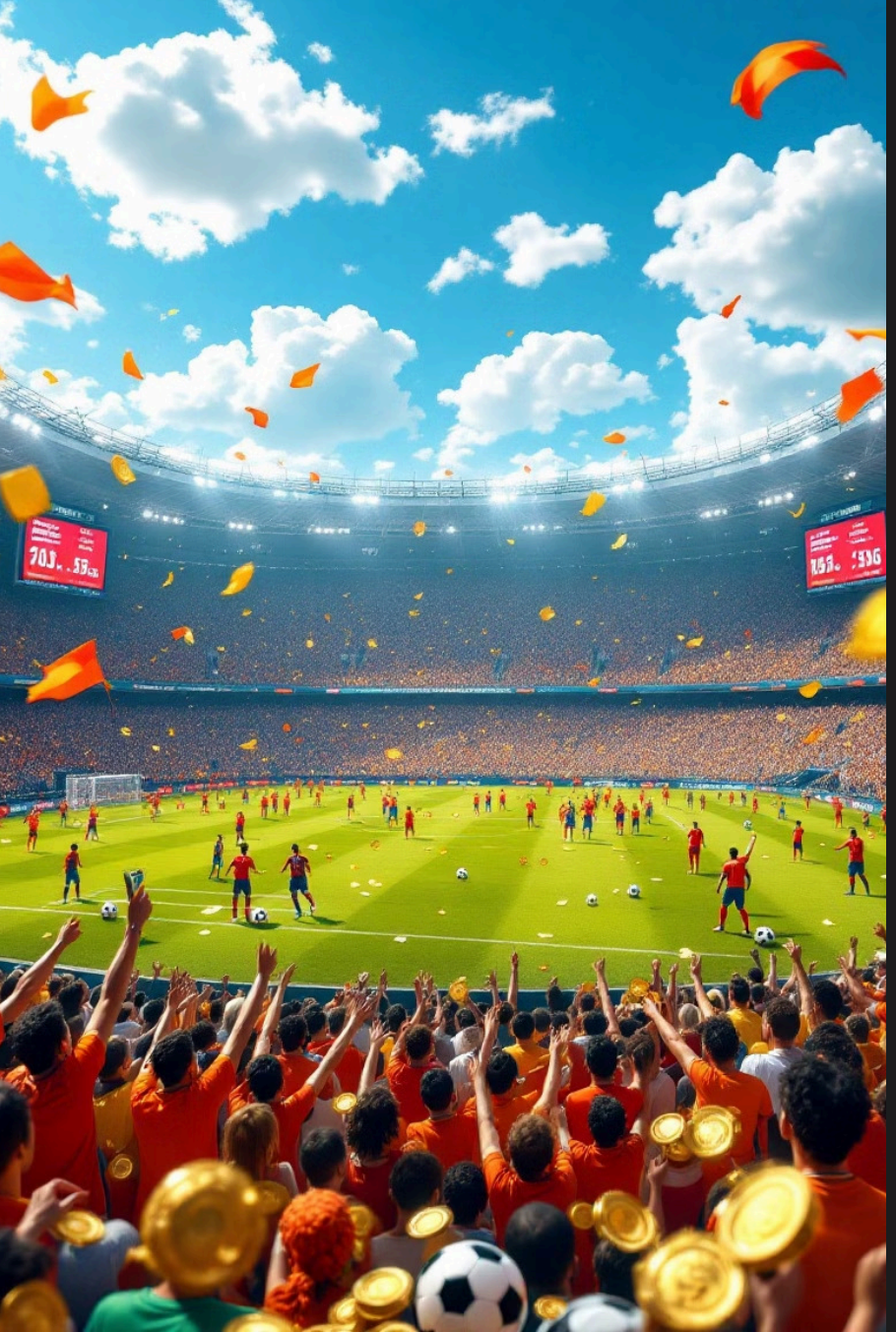
OctoBet:

Predicción de Goles anuales con ML

por Fran Martinez

Made with Gamma



Propuesta de valor

1 **Objetivo**

Prediccion de goles anuales de jugadores

2 **Modelo de regresión**

Target: goles anuales

3 **Aplicaciones**

Apuestas e información deportiva, análisis deportivo y scouting.

Made with Gamma

Etapas del análisis

1. **Primera aproximación**
   EDA: datos fifa y primeros modelos
2. **Cambio de enfoque**
   Nuevas busquedas de datos y web scraping
3. **Feature Engineering**
   Nuevo EDA y primeros modelos
4. **Modelado**
   Implementación modelos, metricas e hiperparametrización
5. **Streamlit**
   Propuesta de valor con Front

Made with Gamma

Correlaciones datos Fifa-goles

	id_fifa	rating	tirosXp	goles_esperados	value_eur	shooting	pace	attacking_crossing	attacking_finishing	attacking_heading_accuracy	attacking_volleys	skill_dribbling	skill_curve	skill_fk_accuracy	skill_ball_control	movement_acceleration	movement_sprint_speed	movement_agility	power_shot_power	power_jumping	power_stamina	power_strength	power_long_shots	mentality_aggression	mentality_positioning	mentality_vision	mentality_penalties	mentality_composure	goles
value_eur	-0.18	0.13	0.1	0.13	1	0.45	0.41	0.4	0.42	0.2	0.39	0.49	0.41	0.32	0.59	0.37	0.42	0.36	0.44	0.14	0.5	0.18	0.42	0.31	0.47	0.54	0.32	0.63	0.11
shooting	-0.29	0.047	0.18	0.21	0.45	1	0.37	0.58	0.98	-0.032	0.91	0.8	0.76	0.71	0.77	0.43	0.31	0.58	0.87	-0.092	0.28	-0.12	0.94	-0.12	0.89	0.78	0.81	0.57	0.2
pace	-0.069	0.046	0.089	0.049	0.41	0.37	1	0.48	0.41	-0.22	0.36	0.54	0.42	0.24	0.4	0.96	0.97	0.73	0.23	0.024	0.26	-0.29	0.29	-0.16	0.47	0.33	0.26	0.25	0.042
attacking_crossing	-0.31	0.065	0.034	0.041	0.4	0.58	0.48	1	0.53	-0.21	0.52	0.75	0.79	0.64	0.7	0.54	0.41	0.62	0.53	-0.085	0.42	-0.27	0.6	0.04	0.67	0.69	0.4	0.5	0.052
attacking_finishing	-0.23	0.042	0.2	0.22	0.42	0.98	0.41	0.53	1	-0.064	0.87	0.77	0.69	0.64	0.72	0.45	0.34	0.58	0.76	-0.12	0.21	-0.17	0.87	-0.21	0.87	0.73	0.79	0.5	0.21
attacking_heading_accuracy	-0.34	0.046	0.0085	0.095	0.2	-0.032	-0.22	-0.21	-0.064	1	-0.015	-0.18	-0.18	-0.22	0.0081	-0.32	-0.12	-0.35	0.099	0.5	0.21	0.69	-0.044	0.43	-0.083	-0.11	0.022	0.26	0.079
attacking_volleys	-0.3	0.035	0.17	0.2	0.39	0.91	0.36	0.52	0.87	-0.015	1	0.73	0.72	0.66	0.7	0.39	0.3	0.54	0.77	-0.052	0.22	-0.11	0.84	-0.12	0.81	0.7	0.78	0.53	0.19
skill_dribbling	-0.27	0.066	0.12	0.13	0.49	0.8	0.54	0.75	0.77	-0.18	0.73	1	0.81	0.66	0.87	0.59	0.47	0.73	0.67	-0.14	0.36	-0.26	0.77	-0.061	0.84	0.82	0.61	0.63	0.13
skill_curve	-0.33	0.062	0.094	0.11	0.41	0.76	0.42	0.79	0.69	-0.18	0.72	0.81	1	0.79	0.78	0.49	0.34	0.64	0.71	-0.1	0.38	-0.24	0.78	-0.026	0.76	0.79	0.61	0.59	0.11
skill_fk_accuracy	-0.3	0.032	0.11	0.098	0.32	0.71	0.24	0.64	0.64	-0.22	0.66	0.66	0.79	1	0.65	0.32	0.17	0.5	0.65	-0.096	0.25	-0.24	0.75	-0.073	0.62	0.7	0.64	0.48	0.1
skill_ball_control	-0.4	0.077	0.1	0.14	0.59	0.77	0.4	0.7	0.72	0.0081	0.7	0.87	0.78	0.65	1	0.44	0.34	0.62	0.71	-0.01	0.47	-0.08	0.76	0.14	0.77	0.86	0.58	0.78	0.13
movement_acceleration	-0.07	0.043	0.092	0.044	0.37	0.43	0.96	0.54	0.45	-0.32	0.39	0.59	0.49	0.32	0.44	1	0.87	0.79	0.26	-0.051	0.25	-0.4	0.35	-0.21	0.52	0.38	0.3	0.24	0.039
movement_sprint_speed	-0.068	0.046	0.081	0.049	0.42	0.31	0.97	0.41	0.34	-0.12	0.3	0.47	0.34	0.17	0.34	0.87	1	0.63	0.2	0.085	0.26	-0.18	0.22	-0.11	0.4	0.26	0.21	0.24	0.041
movement_agility	-0.093	0.02	0.096	0.057	0.36	0.58	0.73	0.62	0.58	-0.35	0.54	0.73	0.64	0.5	0.62	0.79	0.63	1	0.42	-0.087	0.28	-0.46	0.54	-0.2	0.65	0.58	0.42	0.43	0.068
power_shot_power	-0.37	0.051	0.13	0.17	0.44	0.87	0.23	0.53	0.76	0.099	0.77	0.67	0.71	0.65	0.71	0.26	0.2	0.42	1	-0.0037	0.37	0.067	0.86	0.097	0.71	0.72	0.67	0.6	0.15
power_jumping	-0.26	0.0068	0.025	0.021	0.14	-0.092	0.024	-0.085	-0.12	0.5	-0.052	-0.14	-0.1	-0.096	-0.01	-0.051	0.085	-0.087	0.0037	1	0.25	0.44	-0.079	0.36	-0.1	-0.085	0.041	0.16	0.014
power_stamina	-0.32	0.092	0.02	0.032	0.5	0.28	0.26	0.42	0.21	0.21	0.22	0.36	0.38	0.25	0.47	0.25	0.26	0.28	0.37	0.25	1	0.28	0.32	0.52	0.36	0.44	0.15	0.51	0.026
power_strength	-0.24	0.04	-0.0079	0.055	0.18	-0.12	-0.29	-0.27	-0.17	0.69	-0.11	-0.26	-0.24	-0.24	-0.08	-0.4	-0.18	-0.46	0.067	0.44	0.28	1	-0.1	0.56	-0.17	-0.16	-0.096	0.19	0.037
power_long_shots	-0.33	0.049	0.15	0.18	0.42	0.94	0.29	0.6	0.87	-0.044	0.84	0.77	0.78	0.75	0.76	0.35	0.22	0.54	0.86	-0.079	0.32	-0.1	1	-0.054	0.82	0.79	0.73	0.58	0.17
mentality_aggression	-0.34	0.03	-0.085	0.033	0.31	-0.12	-0.16	0.04	-0.21	0.43	-0.12	-0.061	0.026	0.073	0.14	-0.21	-0.11	-0.2	0.097	0.36	0.52	0.56	-0.054	1	-0.079	0.097	-0.17	0.34	-0.038
mentality_positioning	-0.27	0.048	0.16	0.19	0.47	0.89	0.47	0.67	0.87	-0.083	0.81	0.84	0.76	0.62	0.77	0.52	0.4	0.65	0.71	-0.1	0.36	-0.17	0.82	-0.079	1	0.79	0.67	0.55	0.18
mentality_vision	-0.34	0.063	0.098	0.12	0.54	0.78	0.33	0.69	0.73	-0.11	0.7	0.82	0.79	0.7	0.86	0.38	0.26	0.58	0.72	-0.085	0.44	-0.16	0.79	0.097	0.79	1	0.58	0.7	0.11
mentality_penalties	-0.26	0.028	0.18	0.21	0.32	0.81	0.26	0.4	0.79	0.022	0.78	0.61	0.61	0.64	0.58	0.3	0.21	0.42	0.67	-0.041	0.15	-0.096	0.73	-0.17	0.67	0.58	1	0.44	0.21
mentality_composure	-0.47	0.078	0.066	0.11	0.63	0.57	0.25	0.5	0.5	0.26	0.53	0.63	0.59	0.48	0.78	0.24	0.24	0.43	0.6	0.16	0.51	0.19	0.58	0.34	0.55	0.7	0.44	1	0.11
goles	-0.068	0.44	0.6	0.9	0.11	0.2	0.042	0.052	0.21	0.079	0.19	0.13	0.11	0.1	0.13	0.039	0.041	0.068	0.15	0.014	0.026	0.037	0.17	-0.038	0.18	0.11	0.21	0.11	1

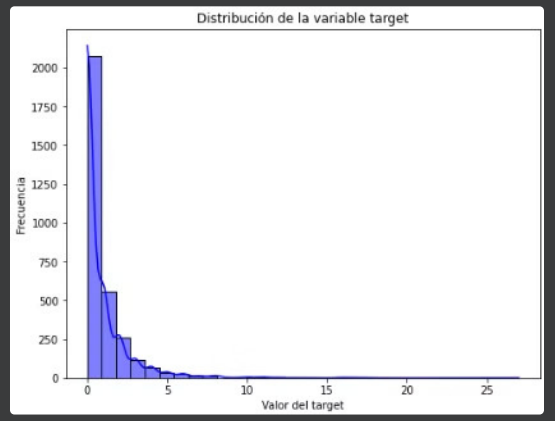
Made with Gamma

Web scraping y nuevo enfoque

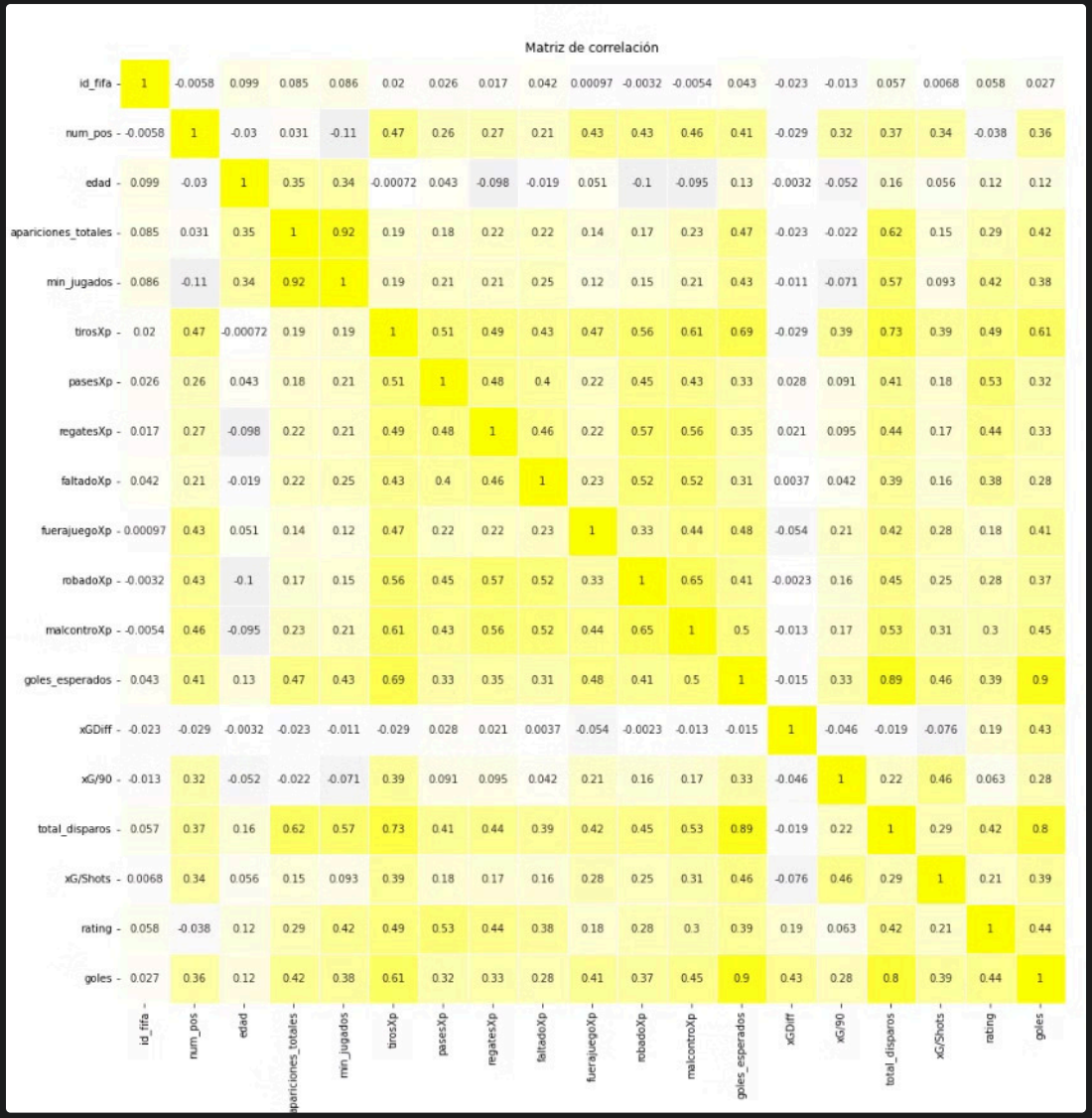
Obtención

Obtención de nuevos datos des stats reales de Jugadores del 2022

Distribución target



Made with Gamma



Made with Gamma

Feature Engineering

1

Limpieza traduccion y orden del dataset

Entender el nuevo dataset y poder ver el peso de determinadas features

2

Tratamiento de los nan values

Eliminar y completar nan values

3

Generación de patrones y mapeados

Creación de grupos con diferente peso

Made with Gamma

Modelos utilizados

Modelos de regresión simples

Regresión lineal	Modelo simple que ha planteado un muy buen rendimiento
Regresión polinómica	Grado 2 y 3
Ridge y Laso y ElasticNet	ElasticNet ganador. Params: L1_ratio y alpha

Made with Gamma

Modelos utilizados

Arboles de decisión

Arboles de decisión	Modelo simple
Random Forest	Modelo que ha dado muy buenos resultado. Feature importances.
XGBOOST	Modelo que ha dado peor rendimiento del esperado
LightGBM	Gradient busting decision tree, buen resultado y sobre todo entrenamiento rapidísimo

Made with Gamma

Modelos utilizados

Red Neuronal y modelo no supervisado

Red Neuronal	Modelo muy potente que siempre ha estado en el top de rendimiento
PCA (no supervisado)	Reducción dimensionalidad con grandes volumenes de features

Made with Gamma

Pipeline & Grid Search con CV

Parámetros utilizados

Escaladores	StandardScaler y MinMaxScalerRe
Reducción dimensionalidad	PCA
XGboost	• n_estimators: 100 • max_depth: 3, 6, 10 • learning_rate: 0.01, 0.1 • subsample: 0.8 • colsample_bytree: 0.6, 0.8
LightGBM	• n_estimators: 100 • max_depth: 3, 6, 10 • learning_rate: 0.01, 0.1 • subsample: 0.8 • colsample_bytree: 0.6, 0.8 • num_leaves: 31, 50, 100
Red Neuronal	2 capas densas (64 y 32 neuronas) con ReLU; loss: mean_absolute_error; 50 épocas, batch size 32.

Made with Gamma

Validación y Evaluación de Modelos

Utilizados: R2 y MAE

Jugando con 3, 4, 7, 8 y las 17 features

Con todas utilicé el PCA de 5, 6 y 10

Made with Gamma

MAE

'Contenido de ../models/r2/3features/reultados.csv:'

	Model	Best_params	MAE	R-squared
0	Linear Regression	{'classifier': LinearRegression(), 'scaler': StandardScaler()}	0.461006	0.813920
1	Polynomial Regression_2	{'classifier': LinearRegression(), 'polynomial__degree': 2, 'scaler': StandardScaler()}	0.411557	0.834132
2	Polynomial Regression_3	{'classifier': LinearRegression(), 'polynomial__degree': 3, 'scaler': StandardScaler()}	0.411480	0.835129
3	Decision Tree	{'classifier__max_depth': 5, 'classifier__min_samples_leaf': 5, 'classifier__min_samples_split': 10, 'scaler': StandardScaler()}	0.439788	0.775243
4	Random Forest	{'classifier__max_depth': 10, 'classifier__min_samples_leaf': 2, 'classifier__min_samples_split': 5, 'classifier__n_estimators': 100, 'scaler': StandardScaler()}	0.418154	0.799667
5	XGBoost (Pipeline)	{'classifier__colsample_bytree': 0.8, 'classifier__learning_rate': 0.05, 'classifier__max_depth': 3, 'classifier__n_estimators': 200, 'classifier__subsample': 0.8, 'scaler': MinMaxScaler()}	0.421111	0.824559
6	LightGBM (Pipeline)	{'classifier__colsample_bytree': 0.6, 'classifier__learning_rate': 0.1, 'classifier__max_depth': 3, 'classifier__n_estimators': 100, 'classifier__num_leaves': 31, 'classifier__subsample': 0.8, 'scaler': StandardScaler()}	0.496765	0.766169
7	Neural Network	-	0.399525	0.841928

'\n----------\n'

'Contenido de ../models/r2/4features/reultados.csv:'

	Model	Best_params	MAE	R-squared
0	Linear Regression	{'classifier': LinearRegression(), 'scaler': MinMaxScaler()}	0.455401	0.815513
1	Polynomial Regression_2	{'classifier': LinearRegression(), 'polynomial__degree': 2, 'scaler': None}	0.410564	0.841785
2	Polynomial Regression_3	{'classifier': LinearRegression(), 'polynomial__degree': 3, 'scaler': StandardScaler()}	0.402301	0.838908
3	Decision Tree	{'classifier__max_depth': 5, 'classifier__min_samples_leaf': 5, 'classifier__min_samples_split': 10, 'scaler': StandardScaler()}	0.439824	0.775312
4	Random Forest	{'classifier__max_depth': 5, 'classifier__min_samples_leaf': 2, 'classifier__min_samples_split': 5, 'classifier__n_estimators': 10, 'scaler': StandardScaler()}	0.445230	0.786889
5	XGBoost (Pipeline)	{'classifier__colsample_bytree': 0.8, 'classifier__learning_rate': 0.1, 'classifier__max_depth': 6, 'classifier__n_estimators': 10, 'classifier__subsample': 0.8, 'scaler': MinMaxScaler()}	0.664833	0.626424
6	LightGBM (Pipeline)	{'classifier__colsample_bytree': 0.8, 'classifier__learning_rate': 0.1, 'classifier__max_depth': 10, 'classifier__n_estimators': 10, 'classifier__num_leaves': 50, 'classifier__subsample': 0.8, 'scaler': StandardScaler()}	0.633522	0.669128
7	Neural Network	-	0.394072	0.851059

Made with Gamma

MAE

```
ontenido de ../models/mae/3features/reultados.csv:'
```

	Model	Best_params	MAE	R-squared
	Linear Regression	{'classifier': LinearRegression(), 'scaler': StandardScaler()}	0.461006	0.813920
	Polynomial Regression_2	{'classifier': ElasticNet(), 'classifier__alpha': 0.1, 'classifier__l1_ratio': 0.1, 'polynomial__degree': 2, 'polynomial__interaction_only': True, 'scaler': None}	0.421791	0.821169
	Polynomial Regression_3	{'classifier': ElasticNet(), 'classifier__alpha': 0.1, 'classifier__l1_ratio': 0.1, 'polynomial__degree': 3, 'polynomial__interaction_only': False, 'scaler': None}	0.419054	0.830546
	Decision Tree	{'classifier__max_depth': 5, 'classifier__min_samples_leaf': 5, 'classifier__min_samples_split': 5, 'scaler': StandardScaler()}	0.439788	0.775243
	Random Forest	{'classifier__max_depth': 5, 'classifier__min_samples_leaf': 2, 'classifier__min_samples_split': 5, 'classifier__n_estimators': 100, 'scaler': None}	0.400956	0.824892
	XGBoost (Pipeline)	{'classifier__colsample_bytree': 0.8, 'classifier__learning_rate': 0.2, 'classifier__max_depth': 2, 'classifier__n_estimators': 200, 'classifier__subsample': 0.8, 'scaler': MinMaxScaler()}	0.422047	0.823115
	LightGBM (Pipeline)	{'classifier__colsample_bytree': 0.6, 'classifier__learning_rate': 0.1, 'classifier__max_depth': 3, 'classifier__n_estimators': 100, 'classifier__num_leaves': 31, 'classifier__subsample': 0.8, 'scaler': None}	0.477647	0.780789
	Neural Network	-	0.358700	0.832047

```
n----------\n'
```

```
ontenido de ../models/mae/4features/reultados.csv:'
```

	Model	Best_params	MAE	R-squared
	Linear Regression	{'classifier': LinearRegression(), 'scaler': MinMaxScaler()}	0.456085	0.816297
	Polynomial Regression_2	{'classifier': ElasticNet(), 'classifier__alpha': 0.1, 'classifier__l1_ratio': 0.1, 'polynomial__degree': 2, 'polynomial__interaction_only': True, 'scaler': None}	0.426487	0.824395
	Polynomial Regression_3	{'classifier': ElasticNet(), 'classifier__alpha': 0.1, 'classifier__l1_ratio': 0.1, 'polynomial__degree': 3, 'polynomial__interaction_only': True, 'scaler': None}	0.401740	0.839252
	Decision Tree	{'classifier__max_depth': 5, 'classifier__min_samples_leaf': 5, 'classifier__min_samples_split': 5, 'scaler': StandardScaler()}	0.439788	0.775243
	Random Forest	{'classifier__max_depth': 10, 'classifier__min_samples_leaf': 5, 'classifier__min_samples_split': 5, 'classifier__n_estimators': 100, 'scaler': None}	0.393147	0.838696
	XGBoost (Pipeline)	{'classifier__colsample_bytree': 0.8, 'classifier__learning_rate': 0.2, 'classifier__max_depth': 3, 'classifier__n_estimators': 50, 'classifier__subsample': 0.8, 'scaler': None}	0.421612	0.814497
	LightGBM (Pipeline)	{'classifier__colsample_bytree': 0.8, 'classifier__learning_rate': 0.1, 'classifier__max_depth': 3, 'classifier__n_estimators': 100, 'classifier__num_leaves': 31, 'classifier__subsample': 0.8, 'scaler': None}	0.408596	0.828967
	Neural Network	-	0.353099	0.839450

```
n----------\n'
```

```
ontenido de ../models/mae/5features/reultados.csv:'
```

	Model	Best_params	MAE	R-squared
	Linear Regression	{'classifier': LinearRegression(), 'scaler': MinMaxScaler()}	0.450837	0.817580
	Polynomial Regression_2	{'classifier': ElasticNet(), 'classifier__alpha': 0.05, 'classifier__l1_ratio': 0.05, 'polynomial__degree': 2, 'scaler': StandardScaler()}	0.426913	0.838441
	Polynomial Regression_3	{'classifier': ElasticNet(), 'classifier__alpha': 0.05, 'classifier__l1_ratio': 0.05, 'polynomial__degree': 3, 'scaler': None}	0.403519	0.843977
	Decision Tree	{'classifier__max_depth': 5, 'classifier__min_samples_leaf': 5, 'classifier__min_samples_split': 5, 'scaler': StandardScaler()}	0.438632	0.775698
	Random Forest	{'classifier__max_depth': 10, 'classifier__min_samples_leaf': 2, 'classifier__min_samples_split': 5, 'classifier__n_estimators': 100, 'scaler': StandardScaler()}	0.407281	0.799386
	XGBoost (Pipeline)	{'classifier__colsample_bytree': 0.8, 'classifier__learning_rate': 0.05, 'classifier__max_depth': 4, 'classifier__n_estimators': 200, 'classifier__subsample': 0.8, 'scaler': StandardScaler()}	0.409604	0.827554
	LightGBM (Pipeline)	{'classifier__colsample_bytree': 0.8, 'classifier__learning_rate': 0.1, 'classifier__max_depth': 6, 'classifier__n_estimators': 100, 'classifier__num_leaves': 50, 'classifier__subsample': 0.8, 'scaler': StandardScaler()}	0.409482	0.820816
	Neural Network	-	0.357442	0.840810

Made with Gamma

Modelo ganador

5 Features

Regresión polinómica de grado 3

Ha resultado ser el que mejor generaliza

Red Neuronal

Mejores métricas

Made with Gamma

Conclusiones y Próximos Pasos

- Añadir las variables del fifa con modelos no supervisados
- Modelo de ML que prediga las apariciones de los jugadores
- Incluir datos de años anteriores y nutrir con más información
- Hacer lo mismo con las asistencias

Made with Gamma

FIN

Muchas Gracias

Made with Gamma