

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN



UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS



FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Materia

Datos Masivos

Profesora

M.C. Mayra Cristina Berro (McKinney, 2010)nes Reyes

Tarea de la clase 22-06-2023

Muestreo

Alumna

I.M. María Luisa Argáez Salcido

Matrícula

2173261

Fecha

29-06-2023

Índice

Comparación del rendimiento de un modelo de clasificación de emociones positivas y negativas utilizando distintos muestreos	3
Instrucción.....	3
Introducción	3
Análisis exploratorio de datos	3
Metodología.....	4
Preprocesamiento y modelado	5
Resultados.....	5
Conclusiones	6

Comparación del rendimiento de un modelo de clasificación de emociones positivas y negativas utilizando distintos muestreos

Instrucción

Se seleccionó la opción 1 como la tarea que se entregará. La opción 1 menciona:

Utilizando una de las bases de datos proporcionadas, seguir un objetivo y generar un muestreo probabilístico, y uno sesgado intencionalmente o híbrido. Compara los resultados y menciona tus conclusiones.

Introducción

La evaluación del rendimiento de un modelo es de las piezas finales y más importantes al realizar un análisis predictivo, por tanto, encontrar diferentes maneras de maximizar su rendimiento es un paso importante. El rendimiento de un modelo de clasificación, por lo general se ve reflejado en tiempo y en la matriz de confusión, así como las diversas métricas que se derivan de esta.

Este documento tiene por principal objetivo realizar una comparación entre dos métodos de muestreo, de tal manera que ambos utilizaran un mismo preprocesamiento de datos y modelo de clasificación, con el fin de comparar métricas de desempeño como tiempo y matrices de confusión para conocer cuál se desempeñó mejor.

Análisis exploratorio de datos

A partir de aquí se cambió la base de datos ya que la anterior solo tenía 175 muestras y no era la mejor opción para el objetivo de este trabajo, por tanto, se utilizó la base de datos Sentiment140 (Kaggle, 2023) que contiene diferentes mensajes de twitter, llamados comúnmente “tweets”. Dicha base de datos con seis columnas y 1, 600,000 filas, las cuales representan lo siguiente:

- Tipo de emoción, 0 si es negativa o 4 si es positiva, o, en otras palabras, si el texto del *tweet* expresa una emoción negativa o positiva.
- Id del mensaje o tweet, funciona como identificador del mensaje.
- Fecha. Es la fecha en la que se realizó el mensaje.
- Consulta. Si es mediante una consulta o no.
- Usuario, indica el nombre del usuario.
- Texto, es el mensaje que expreso el usuario y se clasificó como positivo o negativo.

Para cada tipo de emoción, negativa o positiva, son 800,000 muestras, lo cual indica una base de datos balanceada.

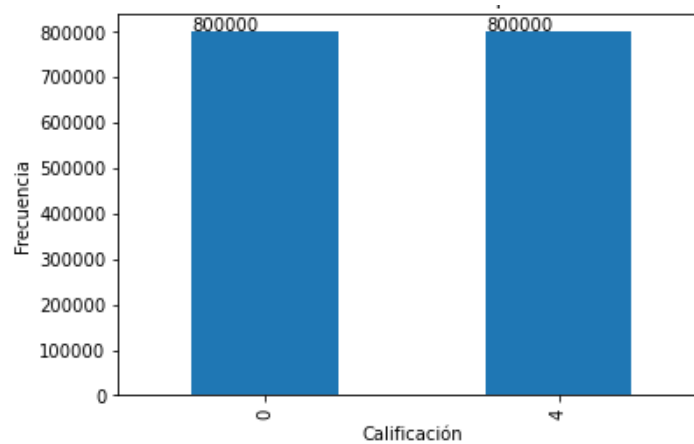


Ilustración 1: Conteo de muestras por tipo de emoción

En el siguiente gráfico de nube de palabras, se muestran aquellas palabras con más frecuencia.



Ilustración 2: Nube de palabras más comunes en los mensajes de tweet

Metodología

Una vez realizado un análisis exploratorio de los datos, se plantea los métodos que utilice para el muestreo de cada uno.

Cabe destacar que el procesador de mi computadora no fue capaz de procesar el código con ninguno de este porcentaje:

Porcentaje de muestras de la base de datos	Tipo de error
10%	MemoryError: Unable to allocate 395. GiB for an array with shape (64000, 207180) and data type int64
5%	MemoryError: Unable to allocate 36.4 GiB for an array with shape (64000, 76363) and data type int64
2%	Si fue posible continuar

Cuando se presenta este tipo de error es importante tomar muestras de los datos, ya que sin esta alternativa no sería posible obtener un resultado.

A continuación, se describirá cada tipo de método de muestreo.

- Método 1 (muestreo sesgado)

Se tomaron los primeros 16000 muestras de cada grupo, emoción positiva o negativa. Después tomó el 80% y 20% de los datos para entrenamiento y prueba, de igual manera se tomaron los primeros 12800 para entrenamiento y los restantes para prueba, así para cada grupo. Cabe resaltar que no fueron seleccionados aleatoriamente, solo los primeros para entrenamiento y los restantes para prueba.

- Método 2 (muestreo aleatorio)

Se tomó para cada grupo una muestra aleatoria con reemplazo del 2% de las muestras de la base de datos, de tal manera que se utilizó 16000 para cada grupo.

Para los subconjuntos de entrenamiento (80%) y prueba (20%) también se realizó un muestreo aleatorio pero esta vez sin reemplazo.

Para ambos muestreos se utilizó la función de *pd.sample* (McKinney, 2010).

En la tabla 1, se observa que el número de muestras es la misma, sin embargo, la selección es diferente.

Tabla 1: Número de muestras por tipo de método

Numero de observaciones					
Método	Entrenamiento		Prueba		Total
1	x	y	x	y	16000
	12800	12800	3200	3200	
2	x	y	x	y	16000
	12800	12800	3200	3200	

Preprocesamiento y modelado

El preprocesamiento y el modelo fue el mismo para cada método. El preprocesamiento consistió en la *tokenización* de datos y la eliminación de las palabras que no aportan mucho al mensaje en cuestión.

Para el modelo, se utilizó una regresión logística ya que solo son dos clases, la emoción positiva y negativa.

Resultados

En la tabla 2 se observa que el método 1 tuvo un desempeño en general menor en todas las métricas menos en la de validación, la cual consistió en dar un mensaje al modelo y analizar si lo clasificaba correctamente, en este caso tanto en el método 1 y 2 fue clasificado correctamente.

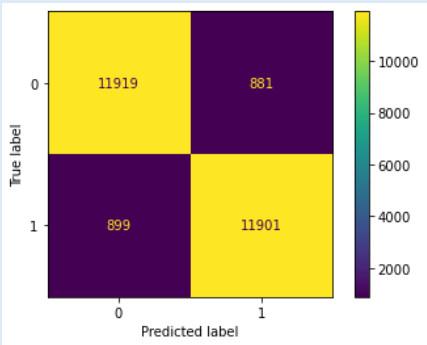
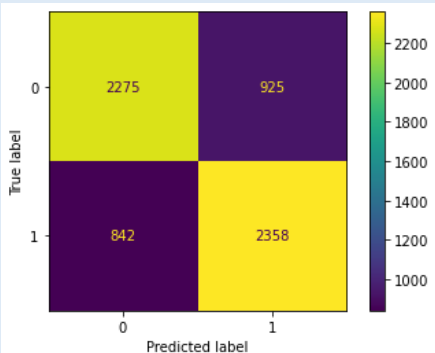
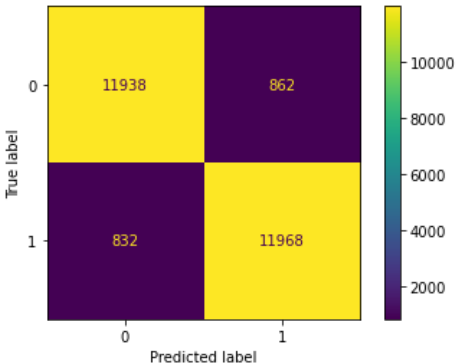
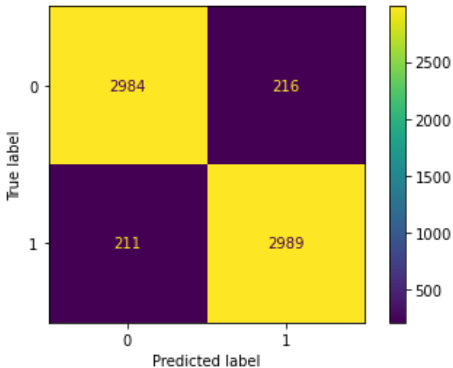
Respecto al tiempo de ejecución, el método 2 resulto ligeramente más rápido que el método 1 con una diferencia de 0.12940 segundos.

Por otro lado, tanto en la matriz de confusión y en la exactitud se observa que su desempeño en el conjunto de test es mejor por 20.93% en el método 2. Las matrices de confusión para entrenamiento y prueba también se observan en la tabla 3.

Tabla 2: Resultados de rendimiento del modelo por tipo de método de muestreo

Método	Tiempo total de ejecución	Matriz de confusión test	Exactitud	Validación
1	7.1165 segundos	[[2275, 925], [842, 2358]]	0.7239	ok
2	6.9871 segundos	[[2984, 216], [211, 2989]]	0.9332	ok

Tabla 3: Matriz de confusión para cada conjunto y método

Método	Entrenamiento	Prueba
1		
2		

Conclusiones

Se concluye que el tipo de muestro para seleccionar una muestra significativa de la base de datos cuando es muy grande para procesadores comunes, es de vital importancia ya que sin esta técnica no sería posible realizar este tipo de análisis.

Por otra parte, es importante seleccionar aleatoriamente las muestras de entrenamiento y validación para que el rendimiento del modelo sea lo más cercano a la realidad de igual manera se trate de que cada muestra tenga la misma posibilidad de ser seleccionada y, por tanto, se evite caer en sesgos.

Finalmente, el muestreo es una técnica crucial cuando se trabaja con datos masivos ya que ayuda a optimizar y a obtener resultados certeros.

Bibliografía

Kaggle. (2023, 06 29). Retrieved from <https://www.kaggle.com/datasets/kazanova/sentiment140>

McKinney, W. (2010). Data structures for statistical computing in python. *Science Conference*, V9, 56-61.