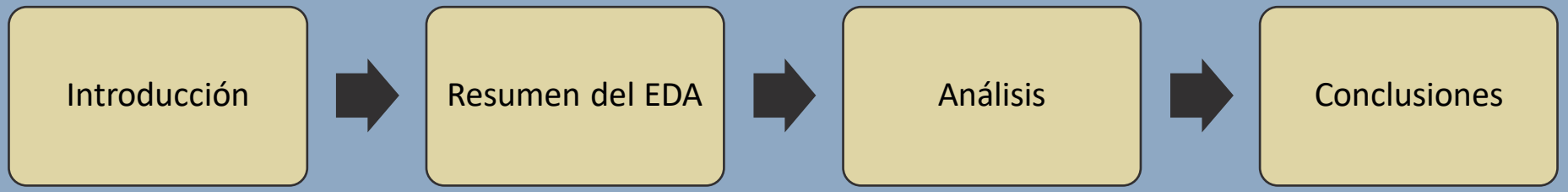
¿Se puede predecir la calidad del vino?

Francisco Canet

INDICE DE CONTENIDOS



- Describir el tema
- Utilidad de un modelo de ML

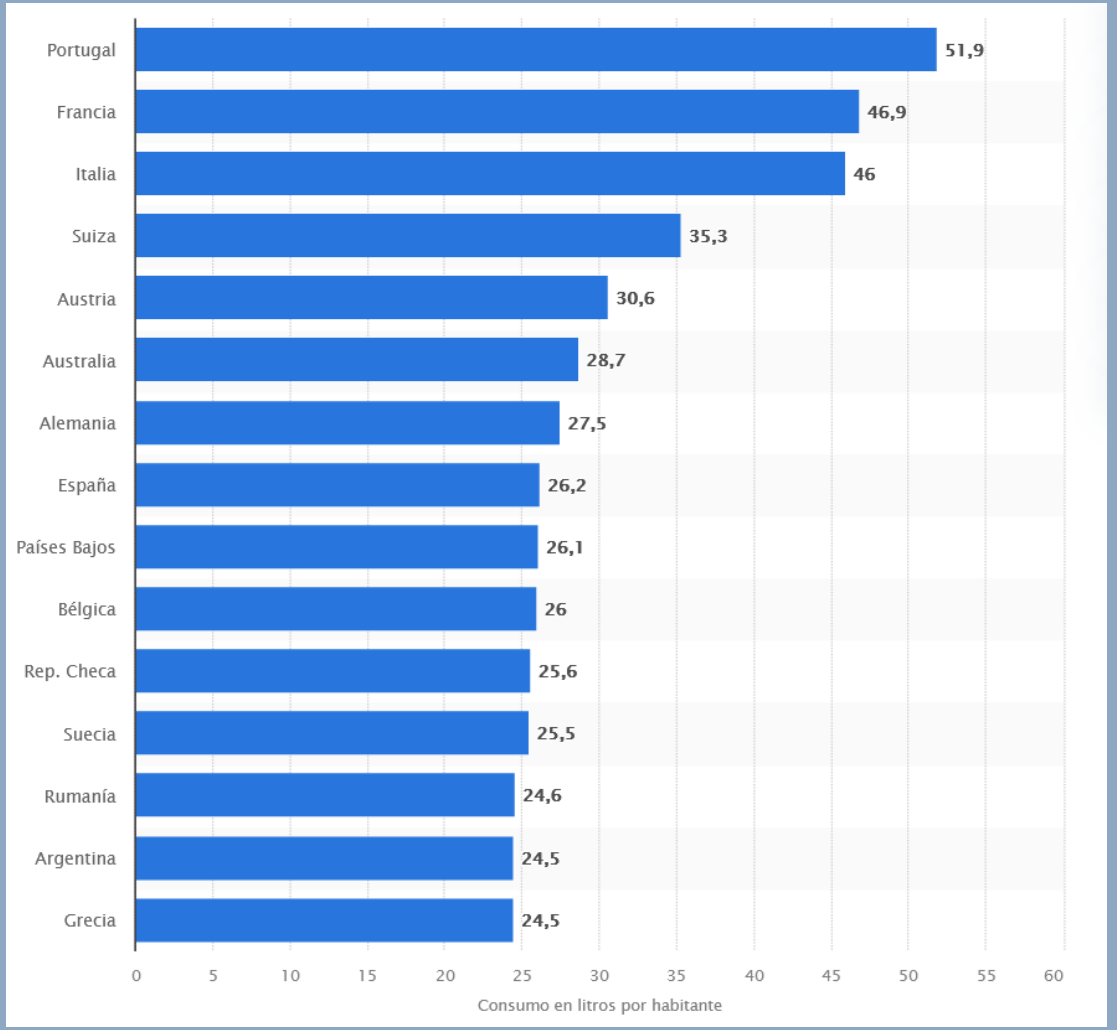
- Describir la fuente y las variables
- Variables importantes

- Distintas aproximaciones
- Resultados
- Feature importance

INTRODUCCIÓN

- Alguna vez considerado como un bien de lujo, hoy en día el vino es esta presente en nuestras vidas…
  - Parte de la dieta mediterránea
  - Celebraciones
  - Comidas
- El consumo per cápita de vino en 2021 en España fue de 26.5 litros.
- Representan un sector económico importante
  - España es el 2do mayor productor a nivel mundial.

Mayores consumidores de vino en el mundo (per cápita)



https://es.statista.com/estadisticas/503596/paises-del-mundo-con-mayor-consumo-per-capita-de-vino/

¿POR QUÉ ES IMPORTANTE EVALUAR LA CALIDAD DEL VINO?

- Mejorar la producción e identificar los factores más influyentes
- Para estratificar los vinos por ejemplo en marcas premium, útil para establecer precios
- Para obtener una certificación DOP y también evitar adulteraciones ilegales

DURANTE LA EVALUACION DE CALIDAD…

- Se toman mediciones fisicoquímicas de rutina como cantidad de alcohol, pH, etc.
- Prueba sensorial realizada por expertos
- El sentido del gusto es complejo, por lo que clasificar a los vinos es un todo un reto.
- A esto hay que sumar que las relaciones entre las variables fisicoquímicas y el sabor que producen es muy compleja
- Un modelo de ML que ayude a predecir la calidad del vino es muy útil.

DATOS – MUESTRAS Y VARIABLES

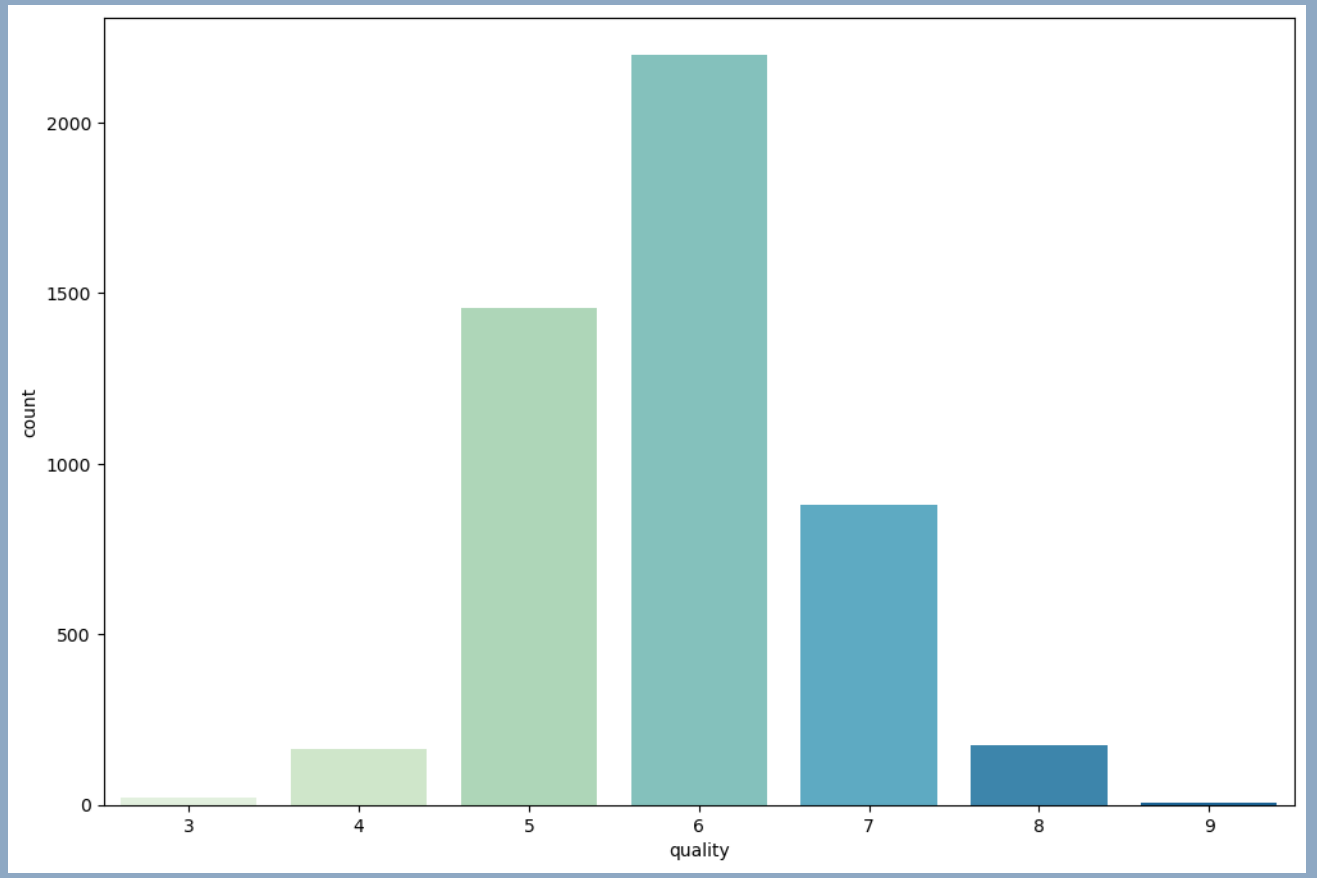
- Muestras de vino blanco con DOP (“Vinho Verde”) (n = 5000)
- Datos sobre calificaciones sensoriales (variable dependiente). 1-10.
- Datos sobre 11 variables fisicoquímicas (variables independientes)
- Fuente de los datos: http://www3.dsi.uminho.pt/pcortez/wine/
- Publicación original: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In *Decision Support Systems*, Elsevier, 47(4):547-553, 2009

DATOS – VARIABLES INDEPENDIENTES

Variable	¿Qué aporta?	¿Inconvenientes?
Acidez fija	Sabor fresco	
Acidez volátil		Sabor a vinagre
Acido cítrico	Sabor fresco	Microorganismos indeseados
pH	Sabor fresco	
Azucares residuales	Sabor dulce	
Alcohol	Equilibrio entre dulce y acido	
Cloruros	Sabor salado	
Dióxido de azufre libre	Protección frente a oxidación y microorganismos indeseados	Alergias
Dióxido de azufre total		
Densidad	Relacionada con el contenido de alcohol	
Sulfatos	Fertilizante	

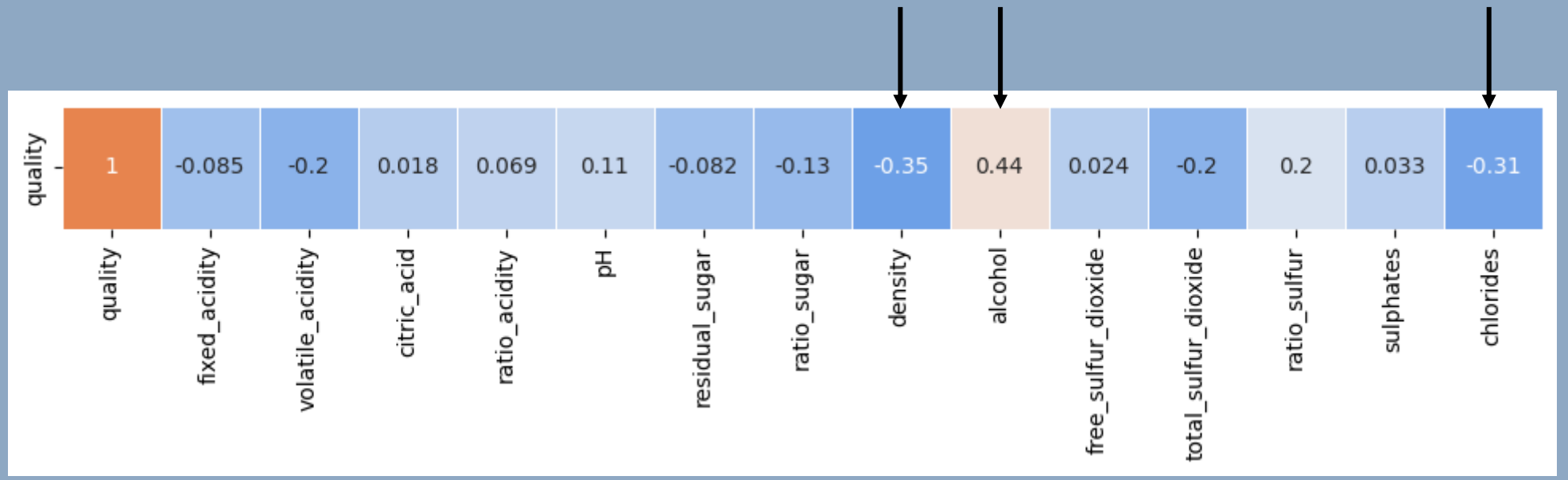
RESUMEN EDA

- La variable de respuesta esta desbalanceada.
- Los vinos muy buenos (9) y muy malos (3) están poco representados.
- Los vinos de 5 y 6 representan el 73%.



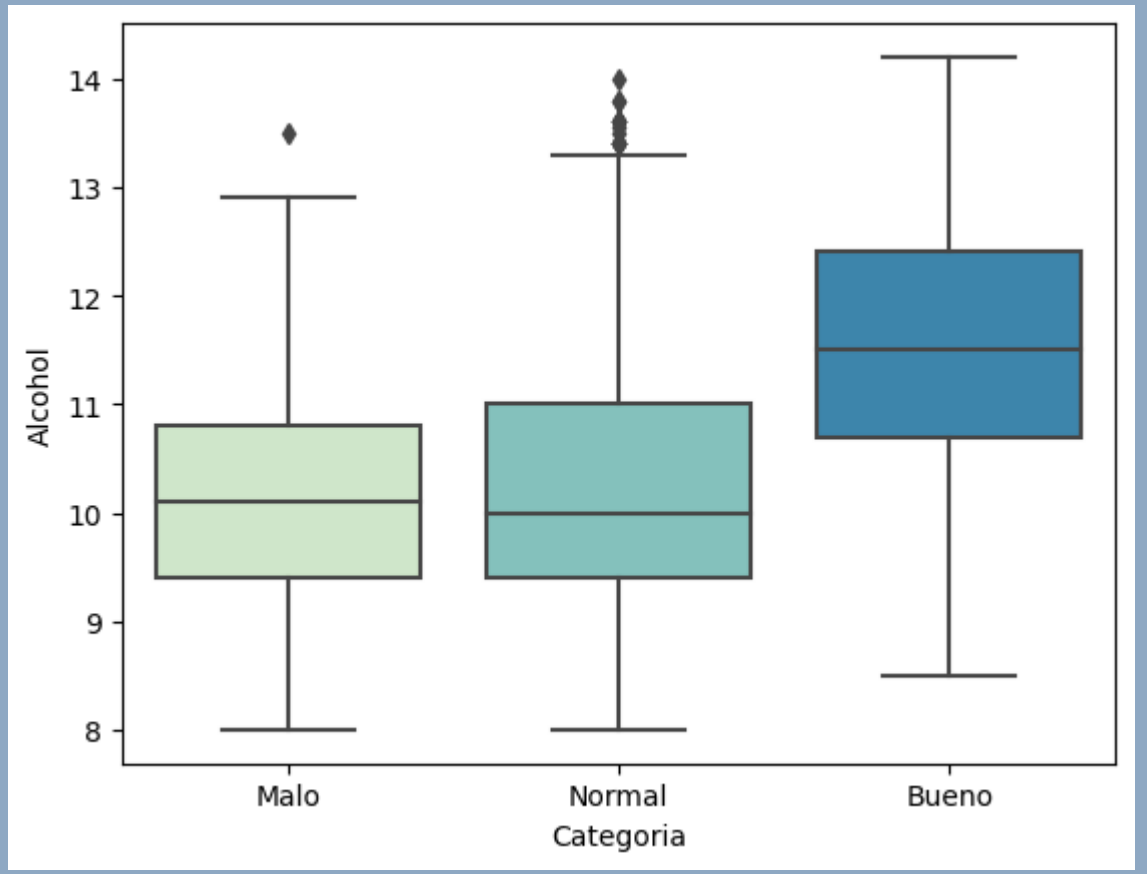
RESUMEN EDA

- Algunas variables fisicoquímicas se correlacionaban (de forma lineal) con la calidad del vino:
  - Contenido de alcohol (positiva)
  - Densidad (Negativa)
  - Cantidad de cloruros (Negativa)



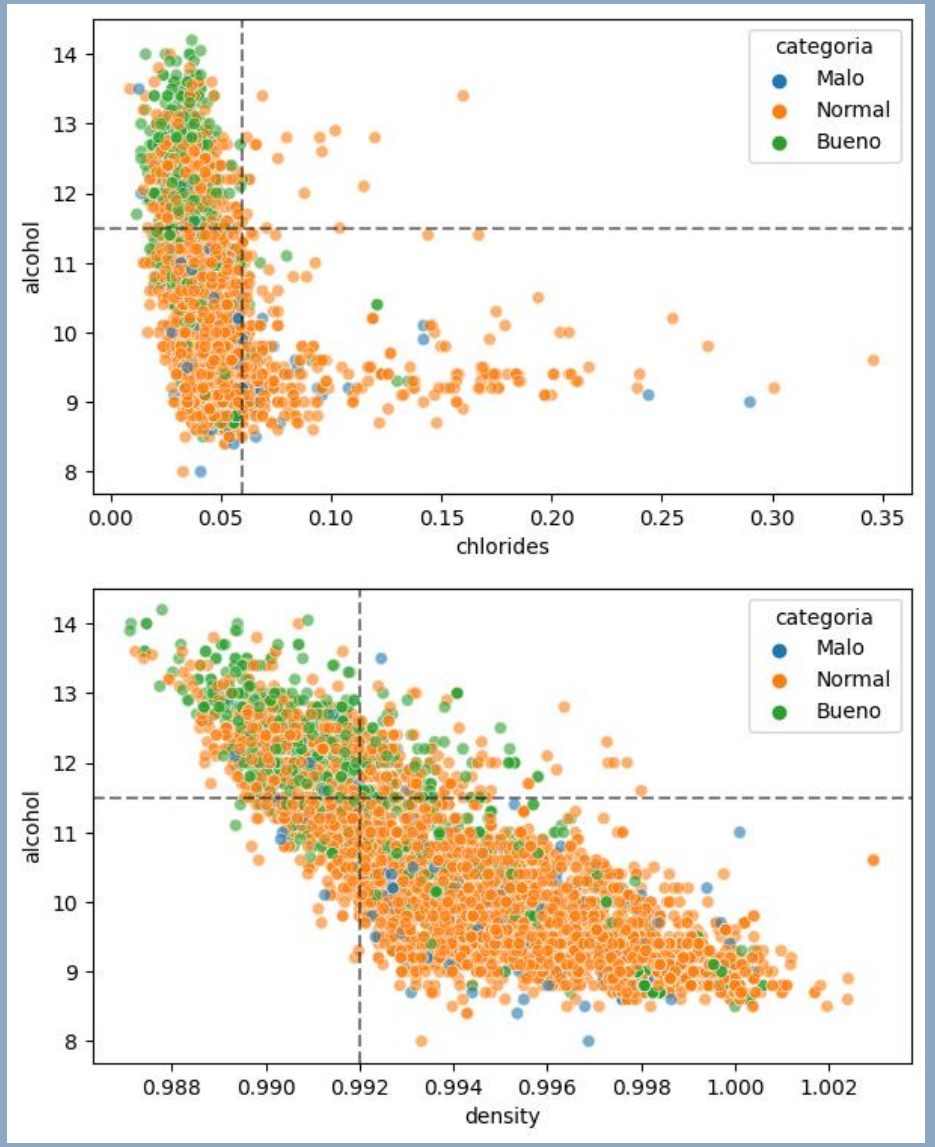
RESUMEN EDA

- Discretizamos la calidad para ver si existía algún patrón.
- Intervalos:
  - 3-4: Malo
  - 5-6: Normal
  - 7-9: Bueno



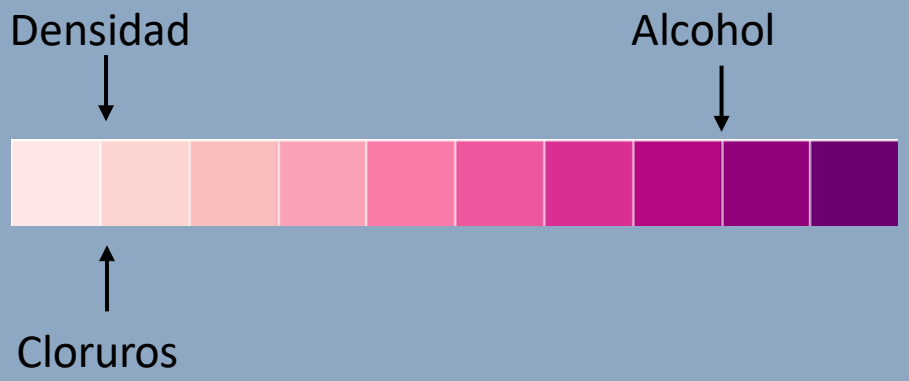
RESUMEN EDA

- La cantidad de alcohol y cloruros y la densidad parecían afectar a la calidad del vino.

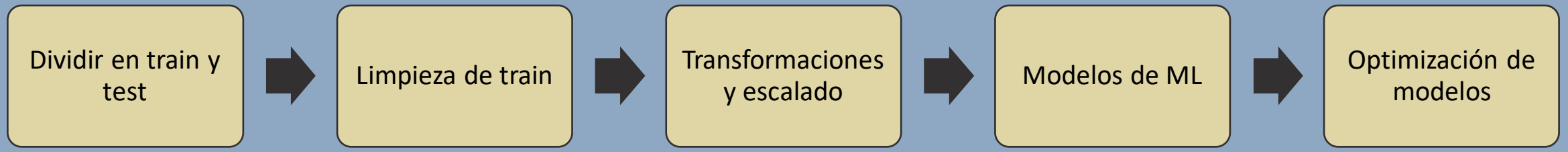


CONCLUSION DEL EDA

Sí seria posible predecir la calidad del vino en función de algunas variables fisicoquímicas usando modelos de ML



FLUJO DE TRABAJO



- 20% para test

- 3 vinos con cantidades de azúcar muy elevadas

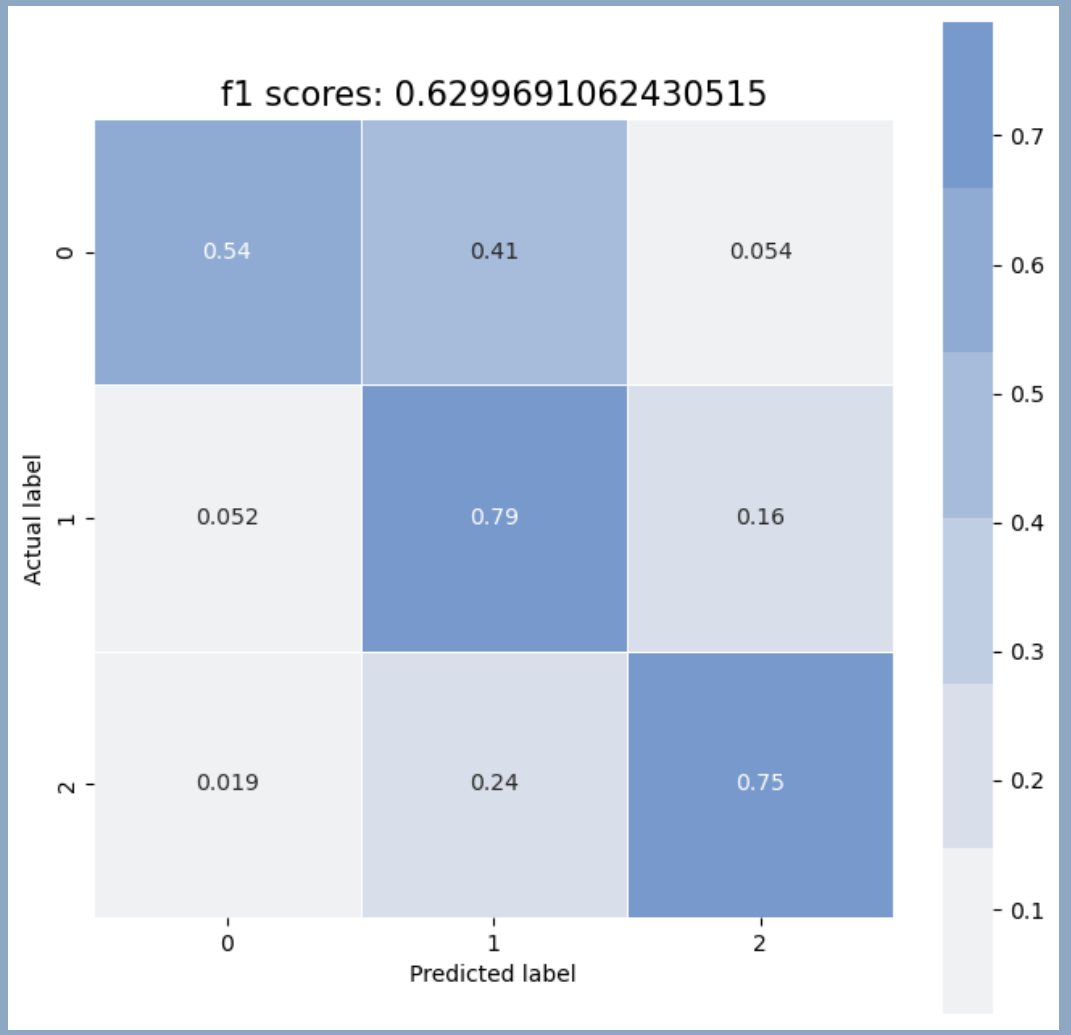
- Yeo-Johnson
- Standard scaler

- Log Regression
- SVM
- Random Forest

- Grid Search
- Random Forest
- Cat boost

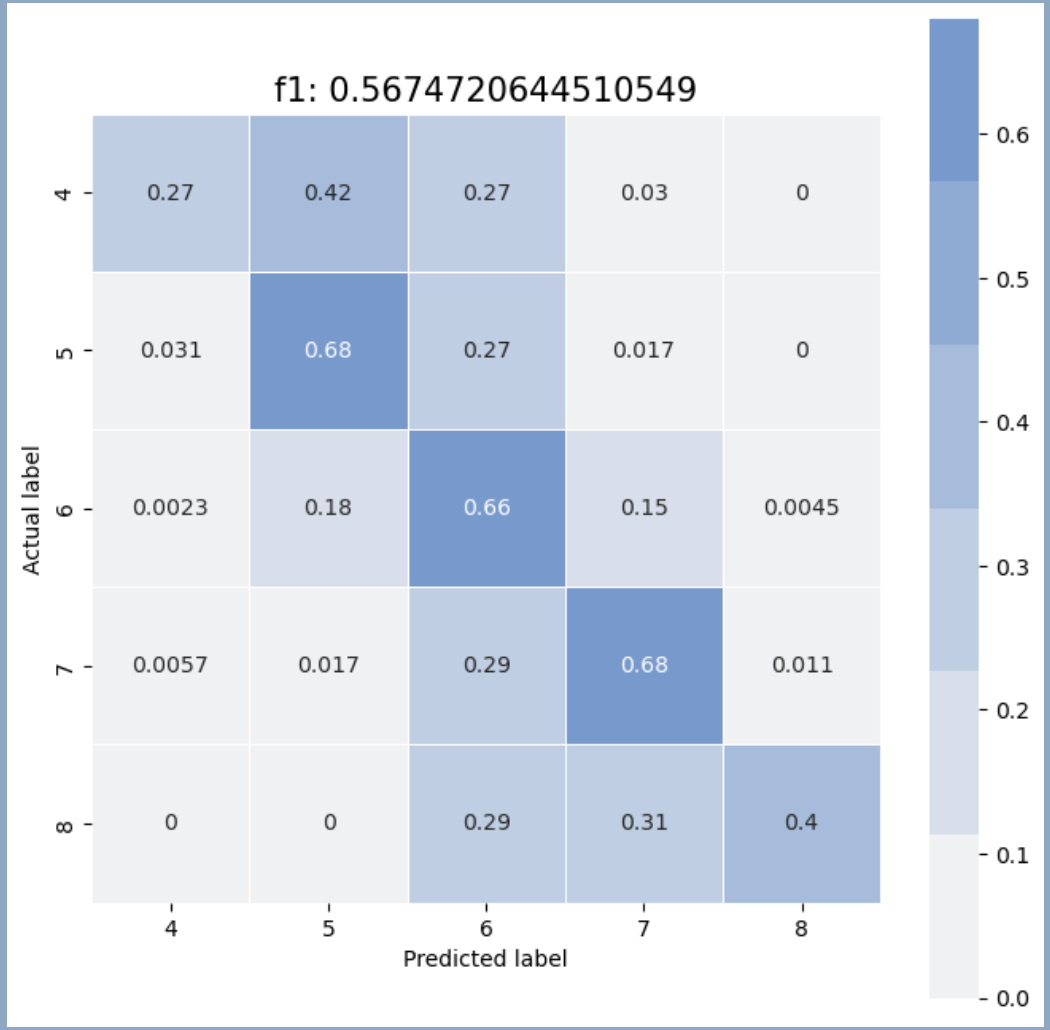
ANALISIS: CALIDAD DISCRETIZADA

- Problema de clasificación en vinos malos, normales o buenos.
- Métrica: F1-score = 0.62
- Random Forest (Class weights = Balanced subsample)
- Problema con los vinos malos
- SMOTE no funcionó mejor que Class weights. F1-score = 0.56

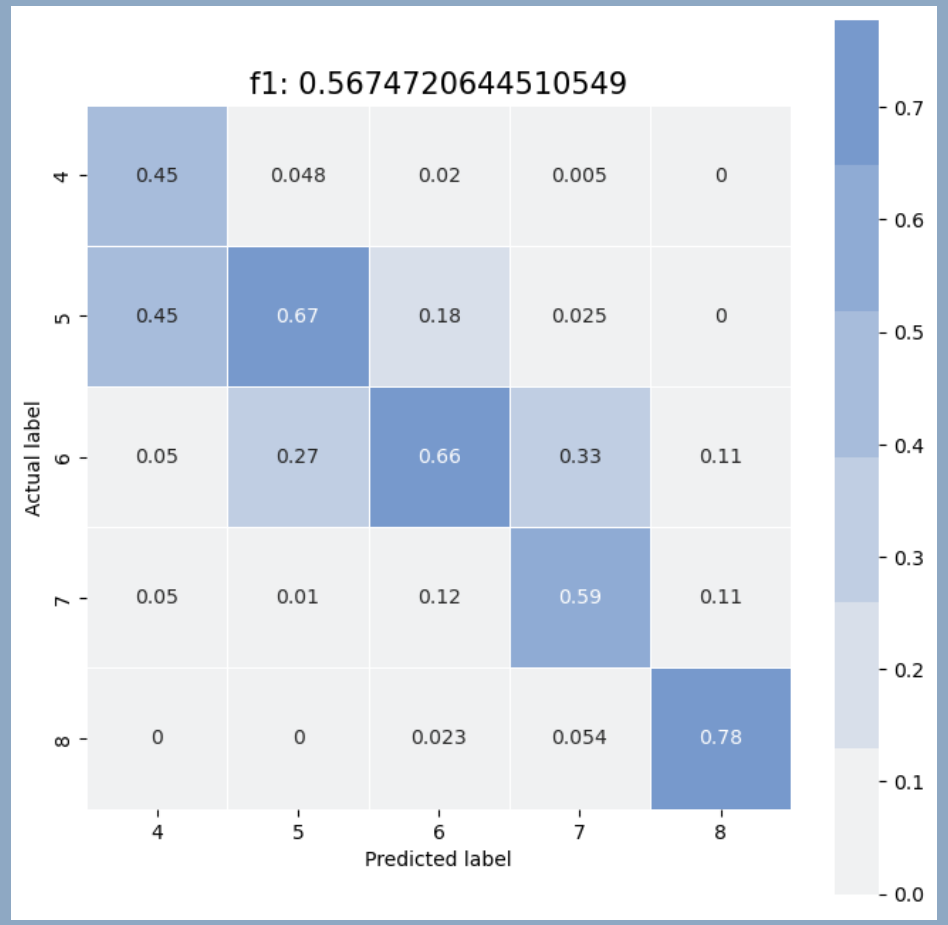


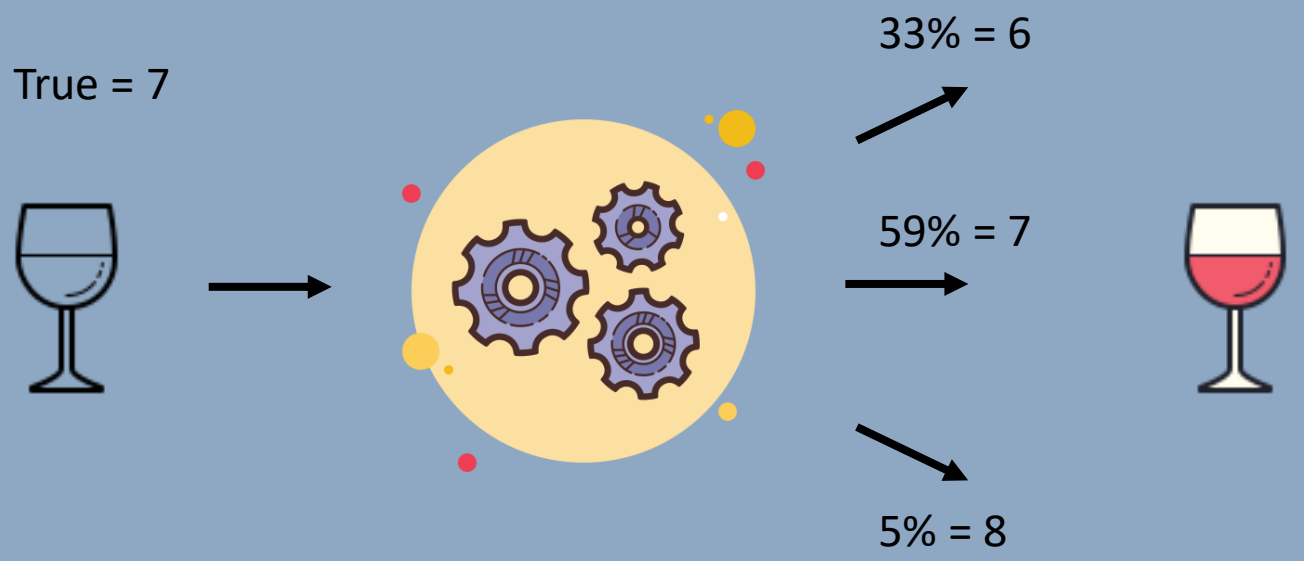
ANALISIS: SIN DISCRETIZAR CALIDAD

- Problema de clasificación: calidad de 4-8.
- Eliminé los vinos con calificaciones de 3 (n = 20) y 9 (n = 5).
- F1-score = 0.56
- Precisión = 0.63
- Random Forest fue el mejor
- Vemos como la mayoría de vinos se clasifican bien, o como mucho una categoría por arriba o por abajo.
- Los vinos de 4 y 8 son los que peor se clasifican



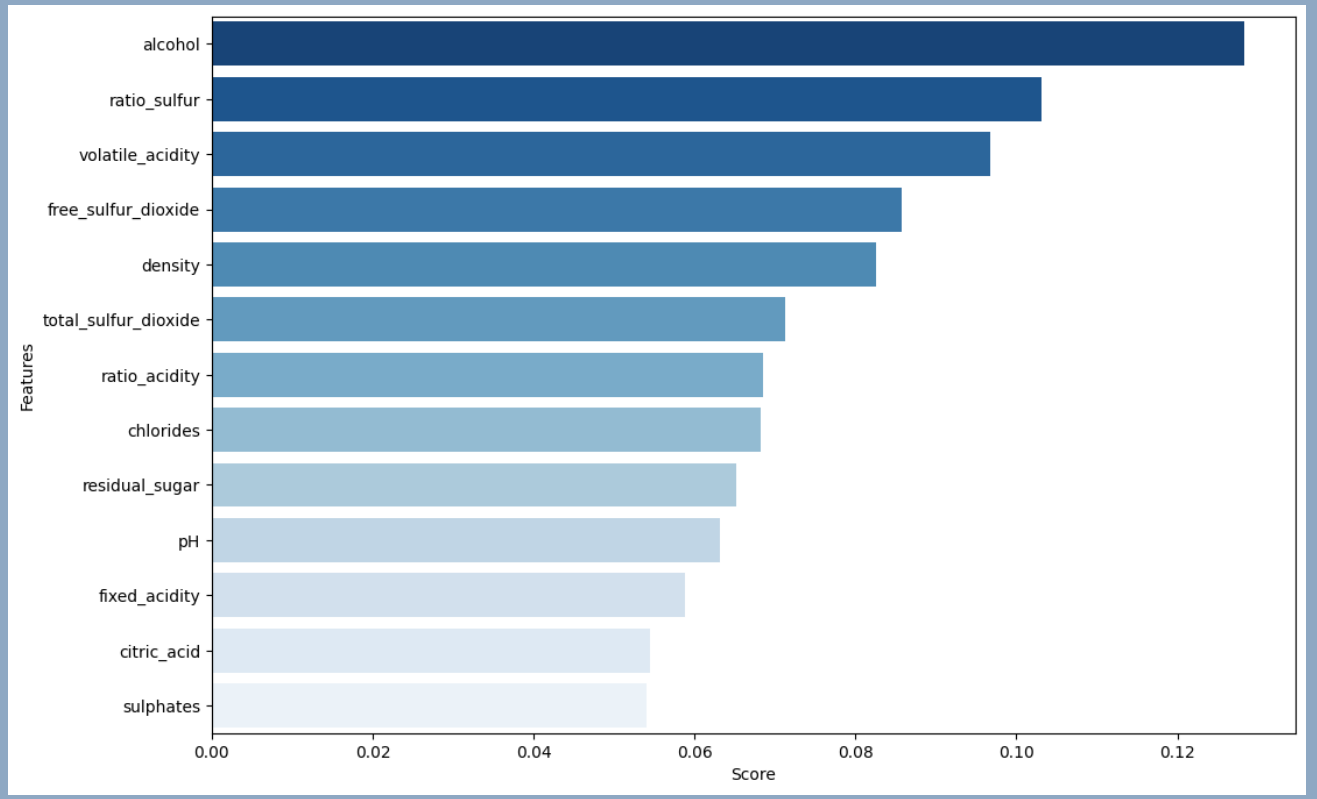
ANALISIS: SIN DISCRETIZAR CALIDAD





FEATURE IMPORTANCE

- El contenido de alcohol fue la variable más importante para predecir la calidad del vino.
- Ratio sulfur (variable creada a partir de otras dos) fue la segunda más importante.



CONCLUSIONES

- Aunque la puntuación del modelo con la calidad discretizada (F1-score = 0.62) fue mayor en comparación al modelo sin discretizar (F1-score = 0.56), el segundo es más informativo.
- El segundo modelo de ML es mas útil para ayudar al proceso de clasificación de los vinos ya que se equivoca ± 1 una clase, pero no más.

ANEXOS

	validation_metric_mean	validation_metric_std	training_metric_mean	training_metric_std
log_reg	0.316502	0.011812	0.326225	0.007914
lsvc	0.371761	0.017120	0.386485	0.006151
rbf_svc	0.400877	0.013821	0.518122	0.005182
sig_svc	0.196866	0.009376	0.188256	0.008430
rand_forest	0.492196	0.018095	0.734436	0.005272

	validation_metric_mean	validation_metric_std	training_metric_mean	training_metric_std
ada_boost_tree	0.319251	0.027519	0.337402	0.019691
light_gbm	0.408732	0.021819	0.529236	0.016446
xg_boost	0.406303	0.027796	0.614670	0.013146
cat_boost	0.555003	0.036632	1.000000	0.000000

	validation_metric_mean	validation_metric_std	training_metric_mean	training_metric_std
opt random forest	0.575638	0.029664	0.939856	0.002372
opt catboost	0.559060	0.036551	0.999202	0.000291

Métrica/Calificación	4	5	6	7	8
Precisión	0.45	0.67	0.66	0.59	0.78
F1-score	0.34	0.68	0.67	0.63	0.53