Review sentiment prediction

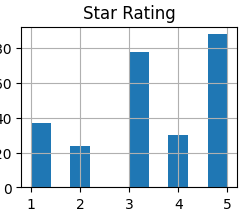
257 filas

Clasificación. Nos interesa el accuracy en fit porque todas las clases tienen la misma importancia.

También miraremos el recall de cada clase: la cantidad de instancias correctas que captura (FN) y la precisión: la pureza de la clase (FP). El F1-score resume esto 2\* p\*r (p+r)

Target:

Star rating: 5 clases. No muy desbalanceado.



Features numéricas:

Sentiment Value

Distribución normal

word count: numérica exponencial. Al aplicarle min max nos comprimirá el pico (mayoría de los datos en valores muy próximos a 0. Y la cola será todo de 0.5 en adelante)

No debería tener una influencia grande.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

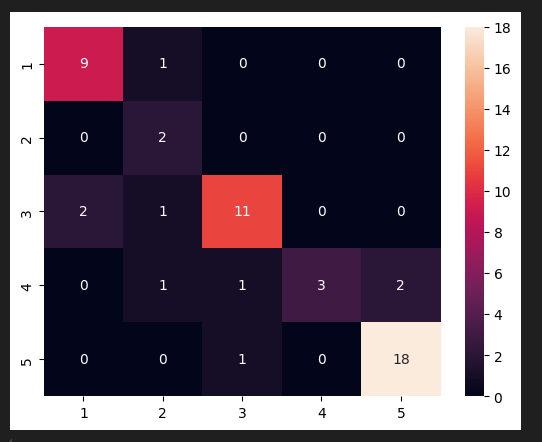
Train test split: 0.2

205 - 52

Pequeña contaminación en test (-0.01). El deterioro no es significativo y KNN no tiene problemas con valores negativos.

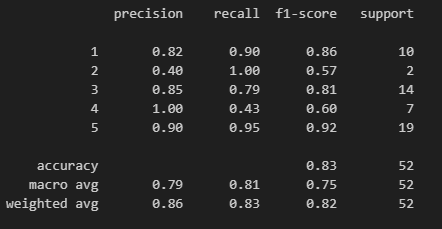
Baseline MinMax KNN = 5

acc = 0.82



Confunde los 4s (precision 0.43). 1/3 con 5s (no muy malo) y ¼ con 3s (no muy malo) y ¼ con 2s (peor). Teníamos muy pocos 4s en la muestra.

Hay un poco de leakage de 3s a 2s y 1s.



Macro avg es la suma ponderada de precision y recall. El f1-score nos resume precision y recall con la media geométrica.

Macro precission average (precisión ponderada).

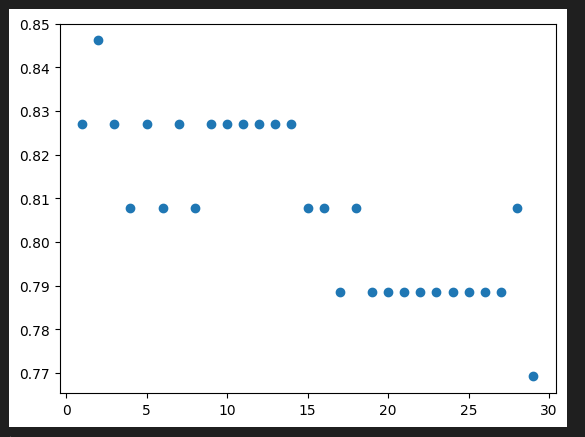
Las predicciones no parecen malas con esos dos valores si supuonemos que wordcount no es importante:

sentiment 2 -> 5 stars

sentiment -3 -> 1 star.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Método un fit(train) eval (test). Barrido en K de 1 a 30. Accuracy.

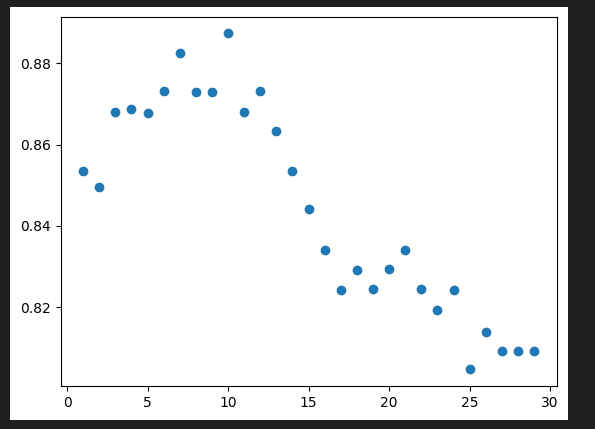


Óptimo 2 pero me quedo con 3 porque el deterioro no es significativo y será más estable. Generalizará mejor.

PROBLEMA test set demasiado pequeño.

Si hacemos un GridSearch casero fit(train) cross\_val(train) o un GridSearch con sklearn:

K óptimo = 10.



k= 10

Acc Train (cv) = 0.88

Acc Test = 082

Acc full (cv) = 0.81

k = 2

Acc full (cv) = 0.77

(buena estimación del error)

Gracias al cross-validation hemos elegido un k mejor. Al método realmente.