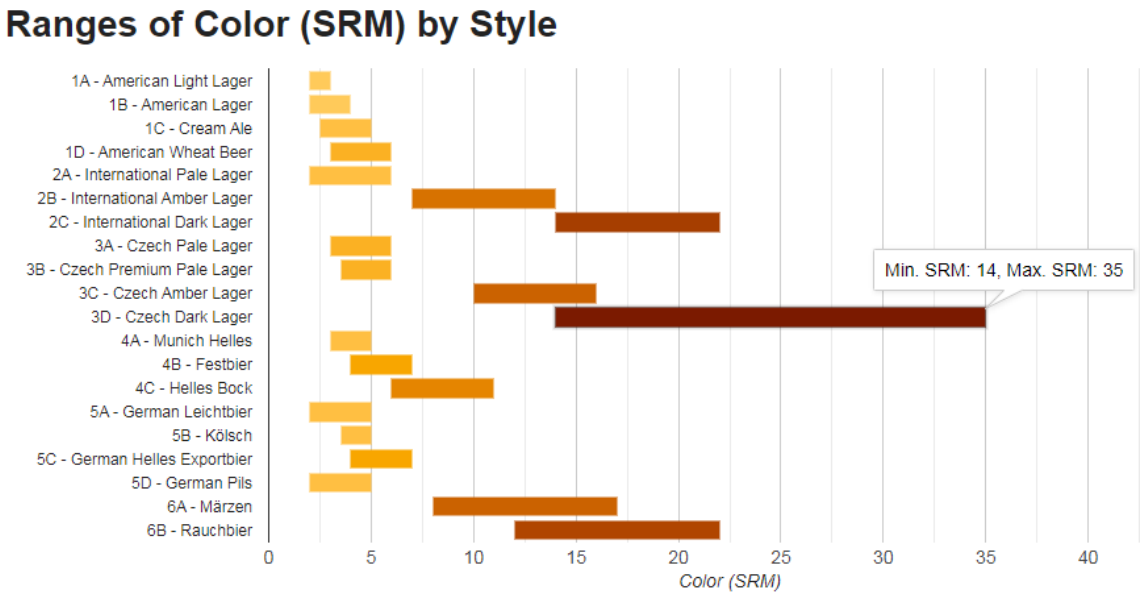
CERTEZA EN TU CERVEZA

Modelos predictivos de estilos de cerveza – *Machine Learning*

DATOS Y VARIABLES

Brewersfriend.com



75 000 cervezas

175 estilos

Niveles de alcohol de 0% al 53% vol.

Niveles de amargor de 0 a 3409 IBU

Escala de color de 1 a 186 SRM

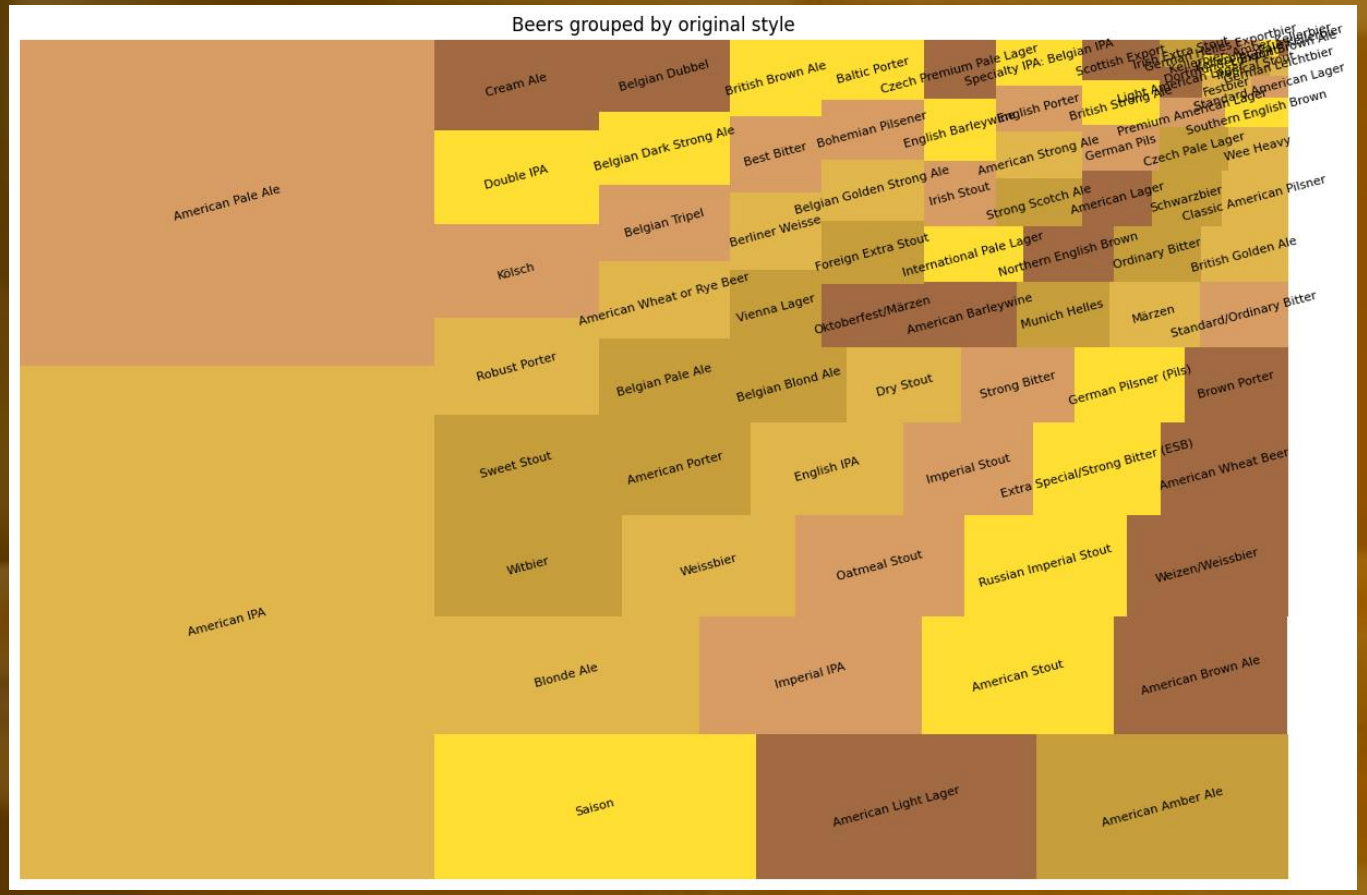
23 variables iniciales

LAS UNIDADES DE MEDIDA

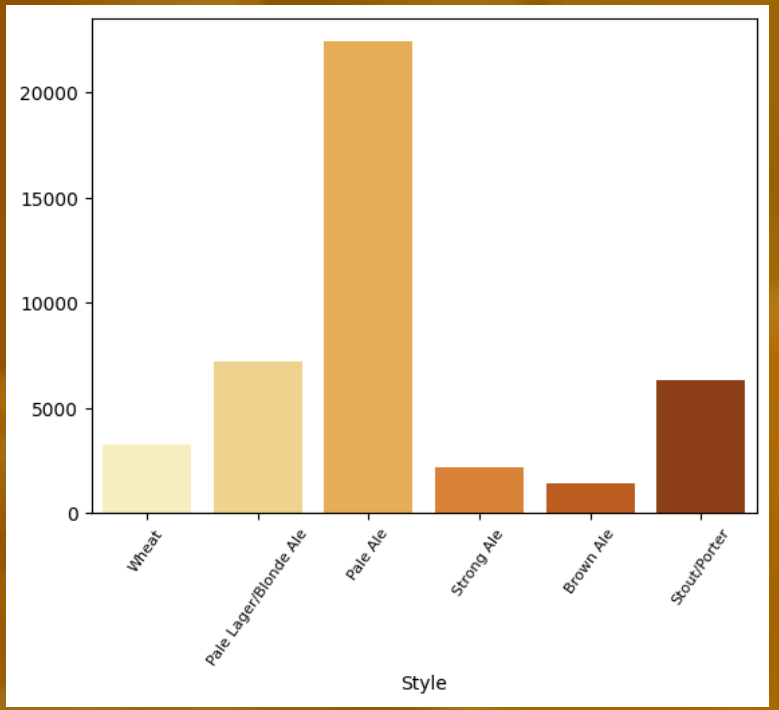
- Alcohol by volume (ABV):
  - *Pale ales* y *stouts* tienen más por lo general, pero no hay un estándar fijo
- International Bitterness Units (IBU):
  - Suave: hasta20 IBU;
  - Algo amarga - de 20 a 30 IBU;
  - Muy Amarga - de 30 a 60 IBU;
  - 'Extrema' - más de 60 IBU.
- Standard Reference Method (SRM):
  - Rubia - de 1 a 15 SRM;
  - Tostada - de 12 a 22 SRM;
  - Oscura - de 23 a 35 SRM;
  - Negra - de 35 a 50 SRM (u 80 EBC)

LIMPIEZA DE LOS NOMBRES DE ESTILOS

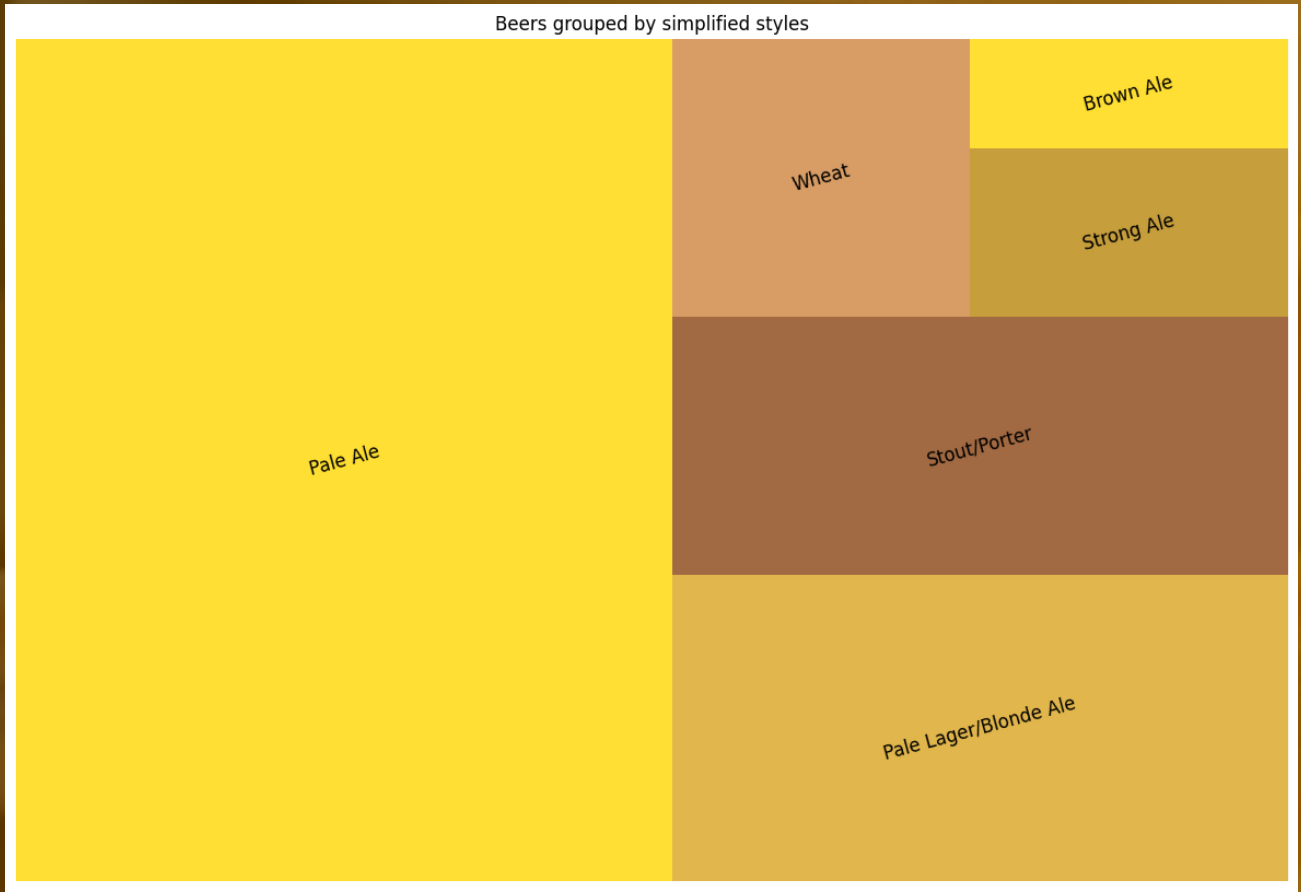
Mapeo exhaustivo de los 175 estilos de cerveza hasta reducirlos a 6:



- Claramente, habrá que equilibrar las clases para el entrenamiento

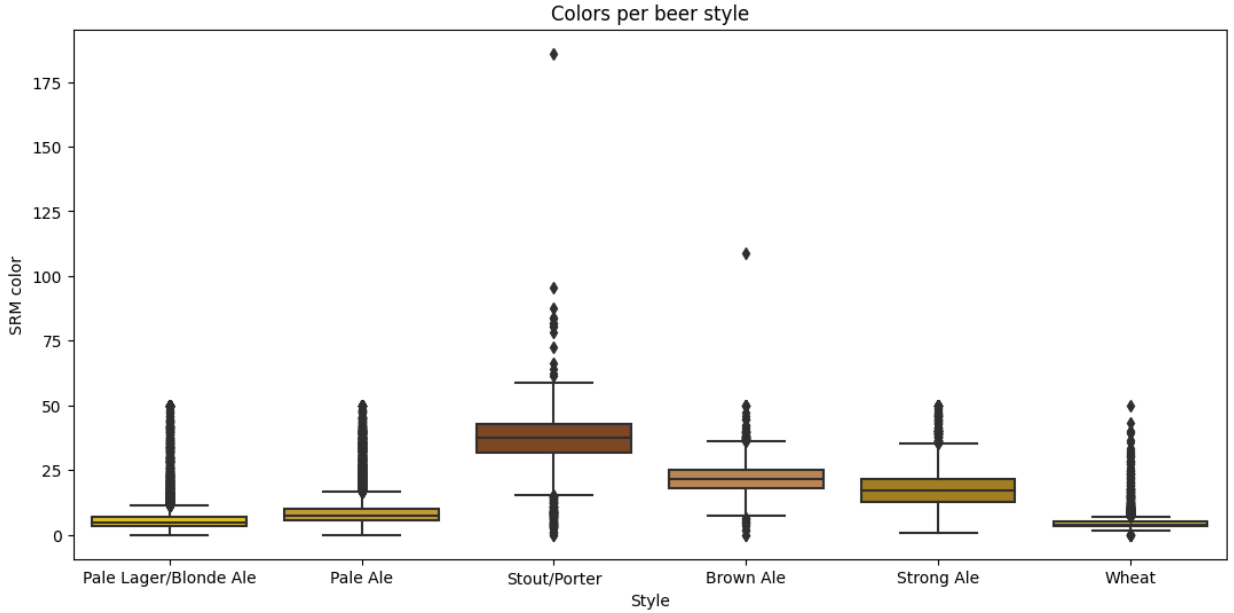


LIMPIEZA DE LOS NOMBRES DE ESTILOS II



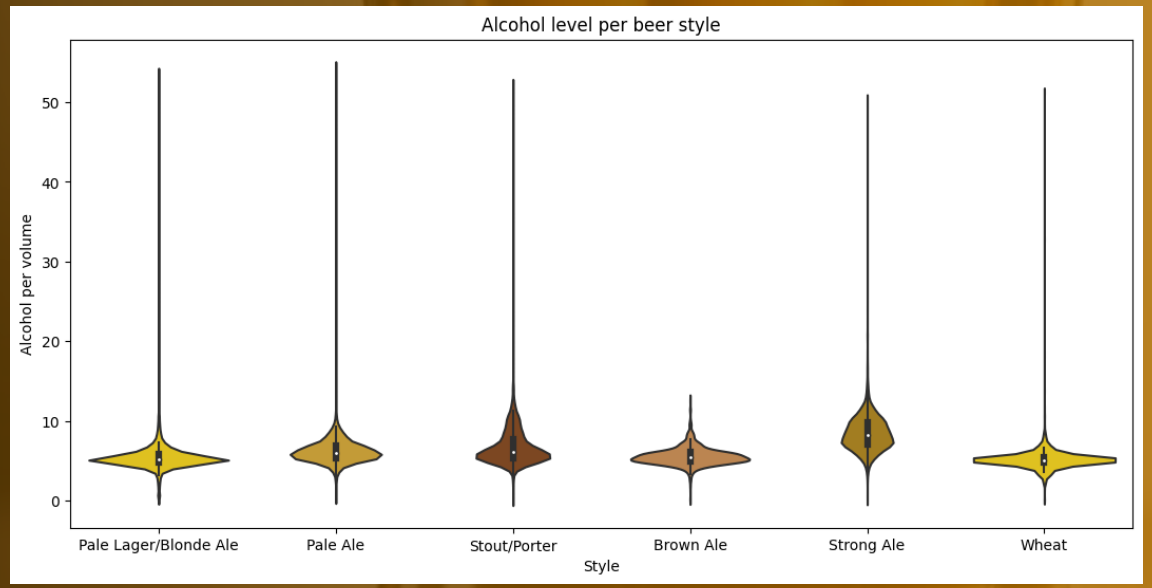
- (Pruebas con ordenación por color y por amargor)

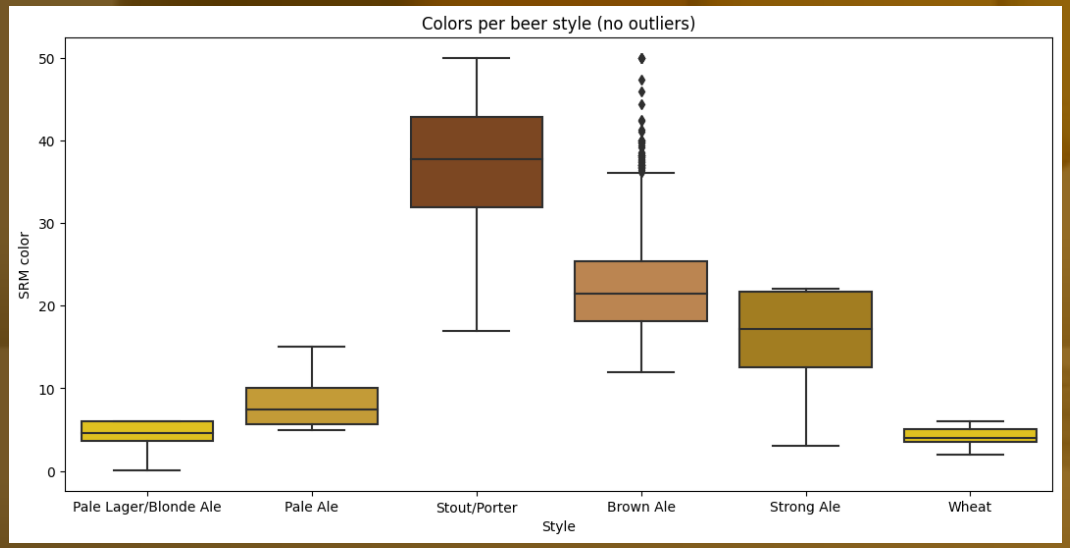
Style	Code
Wheat	0
Pale Lager/Blonde Ale	1
Pale Ale	2
Strong Ale	3
Brown Ale	4
Stout/Porter	5

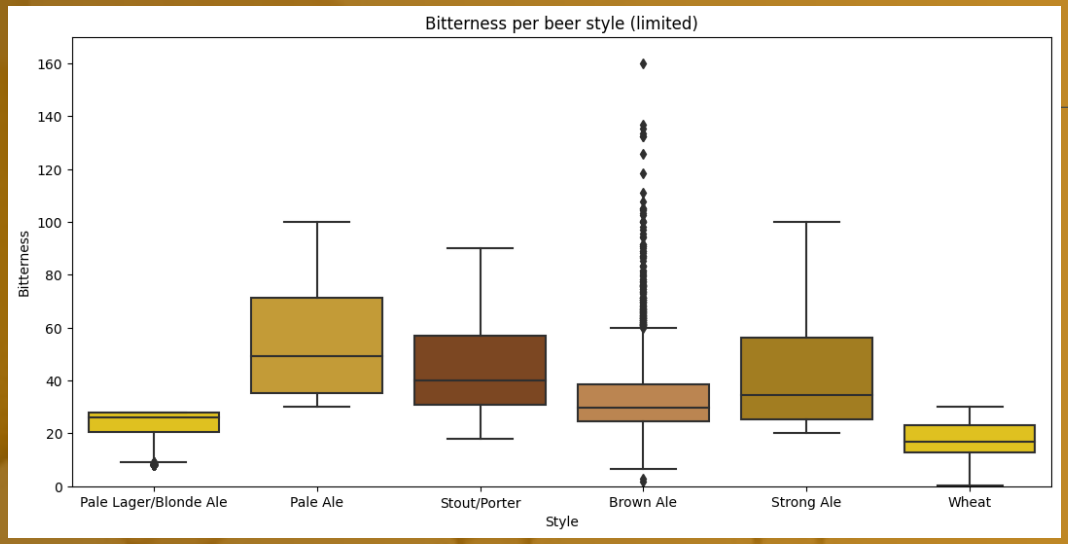


TRANSFORMACIÓN DE VARIABLES

- Exploración de los datos iniciales y eliminación de ruido, variables inútiles y valores nulos:
  - Eliminación de estilos minoritarios (ediciones limitadas, sidras y cervezas sin alcohol)

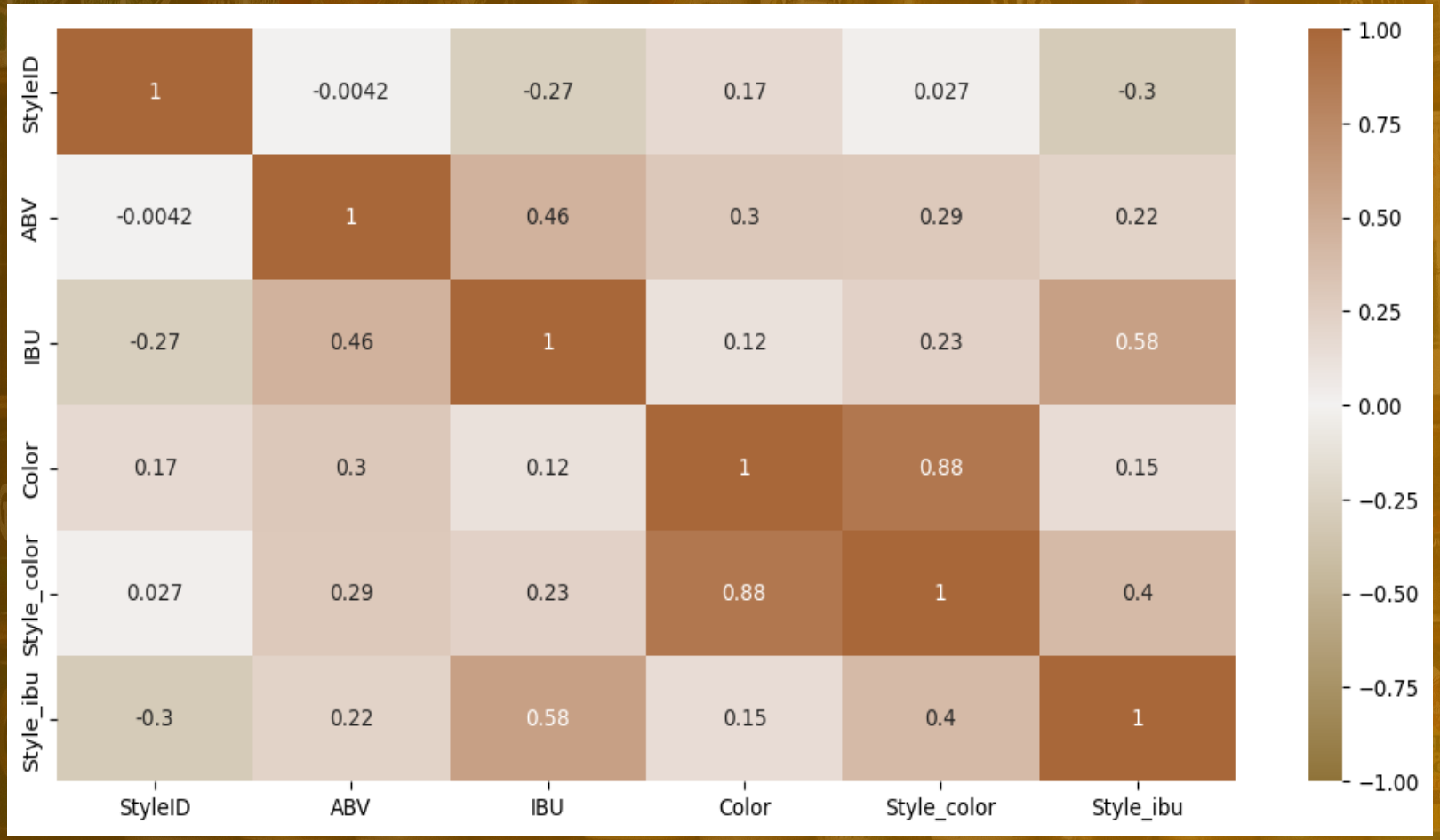






RESULTADOS DE *FEATURE ENGINEERING* Y LIMPIEZA

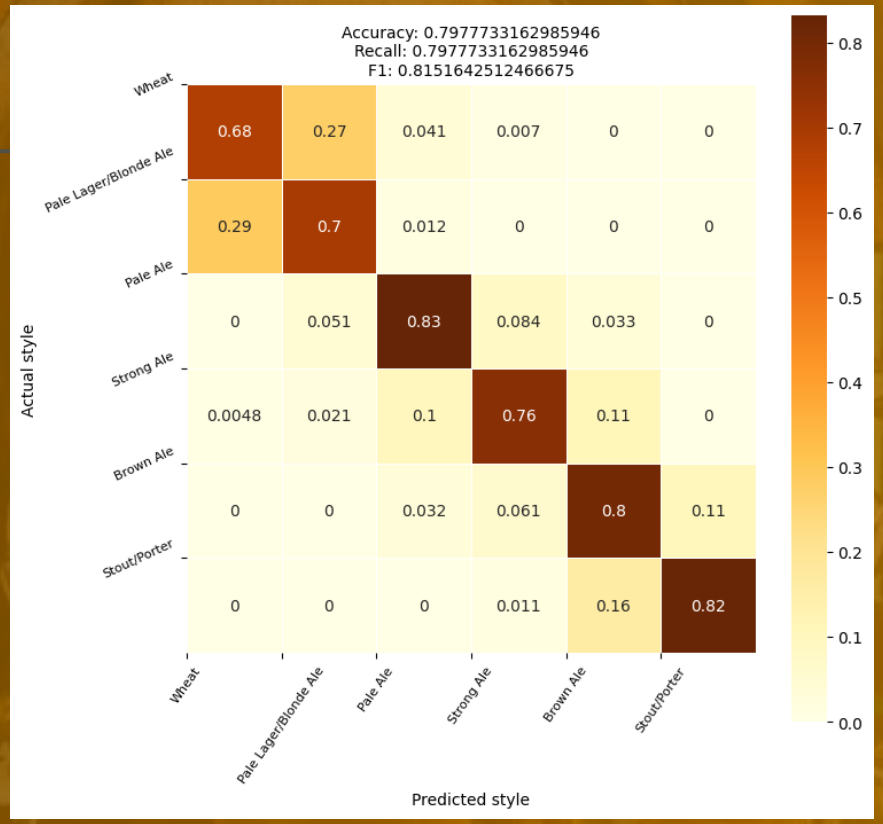
CORRELACIONES TRAS LA LIMPIEZA DE DATOS



MODELO DE PRUEBA: **REGRESIÓN LOGÍSTICA**

- Con y sin escalado de los datos (con dos *scalers* distintos)
- Separación en *train* y *test*
- Definición de X e y:
  - X = [['ABV', 'IBU', 'Color']]
  - y = ['Style']
- Utilización de RandomUnderSampler en todos los casos

Accuracy score	0.79
Precision score	0.84
F1 score	0.81
Recall score	0.79



PIPELINES Y GRID SEARCHES:

Primera prueba:

- Escalado de los datos con ColumnTransformer y MinMaxScaler
- SelectKBest
- Clasificadores:
  - RandomForest (profundidad)
  - SVC (C)
  - Gradient Boosting (profundidad)
  - KNN (neighbors 1-10)
  - Regresión logística
- (Refit: F1 weighted)

Segunda prueba:

- Sin escalado de los datos

GRID SEARCH INSIGHTS:

Mejor estimador:

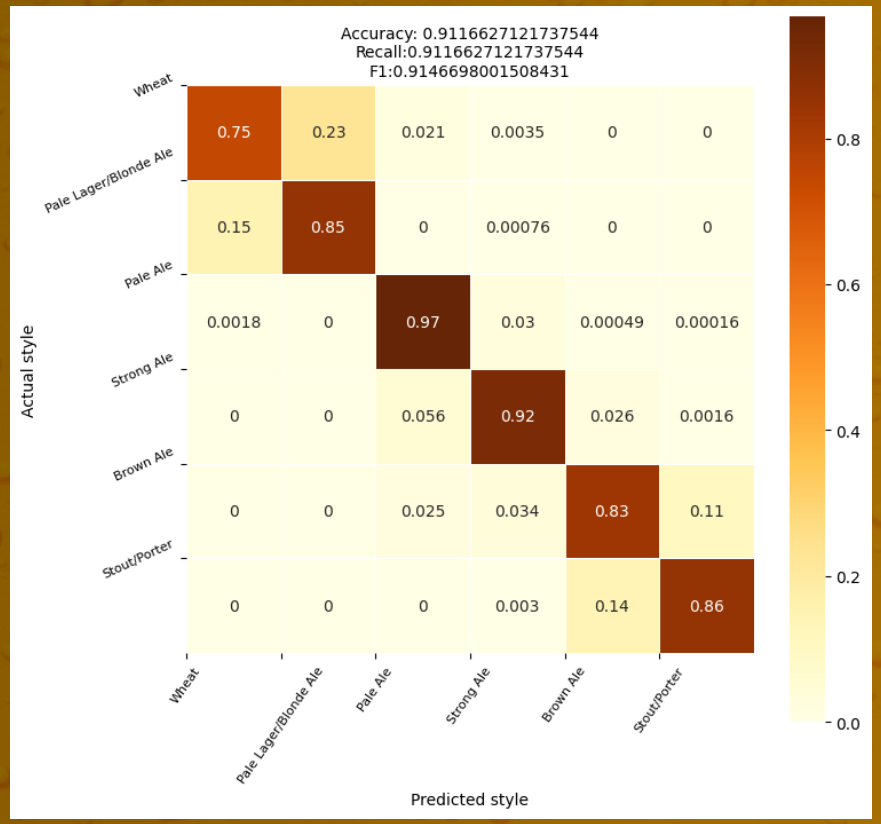
- Gradient Boosting con profundidad =3 y k=3 (cv=5)

Métricas:

F1(weighted) – Combinación de ambas métricas:

Precision – De las identificadas como PA, cuántas lo son

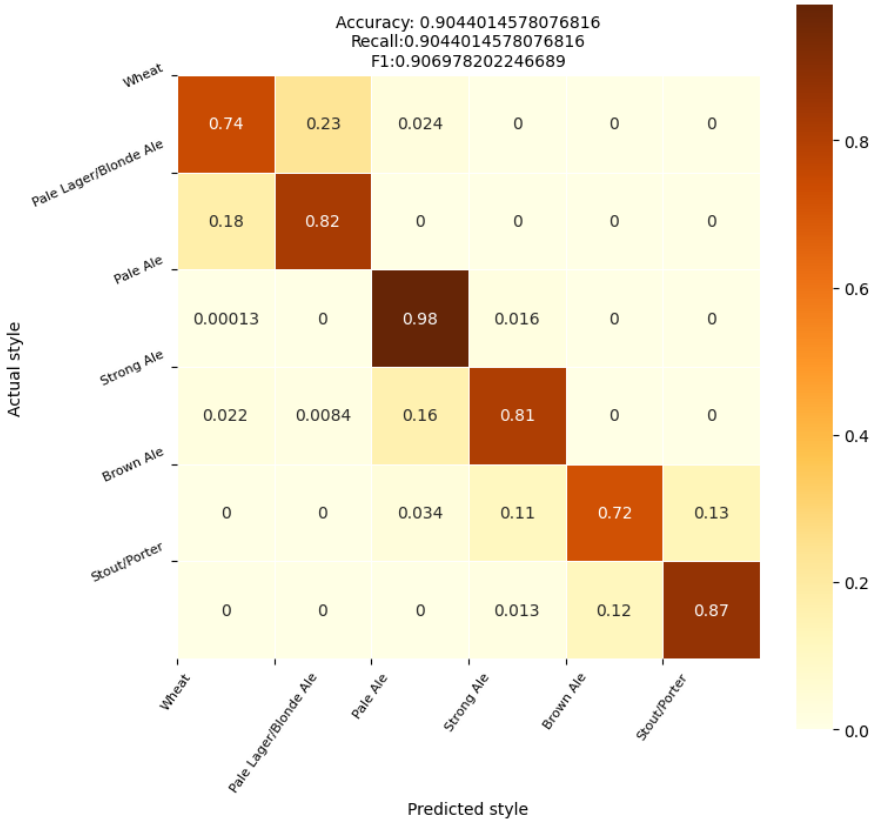
Recall – De todas las PA, cuántas identificó



Mejor *score*: 0,86

Accuracy score	0.91
Precision score	0.92
F1 score	0.91
Recall score	0.91

PIPELINES Y GRID SEARCHES II:



- Tercera prueba:
  - Sin escalado de los datos
  - Sabemos que SelectKBest = 3
  - Clasificadores:
    - SVC con nuevos parámetros (C y kernel)
    - Adaboost (learning rate)
      - Estimator: Decision Tree
        - Profundidad y criterios

3rd GRID SEARCH INSIGHTS:

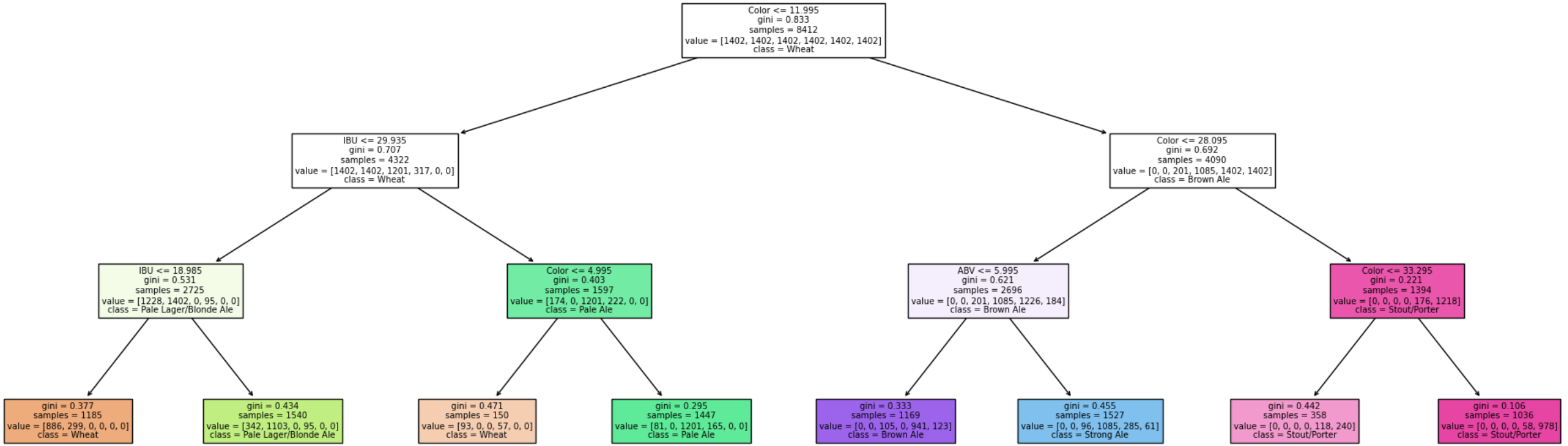
Mejor estimador:

- Adaboost (con Decision Tree)
  - profundidad=3, l.rate=0.01, criterion=gini

Accuracy score	0.90
Precision score	0.90
F1 score	0.90
Recall score	0.90

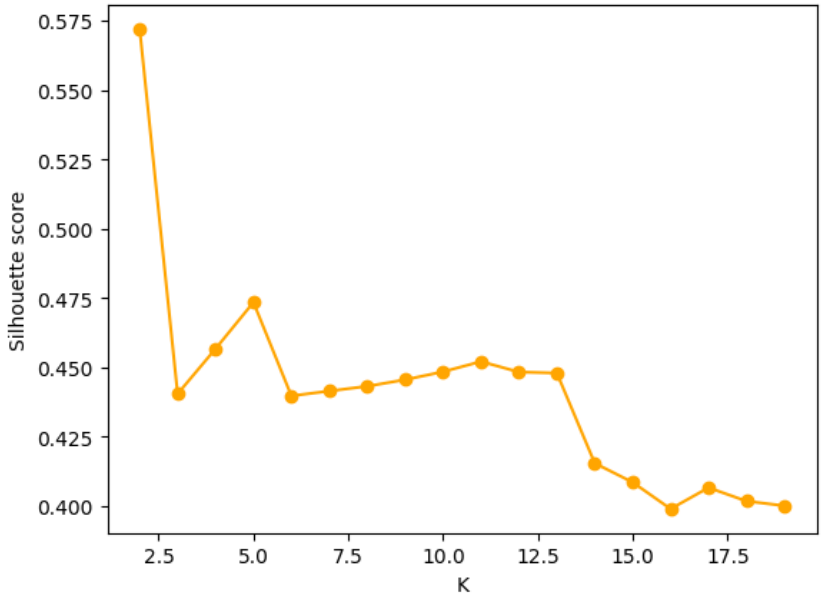
Mejor *score*: 0,81

ADABOOST DECISION TREE



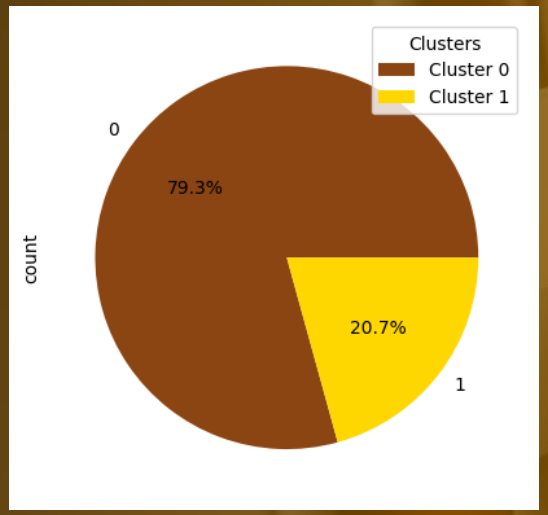
APRENDIZAJE NO SUPERVISADO: **CLUSTERING**

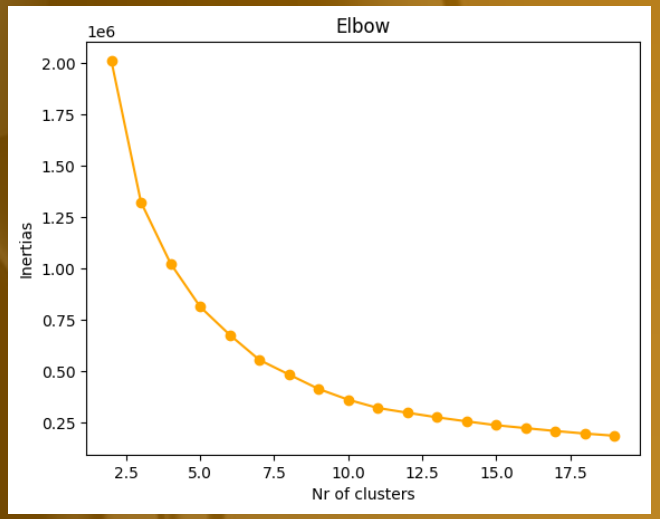
Silhouette score



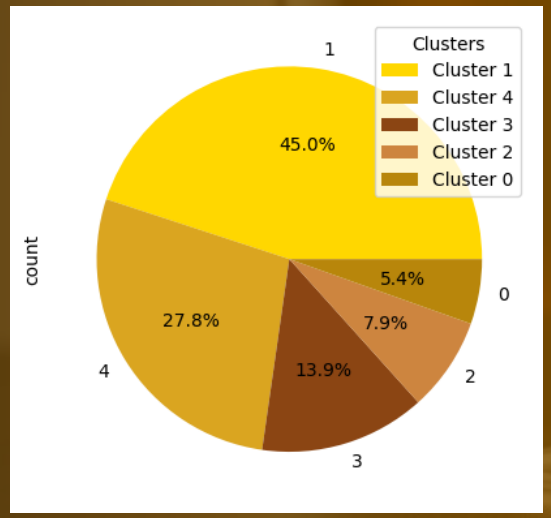
- Mejores valores de k: 2 (y 5)

k=2





k=5



PIPELINES Y GRID SEARCHES III:

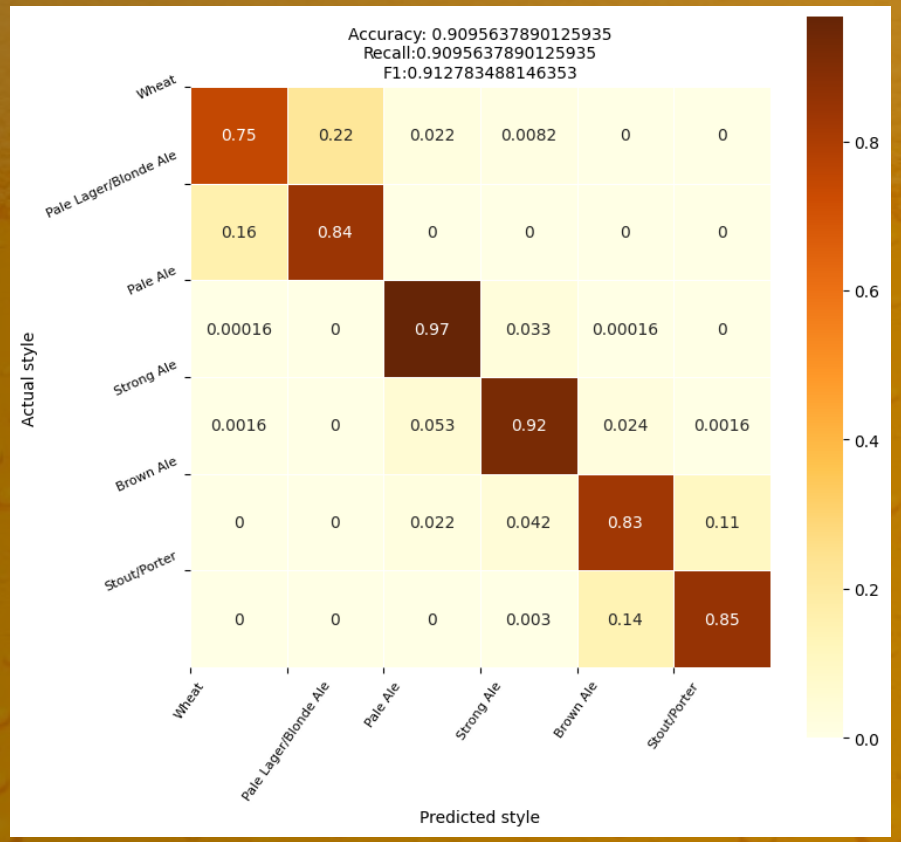
Últimas pruebas:

- Hiperparametrización del Gradient Boosting (mejor estimador)
- Randomized Search (cv=3) /(y GS)
- Sin escalado de los datos

RANDOMIZED SEARCH INSIGHTS:

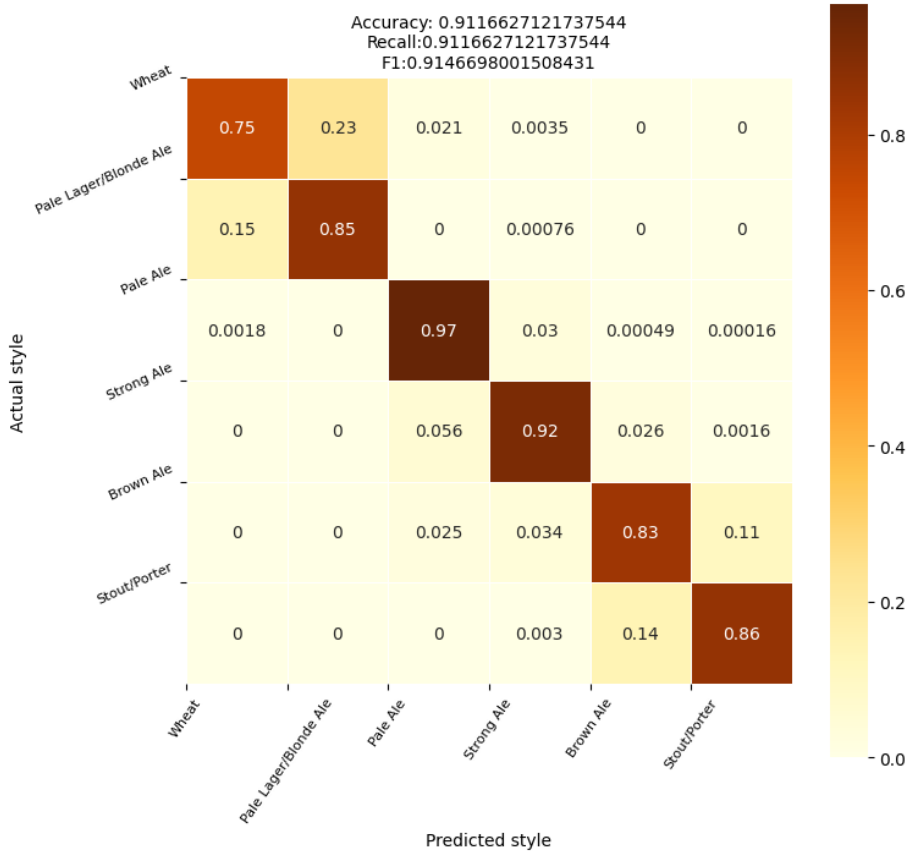
Gradient Boosting

- profundidad =2
- k=3 (cv=5)
- max leaf nodes: 3
- min simples leaf: 3
- n estimators: 200
- criterion: Friedman MSE
- min simples Split: 2



Mejor *score*: 0,86

MODELO FINAL



Model	Gradient Boosting
Max depth	3
SelectKBest	3
Best score	(0.86)
Accuracy score	0.91
Precision score	0.92
F1 score	0.91
Recall score	0.91

RESULTADOS Y PRÓXIMOS PASOS

Obtención de nuevos datos

Si bien el modelo falla más al clasificar las **cervezas rubias**, si hubiera una variable que indicase si es **de trigo o cebada** se mitigaría este error.

Modelos con NN

El próximo paso sería probar a entrenar una red neuronal.

RandomUnderSampler

He utilizado este método para equilibrar las clases, pero se debería probar con otros como *stratify* u otro método para asignar los estilos de forma más uniforme.

Restructurar los datos originales

Se podría clasificar los 175 estilos iniciales de muchas otras maneras hasta dar con la forma óptima (en la que se eliminen menos).

Escala de color

Aunque en la fuente de datos consta la escala SRM, sospecho que algunos usuarios introdujeron el valor en escala EBC, por lo que habría que convertir ciertos datos.

Enlace a Streamlit

¡SALUD!