CRISP-DM: Clasificación de SPAM del SMS usando ML

1. Comprensión del Negocio

Un ejemplo ilustrativo de la metodología vamos a considerar el caso de una organización telefónica especializada en SMS que pretende mejorar la precisión en el mensaje que recibe sus usuarios, si es un tipo SPAM o no.

Para ello se procederá a realizar una captura contextos del mensaje y se medirá de forma automatizada una serie de atributos sobre el contexto que nos permita clasificar el contexto si es **SPAM** (mensaje de promoción) o **HAM** (mensaje enviado por otro usuario/compañía).

El **objetivo final** es establecer un sistema automatizado que permita clasificar automáticamente un mensaje del usuario y complemente el contexto, mejorando la precisión mensaje; Usando los métodos **Naive Bayesian**, **SVM** y **Random Forest Classifier**

Este proyecto se utilizamos la herramienta **Jupyter-Notebook Python**, las librerías que necesitan son **Sklearn**, **Gensim**, y **Pandas**

2. Comprensión y Análisis de Calidad de los datos

Usamos el dato público del conjunto de datos del SPAM en SMS, que no es puramente limpio. El dato consiste en dos columnas una de la clasificación (SPAM o HAM) y el otro del contexto.

```
ham What you doing?how are you?
ham Ok lar... Joking wif u oni...
ham dun say so early hor... U c already then say...
ham MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*
ham Siva is in hostel aha:-.
ham Cos i was out shopping wif darren jus now n i called him 2 ask wat present he wan lor. Then he started guessing who i was wif n he finally guessed darren lor.
spam FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! ubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop
spam Sunshine Quiz! Win a super Sony DVD recorder if you canname the capital of Australia? Text MQUIZ to 82277. B
spam URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU
```

Ilustración 1 ejemplos del dato

Se procede a realizar un análisis estadístico básico. A continuación, se muestra el cálculo de la media, la desviación estándar y los cuartiles para las primeras nueve variables numéricas.

	len								n_words							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
class																
ham	4825.0	71.775337	58.783939	2.0	33.0	52.0	93.0	910.0	4825.0	14.726010	11.974647	1.0	7.0	11.0	19.0	175.0
spam	747.0	139.180723	29.067007	13.0	133.0	149.0	158.0	224.0	747.0	24.463186	6.001311	2.0	22.0	26.0	28.0	37.0

Ilustración 2 Análisis estadístico básico

Usando **Deep Learning**

	acc	loss	val_acc	val_loss
count	28.000000	28.000000	28.000000	28.000000
mean	0.940086	0.145030	0.949952	0.139811
std	0.094867	0.169571	0.056867	0.132794
min	0.528290	0.010197	0.852018	0.046070
25%	0.905533	0.022527	0.905269	0.053173
50%	0.983300	0.055283	0.984305	0.061012
75%	0.993519	0.257297	0.986547	0.249363
max	0.997757	0.687664	0.991031	0.449868

Ilustración 3 Análisis estadístico básico (Deep Learning)

3. Preparación de los datos

Como ejemplo de exploración visual usamos grafico del Función de distribución (FDA) y Diagrama de caja (boxplot) para demostrar la distribución del SMS.

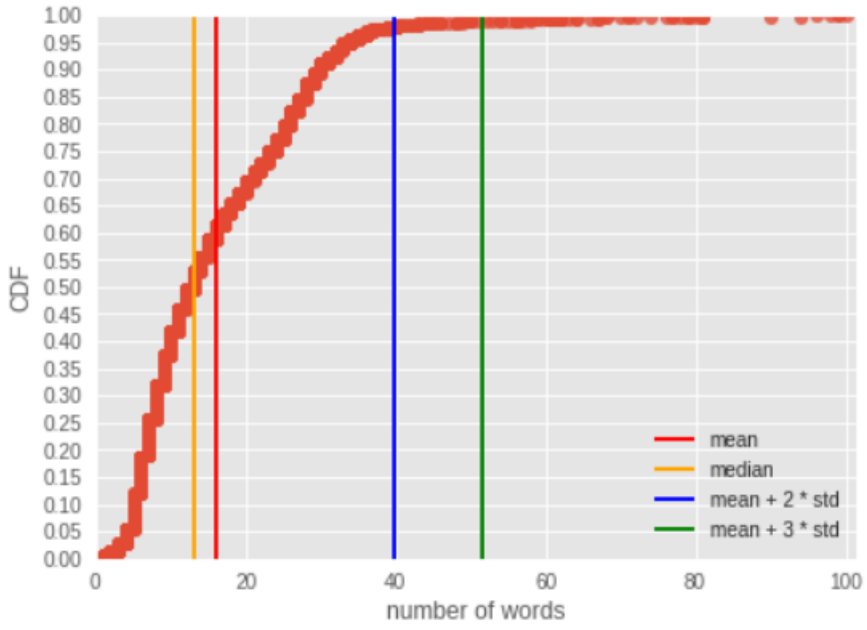

Ilustración 4 números de palabras

Según la ilustración 4:

- 50% del corpus consta de contextos que tienen menos de 13 palabras.
- 90% del corpus consta de contextos que tienen menos de aproximadamente 32 palabras.

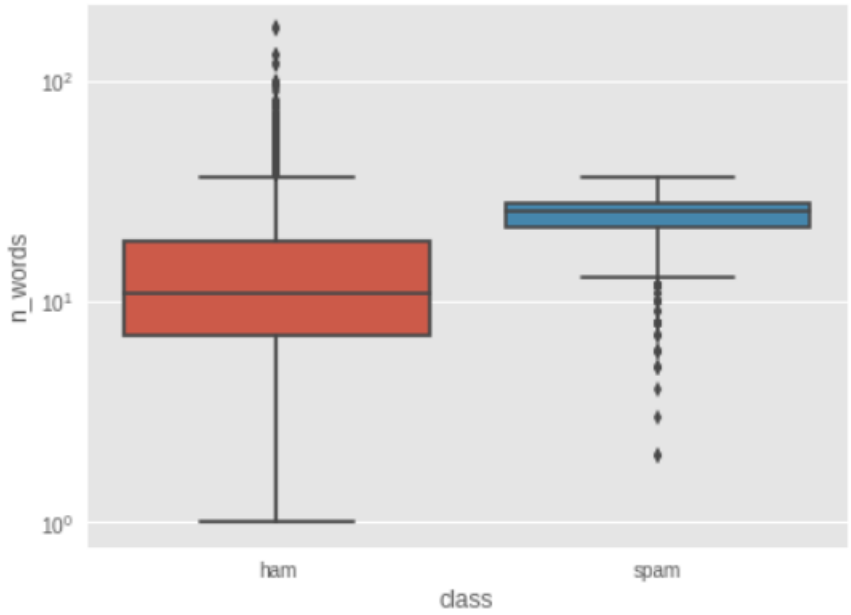

Ilustración 5 clasificación del SMS por números de palabras

Según ilustración 5:

- Las clases (HAM, SPAM) tienen una distribución identificada de forma única, claramente.

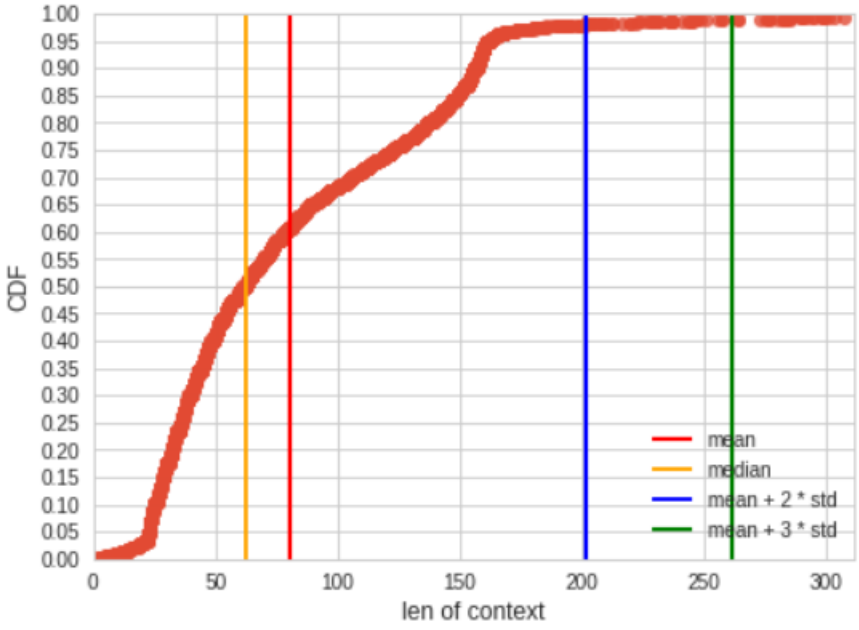

Ilustración 6 Longitudes del contexto

Según ilustración 6:

- El 50% del corpus consta de contextos cuya longitud es inferior a aproximadamente 62.
- El 90% del corpus consta de contextos cuya longitud es inferior a aproximadamente 155.

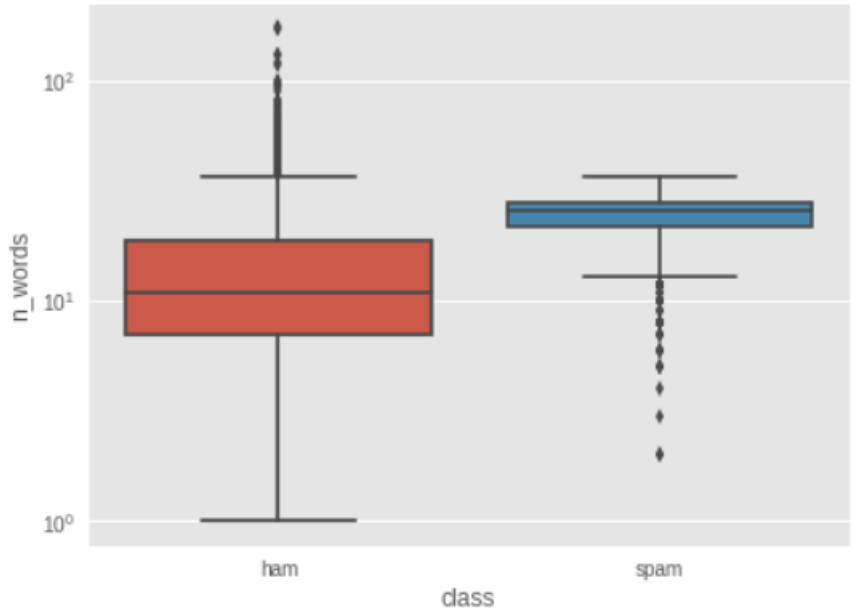

Ilustración 7 clasificación del SMS por longitudes de contexto

Según ilustración 7:

- Las clases (HAM, SPAM) tienen una distribución identificada de forma única, claramente.

4. Modelado

Nos encontramos ante un problema de clasificación. Aplicaremos un método de aprendizaje supervisado, concretamente el clasificador **GridSearchCV**, para usar los hiper parámetros óptimos posteriormente en la etapa de prueba.

Después de seleccionar los hiper parámetros óptimos, cada modelo correspondiente a cada tubería se guardó en formato binario como un archivo **Pickle**. Para probar cualquier instancia en cualquier modelo, esos archivos binarios se pueden usar después de aplicar la deserialización, fácilmente

5. Predicción y Evaluación del Modelo

Verificamos nuestra evaluación de desempeño utilizando una matriz de confusión. Usamos diferentes métricas, como **precisión**, **recuperación**, **puntaje f1** y **exactitud**.

También nos ocupamos del número de Verdadero Positivo (TP), Verdadero Negativo (TN), Falso Positivo (FP), Falso Negativo (FN) en la matriz de confusión.

NB está dando los mejores resultados de acuerdo con la métrica de precisión en las etiquetas de spam además de los datos de prueba. Porque su SMS de HAM no está etiquetado como SPAM en absoluto. La clasificación está funcionando completamente bien si nos enfocamos en el número de FP (HAM vs. SAPM). Nadie no quiere encontrar el SMS relevante (HAM) en su casilla de correo no deseado. Por lo tanto, la precisión del spam es importante para ese caso porque tiene la mejor métrica de precisión (1.00).

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- F1 = 2 * Precision * Recall / (Precision + Recall)

Ilustración 8 Formularios para SPAM vs HAM

```
'--------------------------------------NB--------------------------------------'
--------------------Testing Performance--------------------
             precision    recall  f1-score   support

        ham       0.95      1.00      0.97       965
       spam       1.00      0.63      0.77       149

avg / total       0.95      0.95      0.95      1114

acc:  0.950628366248
--------------------Training Performance--------------------
             precision    recall  f1-score   support

        ham       0.96      1.00      0.98      3860
       spam       1.00      0.75      0.86       598

avg / total       0.97      0.97      0.97      4458
```

Ilustración 9 Evaluación por Naive Bayes classifiers

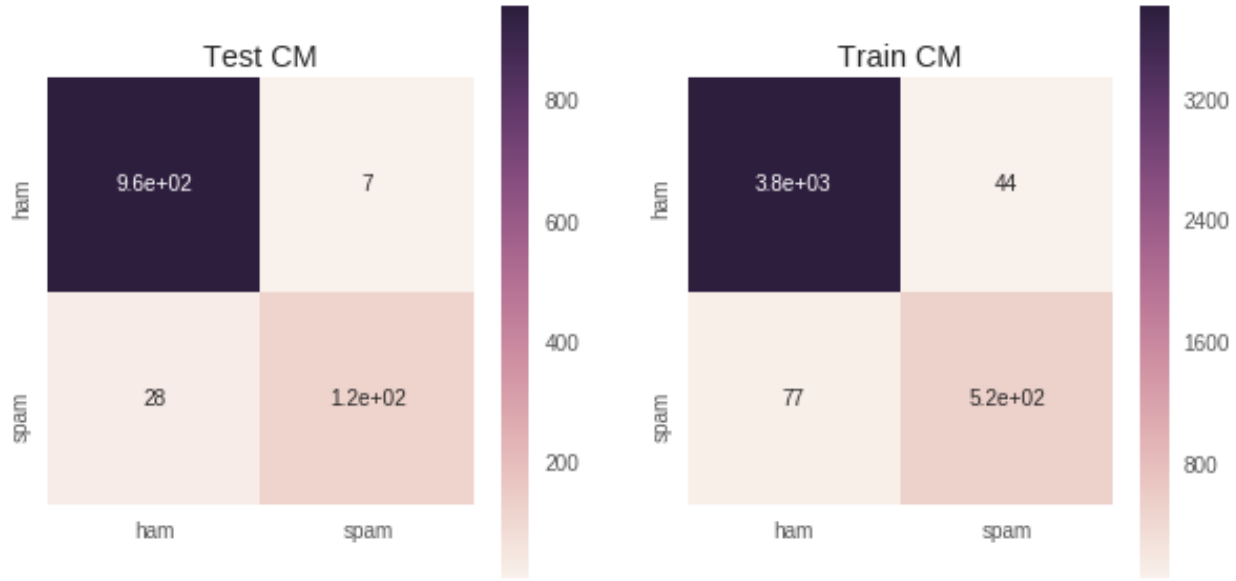

Ilustración 10 Matriz de confusión para NB

```
'----------------------------------------SVM----------------------------------------'
--------------------Testing Performance--------------------
             precision    recall  f1-score   support

        ham       0.97      0.99      0.98       965
       spam       0.95      0.81      0.87       149

avg / total       0.97      0.97      0.97      1114

acc:  0.968581687612
--------------------Training Performance--------------------
             precision    recall  f1-score   support

        ham       0.98      0.99      0.98      3860
       spam       0.92      0.87      0.90       598

avg / total       0.97      0.97      0.97      4458

acc:  0.97285778376
```

Ilustración 11 Evaluación por Máquinas de vectores de soporte

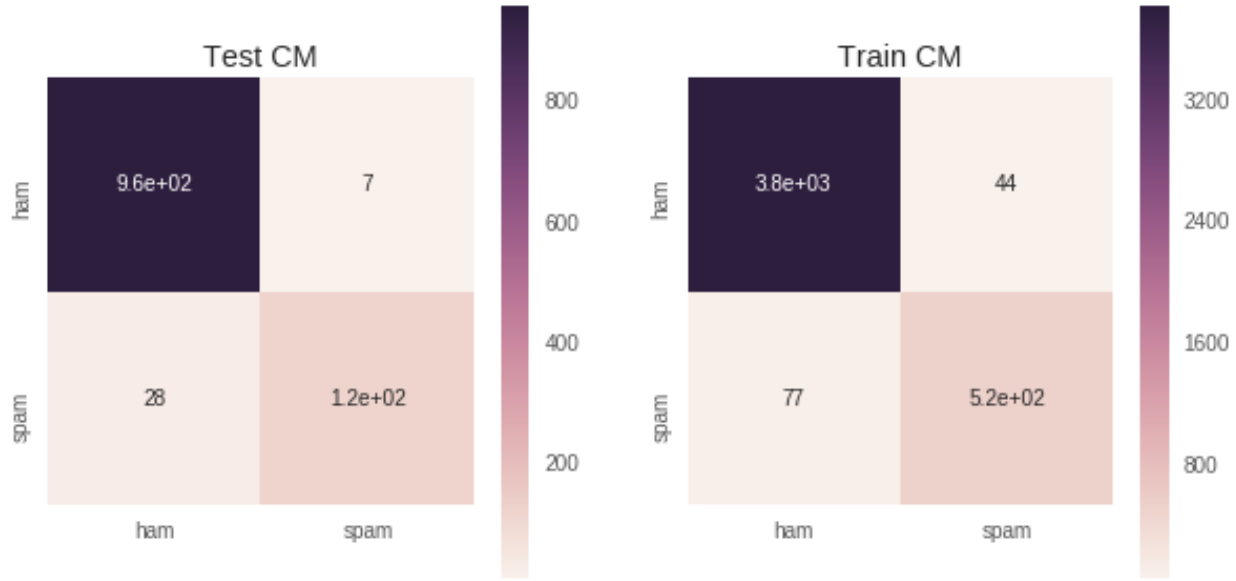

Ilustración 12 Matriz de confusión para SVM

```
'----------------------------------------RFT----------------------------------------'
--------------------Testing Performance--------------------
             precision    recall  f1-score   support

        ham       0.96      0.99      0.97       965
       spam       0.95      0.70      0.81       149

avg / total       0.96      0.96      0.95      1114

acc:  0.955116696589
--------------------Training Performance--------------------
             precision    recall  f1-score   support

        ham       1.00      1.00      1.00      3860
       spam       1.00      0.98      0.99       598

avg / total       1.00      1.00      1.00      4458

acc:  0.99730820996
```

Ilustración 13 Evaluación por Random Forest

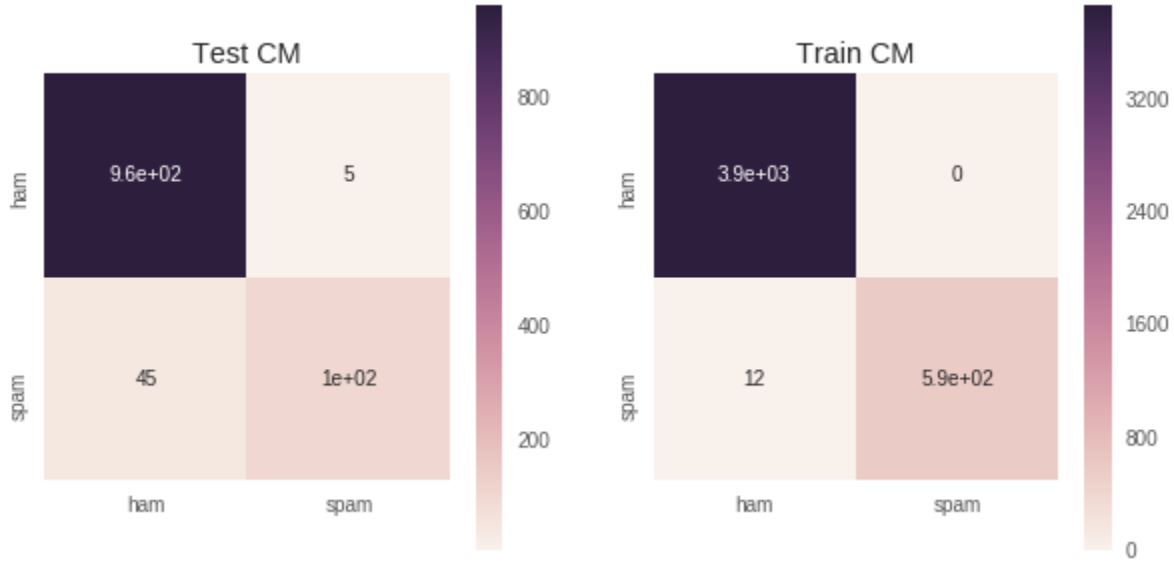

Ilustración 14 Matriz de confusión para Random Forest

6. Despliegue e implementación

Para despliegue solo necesita entra **Jupyter-Notebook Python** y luego subir el archivo *model.ipynb*, luego de eso le generaría un archivo **Pickle**, y del **dataset**

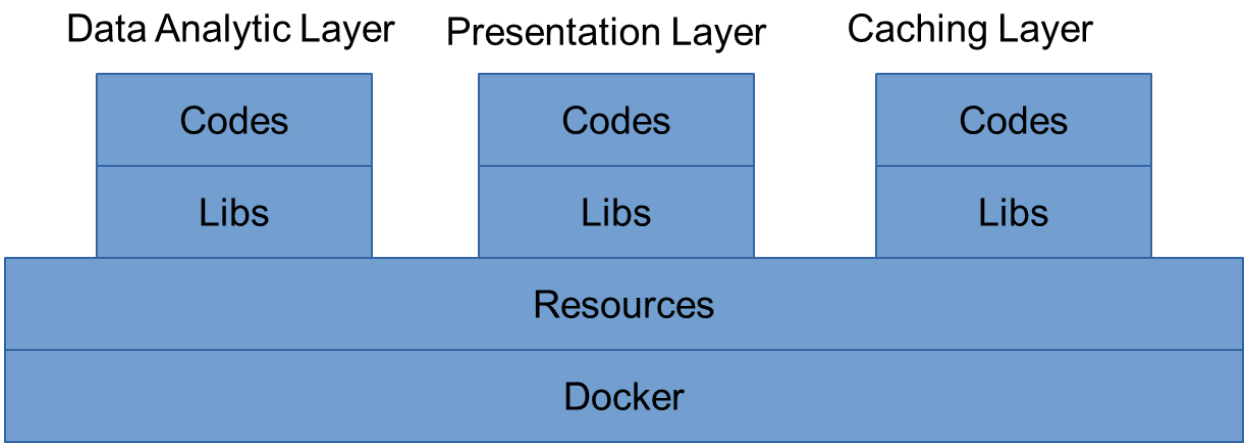

Ilustración 15 Arquitectura del proyecto

7. Gestión del Modelo

El modelo se genera por cada modificación en repositorio ya que es un proyecto **Open Source**, todas las personas tienen permiso de contribuirlo.

8. Referencia

Repositorio en Github