

INTRODUCCION AL  
APRENDIZAJE AUTOMATICO

"GACHI, PACHI Y LOS HONGOS"

¿ME LO PUEDO COMER?

AHORRATE UN MAL VIAJE

CASTILLA VALENTÍN  
ESTEVEZ SALA FRANCISCO  
VÁZQUEZ LUCÍA



# ÍNDICE

01 PRESENTACIÓN DEL PROYECTO

02 DATASET

03 EXPLORACIÓN

04 LIMPIEZA

05 SELECCIÓN DEL MODELO

06 FINE TUNNING

07 MÉTRICAS FINALES

08 COMENTARIOS Y  
CONCLUSIÓN



# PROYECTO

LA INGESTA DE HONGOS VENENOSOS TIENE GRANDES CONSECUENCIAS, QUE VAN  
DESDE EFECTOS ALUCINOGENOS HASTA, EN CASOS EXTREMOS, LA MUERTE  
PARA EVITAR ESTO, SE BUSCA DESARROLLAR UN CLASIFICADOR  
QUE PERMITA IDENTIFICAR SI UN HONGO ES VENENOSO O  
COMESTIBLE

# DATASET

TIENE 61069 INSTANCIAS Y 21 COLUMNAS (20 FEATURES + TARGET)

NUESTRO DATASET LO ENCONTRAMOS EN EL REPOSITORIO DE MACHINE LEARNING DE LA UNIVERSIDAD UC IRVINE. SE BASA EN UNA SERIE DE MUESTRAS HIPOTETICAS BASADAS EN 173 ESPECIES DISTITAS DE HONGOS CON SOMBRERO. ESTA BASADO EN UN DATASET DE J. SCHLIMMER, CON MUESTRAS REALES DE HONGOS DE LA FAMILIA AGARICUS Y LEPIOTA.

SE INDICA EN LA DESCIPCION DEL DATASET, QUE LAS MUESTRAS FUERON CLASIFICADAS COMO COMESTIBLES, INDETERMINADOS Y NO RECOMENDADOS Y VENENOSOS, Y QUE ESTAS ULTIMAS DOS CLASES FUERON JUNTADAS DENTRO DE LA CLASE VENENOSA.

# TIPOS DE DATOS EN EL DATASET INICIAL

## COLUMNAS SIN DATOS FALTANTES

12 (11 FEATURES + TARGET)

	Tipo de dato
class	object
cap-diameter	float64
cap-shape	object
cap-color	object
does-bruise-or-bleed	object
gill-color	object
stem-height	float64
stem-width	float64
stem-color	object
has-ring	object
habitat	object
season	object

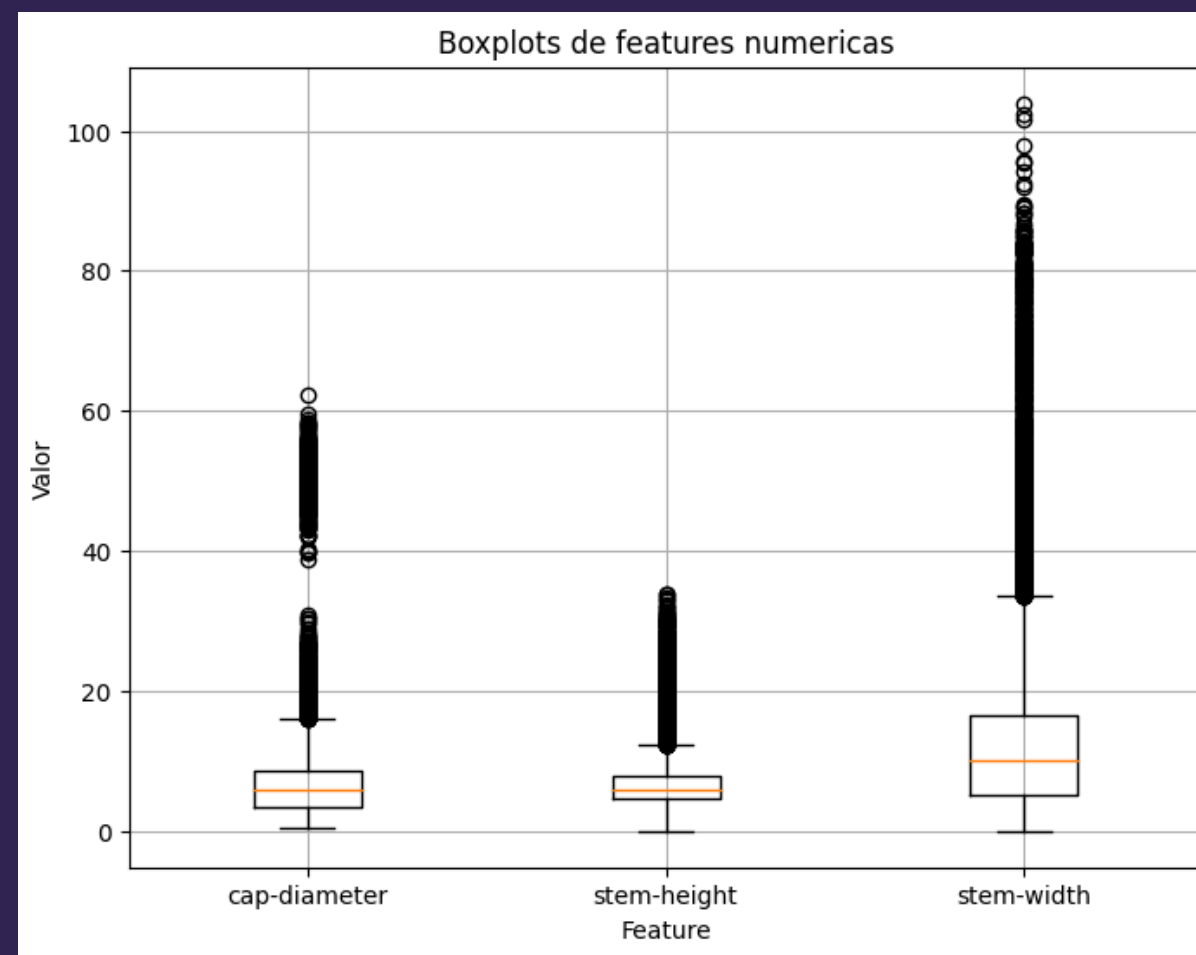
## COLUMNAS CON DATOS FALTANTES

9 (FEATURES)

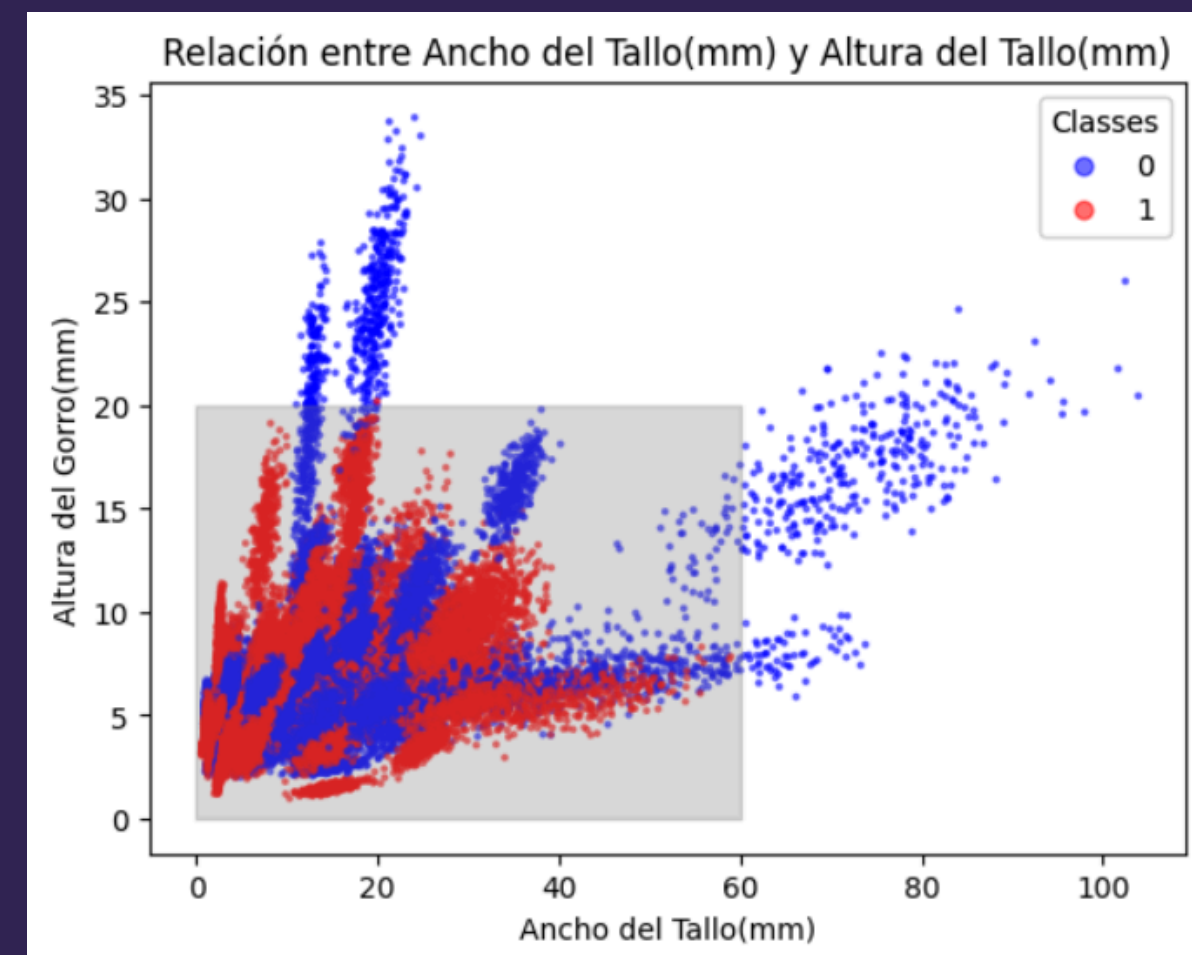
	Tipo de dato	Datos Faltantes	Datos No Faltantes	% datos faltantes
cap-surface	object	14120	46949	23.1214
gill-attachment	object	9884	51185	16.185
gill-spacing	object	25063	36006	41.0405
stem-root	object	51538	9531	84.3931
stem-surface	object	38124	22945	62.4277
veil-type	object	57892	3177	94.7977
veil-color	object	53656	7413	87.8613
ring-type	object	2471	58598	4.04624
spore-print-color	object	54715	6354	89.5954



# EXPLORACIÓN



REPRESENTACION DE LA DISPERSION DE LOS VALORES EN LAS COLUMNAS NUMERICAS



PRUEBA DE QUE EXISTE ALGUN TIPO DE SEPARACION POSIBLE ENTRE LAS CLASES DEPENDIENDO DE LOS ATRIBUTOS QUE SE UTILICEN

LIMPIEZA



**¿LA COLUMNA TIENE MAS DE UN 5% DE DATOS  
FALTANTES?**





# ¿LA FILA QUEDO CON ALGÚN DATO FALTANTE?



# ¿HAY DATOS QUE NO TIENEN SENTIDO?



# TRATAMOS TODAS LAS VARIABLES CATEGORICAS USANDO VARIABLES DUMMY Y ONE HOT


CON DUMMIES TRATAMOS EL TARGET Y OTRAS 3 COLUMNAS

EN LA UNICA VARIABLE CATEGORICA ORDINAL (SEASON) REEMPLAZAMOS LOS VALORES POR NÚMEROS

CON ONE HOT TRATAMOS LAS 6 COLUMNAS RESTANTES

## ESTO ULTIMO NOS CAUSO VARIOS PROBLEMAS AL MOMENTO DE PLANTEAR UN MODELO

### ¿POR QUE?

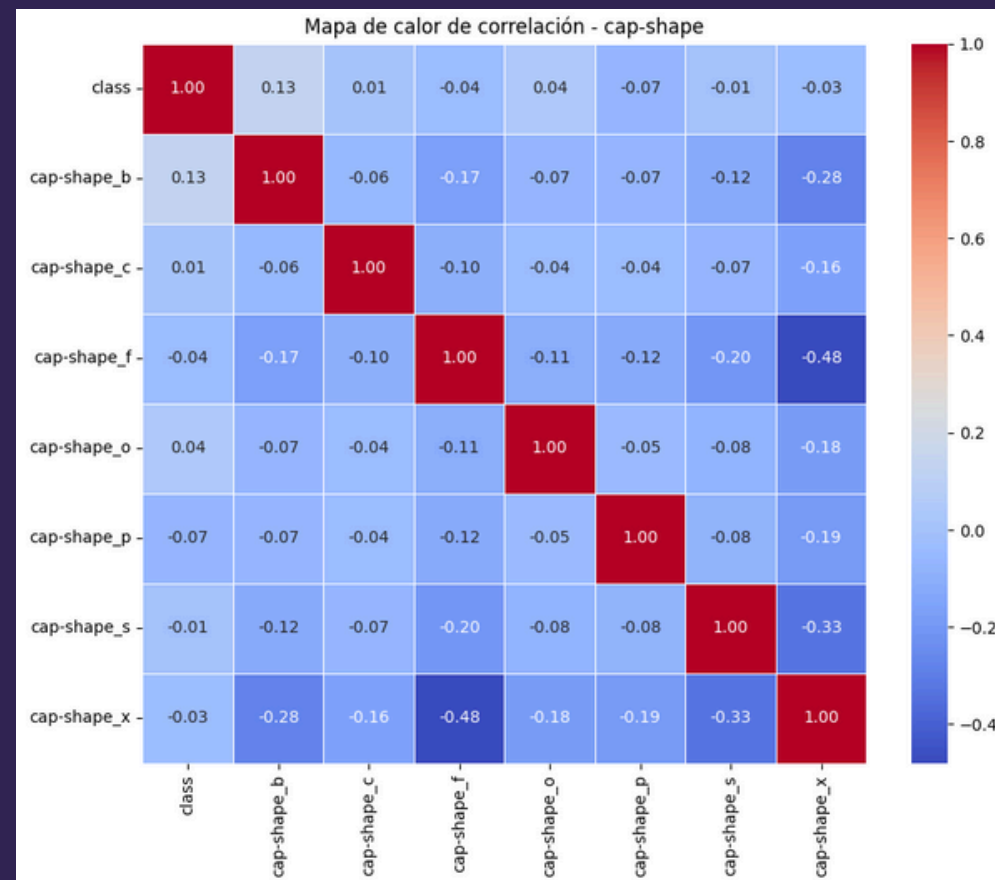


ENTRE LAS VARIBALES NUMERICAS, LAS VARIABLES QUE A LAS QUE PODEMOS APLICAR DUMMIES Y LAS CATEGORICAS ORDINALES, NOS QUEDA UN DATASET DE 57.539 INSTANCIAS Y 7 COLUMNAS (6 FEATURES + TARGET)

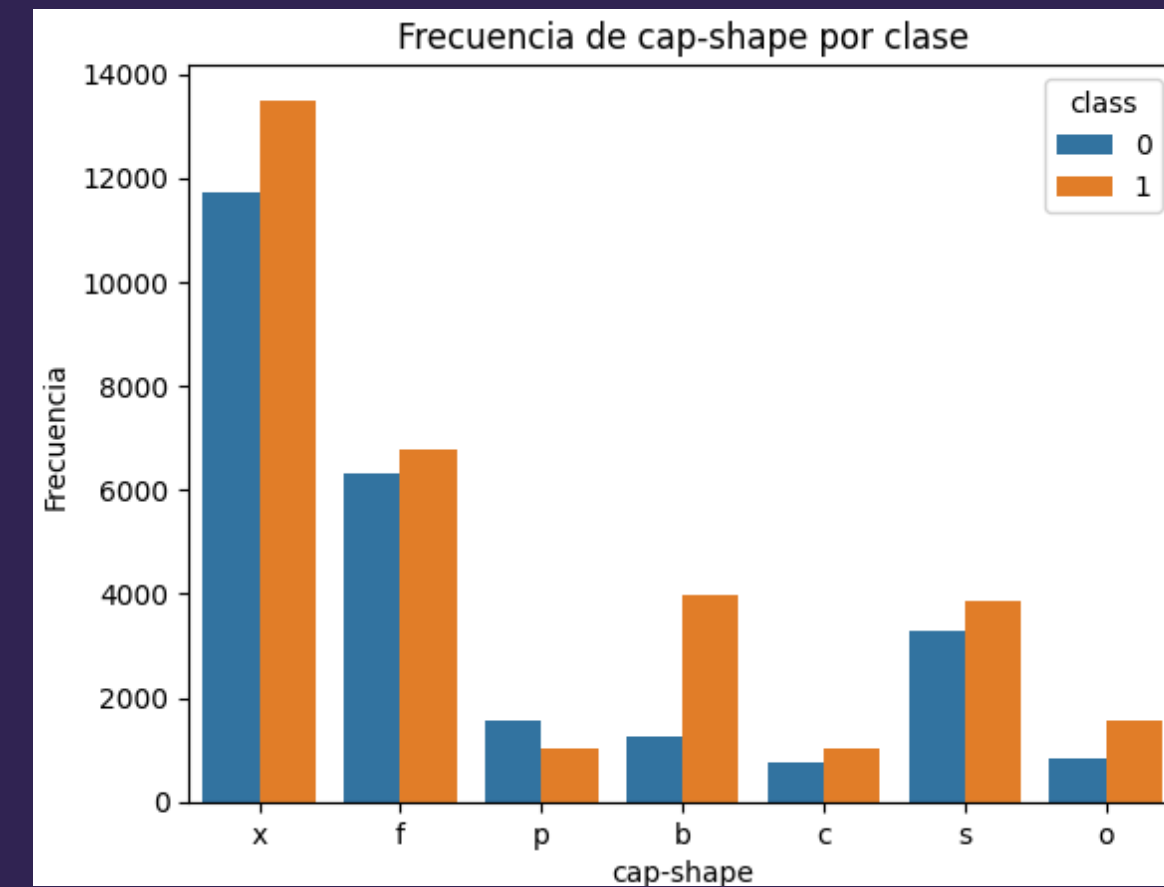
PERO AL APLICAR ONE-HOT AL RESTO DE LAS COLUMNAS COMPLETAS (LAS CATEGORICAS NOMINALES) NOS QUEDA UN DATASET DE 57.539 INSTANCIAS Y 66 COLUMNAS (65 FEATURES + TARGET)



# EXPLORACIÓN PTE. 2



REPRESENTACION DE LA CORRELACION ENTRE  
UN FEATURE Y EL TARGET



REPRESENTACION DE LA CANTIDAD DE  
INSTANCIAS DE CADA GRUPO DENTRO DE UN  
FEATURE QUE PERTENECEN A CADA CLASE



# MODELO

DECIDIMOS PLANTEAR DOS MODELOS INCIALES, UNO CON CADA DATASET, Y COMPARARLOS

ELEGIMOS USAR RANDOM FOREST

PENSAMOS QUE DENTRO DE LOS CLASIFICADORES ES EL QUE MAS SE ADAPTA AL TIPO DE DATASET CON LOS QUE ESTAMOS TRABAJANDO

# MODELO DATASET TRABAJADO

Matriz de confusión	Pred Positiva	Pred Negativa
Verdaderos Positivos	6663	993
Verdaderos negativos	964	8642

## Métricas de evaluación

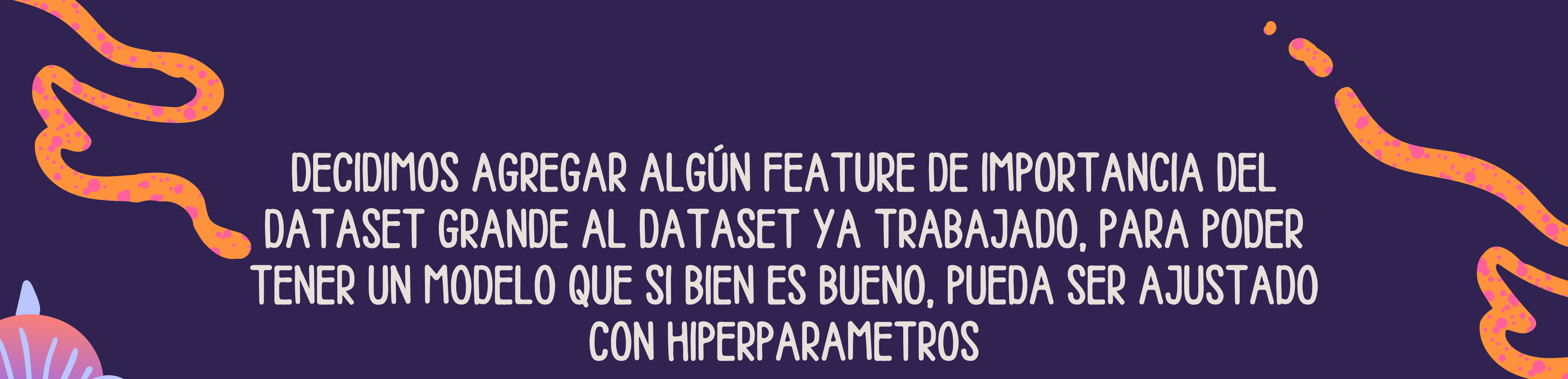
- Precision: 0.89
- Recall: 0.88
- F1-score: 0.89

# MODELO DATASET GIGANTE

Matriz de confusión	Pred Positiva	Pred Negativa
Verdaderos Positivos	7630	26
Verdaderos negativos	30	9756

## Métricas de evaluación

- Precision: 0.99
- Recall: 0.99
- F1-score: 0.99



DECIDIMOS AGREGAR ALGÚN FEATURE DE IMPORTANCIA DEL DATASET GRANDE AL DATASET YA TRABAJADO, PARA PODER TENER UN MODELO QUE SI BIEN ES BUENO, PUEDA SER AJUSTADO CON HIPERPARAMETROS

ESTA FUE: **STEM-COLOR**

Precision: 0.956

Recall: 0.959

F1-score: 0.958



# FINE TUNNING



PARA AJUSTAR LOS HIPERPARÁMETROS USAMOS EL MÉTODO GRID  
SEARCH. LA COMBINACIÓN DE HIPERPARÁMETROS CON MEJOR  
RESULTADO FUE LA SIGUIENTE:

```
{'max_depth': None, 'min_samples_leaf': 1,  
'min_samples_split': 5, 'n_estimators': 100}
```



# MÉTRICAS FINALES

Matriz de confusión	Pred Positiva	Pred Negativa
Verdaderos Positivos	7398	409
Verdaderos negativos	377	9078

## Métricas de evaluación

- Precision: 0.956
- Recall: 0.96
- F1-score: 0.958

# CONCLUSIONES

SI BIEN LAS METRICAS FUERON SIMILARES ANTES Y  
DESPUES DEL GRID SEARCH Y LA VALIDACION  
CRUZADA, ESTE SEGUNDO CLASIFICADOR ES MAS  
FUERTE PORQUE NOS ASEGURAMOS QUE LAS  
METRICAS NO SE RELACIONAN CON EL GRUPO DE  
TRAIN-TEST PARTICULAR, Y ADEMAS, MEJORO LA  
EXHAUSTIVIDAD, QUE ES LO MAS IMPORTANTE EN  
NUESTRO CASO PARTICULAR

\*Se inventa la máquina  
del tiempo\*



Yo:

NOOOO  
GACHI Y PACHI,  
NO USEN DATASETS  
SINTÉTICOS



GRACIAS POR ESCUCHARNOS

¿PREGUNTAS?

