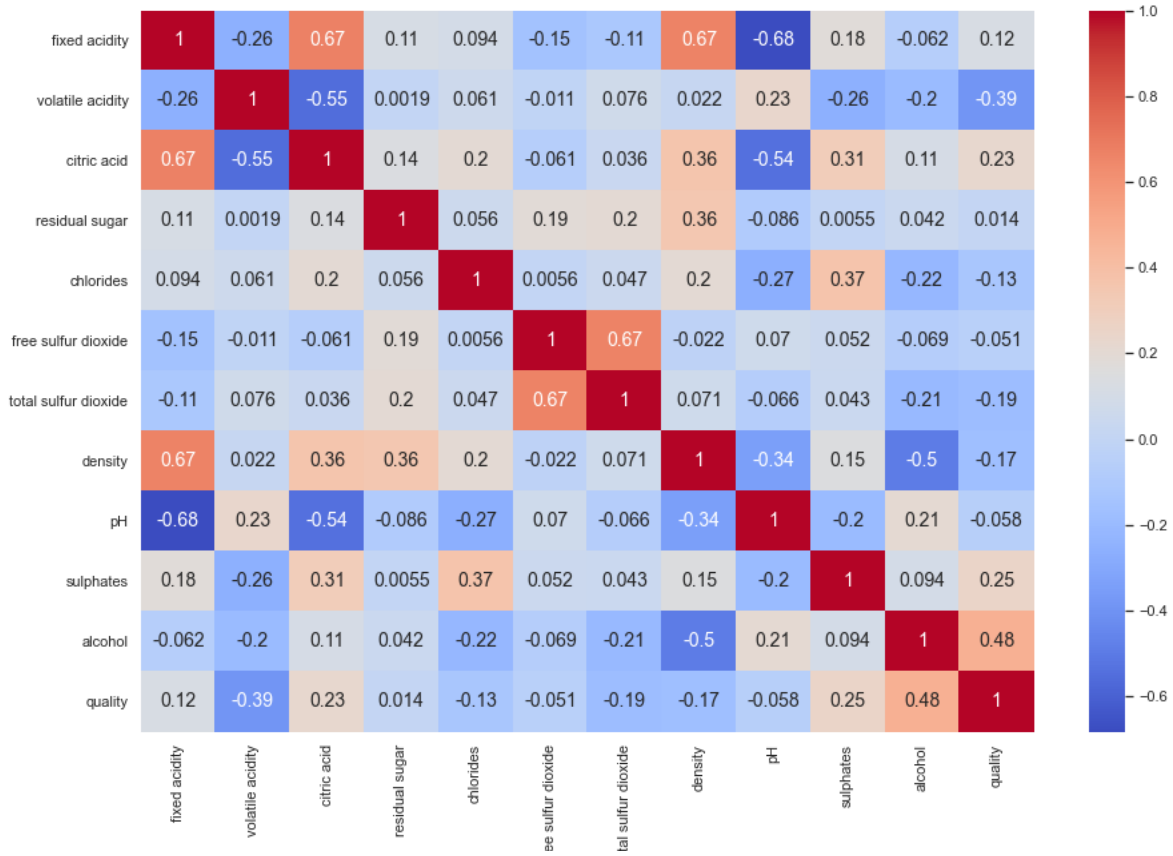


Figure 6: Mapa de calentamiento de Correlación de las variables

Con base en el análisis anterior , procedimos a realizar una eliminando los valores atípicos mediante el uso de $Z - score$ La mayoría de las veces es importante eliminar los valores atípicos, ya que lo más probable es que afecten al rendimiento de los modelos de aprendizaje de máquinas.Sin embargo , es necesario ser prudente en la eliminación de estos porque probablemente hay algo más que está pasando y necesita ser inspeccionado más a fondo. Para encontrar y eliminar los valores atípicos, usé el $z-score$. se define mediante la ecuación:

$$\frac{X - \mu}{\sigma}$$

Su interpretación es tomar el punto de datos, restar la media de la población y dividirla por la desviación estándar. Representa cuántas desviaciones estándar de un punto de datos se alejan de la media. Los puntos de datos que están demasiado lejos de la media se consideran atípicos. En la mayoría de los casos el umbral para la detección de valores atípicos es o bien $z - score > 3$ o $z - score < -3$. Del Análisis de datos atípicos inicialmente teníamos en el dataframe de 1599 datos y al realizar la depuración quedaron 1451 para las 12 variables (11 predicativas y 1 dependiente).A continuación se muestra el mapa de calentamiento de las nuevas variables.

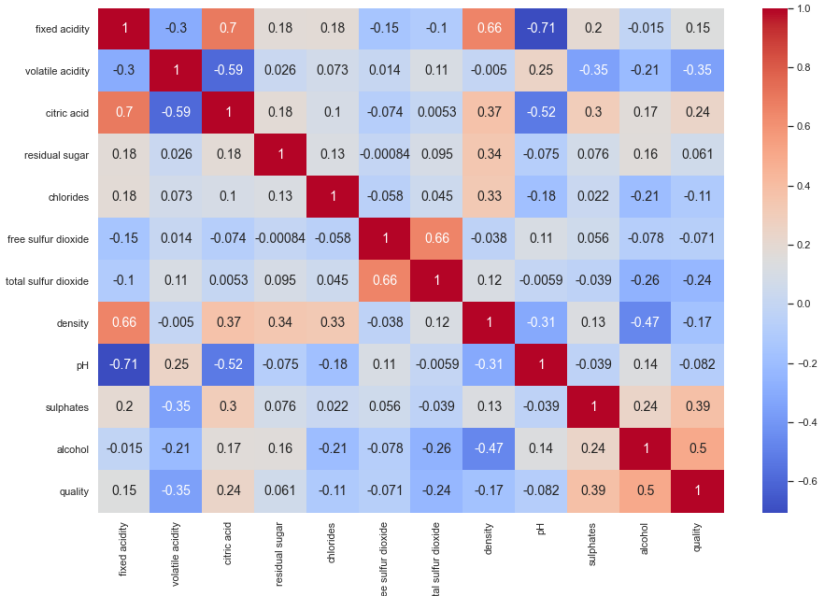


Figure 7: Mapa de calentamiento de Correlación de las variables

Ahora mostramos, hacemos una estandarización de datos y aplicamos Entrenamiento y prueba del modelo -divisió el grupo de entrenamiento 70% y grupo de prueba 30%. El balance de las variables se muestra en la siguiente gráfica. Se muestras que la data esta balanceado para los valores de calidad o cluster $3, 4, 5, 6.7, 8$.

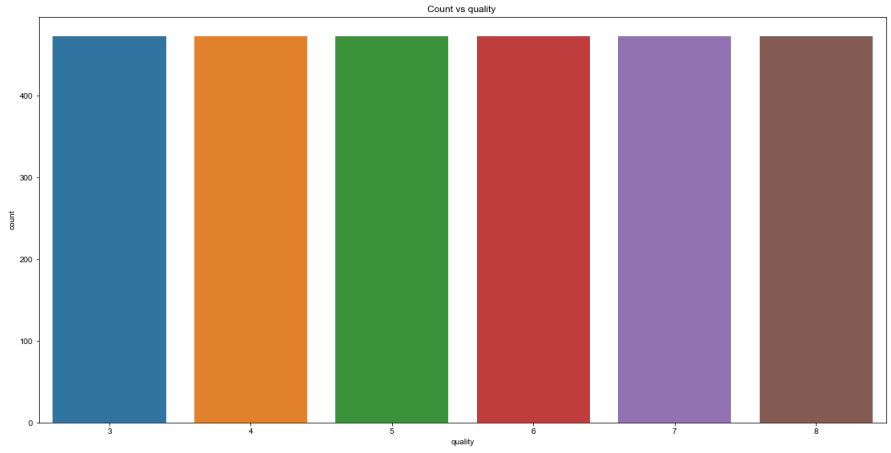


Figure 8: Balance de Variables

Con el los datos balanceados , procedemos a realizar la aplicación del algoritmo de K-Nearest Neighbors y buscar la clasificación de los vinos de acuerdo a su valor de calidad. El algoritmo requiere conocer un K-Numero de vecinos para realizar el conteo de estos y proceder la asignación a cluster de calidad $(3, 4, 5, 6, 7, 8)$.Lo que hace este algoritmo es que toma un punto de datos y selecciona el número K de observaciones en los datos de entrenamiento que son las más cercanas al punto de datos y luego predice la respuesta del punto de datos con respecto al valor de respuesta más popular de los vecinos K-cercanos.

Como no conocemos el valor de k, creamos una gráfica para ver cómo la precisión cambiaba con el número de K.

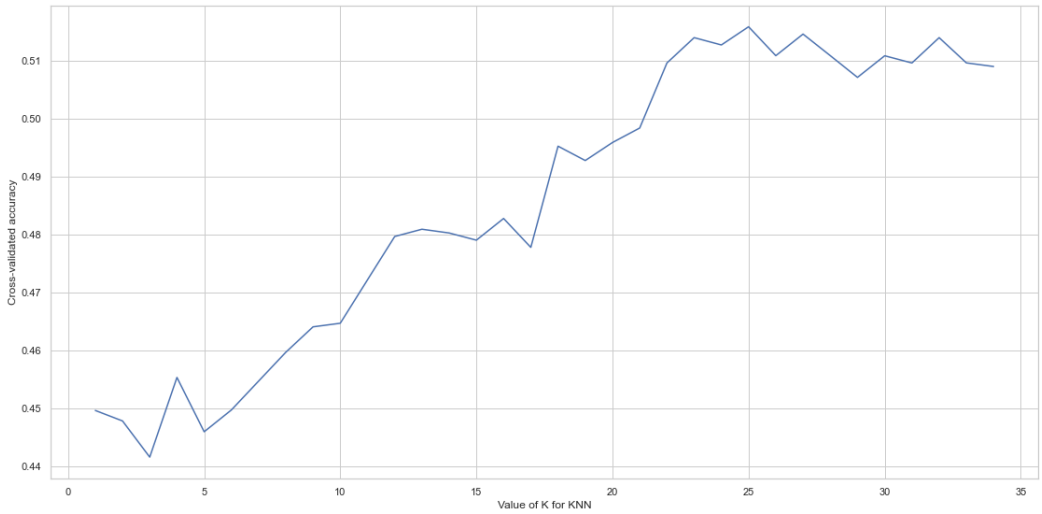


Figure 9: Calculo de K

La gráfica muestra que el valor donde K con mayor presión para un rango de observaciones de 1 a 35 esta alrededor de un valor de $K = 23$. Con este valor procedemos a validar el algoritmo de clasificación KNN.

De la matriz de confusión , tenemos que el modelo solo puede predecir las variables que estan dentro del los cluster o valores de calidad de $5, 6, 7$.

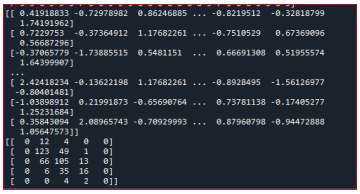
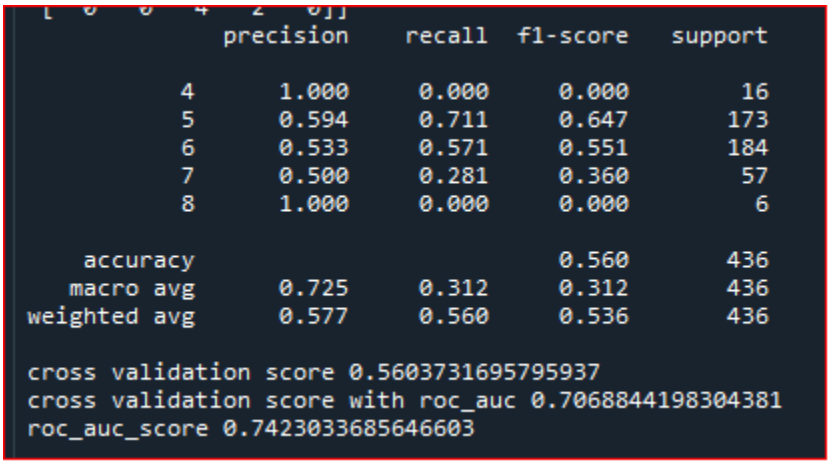


Figure 10: Matriz de Confusión

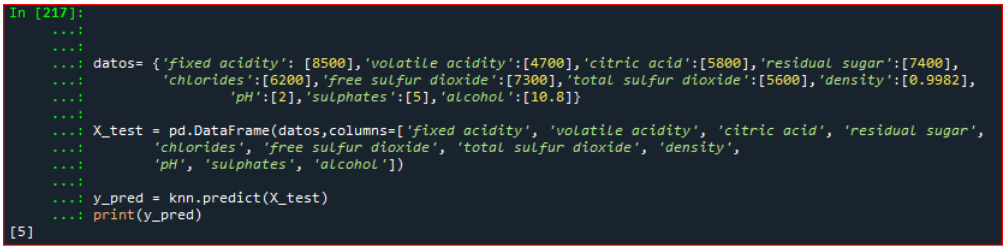
Finalmente,analizamos el reporte del método de clasificación, se observa inmediatamente que las clases 4 y 8 no se han tenido en cuenta en el entrenamiento porque sus resultados de recuerdo son cero. Esto significa que, de todos los miembros de la clase 4 y 8, no predijo ninguno de ellos correctamente. Por lo tanto, no sería un buen modelo para nuestro conjunto de datos. Igualmente los parámetros de validación son valores bajos.



```
[  0   0   4   2   0]]
              precision    recall  f1-score   support

           4      1.000     0.000     0.000        16
           5      0.594     0.711     0.647       173
           6      0.533     0.571     0.551       184
           7      0.500     0.281     0.360        57
           8      1.000     0.000     0.000         6

    accuracy                          0.560       436
   macro avg      0.725     0.312     0.312       436
weighted avg      0.577     0.560     0.536       436

cross validation score 0.5603731695795937
cross validation score with roc_auc 0.7068844198304381
roc_auc_score 0.7423033685646603
```

Figure 11: Resultados

A modo de ejemplo, tomamos un tipo de vino para clarificarlo de acuerdo a las variables y con valore distintos, los resultados se muestran debajo.



```
In [217]:
     ...:
     ...:
     ...: datos= {'fixed acidity': [8500],'volatile acidity':[4700],'citric acid':[5800],'residual sugar':[7400],
     ...:         'chlorides':[6200],'free sulfur dioxide':[7300],'total sulfur dioxide':[5600],'density':[0.9982],
     ...:                 'pH':[2],'sulphates':[5],'alcohol':[10.8]}
     ...:
     ...: X_test = pd.DataFrame(datos,columns=['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
     ...:        'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density',
     ...:        'pH', 'sulphates', 'alcohol'])
     ...:
     ...: y_pred = knn.predict(X_test)
     ...: print(y_pred)
[5]
```

Figure 12: Resultados pruebas