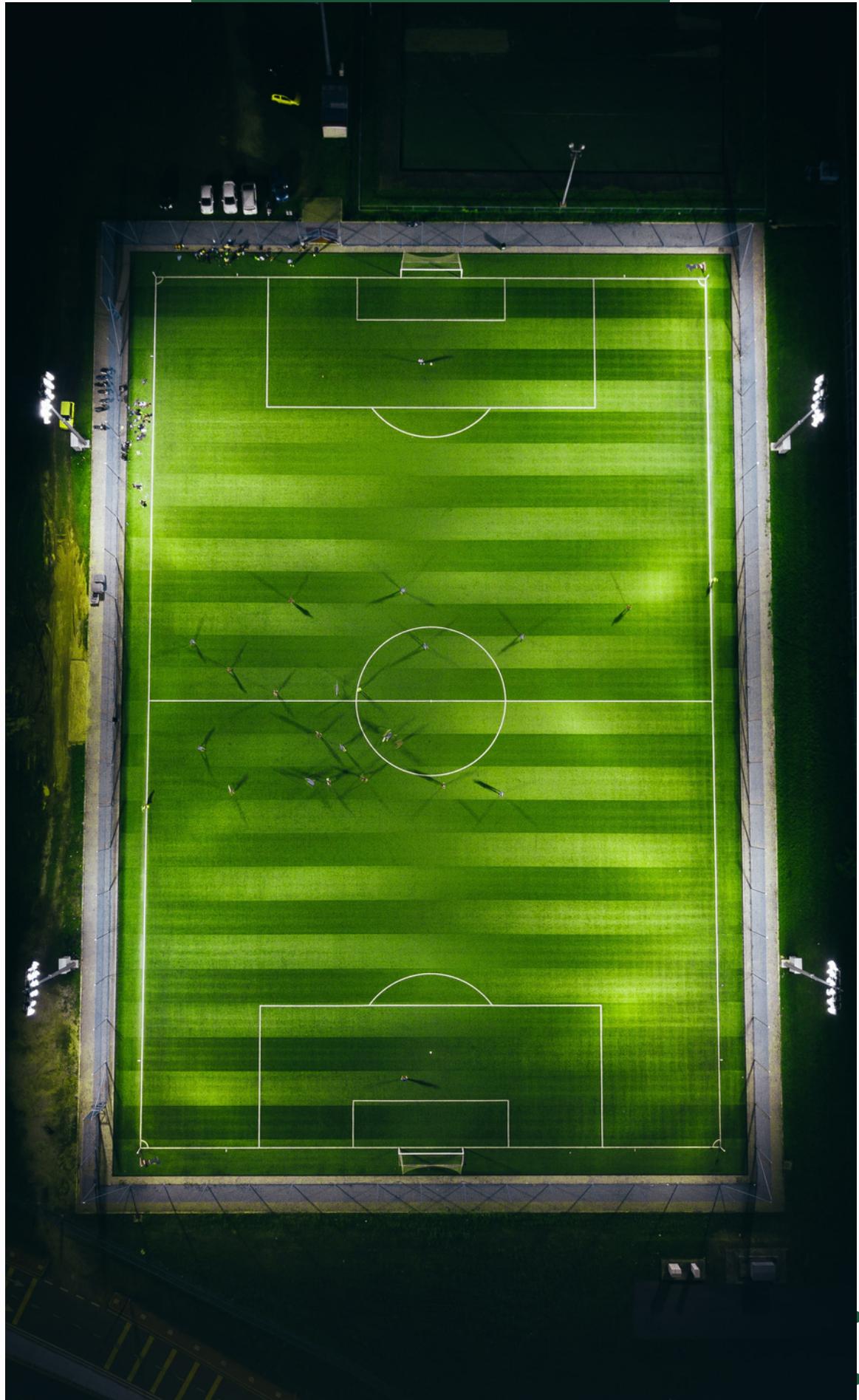




# PROYECTO DATA SCIENCE

## LEONARDO MESAGLIO

"Reconocimiento de patrones y tendencias en los torneos locales Argentinos (2015/2022)"



# Indice:

- 03** Resumen
- 04** Objetivos y metas
- 05** Hipotesis e insights
- 06** Descripcion del Dataset
- 07** Analisis Multivariado
- 08** Heatmap
- 09** Selección del modelo
- 10** Modelo de Clasificación
- 11** Metricas: Clasificación
- 12** Precisión y Desviación.
- 13** Curva ROC
- 14** Conclusiones



# Resumen:

Breve descripción  
del resumen del  
proyecto

Este proyecto parte de un análisis de un set de datos que contiene datos de los torneos argentinos de fútbol desde el año 2015 hasta el año 2022. Contiene información detallada de 2821 partidos de primera división.

El dataset fue obtenido de la página:

[https://www.kaggle.com/datasets/camussonif/argentinian-football-results-20152022?select=afa\\_2015\\_2022\\_spa.csv](https://www.kaggle.com/datasets/camussonif/argentinian-football-results-20152022?select=afa_2015_2022_spa.csv)

De acuerdo al creador del dataset, la información obtenida y plasmada en el dataset fue obtenida de las web:

<https://www.promiedos.com.ar>

<https://www.transfermarkt.com>

<https://www.oddsportal.com>



# Objetivos

Detalle los objetivos del proyecto:

## Finalidad

El objetivo del análisis es mediante la relación entre las diferentes variables llegar a conclusiones que puedan ayudar en el aspecto directivo a desarrollar estrategias futbolísticas a los fines de lograr mejores resultados deportivos.

## Desarrollo

El análisis puede ayudar a todos los equipos técnicos a identificar las variables que tienen el mayor impacto en el resultado y así tener mejor información para desarrollar una estrategia ya sea de juego, para futuras compras de jugadores, etc.

## Eficiencia

Las diversas conclusiones a las que se arribaran con el análisis llevarán a la eficiencia de las decisiones deportivas de los equipos analizados.



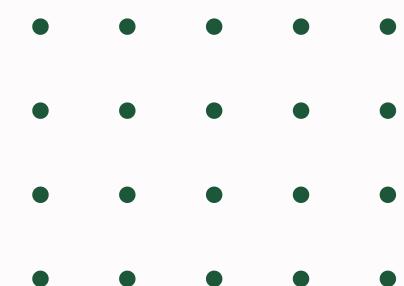


## HIPOTESIS E INSIGHTS:

La hipótesis del proyecto es que a medida que aumente la posesión (ya sea local o visitante), o los tiros al arco (local o visitante), deberían aumentar la cantidad de goles del Equipo en los partidos y por ende la cantidad de victorias.

Esta hipótesis se basa en la idea de que la posesión del balón y los tiros al arco son indicadores clave del desempeño del equipo en un partido de fútbol y que de dichas variables podemos realizar un modelo que a partir de la relación entre esos factores pueda realizar una predicción lo mas precisa posible.

Este proyecto comenzó con la idea de realizar mediante una regresión logística predicciones acerca de resultados deportivos. Sin embargo y en base a los resultados obtenidos, tuve que redirigir el proyecto hacia un modelo de clasificación a los fines de obtener un modelo fiable.





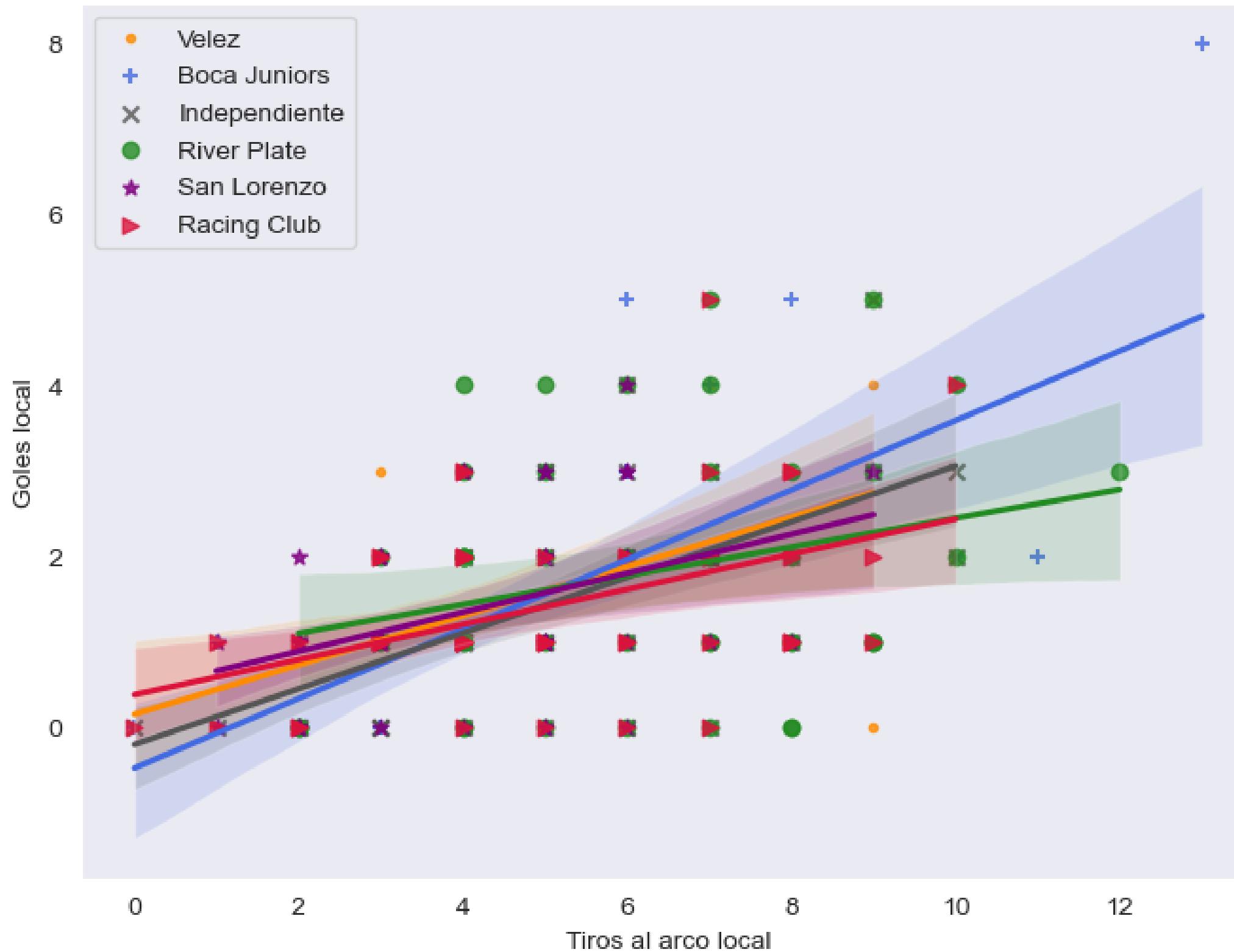
## Descripcion del dataset

El dataframe posee en su mayoria variables cuantitativas que analizan en profundidad las situaciones dadas en cada uno de los partidos de la liga. Es un dataset que consta de 34 columnas.

Para mejor compresión a primera vista, la primer columna indica el torneo o campeonato en el cual se disputo el partido, la segunda el equipo local, la tercera el equipo visitante y las restantes indican las diferentes situaciones dentro del partido indicando si fueron situaciones para el local o para el visitante.

	torneo	equipo_local	equipo_visitante	goles_local	goles_visitante	posesion_local	posesion_visitante	tiros_arco_local	tiros_arco_visitante
1536	Campeonato 2018/19	River Plate	Belgrano	0	0	79.0	21.0	7.0	0.0
1550	Campeonato 2018/19	River Plate	Argentinos	0	0	67.0	33.0	4.0	4.0
1573	Campeonato 2018/19	River Plate	San Martin (SJ)	4	1	65.0	35.0	4.0	6.0
1621	Campeonato 2018/19	River Plate	Def y Justicia	0	1	66.0	34.0	4.0	4.0
1637	Campeonato 2018/19	River Plate	Aldosivi	1	0	51.0	49.0	3.0	1.0

Relación entre Tiros al arco y goles por equipo



## Grafico analisis Multivariado:

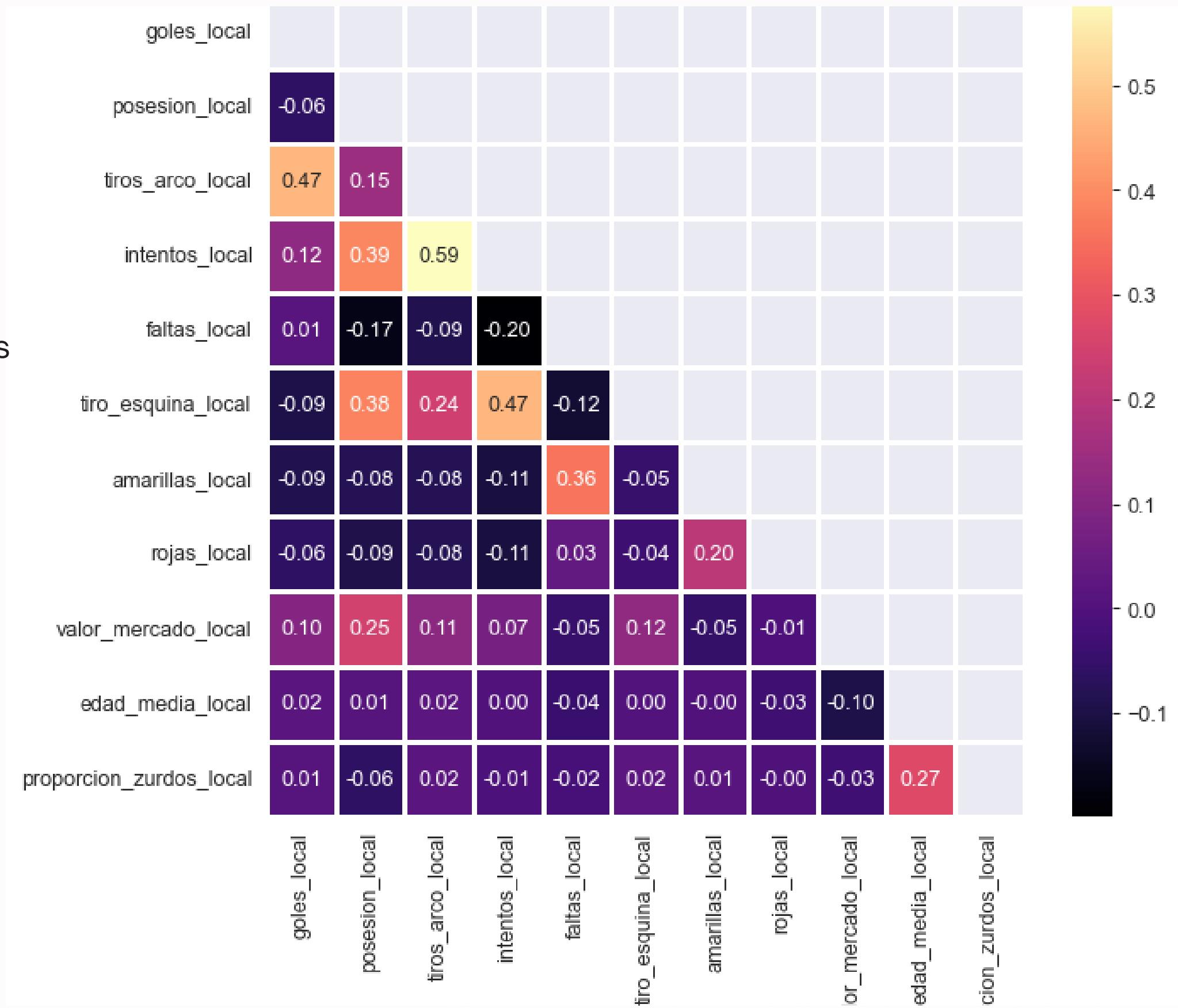
Grafico demostrando la correlación trazando una recta por cada uno de los equipos, con color e identificacion para demostrar que independientemente del equipo existe la correlación entre tiros al arco y goles.

El gráfico parecía mostrar una correlación evidente entre las variables. Sin embargo al analizar mas profundamente veremos que esto es en principio engañoso a los fines de realizar un modelo.

# Heatmap: Busqueda de correlaciones.

Podemos ver que existe una correlación moderada (0,47) entre Tiros al arco y goles. Es decir que a mas tiros al arco se anotarán mas goles. Razón por la cual, en principio podríamos realizar un modelo de regresión lineal para predecir cantidad de goles en un partido.

**Sin embargo, aun desconocemos si esta correlación sera suficiente a los fines de una regresión lineal.**



# SELECCION DEL MODELO

## Regresion lineal:

como anticipé anteriormente intenté mediante regresión lineal buscar la forma de predecir la cantidad de goles que podían hacerse en un partido. Sin embargo a la hora de los resultados, los mismos no fueron para nada alentadores.

Error cuadrado medio (MSE): 0.8151494427229604

Error Absoluto Medio (MAE): 0.718590911885195

coeficiente r<sup>2</sup>: 0.3114380769517333

El coeficiente de determinación, indica la proporción de la variabilidad total de los datos que el modelo es capaz de explicar. En este caso, el valor de 0.31 significa que el modelo solo puede explicar aproximadamente el 31% de la variabilidad de los datos.



**Razon por la cual es necesario redigir el analisis hacia un modelo de clasificacion para lograr un modelo mas fiable.**

# Modelo de clasificación

Como metodo para solucionar el problema, debido a las dificultades de realizar predicciones en cuestiones deportivas, intentaré tomar un rumbo distinto eligiendo otro modelo, dejando de lado la regresión y pasando a un modelo de clasificación.

Intentando en vez de predecir cuantos goles va a realizar un equipo, reducirlo a un modelo mas simple intentando clasificar si hara por lo menos un gol en el partido.



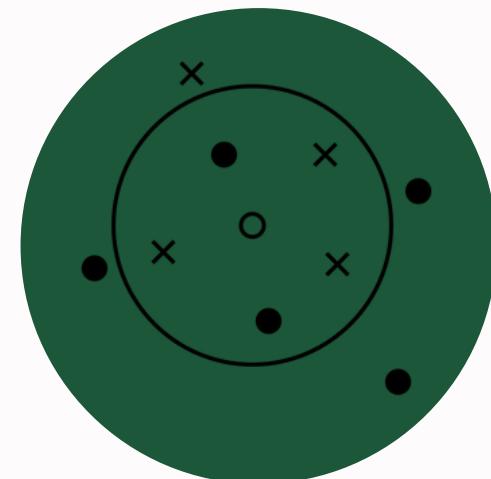
## Modelos de clasificación

Al cambiar a modelos de clasificación, las métricas de presición son mas adecuadas y fiables a los fines del análisis.



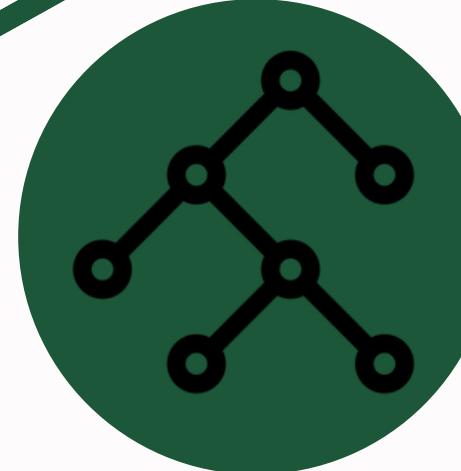
**Logistic Regression**

Precisión del modelo: 0.83



**K-Nearest Neighbors Classifier**

Precisión del modelo: 0.75



**Random Forest**

Precisión del modelo: 0.81

# Random Forest



## Precisión media: 0.79

EL modelo obtuvo una precisión del 79.86% en las diferentes divisiones de los datos durante el proceso de validación cruzada

## Desviación estándar: 0.010

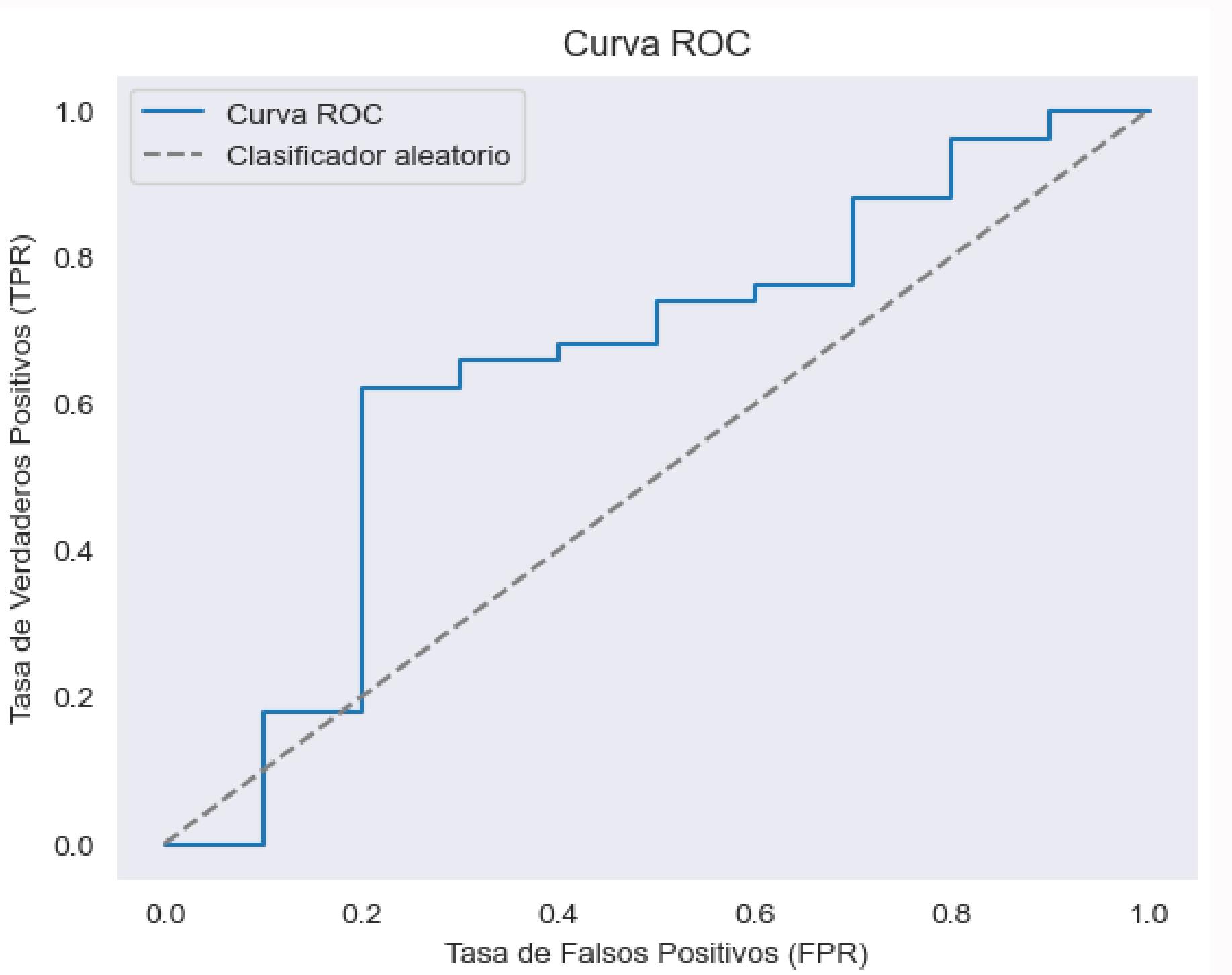
Una desviación estándar baja indica que los resultados tienen poca variación y que el modelo tiene un rendimiento consistente en todas las divisiones de los datos.

# CURVA ROC

sin embargo, no todo lo que brilla es oro:

ROC AUC = 0.648

El valor de `roc_auc` (área bajo la curva ROC) de 0.65 indica que el modelo tiene una capacidad de discriminación razonable, pero no es excelente. Un valor de 0.5 representa una clasificación aleatoria, mientras que un valor de 1.0 indica una clasificación perfecta. En este caso, el valor de 0.66 sugiere que el modelo puede distinguir entre clases positivas y negativas mejor que una clasificación aleatoria, pero aún puede haber margen de mejora para obtener una clasificación más precisa.



# CONCLUSIONES:

- La regresión lineal en cuestiones deportivas es compleja debido a los factores que influyen en los resultados, como el rendimiento individual, tácticas, motivación y lesiones. Además, la falta de linealidad y la alta variabilidad en los datos deportivos presentan desafíos adicionales.
- Si bien las métricas de precisión y desviación estándar pueden indicar un rendimiento aparentemente bueno del modelo, la curva ROC puede arrojar resultados inesperados y desfavorables al tratar de predecir eventos deportivos.
- El modelo de clasificación con una precisión de 0.8 muestra un buen rendimiento al predecir correctamente el 80% de las muestras. Sin embargo, su curva ROC de 0.65 indica que tiene dificultades para discriminar entre clases, lo que sugiere margen de mejora en la capacidad de clasificación.

# MUCHAS GRACIAS

*Por una excelente  
cursada.*

*Mesaglio Leonardo Matias*

**CODER HOUSE**