Luis Daniel Rivero Sosa A01374527

1. Selecciona un dataset que te llame la atención de kaggle.com

El dataset seleccionado se puede encontrar en la siguiente página: https://www.kaggle.com/animeshgoyal9/predict-click-through-rate-ctr-for-a-website/data

El mismo contiene información sobre usuarios que aplicaron a un trabajo dentro de una plataforma digital. Elegí este dataset por que se puede aprovechar con un clasificador de una y dos clases, ya que las clases son: aplicó al trabajo y no aplicó al trabajo.

El dataset contiene 10 campos, los cuales son:

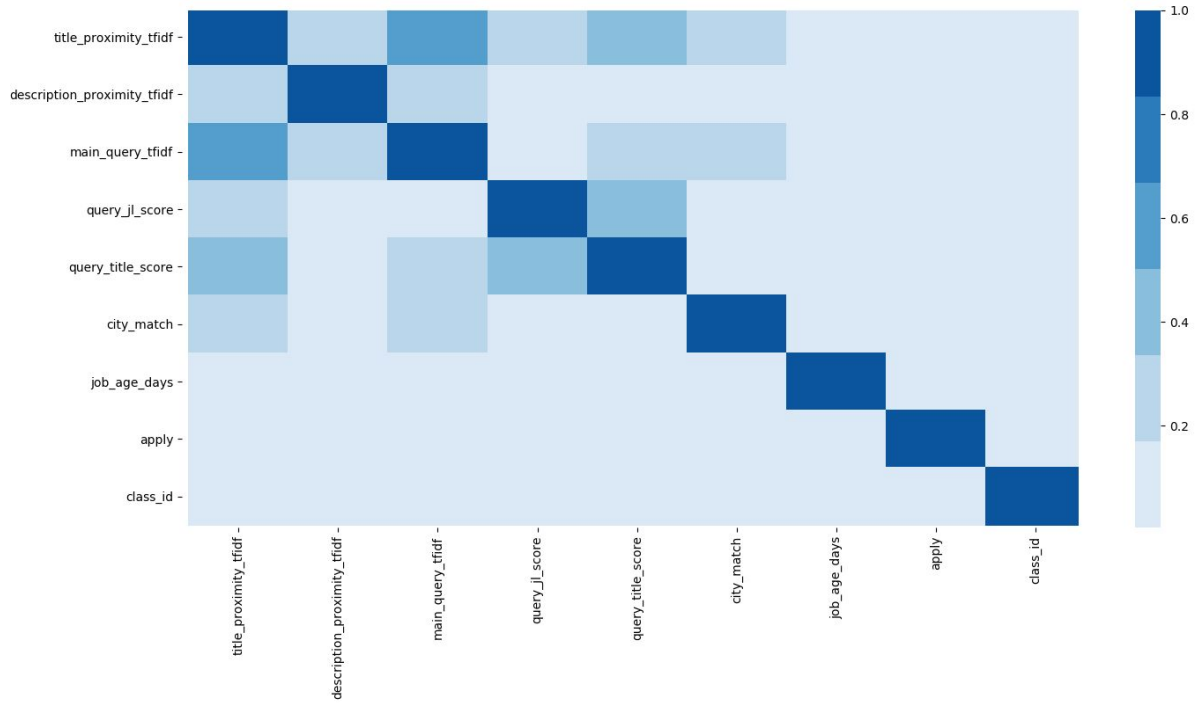
1. title proximity tfidf: Measures the closeness of query and job title.
2. description proximity tfidf: Measures the closeness of query and job description.
3. main query tfidf: A score related to user query closeness to job title and job description.
4. query jl score: Measures the popularity of query and job listing pair.
5. query title score: Measures the popularity of query and job title pair.
6. city match: Indicates if the job listing matches to user (or, user-specified) location.
7. job age days: Indicates the age of job listing posted.
8. apply: Indicates if the user has applied for this job listing.
9. search date pacific: Date of the activity.
10. class id: Class ID of the job title clicked.

2. Determina una hipótesis acerca del dataset (Los clasificadores de una sola clase obtienen un mayor desempeño que los binarios, Se puede obtener un buen desempeño sin tomar en cuenta todas las características, La característica X afecta negativamente el desempeño de las redes neuronales, etc etc).

Mi hipótesis es que los clasificadores binarios tienen un mejor desempeño que los de una sola clase.

3. Realiza el análisis de los datos (correlación, mapas de calor) y determina las características a usar.

Se realizó el mapa de calor con el dataset y el resultado fue el siguiente:



Después de analizar el mapa de calor quité el atributo de “title_proximity_tfidf”

Adicionalmente, elimine los atributos “search_date_pacific” y “class_id” dado que la descripción del dataset en kaggle.com sugiere que no se tomen en cuenta

4. Utiliza al menos 3 diferentes clasificadores, probando con 10-fold cross-validation.

Dado que mi hipótesis es que los clasificadores binarios tienen un mejor desempeño que los de una sola clase, elegí utilizar los siguientes clasificadores:

- MLP (Binario)
- KMeans (Una clase)
- OKCRA (Una clase)

Elegí estos clasificadores tomando como hipótesis que el MLP es el mejor clasificador binario y su desempeño se comparará con Kmeans y OCKRA que son clasificadores de una sola clase.

Los archivos usados para correr las pruebas se encuentran dentro de la carpeta adjunta a este proyecto.

Las pruebas son llamadas desde el archivo llamado “mainProgram.py”. Desde este script se mandan a llamar los archivos: “KM”, “MLP” y “OCKRA”. Donde cada uno realiza el entrenamiento y clasificación por cada clasificador.

El resultado de desempeño por cada clasificador fue la siguiente:

```
-----------------------------------------------
Resultados de MLP:

AUC = 0.5
Resultados individuales = [0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5]
-----------------------------------------------


-----------------------------------------------
Resultados de KMeans:

AUC = 0.499025817232
Resultados individuales = [0.49578925062870516, 0.49819356653529567, 0.48757699933481885, 0.50135844968623555, 0.49193995414771713, 0.50796683745631888,
0.50325862335451055, 0.51540677129345736, 0.4947435802711222, 0.49402413960977598]
-----------------------------------------------


-----------------------------------------------
Resultados de OCKRA:

AUC = 0.508957062952
Resultados individuales = [0.51757841029170437, 0.51087625457936414, 0.4955325083655539, 0.51850509470231609, 0.50029458590521725, 0.52251572088218545,
0.50996265883776237, 0.49432228546997364, 0.51435103218895439, 0.50563207829196855]
-----------------------------------------------
```

Donde se puede observar que el desempeño de todos los clasificadores es prácticamente el mismo, teniendo un AUC de aproximadamente 0.5

5. Intenta obtener el mejor desempeño posible (modificar parámetros, usar un genético, probar otra combinación de atributos)

El mejor desempeño se encontró al usar las siguientes características de cada clasificador:

- MLP:
    - Se usó la siguiente configuración: hidden_layer_sizes=(10, 10, 10). La cual significa que tiene 3 capas y cada una de ellas cuenta con 10 neuronas
- KMeans (Una clase):
    - Se usó una K=2, lo cual significa que se crearon dos grupos
- OKCRA (Una clase):
    - La cantidad de clasificadores usados fueron: 10

6. Realiza el análisis estadístico de tu hipótesis para determinar si se puede aceptar o no (No es necesario que sea aceptada, pero sí que hagas el proceso correcto).

Se realizó el análisis estadístico usando la evaluación de Wilcoxon Signed-Rank. Los resultados fueron los siguientes:

```
Analisis estadistico de MLP vs KM:
Statistics=21.000, p=0.508
Same distribution (fail to reject H0)

Analisis estadistico de MLP vs OCKRA:
Statistics=6.000, p=0.028
Different distribution (reject H0)
```

Donde se rechazó una distribución y se aceptó la otra

Conclusiones:

- Se puede observar que el desempeño de todos los clasificadores es prácticamente el mismo, teniendo un AUC de aproximadamente 0.5. Por lo que no se puede afirmar que los clasificadores binarios tienen un mejor desempeño que los de una sola clase
- A partir de la evaluación de Wilcoxon Signed-Rank se puede ver que:
    - Los clasificadores MLP y OCKRA para este ejemplo particular tienen diferente distribución, por lo que el desempeño de ambos clasificadores puede ser diferente.
    - Los clasificadores MLP y KMeans para este ejemplo particular tienen la misma distribución y la hipótesis nula no puede ser rechazada.