CODERHOUSE

JUNIO 2024

PRESENTACIÓN
PROYECTO FINAL

Presentado por Jorge Mariano Miro

ÍNDICE DE CONTENIDOS

01 PRESENTACIÓN

El dataset seleccionado corresponde a la información de todos los partidos jugados en el circuito ATP desde Enero 2000 hasta principios 2024, en cualquier tipo de superficie, ronda o ambiente. El proyecto consiste en encontrar la mayor probabilidad de acierto cuando se dan ciertas características que avalado por la estadística, no están ajustadas en el valor de la cuota de apuesta.

MOTIVACIÓN

Se ha propuesto avanzar en el análisis de este dataset, buscando los puntos debiles del gambling en las casas de apuestas. Para esto se propone trabajar en primer lugar, en deportes donde no haya una tercera opcion de resultado, lo que permite ser mas eficientes en aciertos. El fin, es lograr obtener resultados que puedan intervenir positivamente en la decision de un apostador.
Este trabajo se analizará por medio de un algoritmo de clasificación, que tan probable es el acierto o proyección de resultado de un partido dada sus características.

AUDIENCIA

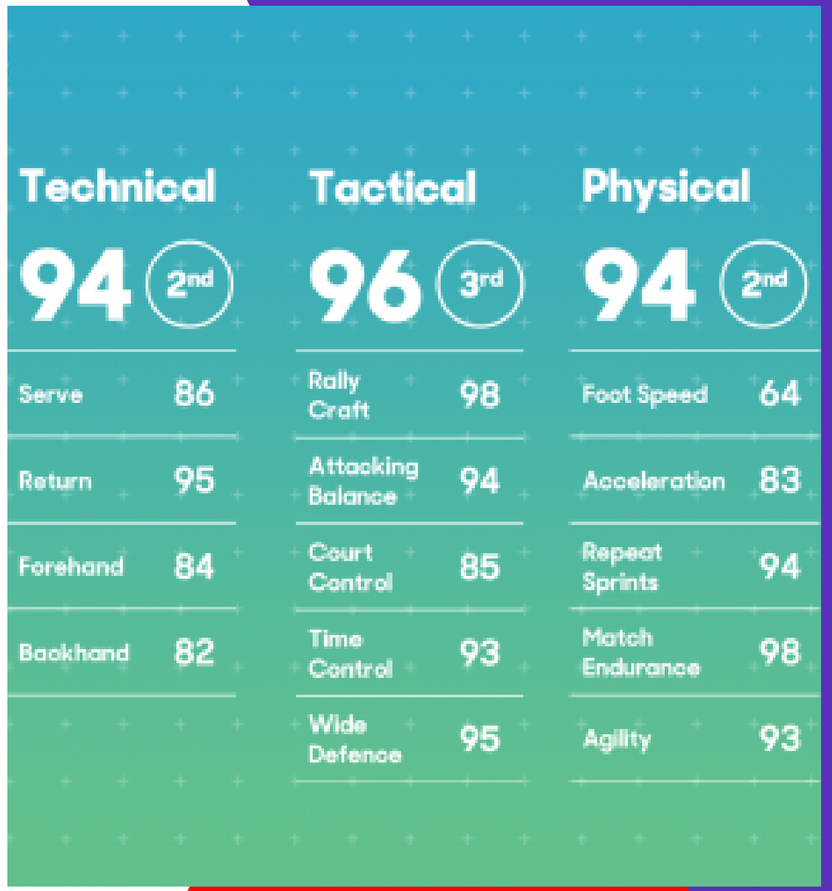
La audiencia de interés en los resultados del presente estudio pueden ser:

- Apostadores
- Casas de Apuestas
- Deportistas

CONTEXTO COMERCIAL

Las casas de apuestas deportivas han emergido como actores clave en el panorama comercial, fusionando emoción y estrategia. En este ecosistema, las cuotas desempeñan un papel crucial al reflejar las percepciones de probabilidad. El tenis, uno de los deportes centrales en esta dinámica, ve sus cuotas ajustarse en tiempo real según el rendimiento de los jugadores y las variables del encuentro.

El gambling, intrínseco a esta industria, ha experimentado un auge significativo, atrayendo a una audiencia ávida de emoción y entretenimiento. Este sector comercial, en constante evolución, ofrece experiencias únicas que trascienden las competiciones, convirtiendo cada evento en una oportunidad para la participación activa. La intersección entre deportes y apuestas crea un mercado vibrante y adaptable, proporcionando a los aficionados una conexión más inmersiva con sus deportes favoritos, sin embargo, suele ser un negocio netamente perdedor a largo plazo.



LINK AL NOTEBOOK

TENIS

02 PREGUNTAS Y OBJETIVOS

Respecto al dataframe, algunas de las preguntas que fueron surgiendo son:

- Es mas alto es el porcentaje de ganador del favorito en primera ronda con respecto al resto del torneo?
- Que variables no estadisticas hay que tener en cuenta? Cansancio acumulado? Vuelta al circuito despues de una lesion? Historial cara a cara?
- Que porcentaje de acierto hay en el favoritismo respecto la diferencia de rankings? Es decir, si el favorito de ranking es 30, juega contra el ranking 35 (diferencia 5) o si el favoritismo de un tenista ranking 20 juega contra un ranking 110 (diferencia 90)
- El favorito estando en desventaja, tiene mas chances de exito, 0-1 en torneo 3 sets, o 1-2 en torneo de 5 sets?
- Que influye mas en el analisis? paridad de rankings, o paridad de cuotas iniciales?
- Que porcentaje de eficacia hay en los favoritos que ganan el primer set, y terminan ganando el partido?
- Cuanto influye la cantidad de puntos ganados del tenista?
- Cuanto influye la cantidad de aces logrados por un tenista?
- Cuanto influye la cantidad de errores no forzados de un tenistas?

OBJETIVOS

Crear un algoritmo, en el que se pueda predecir dar por ganador al jugador favorito en tenis cuando se cumplan ciertos requisitos probabilisticos.

03 ESTRUCTURACIÓN DEL PROYECTO

EXPLORATORY DATA ANALYSIS (EDA)

```
[ ] df.shape
```

```
(61668, 17)
```

```
df.tail(5)
```

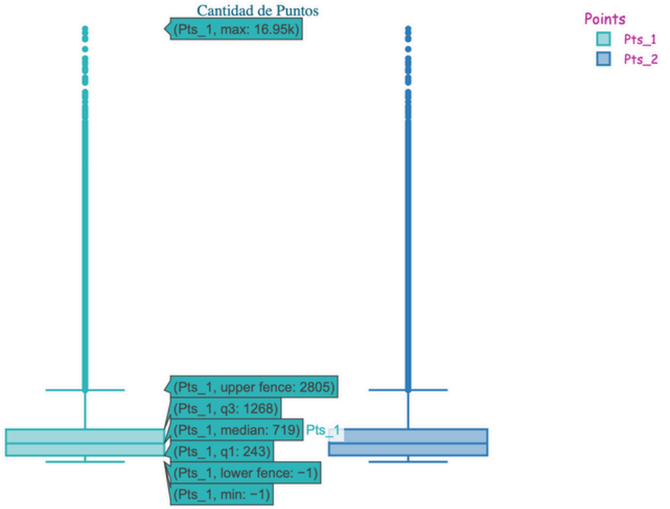
	Tournament	Date	Series	Court	Surface	Round	Best of	Player_1	Player_2	Winner	Rank_1	Rank_2	Pts_1	Pts_2	Odd_1	Odd_2	Score
61663	ASB Classic	2024-01-10	ATP250	Outdoor	Hard	Quarterfinals	3	Muller A.	Daniel T.	Daniel T.	79	74	711	739	1.91	1.91	4-6 7-6 3-6
61664	ASB Classic	2024-01-11	ATP250	Outdoor	Hard	Quarterfinals	3	Shelton B.	Carballes Baena R.	Shelton B.	16	67	2145	796	1.17	5.00	6-4 6-3
61665	ASB Classic	2024-01-12	ATP250	Outdoor	Hard	Semifinals	3	Shelton B.	Daniel T.	Daniel T.	16	74	2145	739	1.22	4.33	5-7 6-7
61666	ASB Classic	2024-01-12	ATP250	Outdoor	Hard	Semifinals	3	Tabilo A.	Fils A.	Tabilo A.	82	35	707	1208	4.00	1.25	6-2 7-5
61667	ASB Classic	2024-01-13	ATP250	Outdoor	Hard	The Final	3	Daniel T.	Tabilo A.	Tabilo A.	74	82	739	707	1.73	2.10	2-6 5-7

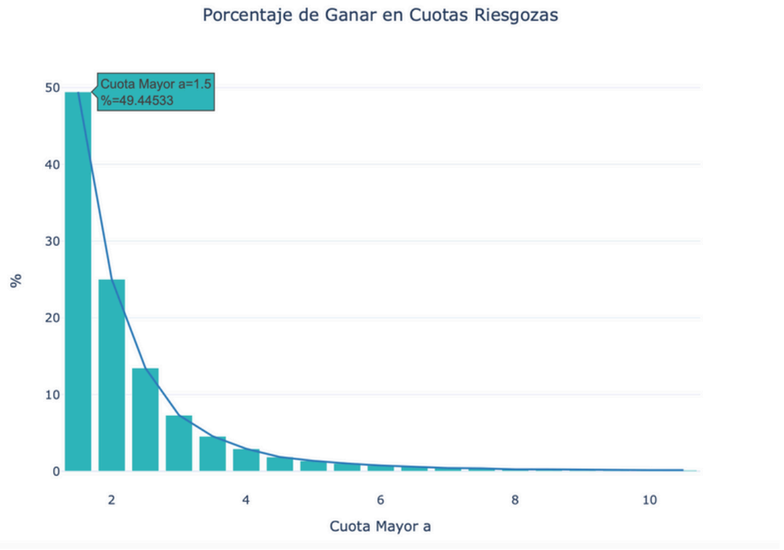
```
# Creo columna con la diferencia entre las cuotas de tenistas
def dif_rank_cuotas(row):
    if row['Cuota_player1'] > row['Cuota_player2']:
        return np.abs(row['Cuota_player1'] - row['Cuota_player2'])
    elif row['Cuota_player1'] < row['Cuota_player2']:
        return np.abs(row['Cuota_player2'] - row['Cuota_player1'])
    else:
        return 0

# Creo la nueva columna 'diferencia cuotas'
df_modelado_tenis['dif_rank_cuotas'] = df_modelado_tenis.apply(dif_rank_cuotas, axis=1)
```

CREACIÓN DE COLUMNAS Y MANIPULACIÓN DE DATOS

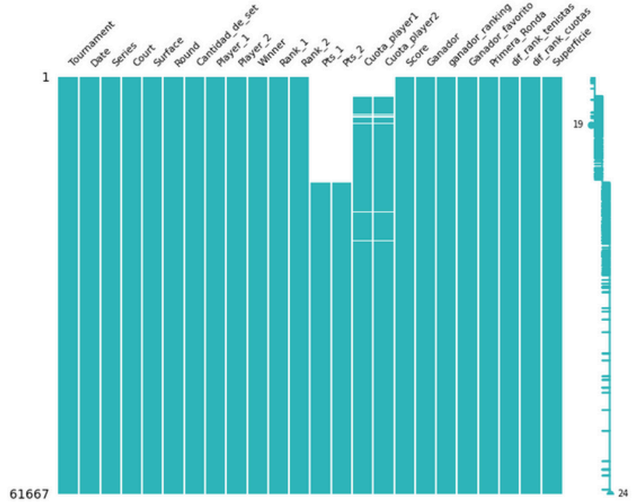
	Date	Cantidad_de_set	Rank_1	Rank_2	Pts_1	Pts_2	Cuota_player1	Cuota_player2	Ganador	ganador_ranking	dif_rank_tenistas
count	61667	61667.000000	61667.000000	61667.00000	61667.000000	61667.000000	61667.000000	61667.000000	61667.000000	61667.000000	61667.000000
mean	2011-06-19 23:20:32.185772032	3.376993	76.028281	75.65844	1093.473349	1099.280815	2.399871	2.392336	1.499992	59.019459	72.540240
min	2000-01-03 00:00:00	3.000000	-1.000000	-1.00000	-1.000000	-1.000000	-1.000000	-1.000000	1.000000	1.000000	1.000000
25%	2005-06-07 00:00:00	3.000000	25.000000	24.00000	-1.000000	-1.000000	1.330000	1.330000	1.000000	17.000000	19.000000
50%	2011-03-25 00:00:00	3.000000	54.000000	54.00000	671.000000	673.000000	1.727000	1.746333	1.000000	42.000000	41.000000
75%	2017-05-11 00:00:00	3.000000	92.000000	92.00000	1205.000000	1210.000000	2.750000	2.750000	2.000000	78.000000	80.000000
max	2024-01-13 00:00:00	5.000000	3390.000000	4915.00000	16950.000000	16950.000000	67.000000	51.000000	2.000000	1890.000000	4911.000000
std	NaN	0.782222	100.877944	101.53385	1707.225335	1730.742427	2.671147	2.624749	0.500004	72.163539	117.809719

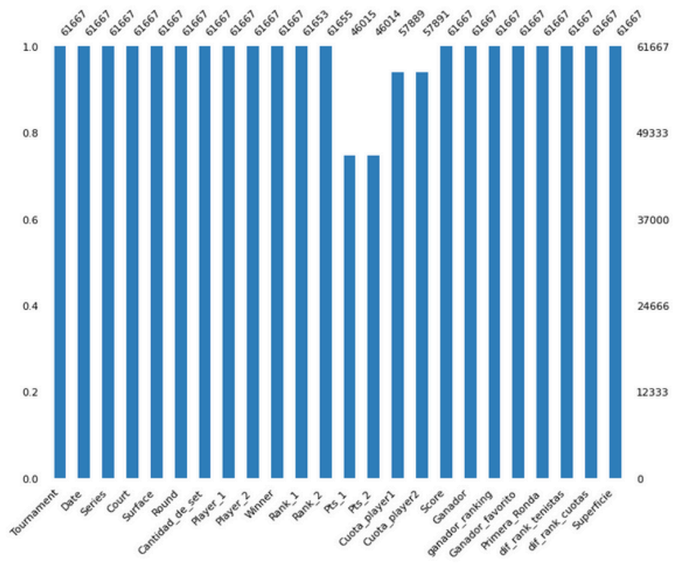
VISUALIZACION DE DATOS

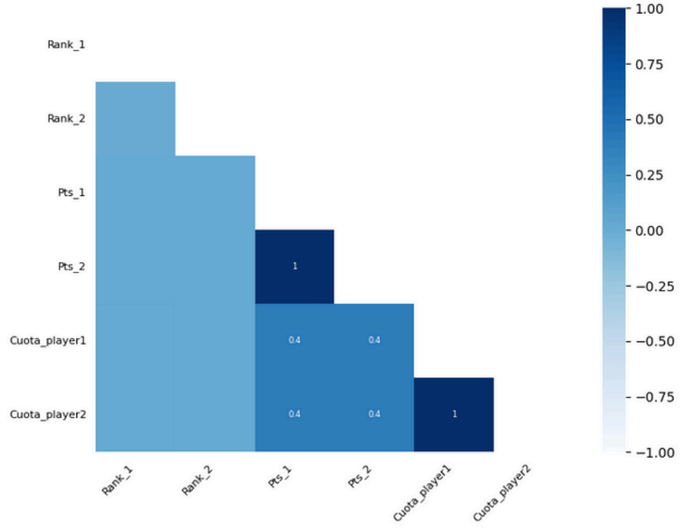




VISUALIZACION DE DATOS NULOS CON MISSINGNO

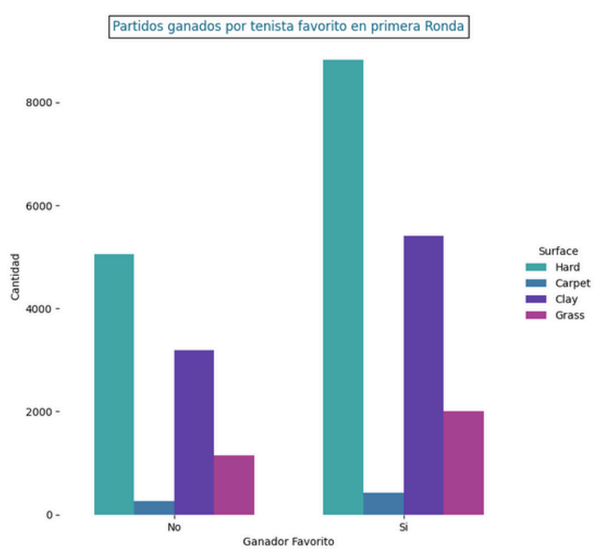


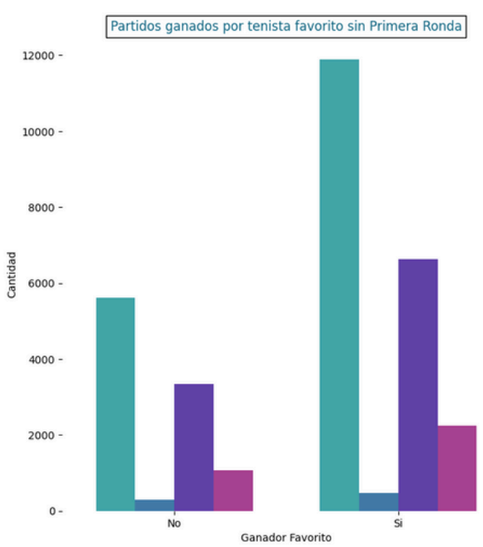


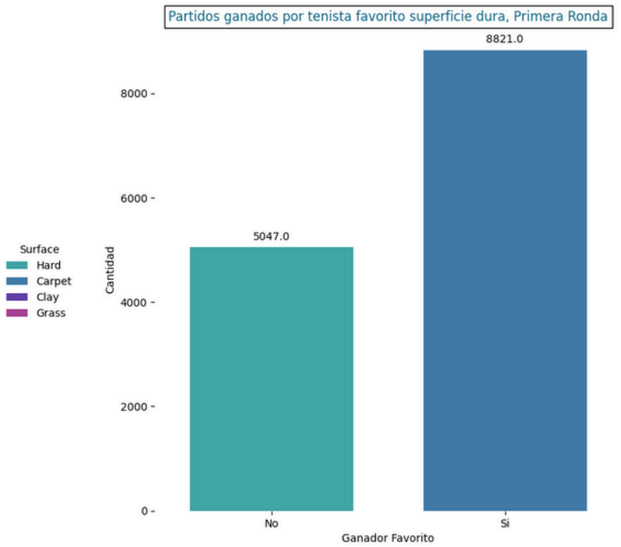


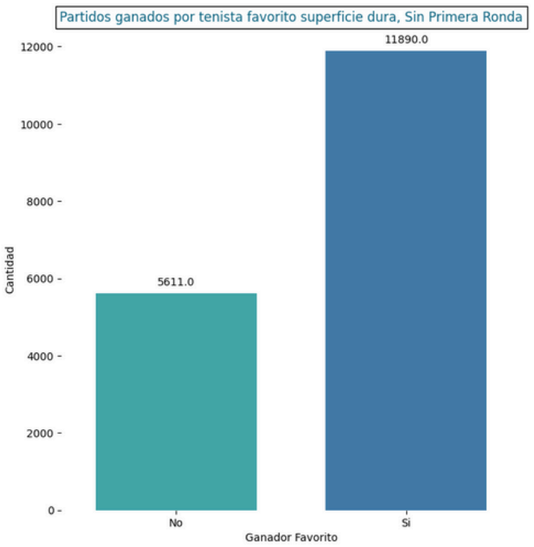
04 HIPOTESIS DEL PROBLEMA ANALÍTICO

VISUALIZACION DE HIPOTESIS 1





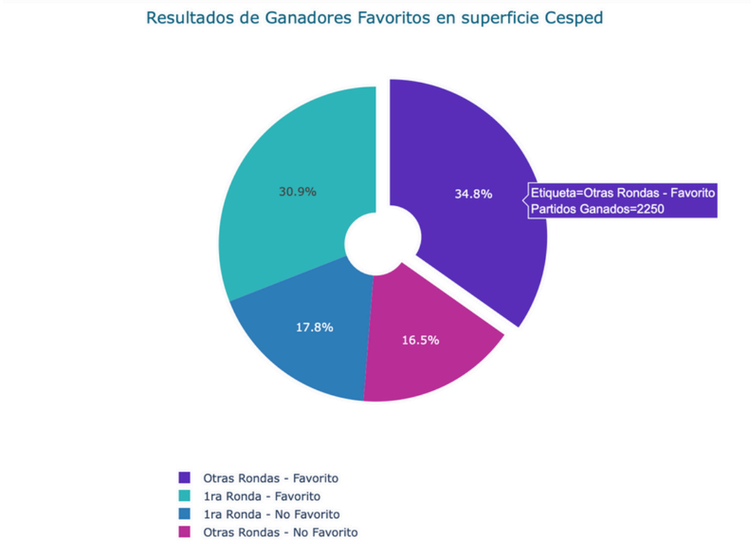


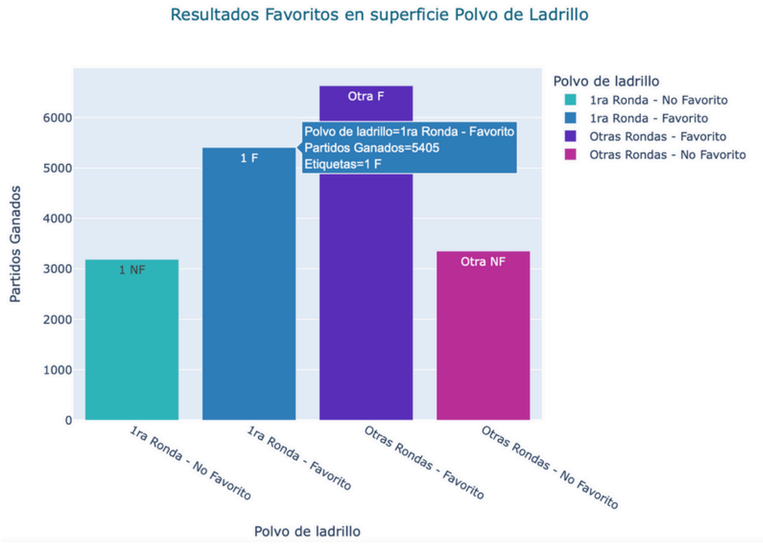


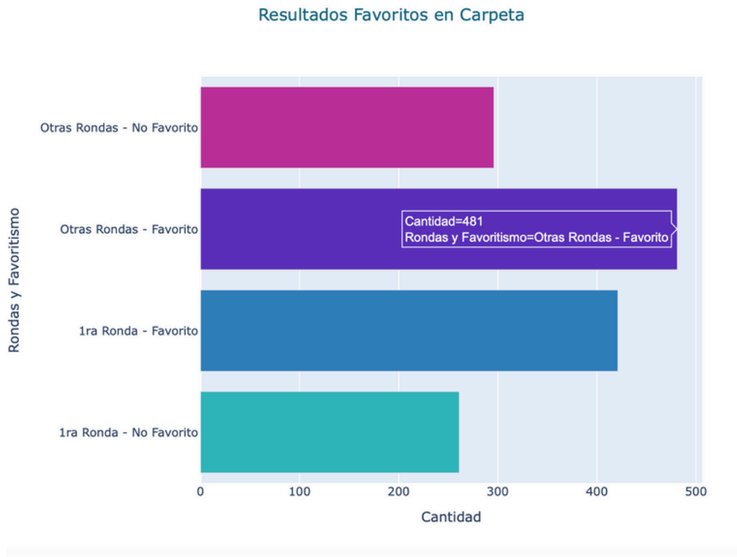
- **Conclusión**

A primera vista en primera ronda suele haber una probabilidad mas alta de que NO gane el favorito. La mayor diferencia con 1ra ronda y sin 1ra ronda, suele darse en superficie dura.

VISUALIZACION DE HIPOTESIS 1



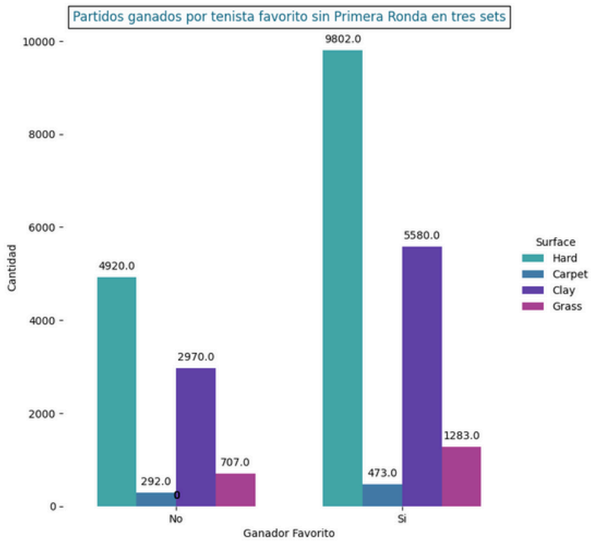


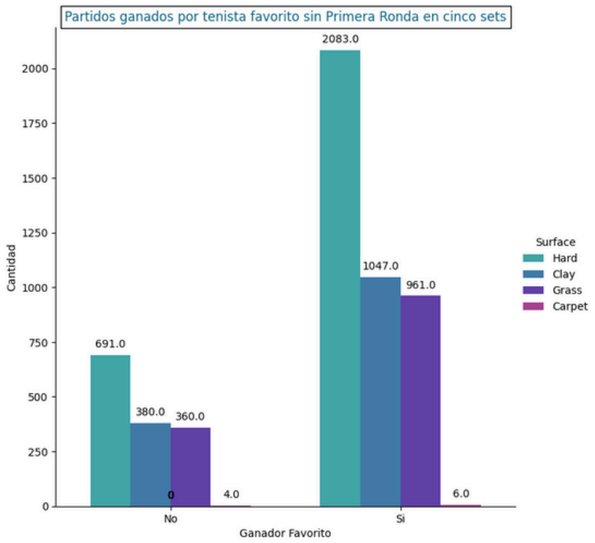


- **Conclusión**

Confirmamos Hipotesis, los Favoritos suelen hacerse mas fuertes pasada la primer Ronda, sin importar la superficie que esten jugando

VISUALIZACION DE HIPOTESIS 2






```
[ ] print(f"El porcentaje de ganar de los favoritos ,en cancha dura, a cinco sets, perdiendo 0-1 es: {hard_porc_vic_0_1_5sets:.2f}%")
    print(f"El porcentaje de ganar de los favoritos, a cinco sets, en cualquier superficie, perdiendo 0-1 es: {porc_vic_0_1:.2f}%")

    print(f"El porcentaje de ganar de los favoritos ,en cancha dura, a cinco sets, perdiendo 1-2 es: {hard_porc_vic_1_2:.2f}%")
    print(f"El porcentaje de ganar de los favoritos ,en cancha dura, a tres sets, perdiendo 0-1 es: {hard_porc_vic_0_1_3sets:.2f}%")

    print(f"El porcentaje de ganar de los favoritos, a cinco sets, en cualquier superficie, perdiendo 1-2 es: {porc_vic_1_2:.2f}%")
    print(f"El porcentaje de ganar de los favoritos, a tres sets, en cualquier superficie, perdiendo 0-1 es: {porc_vic_0_1_3sets:.2f}%")

El porcentaje de ganar de los favoritos ,en cancha dura, a cinco sets, perdiendo 0-1 es: 52.30%
El porcentaje de ganar de los favoritos, a cinco sets, en cualquier superficie, perdiendo 0-1 es: 49.39%
El porcentaje de ganar de los favoritos ,en cancha dura, a cinco sets, perdiendo 1-2 es: 23.12%
El porcentaje de ganar de los favoritos ,en cancha dura, a tres sets, perdiendo 0-1 es: 31.92%
El porcentaje de ganar de los favoritos, a cinco sets, en cualquier superficie, perdiendo 1-2 es: 21.54%
El porcentaje de ganar de los favoritos, a tres sets, en cualquier superficie, perdiendo 0-1 es: 30.73%
```

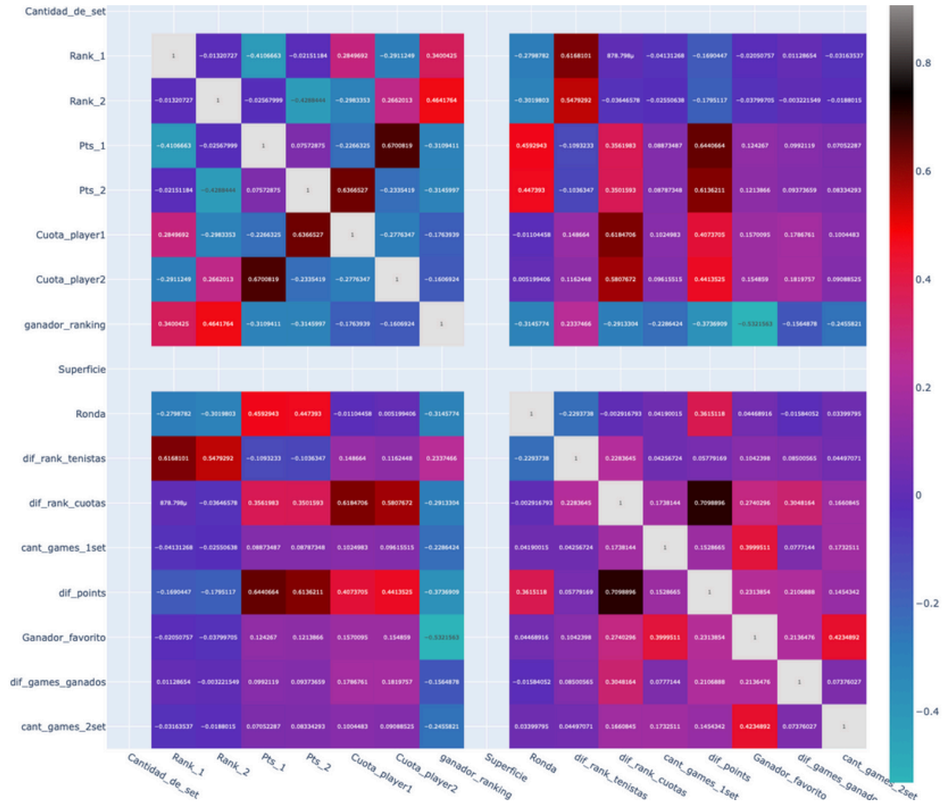
- **Conclusión**

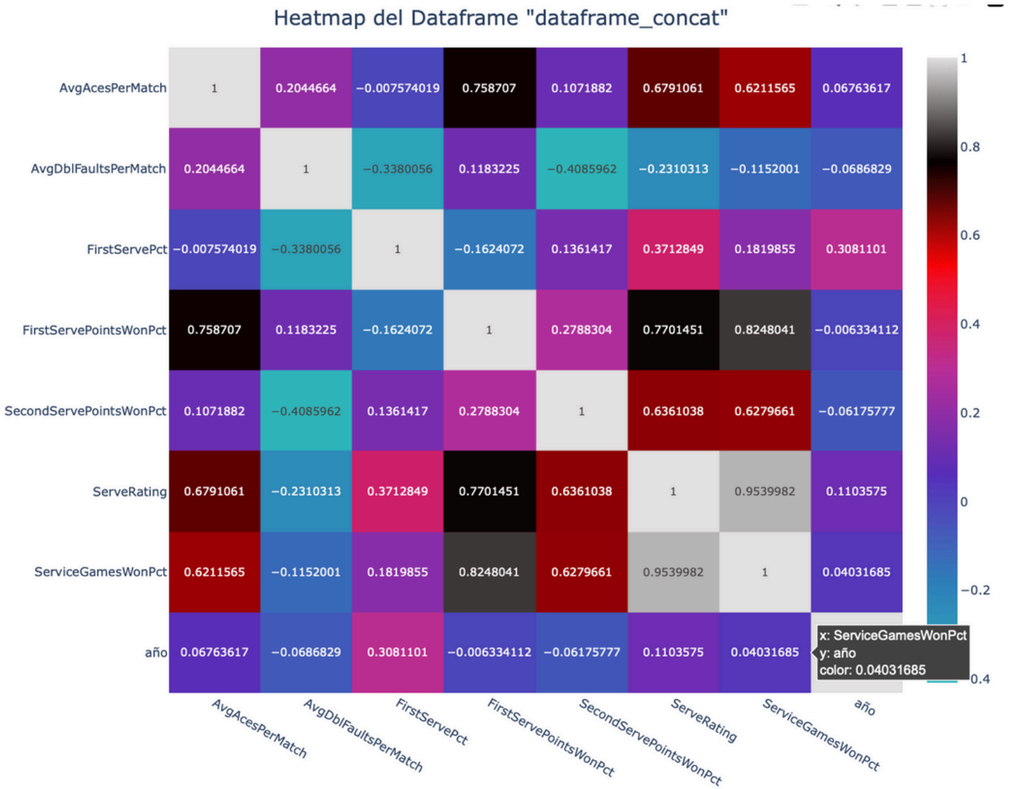
Queda reflejado que a 5 sets perdiendo el primer set los favoritos suelen recuperarse un poco mas de la mitad de las veces en superficie dura, sin importar la diferencia de ranking que haya entre ambos. En el general , sin importar la superficie ese porcentaje baja a 49%, pero hay una gran diferencia relacionandolo con 1-2.

Nuestra conclusion llega a que en cancha dura, o en cualquier superficie perdiendo 0-1 a tres sets sigue siendo mas probables de victoria (casi 32%) que 1-2 en 5 sets (23%)

05 PREPARACIÓN DATAFRAME FINAL

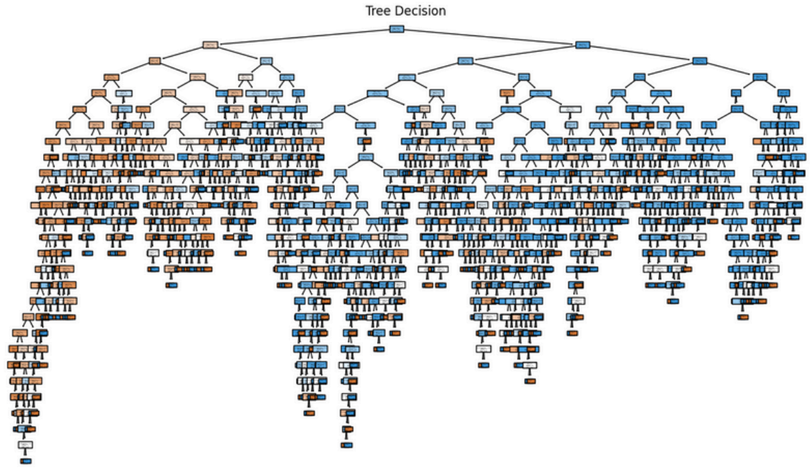
MATRIZ DE CORRELACION ENTRE LOS ATRIBUTOS DE AMBOS DATAFRAMES

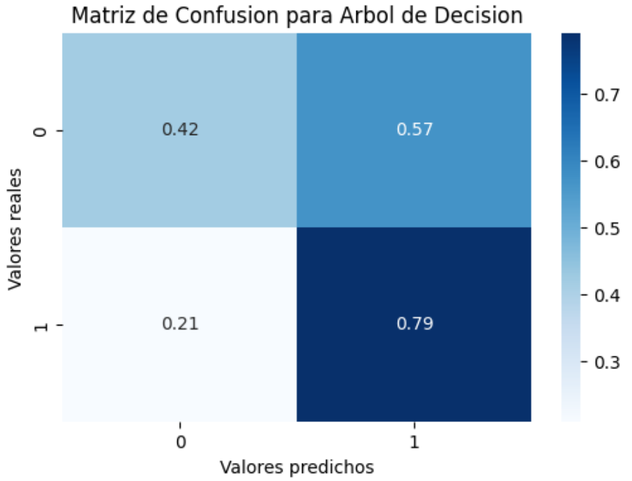


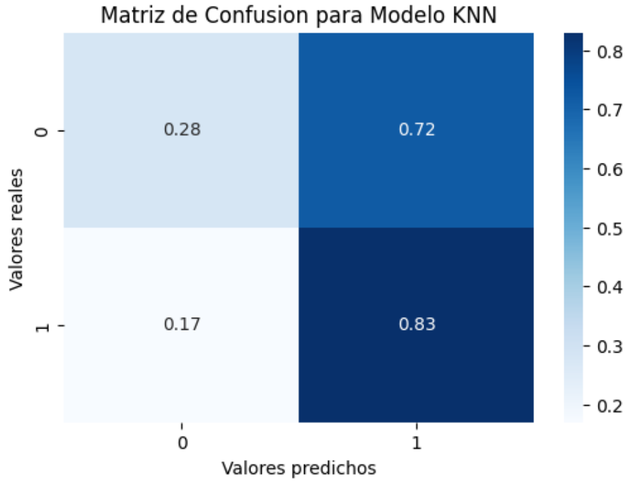


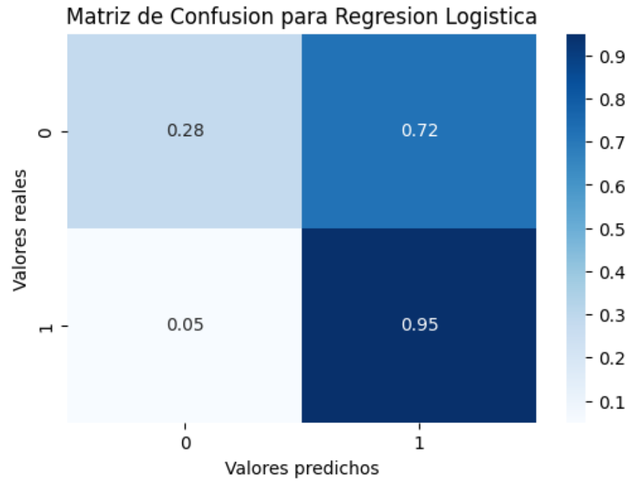
06 PRUEBA Y ELECCIÓN DEL ALGORITMO

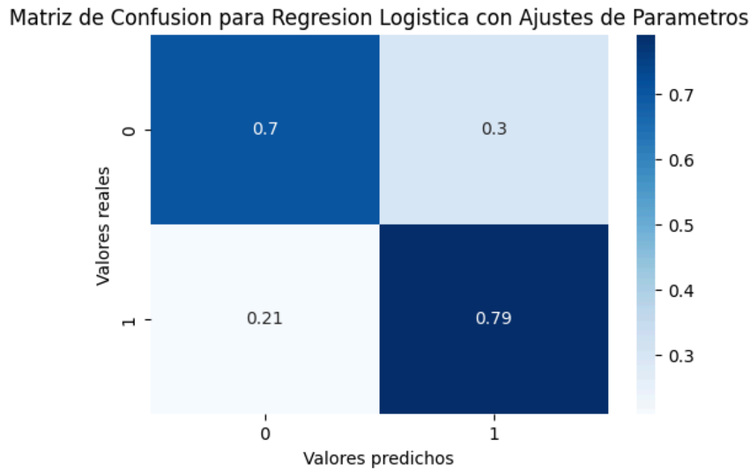
MATRIZ DE CONFUSION DE LOS DIFERENTES MODELOS



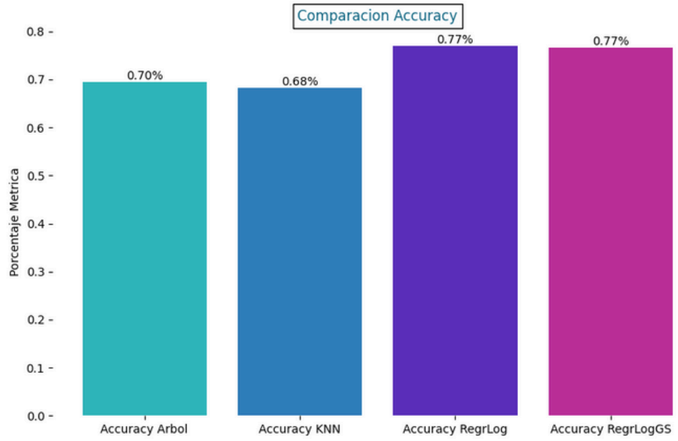


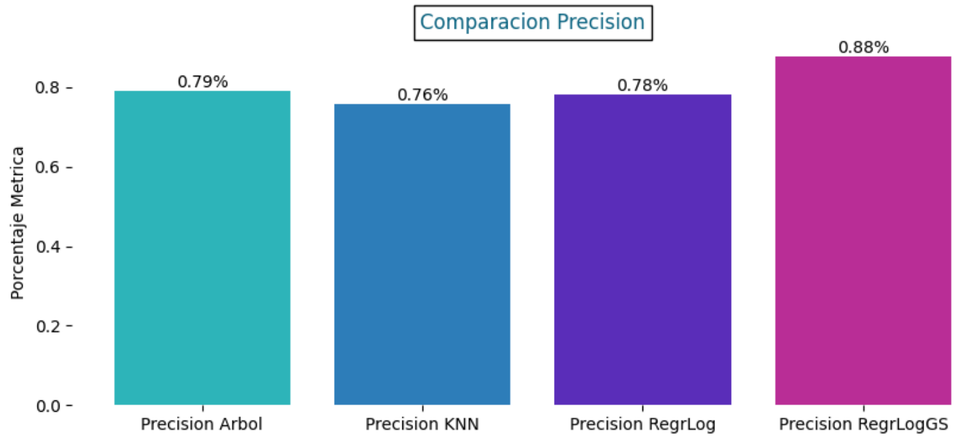


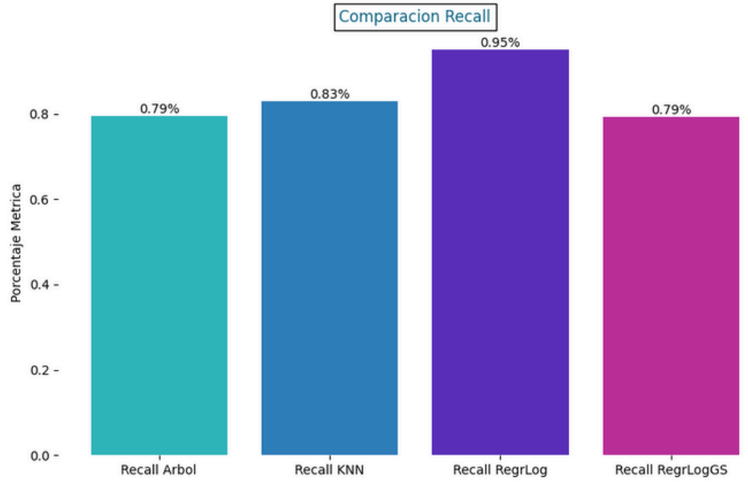


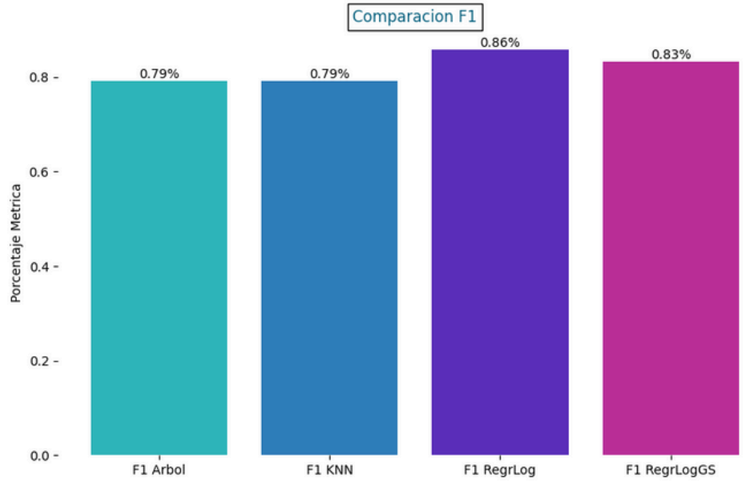


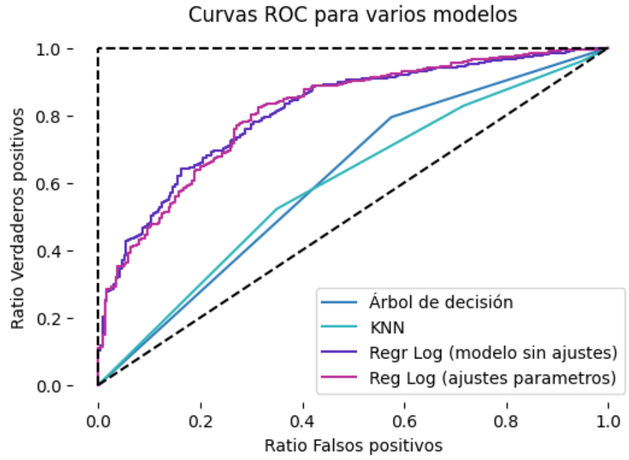
COMPARACION METRICAS DIFERENTES MODELOS



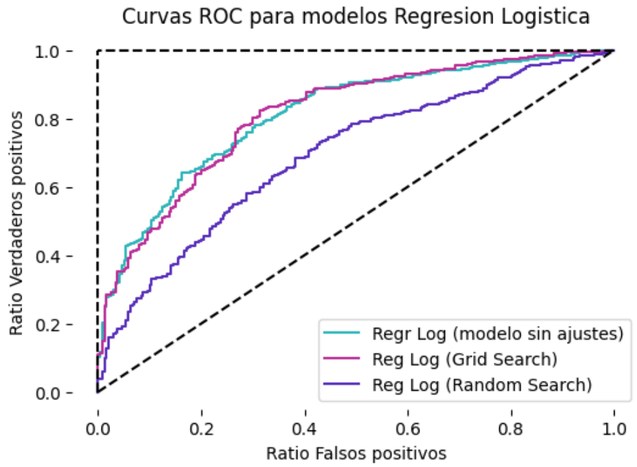


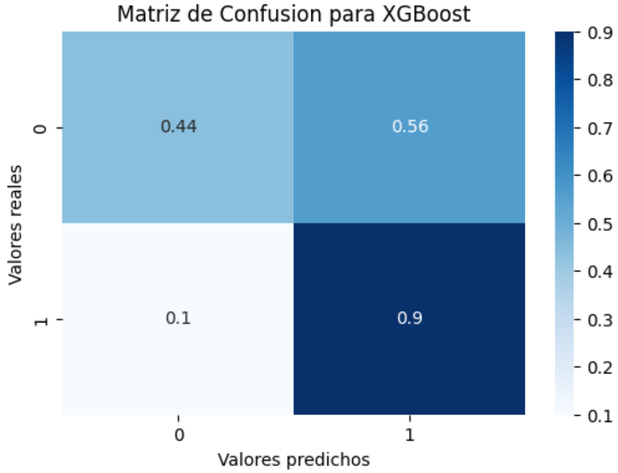


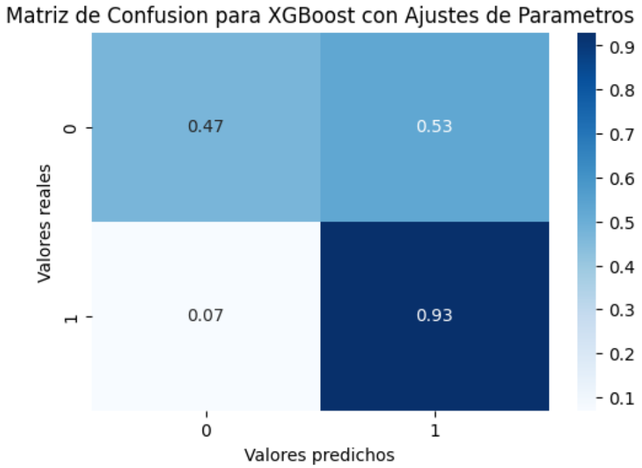


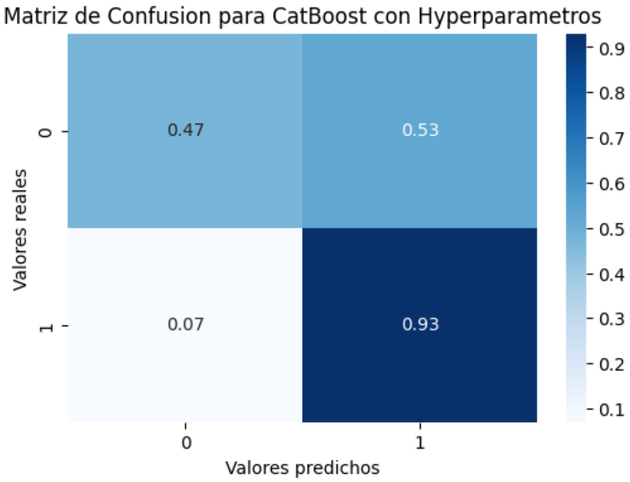


COMPARACION METRICAS DIFERENTES MODELOS









BENCHMARK ENTRE LOS DOS MODELOS ESCOGIDOS