



# Instituto Politécnico Nacional

ESCUELA SUPERIOR DE COMPUTO

LICENCIATURA EN CIENCIA DE DATOS

**Aprendizaje de Máquina e Inteligencia Artificial**

Grupo: 4AM1

## **Práctica IV: Clasificación de Texto**

### **Unidad 2: Aprendizaje Supervisado**

FECHA DE ENTREGA: 27 DE ABRIL DE 2022

**Profesor:**

Joel Omar Gambino Juárez

**Alumno:**

Alcibar Zubillaga Julián - De Luna Ocampo Yanina - Ramirez Mendez Kevin - Sainz Takata Juan  
Pablo Minoru

# Índice

<b>1. Introducción</b>	<b>3</b>
<b>2. Desarrollo.</b>	<b>4</b>
2.1. Experimentos sobre el corpus de entrenamiento. . . . .	4
2.2. Resultados sobre el conjunto de prueba. . . . .	4
<b>3. Código Fuente</b>	<b>4</b>
<b>4. Conclusiones.</b>	<b>6</b>

# Aprendizaje de Máquina e Inteligencia Artificial

## Práctica IV: Clasificación de Texto

27 de abril de 2022

### Resumen

Esta practica sirve como preámbulo introductorio al proceso de clasificación de textos con la finalidad de obtener una polaridad de opinion de los mismos, para esto ocupamos un modelo de aprendizaje automático que nos permita clasificar y tener herramientas para procesar el lenguaje natural en nuestro idioma.

## 1. Introducción

Una tarea común dentro del área del Aprendizaje Automático es la clasificación de textos, se utiliza para procesar grandes volúmenes de texto no estructurado o texto que no tiene un formato predefinido, para derivar conocimientos y patrones. Con esto se puede hacer un análisis de reputación de empresas, productos o personalidades, estudios de impacto de marketing, extracción de opiniones y predicción de tendencias, en esta practica, como ya se menciono, utilizaremos las opiniones para determinar la polaridad, muchas veces esta se necesita obtener con base en reviews o etiquetar documentos por su contenido, a estas tareas se les da solución con herramientas como el NLP y el modelos de clasificación.

Para poder analizar lenguaje natural necesitamos que de alguna forma la maquina analice las características de los diversos textos a analizar. Para esto hacemos uso de el Procesamiento de Lenguaje Natural, esta rama de las ciencias cognitivas tiene como finalidad cerrar la brecha entre la comunicación humana y el entendimiento de las computadoras.

Apoyandose de los siguientes conceptos:

**Tokenizar.** Dividir el texto en palabras o frases.

**Lematizar.** Normalizar las palabras para asignar distintas formas a la palabra canónica con el mismo significado. Por ejemplo, “corriendo” se asigna a “correr”.

Una maquina puede transformar texto a una forma que sea analizable haciendo posible iniciar la clasificación de textos que se quiere lograr.

## 2. Desarrollo.

### 2.1. Experimentos sobre el corpus de entrenamiento.

<i>Representación</i>	<i>Accuracy Promedio</i>	<i>Precisión Promedio</i>	<i>Recall Promedio</i>	<i>F-measure Promedio</i>
Binarizado	0.8776691067755757	0.8691279457409263	0.8776691067755757	0.8719938985571407
Frecuencia	0.862853760981004	0.8482412973105129	0.862853760981004	0.8522888233019662

Tabla 1: Predicción de Polaridad

<i>Representación</i>	<i>Accuracy Promedio</i>	<i>Precisión Promedio</i>	<i>Recall Promedio</i>	<i>F-measure Promedio</i>
Binarizado	0.8830907612262823	0.8846448518184281	0.8830907612262823	0.881237093954998
Frecuencia	0.8832976977402872	0.8860189437816083	0.8832976977402872	0.8815085405125123

Tabla 2: Predicción de Lugar

### 2.2. Resultados sobre el conjunto de prueba.

<i>Representación</i>	<i>Accuracy Promedio</i>	<i>Precisión Promedio</i>	<i>Recall Promedio</i>	<i>F-measure Promedio</i>
¡Mejor Representacion!	0.8879510095994704	0.8802383997455967	0.8879510095994704	0.8830636677687187

Tabla 3: Predicción de Polaridad

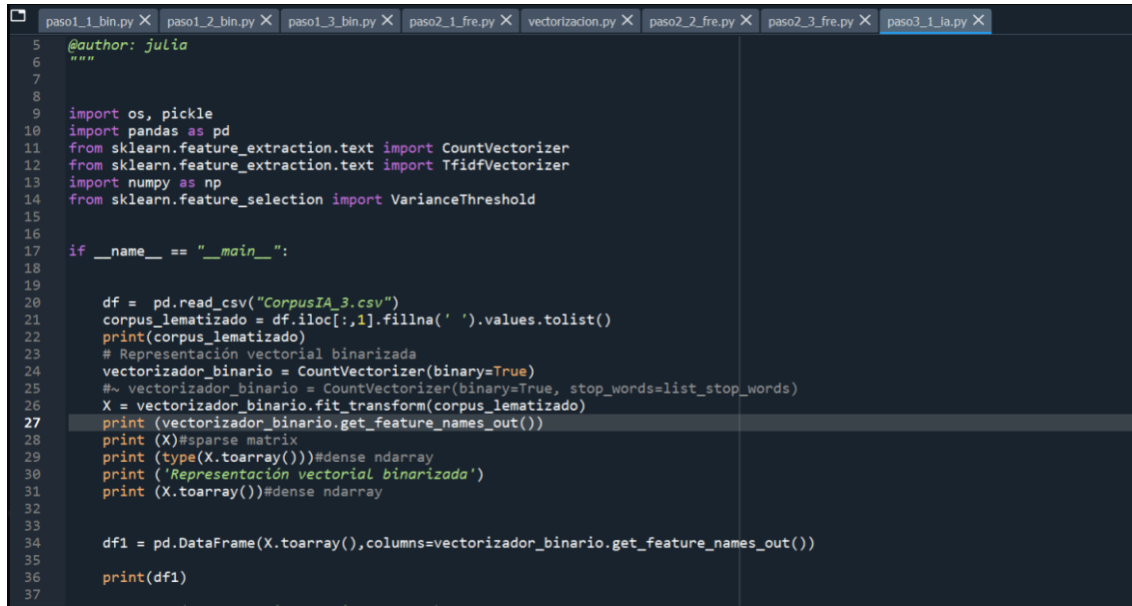
<i>Representación</i>	<i>Accuracy Promedio</i>	<i>Precisión Promedio</i>	<i>Recall Promedio</i>	<i>F-measure Promedio</i>
¡Mejor Representacion!	0.8901026150281364	0.8925976764256331	0.8901026150281364	0.8884187966432017

Tabla 4: Predicción de Lugar

## 3. Codigo Fuente

Procederemos a poner partes del código que consideramos de mayor importancia al momento de realizar la práctica, el análisis y la obtención de los resultados previamente visualizados en las tablas.

En este pedazo de código, podemos observar cómo con las librerías declaradas para el uso de esta práctica, se llama el archivo mediante pandas que utilizaremos, una vez hecho eso se lematiza y se vectoriza.



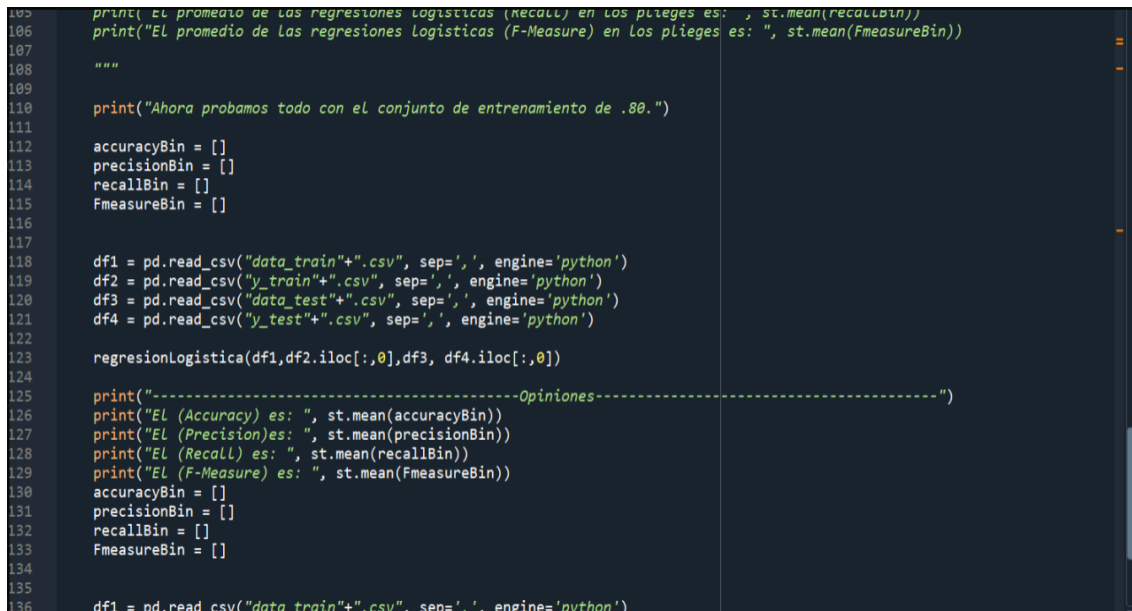
```

5 @author: julia
6 """
7
8
9 import os, pickle
10 import pandas as pd
11 from sklearn.feature_extraction.text import CountVectorizer
12 from sklearn.feature_extraction.text import TfidfVectorizer
13 import numpy as np
14 from sklearn.feature_selection import VarianceThreshold
15
16
17 if __name__ == "__main__":
18
19
20     df = pd.read_csv("CorpusIA_3.csv")
21     corpus_lematizado = df.iloc[:,1].fillna(' ').values.tolist()
22     print(corpus_lematizado)
23     # Representación vectorial binarizada
24     vectorizador_binario = CountVectorizer(binary=True)
25     #~ vectorizador_binario = CountVectorizer(binary=True, stop_words=list_stop_words)
26     X = vectorizador_binario.fit_transform(corpus_lematizado)
27     print(vectorizador_binario.get_feature_names_out())
28     print(X)#sparse matrix
29     print(type(X.toarray()))#dense ndarray
30     print('Representación vectorial binarizada')
31     print(X.toarray())#dense ndarray
32
33
34     df1 = pd.DataFrame(X.toarray(),columns=vectorizador_binario.get_feature_names_out())
35
36     print(df1)
37

```

Figura 1: Código de la lematización.

En esta captura, lo que se visualiza, es la obtención de: precision, accuracy, recall y F-measure, probando todo con el conjunto de entrenamiento. Asimismo, lo que se obtuvo se plasmó en las tablas previas a estas imágenes.



```

106 print("El promedio de las regresiones logisticas (Recall) en los plieges es: ", st.mean(recallBin))
107 print("El promedio de las regresiones logisticas (F-Measure) en los plieges es: ", st.mean(FmeasureBin))
108
109 """
110
111 print("Ahora probamos todo con el conjunto de entrenamiento de .80.")
112
113 accuracyBin = []
114 precisionBin = []
115 recallBin = []
116 FmeasureBin = []
117
118 df1 = pd.read_csv("data_train"+"csv", sep=',', engine='python')
119 df2 = pd.read_csv("y_train"+"csv", sep=',', engine='python')
120 df3 = pd.read_csv("data_test"+"csv", sep=',', engine='python')
121 df4 = pd.read_csv("y_test"+"csv", sep=',', engine='python')
122
123 regresionLogistica(df1,df2.iloc[:,0],df3, df4.iloc[:,0])
124
125 print("-----Opiniones-----")
126 print("EL (Accuracy) es: ", st.mean(accuracyBin))
127 print("EL (Precision)es: ", st.mean(precisionBin))
128 print("EL (Recall) es: ", st.mean(recallBin))
129 print("EL (F-Measure) es: ", st.mean(FmeasureBin))
130 accuracyBin = []
131 precisionBin = []
132 recallBin = []
133 FmeasureBin = []
134
135
136 df1 = pd.read_csv("data_train"+"csv", sep=',', engine='python')

```

Figura 2: Código de valores a analizar.

## 4. Conclusiones.

Con esta práctica pusimos a prueba nuestros conocimientos con los temas observados durante clase en un solo dataset. Entendimos y recordamos las metas que esta tiene, el primero es la posibilidad de ocurrencia de un evento, la existencia o no de varios componentes y el costo o elemento de los mismos y como numero dos, establecer el modelo más ajustado que describa la relación entre la variable contestación y un grupo de cambiantes regresoras.

## Referencias

[Abu-Mostafa] Abu-Mostafa, Y. S. Learning From Data: A Short Course. Recuperado de: <https://work.caltech.edu/lectures.html>. Fecha de consulta: 20/03/2022.

[Scikit-Learn] Scikit-Learn. Linear model: Api reference. [https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear\\_model](https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model). Fecha de consulta: 20/03/2022.