



MACHINE LEARNING FC

DEL BALÓN AL MODELO

Cómo convertir un pase, una falta o una tarjeta en predicción inteligente.

Fernando Sanchez Llorens

UN GOL AL AZAR: MODELANDO EL CAOS DEL FÚTBOL CON CIENCIA DE DATOS

- Predicción en otros deportes
- Métrica gol

¿Por qué es difícil predecir fútbol?



EXPLORATORY DATA ANALYSIS

Continuación del proyecto “El fútbol Bajo la Lupa”



01

Tamaño del dataset

Más de 9.000 partidos de la Premier League desde 2000/2001

02

Distribución de clases

Victoria Local 45% aprox
Empate 25% aprox
Victoria Visitante 30%

03

Datos analizados

Se exploran métricas como XG, posesión de balón, disparos a puertas, faltas, tarjetas, etc.

FEATURE ENGINEERING

PREPARACIÓN DE DATOS

01 Construir base de datos

Utilizar merge para unir 2 bases y obtener fechas.

03 Comprobación de nulos

Se encontraron mucho valores nulos en columnas importantes.

02 Eliminar temporada 2024/2025

Temporada incompleta y sin sentido cronológico.

04 Corrección de datos

Datos erroneos en el dataset.
Se buscaron manualmente y se reemplazaron.

FEATURE ENGINEERING

PREPARACIÓN DE DATOS

01 Creación de features

Se crearon features como save_ratio.

03 Método de cálculo de estadísticas

Dependiendo equipo y localia.

02 Imputaciones y transformaciones

Uso de medianas para completar columnas vacías o con 0.

04 División de dataset

- Train: temporadas anteriores + 1^a mitad de 2023/24.
- Test: 2^a mitad de 2023/24.

MODELOS UTILIZADOS

Random Forest

Basico
Flexible
Estricto
Escalado
Balanced

XGBoost

Flexible
Estricto
Escalado

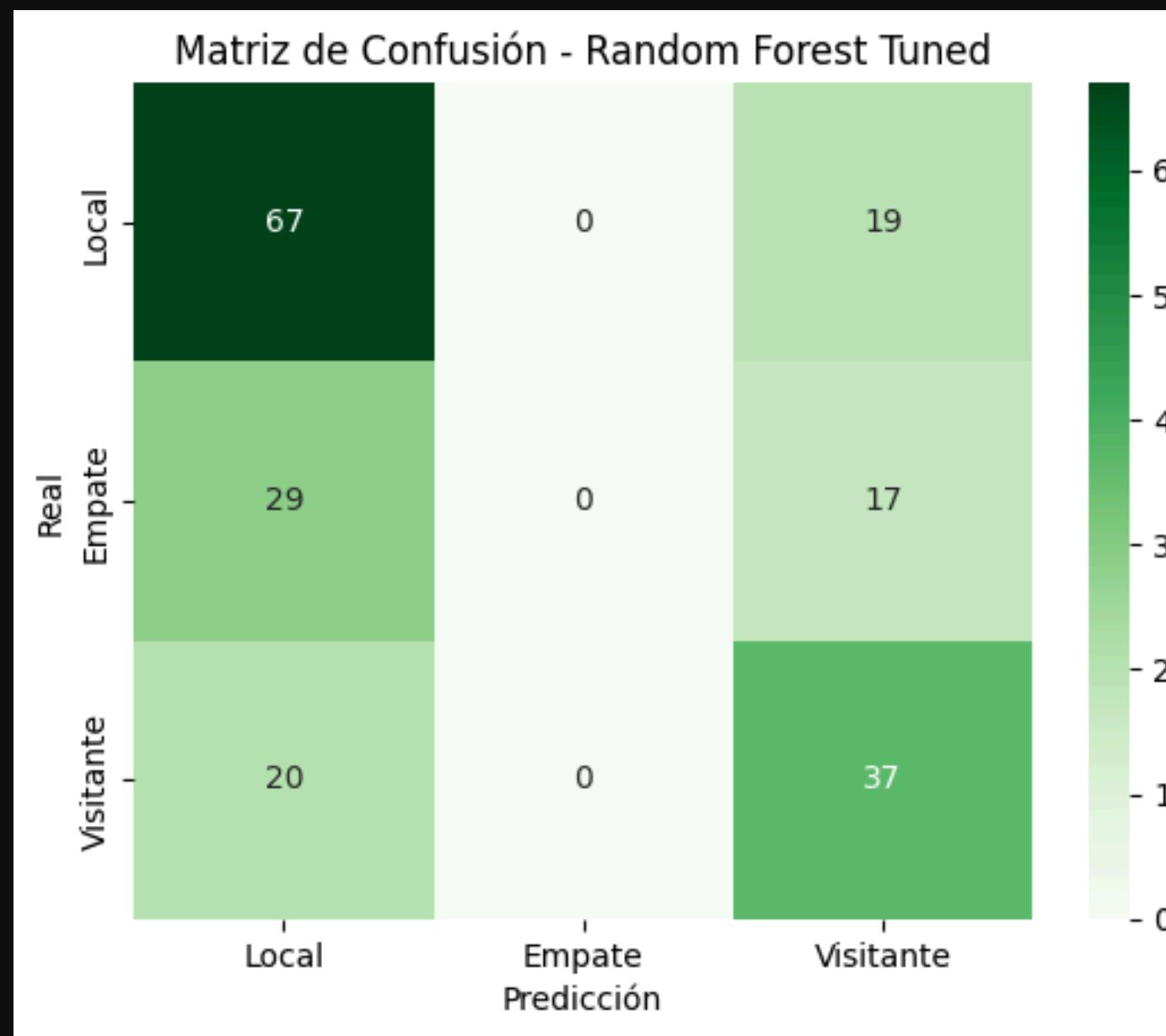
LightGBM

Estricto
Escalado

StackingClassifier

RF
XGB
LGBM
LogisticRegression

OVERFITTING MÁXIMO



TRAIN

Train Accuracy: 0.9721348314606741
 Train F1 Score (macro): 0.9691603085162851

Train Classification Report:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.99 | 0.97 | 0.98 | 4092 |
| 1 | 0.94 | 0.96 | 0.95 | 2204 |
| 2 | 0.97 | 0.98 | 0.97 | 2604 |
| accuracy | | | 0.97 | 8900 |
| macro avg | 0.97 | 0.97 | 0.97 | 8900 |
| weighted avg | 0.97 | 0.97 | 0.97 | 8900 |

TEST

Accuracy: 0.5502645502645502
 F1 Score (macro): 0.4108657019548108

Classification Report:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.58 | 0.78 | 0.66 | 86 |
| 1 | 0.00 | 0.00 | 0.00 | 46 |
| 2 | 0.51 | 0.65 | 0.57 | 57 |
| accuracy | | | 0.55 | 189 |
| macro avg | 0.36 | 0.48 | 0.41 | 189 |
| weighted avg | 0.42 | 0.55 | 0.47 | 189 |

Ajustar Hiperparametros

Retocamos los hiperparametros para afinarlos lo maximo posible.

LUCHA CONTRA EL OVERFITTING

Importancia de las variables

Vimos que el modelo daba un 30%-35% de importancia a los save_ratio.

Poca importancia a otras métricas.



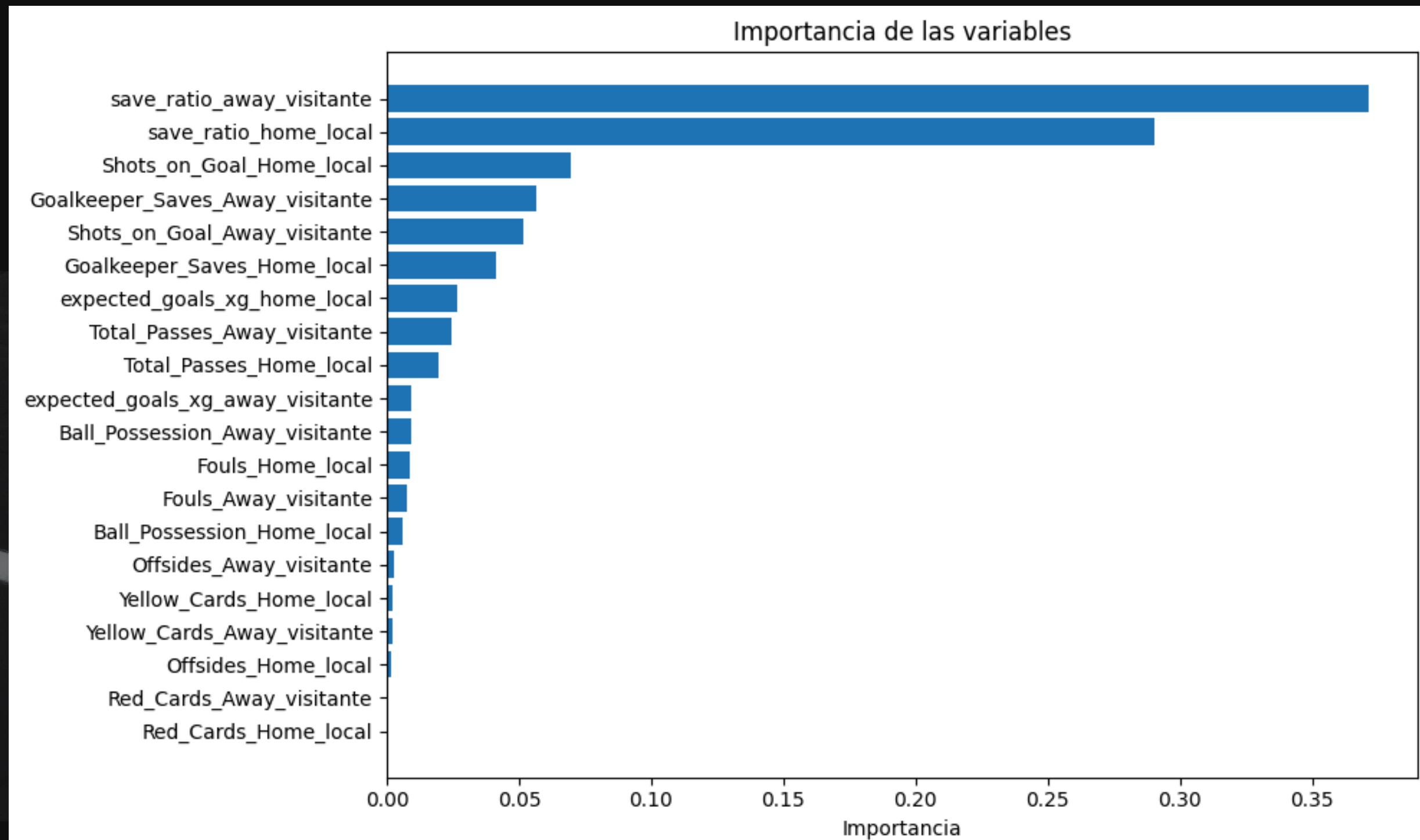
Media de variables importantes

Miramos la media de las variables importantes.

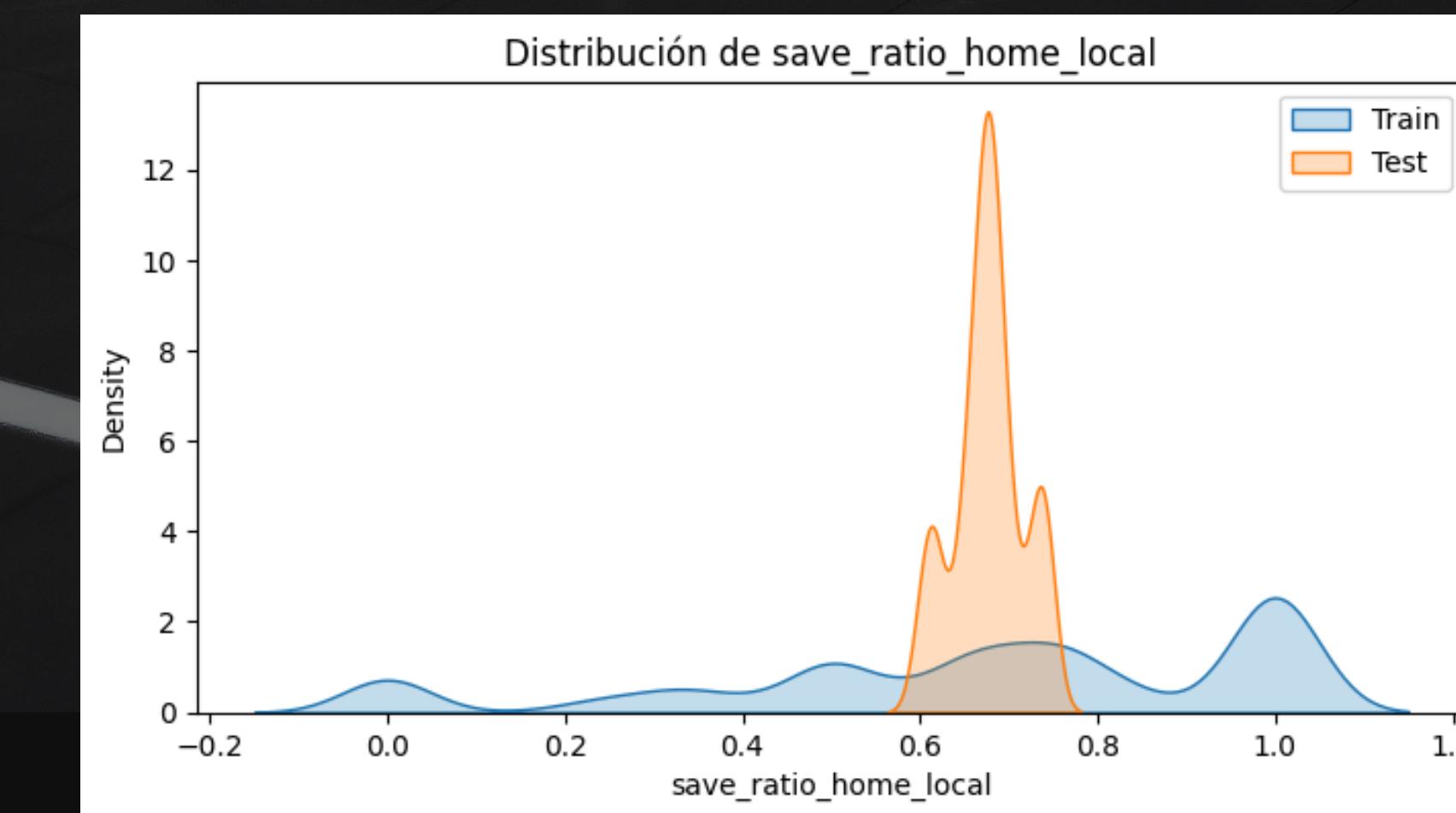
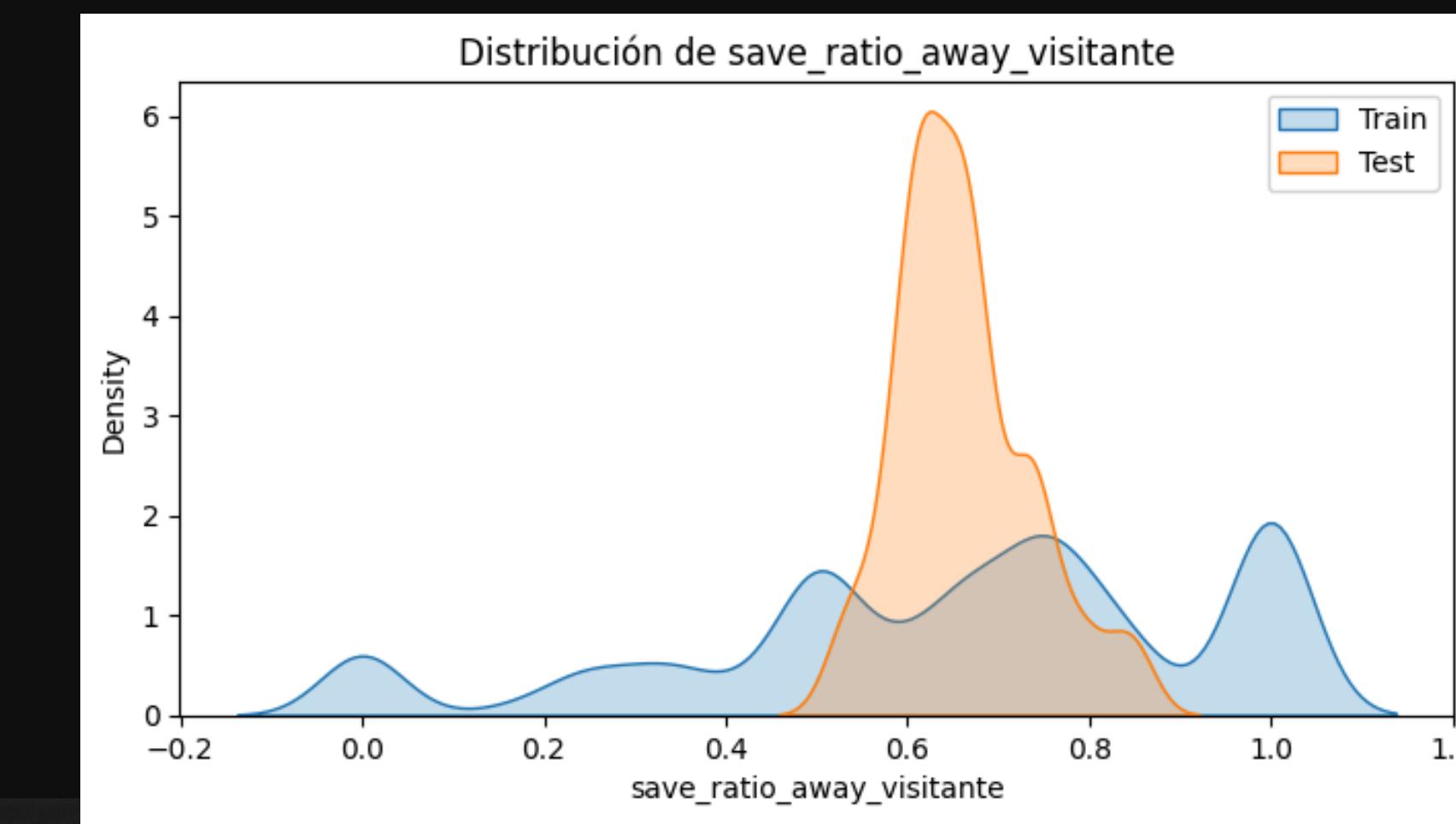
LUCHA CONTRA EL OVERFITTING

```
Media por feature en TRAIN:  
save_ratio_away_visitante      0.651028  
save_ratio_home_local          0.674323  
Shots_on_Goal_Home_local       4.562640  
Shots_on_Goal_Away_visitante   3.779888  
dtype: float64
```

```
Media por feature en TEST:  
save_ratio_away_visitante      0.661338  
save_ratio_home_local          0.676388  
Shots_on_Goal_Home_local       4.644840  
Shots_on_Goal_Away_visitante   4.693122  
dtype: float64
```



LUCHA CONTRA EL OVERFITTING



Eliminar features

Eliminamos las features mas importantes.

Escalamos variables

Tratamos de igualar los pesos o importancias de variables con mucha diferencia numerica

LUCHA CONTRA EL OVERFITTING

Eliminar Features

Eliminamos las features menos importantes

LUCHA CONTRA EL OVERFITTING

Train test split

Sobre el dataset de X

```
Accuracy (val): 0.9668539325842697
```

```
F1 Score (macro - val): 0.9632516037910411
```

```
Classification Report (val):
```

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.97 | 0.98 | 0.98 | 818 |
| 1 | 0.94 | 0.93 | 0.94 | 441 |
| 2 | 0.98 | 0.97 | 0.98 | 521 |
| accuracy | | | 0.97 | 1780 |
| macro avg | 0.96 | 0.96 | 0.96 | 1780 |
| weighted avg | 0.97 | 0.97 | 0.97 | 1780 |

```
ROC AUC Score (val - One-vs-Rest): 0.9978408801085129
```

Cross Validation Score

Para comprobar el modelo en train

```
array([0.95972482, 0.96664    , 0.96922864, 0.97152187, 0.95117853,
       0.95229484, 0.95492594, 0.95660611, 0.99119579, 0.98723376])
```

MODELO FINAL

Random Forest con Escalado

Rendimiento más estable a largo plazo.



```
Accuracy: 0.5767195767195767  
F1 Score (macro): 0.5211544812263794
```

Classification Report:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.64 | 0.69 | 0.66 | 86 |
| 1 | 0.43 | 0.22 | 0.29 | 46 |
| 2 | 0.54 | 0.70 | 0.61 | 57 |
| accuracy | | | 0.58 | 189 |
| macro avg | 0.54 | 0.54 | 0.52 | 189 |
| weighted avg | 0.56 | 0.58 | 0.56 | 189 |



Machine Learning

Modelo eficiente aunque el set de test
no ayudo a la predicción

CONCLUSIONES



Sentimiento

Orgullo con el trabajo realizado y
líneas de investigaciones



Machine Learning FC

GRACIAS

Modelando el Caos del Fútbol con Ciencia de Datos

Fernando Sanchez Llorens

