

PROYECTO MACHINE LEARNING

POR: SILVIA MARTINEZ

"Descorriendo la
Excelencia: Un Viaje
Analítico para Prever la
Calidad de Blancos"



Introducción



Es un estudio sobre el vinho verde , un producto único de la región de Minho (noroeste) de Portugal. De contenido medio alcohólico, es especialmente apreciado por su frescor (especialmente en verano). Este vino representa el 15% de la producción total portuguesa, y alrededor del 10% se exporta, principalmente vino blanco.

En los últimos años, el interés por el vino ha aumentado, lo que ha llevado al crecimiento de la industria vitivinícola. Como consecuencia, las empresas están invirtiendo en nuevas tecnologías para mejorar la producción y venta de vino. La certificación de calidad es un paso crucial para ambos procesos y actualmente depende en gran medida de la cata del vino por parte de expertos humanos.

"Este trabajo tiene como objetivo la predicción de las preferencias de vino a partir de pruebas analíticas objetivas que están disponibles en la etapa de certificación."

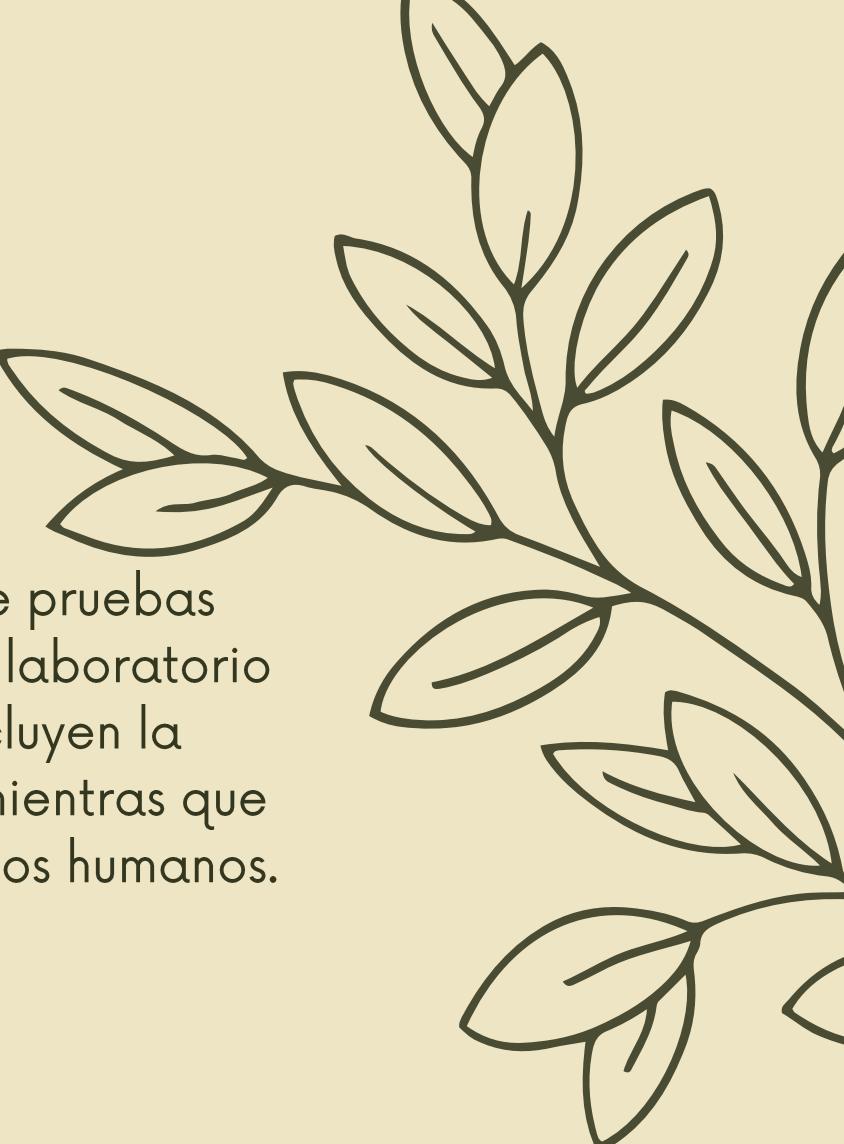
Visión general

Antecedentes

La certificación del vino se evalúa generalmente mediante pruebas fisicoquímicas y sensoriales. Las pruebas fisicoquímicas de laboratorio que se utilizan habitualmente para caracterizar el vino incluyen la determinación de los valores de densidad, alcohol o pH, mientras que las pruebas sensoriales se basan principalmente en expertos humanos.

El propósito del proyecto

Presento un estudio de caso para modelar preferencias gustativas basadas en datos analíticos que, están fácilmente disponibles en el paso de certificación del vino. Construir un modelo de este tipo es valioso no sólo para las entidades de certificación sino también para los productores de vino e incluso para los consumidores. Puede utilizarse para apoyar las evaluaciones de vinos del enólogo, mejorando potencialmente la calidad y la rapidez de sus decisiones. Además, medir el impacto de las pruebas fisicoquímicas en la calidad final del vino es útil para mejorar el proceso de producción.



FASE 1

Recopilación de los datos iniciales.

- El conjunto de datos sobre calidad del vino implica predecir la calidad de los vinos blancos en una escala dada con las medidas químicas de cada vino.
<https://archive.ics.uci.edu/dataset/186/wine+quality>.
- Este dataset contiene un total de 4898 muestras blancas.

FASE 2

Variables del conjunto de datos

- 1 - acidez fija
- 2 - acidez volátil
- 3 - ácido cítrico
- 4 - azúcar residual
- 5 - cloruros
- 6 - dióxido de azufre libre
- 7 - dióxido de azufre total
- 8 - densidad
- 9 - ph
- 10 - sulfatos
- 11 - alcohol
- 12 - calidad (puntuación entre 0 y 10)

FASE 3

El análisis de los datos nos lleva a la predicción de Calidad en cuatro categorías

- 3: 'Deficiente'
- 4: 'Deficiente',
- 5: 'Aceptable',
- 6: 'Bueno'
- 7: 'Bueno'
- 8: 'Excelente',
- 9: 'Excelente'

El análisis de los datos

01

Revisar los datos



```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 4898 entries, 0 to 4897
```

Data columns (total 12 columns):

#	Column	Non-Null Count	Dtype
0	fixed acidity	4898 non-null	float64
1	volatile acidity	4898 non-null	float64
2	citric acid	4898 non-null	float64
3	residual sugar	4898 non-null	float64
4	chlorides	4898 non-null	float64
5	free sulfur dioxide	4898 non-null	float64
6	total sulfur dioxide	4898 non-null	float64
7	density	4898 non-null	float64
8	pH	4898 non-null	float64
9	sulphates	4898 non-null	float64
10	alcohol	4898 non-null	float64
11	quality	4898 non-null	int64

dtypes: float64(11), int64(1)



O2

Analizando datos

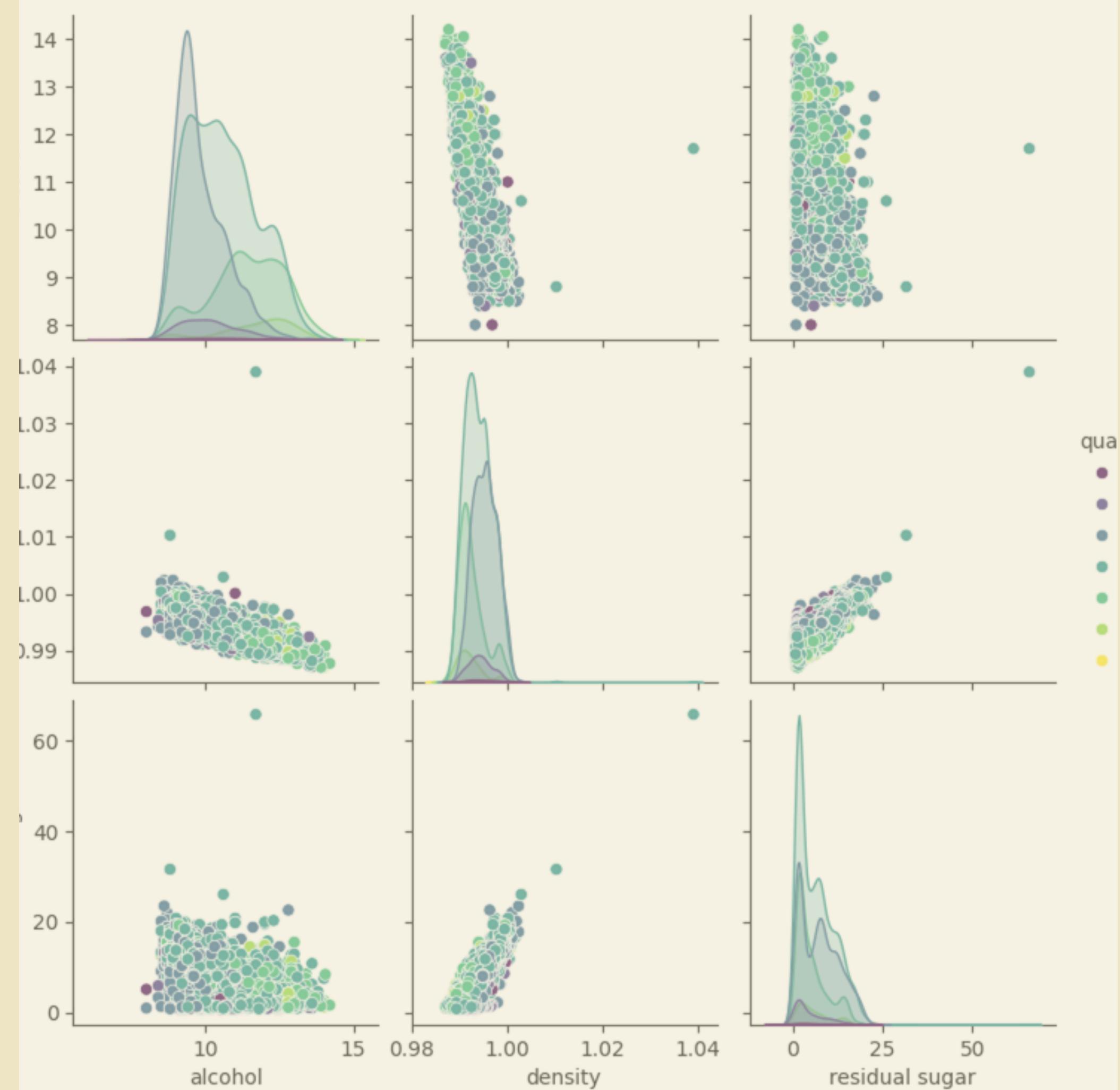
- Hay una correlación positiva moderada (**0.29**) entre "fixed acidity" y "citric acid".
- Existe una correlación negativa moderada (**-0.42**) entre "fixed acidity" y "pH".
- La columna "density" está fuertemente correlacionada con "residual sugar" (**0.84**) y "alcohol" (**-0.78**).
- La columna "alcohol" está fuertemente correlacionada negativamente con "density" (**-0.78**).

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	
fixed acidity	1.000000	0.022697	0.289181	0.089021	0.023086	0.049396	0.091070	0.265331	0.425858	0.017143	-0.120881	-0.113663
volatile acidity	0.022697	1.000000	-0.149472	0.064286	0.070512	-0.097012	0.089261	0.027114	0.031915	0.035728	0.067718	-0.194723
citric acid	0.289181	0.149472	1.000000	0.094212	0.114364	0.094077	0.121131	0.149503	0.163748	0.062331	-0.075729	0.009209
residual sugar	0.089021	0.064286	0.094212	1.000000	0.088685	0.299098	0.401439	0.838966	0.194133	0.026664	-0.450631	-0.097577
chlorides	0.023086	0.070512	0.114364	0.088685	1.000000	0.101392	0.198910	0.257211	0.090439	0.016763	-0.360189	0.209934
free sulfur dioxide	0.049396	0.097012	0.094077	0.299098	0.101392	1.000000	0.615501	0.294210	0.000618	0.059217	-0.250104	0.008158
total sulfur dioxide	0.091070	0.089261	0.121131	0.401439	0.198910	0.615501	1.000000	0.529881	0.002321	0.134562	-0.448892	-0.174737
density	0.265331	0.027114	0.149503	0.838966	0.257211	0.294210	0.529881	1.000000	-0.093591	0.074493	-0.780138	-0.307123
pH	0.425858	0.031915	0.163748	0.194133	0.090439	0.000618	0.002321	0.093591	1.000000	0.155951	0.121432	0.099427
sulphates	0.017143	0.035728	0.062331	-0.026664	0.016763	0.059217	0.134562	0.074493	0.155951	1.000000	-0.017433	0.053678
alcohol	0.120881	0.067718	-0.075729	-0.450631	0.360189	-0.250104	0.448892	0.780138	0.121432	-0.017433	1.000000	0.435575
quality	0.113663	0.194723	-0.009209	-0.097577	0.209934	0.008158	0.174737	0.307123	0.099427	0.053678	0.435575	1.000000

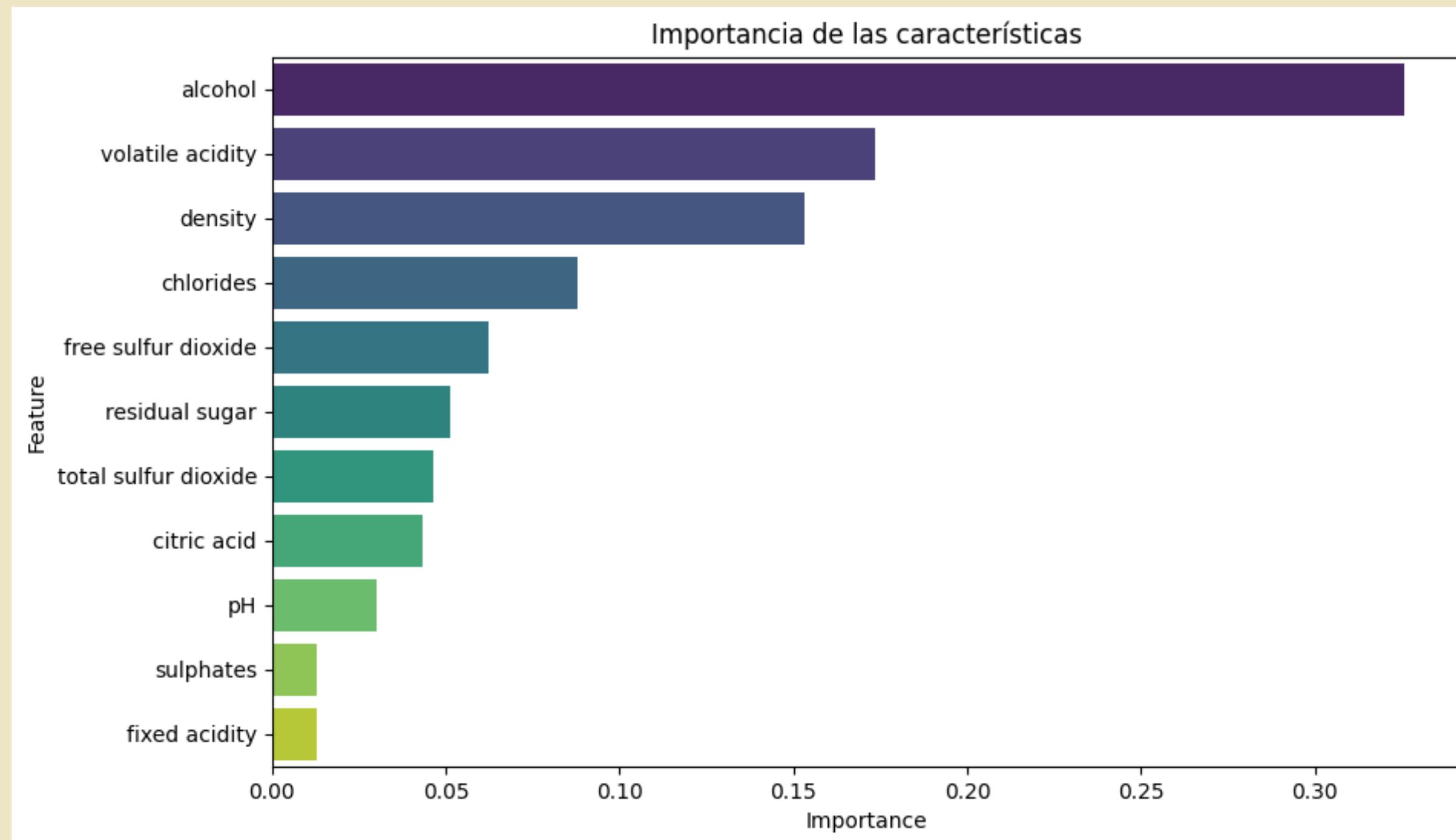
03

Correlaciones no lineales

- **Calidad vs. Alcohol**: 0.462869 - Hay una correlación positiva moderada entre la calidad del vino y el contenido de alcohol. Esto sugiere que vinos de mayor calidad tienden a tener un mayor contenido de alcohol.
- **Calidad vs. Acidez fija**: -0.124636 - Existe una correlación negativa débil entre la calidad y la acidez fija. Esto sugiere que vinos de mayor calidad pueden tener niveles más bajos de acidez fija, aunque la correlación es relativamente débil.
- **Densidad vs. Azúcar residual**: 0.820498 - Hay una correlación positiva fuerte entre la densidad del vino y el contenido de azúcar residual. Esto indica que los vinos con mayor contenido de azúcar residual tienden a tener una mayor densidad.
- **Densidad vs. Alcohol**: -0.760162 - Existe una correlación negativa fuerte entre la densidad y el contenido de alcohol. Esto sugiere que los vinos con mayor contenido de alcohol tienden a tener una menor densidad.



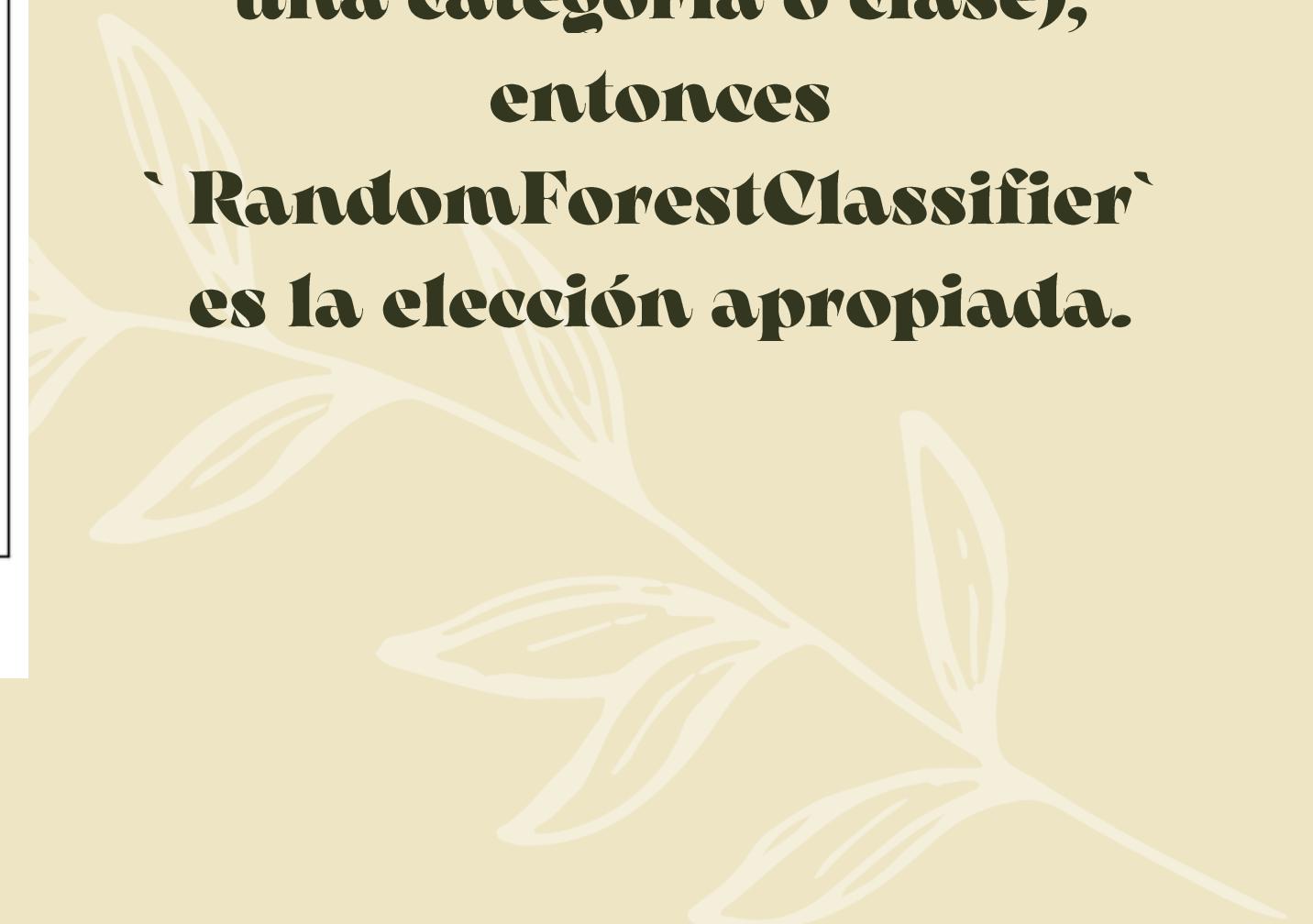
Obtener el `feature importances` de cada variable



****Variable Objetivo:****

- Al tratar con la calidad de vinos, y ser un problema de clasificación (predicción de una categoría o clase), entonces

'RandomForestClassifier' es la elección apropiada.





Primera prueba con RandomForest antes de asignar etiquetas al target

Accuracy: 0.69

Confusion Matrix:

[0	0	4	1	0	0]
[0	6	11	8	0	0]
[0	4	201	82	4	0]
[0	0	66	341	25	0]
[0	0	3	73	112	4]
[0	0	1	12	6	16]

Classification Report:

precision recall f1-score support

3	0.00	0.00	0.00	5
4	0.60	0.24	0.34	25
5	0.70	0.69	0.70	291
6	0.66	0.79	0.72	432
7	0.76	0.58	0.66	192
8	0.80	0.46	0.58	35

accuracy		0.69	980
macro avg	0.59	0.46	0.50
weighted avg	0.69	0.69	0.68

Asignar una etiqueta de calidad con mapeo para poder realizar las predicciones y, al finalizar devolverán como salida "deficiente", "aceptable", "bueno" y "excelente"

3: 'Deficiente', 4: 'Deficiente', 5: 'Aceptable', 6: 'Bueno', 7: 'Bueno', 8: 'Excelente', 9: 'Excelente'

0 a 4 (0)

5 (1)

6 y 7 (2)

8 y 9 (3)

Sobre-muestreo (Over-sampling)

category

2 3078

1 3078

3 3078

0

3078



Primera prueba con RandomForest después de categorizar y hacer sobre-muestreo

Accuracy en datos balanceados: 0.91

Confusion Matrix en datos balanceados:

```
[[590  0 15  0]
 [ 7539 60  0]
 [ 0 147 505  0]
 [ 0  2  0 598]]
```

Classification Report en datos balanceados:

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.99	0.98	0.98	605
1	0.78	0.89	0.83	606
2	0.87	0.77	0.82	652
3	1.00	1.00	1.00	600

accuracy		0.91	2463	
macro avg	0.91	0.91	0.91	2463
weighted avg	0.91	0.91	0.91	2463

Pruebo también con SMOTE pero da peores resultados

Accuracy en datos con SMOTE: 0.87

Confusion Matrix en datos con SMOTE:

```
[[544 199  4  0]
 [156 581  0  0]
 [ 0 207 34  0]
 [ 0  0  0 686]]
```

Classification Report en datos con SMOTE:

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

Aceptable	0.78	0.73	0.75	747
Bueno	0.73	0.79	0.76	737
Deficiente	0.99	0.97	0.98	754
Excelente	1.00	1.00	1.00	686

accuracy		0.87	2924	
macro avg	0.87	0.87	0.87	2924
weighted avg	0.87	0.87	0.87	2924

GridSearch & Pipelines con los datos balanceados y comparación de modelos

En general, parece que el modelo tiene un rendimiento bueno, con altas precisiones, recalls y f1-scores para cada clase. El accuracy del 100% también indica un buen rendimiento general del modelo en el conjunto de datos de prueba.

Mejores parámetros encontrados:

```
{'classifier': DecisionTreeClassifier(), 'classifier__max_depth': 3, 'classifier__min_samples_split': 2}
```

Rendimiento del mejor modelo:

Accuracy en conjunto de prueba: 1.0

Accuracy en datos con over sampler: 1.00

Confusion Matrix en datos con over sampler:

```
[[605  0  0  0]
 [ 0 606  0  0]
 [ 0  0 652  0]
 [ 0  0  0 600]]
```

Classification Report en datos con over sampler:

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	1.00	1.00	1.00	605
1	1.00	1.00	1.00	606
2	1.00	1.00	1.00	652
3	1.00	1.00	1.00	600

accuracy		1.00	2463	
----------	--	------	------	--

macro avg	1.00	1.00	1.00	2463
-----------	------	------	------	------

weighted avg	1.00	1.00	1.00	2463
--------------	------	------	------	------

Accuracy (Precisión):

- **Resultado:** 1.00 (100%)

- **Explicación:** La precisión es la proporción de predicciones correctas en relación con el total de predicciones. Un valor de 1.00 significa que todas las predicciones fueron correctas. Es un resultado excelente.

Confusion Matrix

- **Explicación:**

- La diagonal principal muestra el número de predicciones correctas para cada clase.

- En este caso, no hay errores en la clasificación. Todas las instancias de cada clase se clasificaron correctamente..

Classification Report (Informe de Clasificación):

...

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

...

- **Precision:** La proporción de verdaderos positivos entre todas las instancias predichas positivas. En este caso, es 1.00 para todas las clases, indicando que no hubo falsos positivos.

- **Recall:** La proporción de verdaderos positivos entre todas las instancias reales positivas. También es 1.00 para todas las clases, lo que significa que no hubo falsos negativos.

- **F1-score:** La media armónica entre precisión y recall. Igualmente, es 1.00 para todas las clases.

- **Support:** El número total de instancias de cada clase en el conjunto de datos.

Se incluye un nuevo DataFrame de vino para test, que contiene las mismas columnas. Con esta prueba se intenta resolver si el modelo predice igual de bien con datos externos.

En el Pipeline se han utilizado los siguientes parámetros

```
#Definir los parámetros para cada clasificador
logistic_params = {
    'scaler': [StandardScaler()],
    'classifier': [LogisticRegression(max_iter=1000, solver='liblinear')],
    'classifier_penalty': [l1, l2],
}

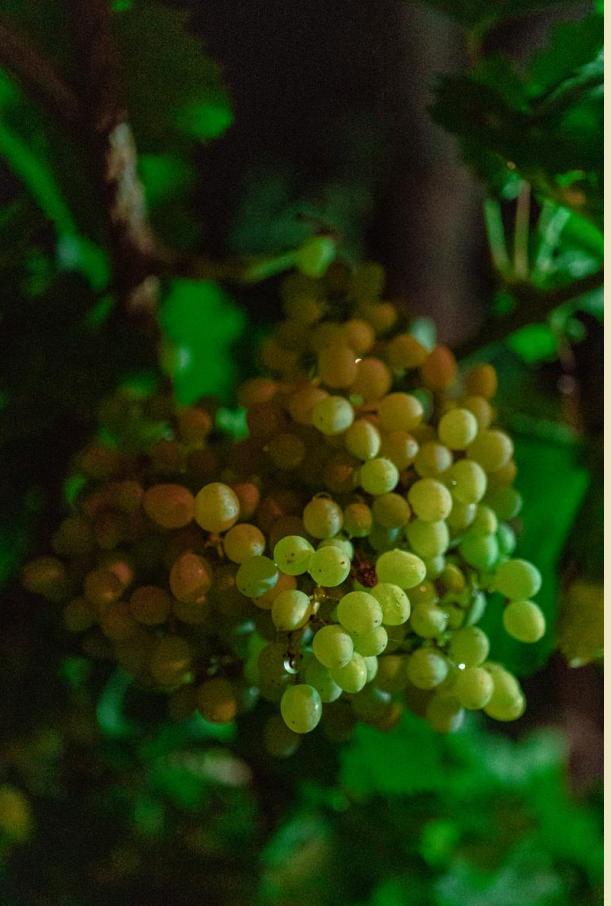
random_forest_params = {
    'classifier': [RandomForestClassifier()],
    'classifier_max_features': [1, 2, 3],
    'classifier_max_depth': [1, 2, 3]
}
xgb_param = {
    'classifier': [XGBRFClassifier()],
    'classifier_n_estimators': [50, 100],
    'classifier_max_depth': [3, 4, 5],
}

decision_tree_param = {
    'classifier': [DecisionTreeClassifier()],
    'classifier_max_depth': [2, 3, 4, 5],
    'classifier_min_samples_split': [2, 4, 6, 8, 10]
}

knn_params = {
    'classifier': [KNeighborsClassifier()],
    'classifier_n_neighbors': [3, 5, 7],
    'classifier_p': [1, 2]
}

#Lista de todos los clasificadores con sus parámetros
search_space = [
    logistic_params,
    random_forest_params,
    decision_tree_param,
    xgb_param,
    knn_params
]

# Se crea el gridsearch indicándole que trabaje con un pipeline y que pruebe todos los parámetros y modelos antes definidos
clf = GridSearchCV(estimator=pipe,
                    param_grid=search_space,
                    cv=10)
```



Mejores parámetros encontrados:
{'classifier': DecisionTreeClassifier(), 'classifier_max_depth': 3, 'classifier_min_samples_split': 2}

Rendimiento del mejor modelo:
Accuracy en conjunto de prueba: 1.0
Accuracy en datos con over_sampler: 1.00

Confusion Matrix en datos con over_sampler:

[[63 0 0 0]
[0 577 0 0]
[0 0 702 0]
[0 0 0 17]]

Classification Report en datos con over_sampler:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	63
1	1.00	1.00	1.00	577
2	1.00	1.00	1.00	702
3	1.00	1.00	1.00	17
accuracy			1.00	1359
macro avg	1.00	1.00	1.00	1359
weighted avg	1.00	1.00	1.00	1359

Se guardan las predicciones en un archivo pickle

1. **DecisionTreeClassifier con Parámetros Óptimos:**

- El modelo Decision Tree ha sido configurado para tener una profundidad máxima de 3 y un mínimo de 2 muestras para dividir un nodo. Esto podría ser para evitar un sobreajuste (overfitting) al conjunto de entrenamiento.

2. **Rendimiento del Modelo:**

- El modelo ha logrado una precisión del 100% tanto en los datos con oversampling, como con el nuevo conjunto de datos. Esto indica que el modelo es capaz de predecir correctamente todas las clases en ambos conjuntos.

3. **Matriz de Confusión y Reporte de Clasificación:**

- La matriz de confusión y el reporte de clasificación confirman que el modelo está prediciendo correctamente todas las clases, ya que no hay errores en la clasificación. Todas las métricas (precision, recall, f1-score) son igual a 1.0 para cada clase.

El modelo parece estar funcionando excepcionalmente bien en este conjunto de datos específico, y con el nuevo conjunto de datos utilizados para TEST. Sin embargo, es importante tener en cuenta que un rendimiento perfecto en el conjunto de entrenamiento y prueba puede ser indicativo de sobreajuste. Además, se debe considerar que el oversampling ha influido en el rendimiento del modelo.

En resumen, el modelo actual ha demostrado ser altamente efectivo en este conjunto de datos específico, sería necesaria la evaluación en conjuntos de datos adicionales para confirmar la generalización del modelo.

Conclusiones

La realización de este estudio puede proporcionar beneficios al sector.

La construcción exitosa de un modelo de predicción para la calidad del vino no solo beneficia a las entidades de certificación, sino que también abre oportunidades valiosas para productores y consumidores en la industria vinícola.

1. **Apoyo a Evaluaciones de Enólogos:**

- El modelo puede servir como una herramienta valiosa para respaldar las evaluaciones de los enólogos. Al mejorar la rapidez y precisión de las decisiones sobre la calidad del vino, se facilita la toma de decisiones durante el proceso de producción.

2. **Mejora del Proceso de Producción:**

- La capacidad del modelo para medir el impacto de las pruebas fisicoquímicas en la calidad final del vino proporciona información crucial para mejorar el proceso de producción. Identificar las variables más influyentes permite ajustes precisos que pueden elevar la calidad del producto.

3. **Marketing Objetivo:**

- La aplicación de técnicas similares para modelar las preferencias del consumidor abre puertas al marketing objetivo. Entender las preferencias en nichos de mercado o mercados rentables permite estrategias de comercialización más efectivas, adaptando productos a las demandas específicas de los consumidores.

4. **Beneficio para los Consumidores:**

- Los consumidores se benefician directamente de la mejora en la calidad del vino y la adaptación de productos a sus preferencias. La confianza en las evaluaciones respaldadas por el modelo puede influir en las decisiones de compra, mejorando la experiencia del consumidor.

5. **Impacto en la Industria Vinícola:**

- En conjunto, la implementación exitosa de este modelo tiene el potencial de impactar positivamente la calidad del vino, la eficiencia en la toma de decisiones y la adaptación a las preferencias del mercado. Esto no solo beneficia a las partes involucradas directamente, sino que también fortalece la posición de la industria vinícola en su conjunto.

En resumen, la construcción y aplicación de este modelo no solo representa un avance técnico, sino que también tiene el poder de transformar y optimizar varios aspectos de la cadena de producción y consumo de vino, generando beneficios significativos para todos los actores involucrados en la industria.



Gracias

Silvia Martinez