Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

Departamento de Ingeniería

TC3006C

Concentracion IA

Grupo: 101

Profesor: Jorge Adolfo Ramirez Uresti

"Análisis y Reporte sobre el desempeño del modelo"

Fecha: 14/09/2025

Alumno:
Jesús Ángel Guzmán Ortega A01799257

Resultados del Algoritmo Random Forest con scikit-learn

Para el análisis sobre el desempeño del set de datos de players_data_light-2024_2025.csv en la implementación utilizando un framework para un modelo de Random Forest utilizando scikit-learn, Se aplicaron dos enfoques distintos:

1. Regresión para predecir el número de goles (Gls) de los jugadores.
2. Clasificación binaria, donde se buscó determinar si un jugador anotó al menos un gol en la temporada (1 = sí, 0 = no)

El dataset utilizado corresponde a la temporada 2024-2025, conteniendo información estadística de desempeño y características de los jugadores.

Preprocesamiento

- Se eliminó únicamente la columna objetivo Gls del conjunto de atributos.

- Todas las variables categóricas fueron transformadas a valores numéricos mediante OrdinalEncoder, utilizando la opción de asignar -1 para categorías desconocidas.

- La división de los datos se realizó con un 80% para entrenamiento y un 20% para prueba.

- Para la clasificación, se generó una nueva variable objetivo binaria (y_class) basada en si el jugador tenía al menos un gol.

Configuración de los modelos

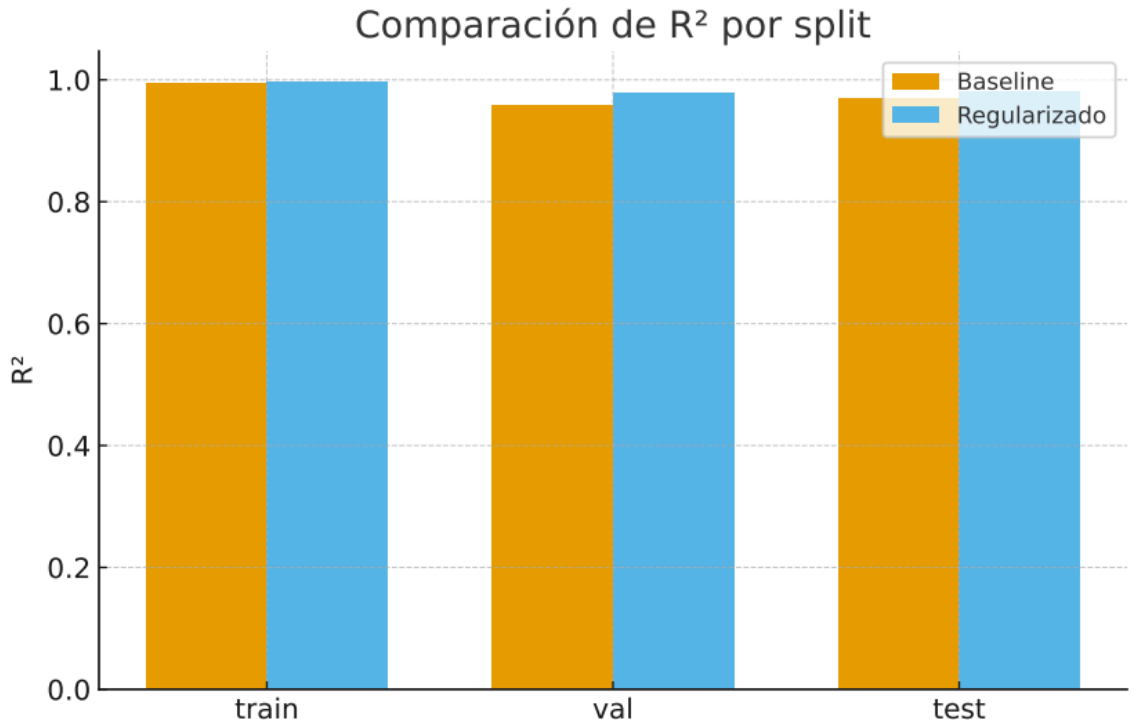
Regresión

- **n_estimators:** 100 árboles
- **max_depth:** 8
- **max_features:** "sqrt"
- **oob_score:** True

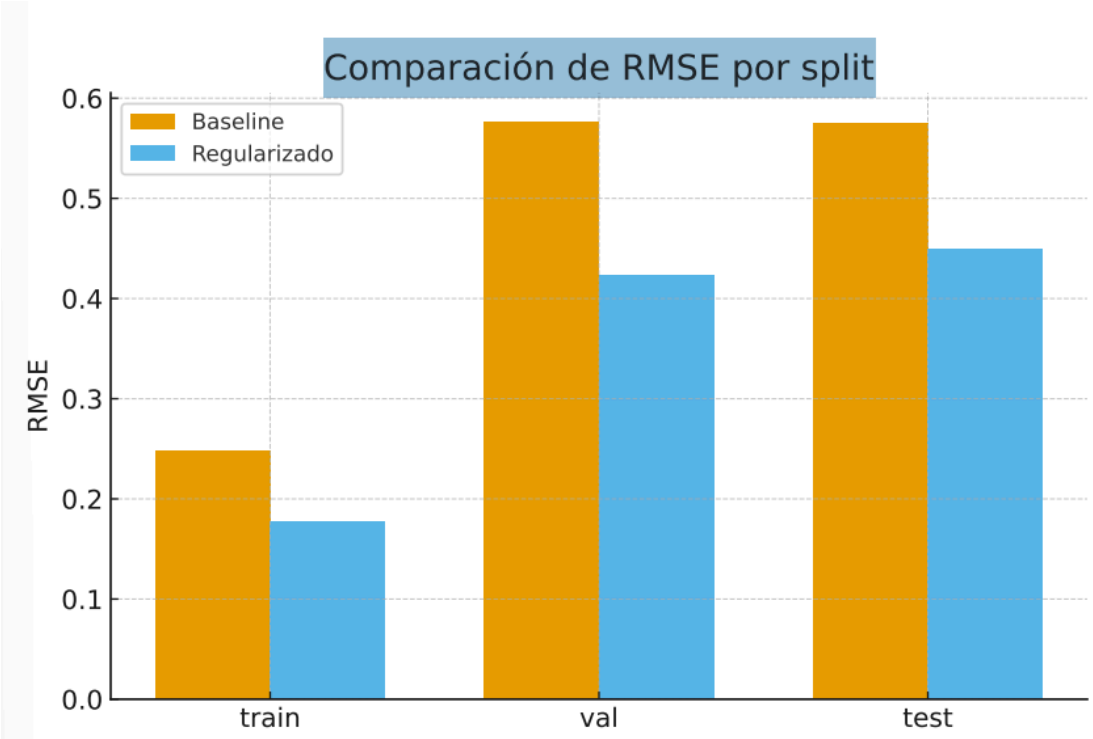
Clasificación

- **n_estimators:** 10 árboles
- **max_depth:** 2
- **max_features:** "sqrt"
- **min_samples_leaf:** 80
- **oob_score:** True

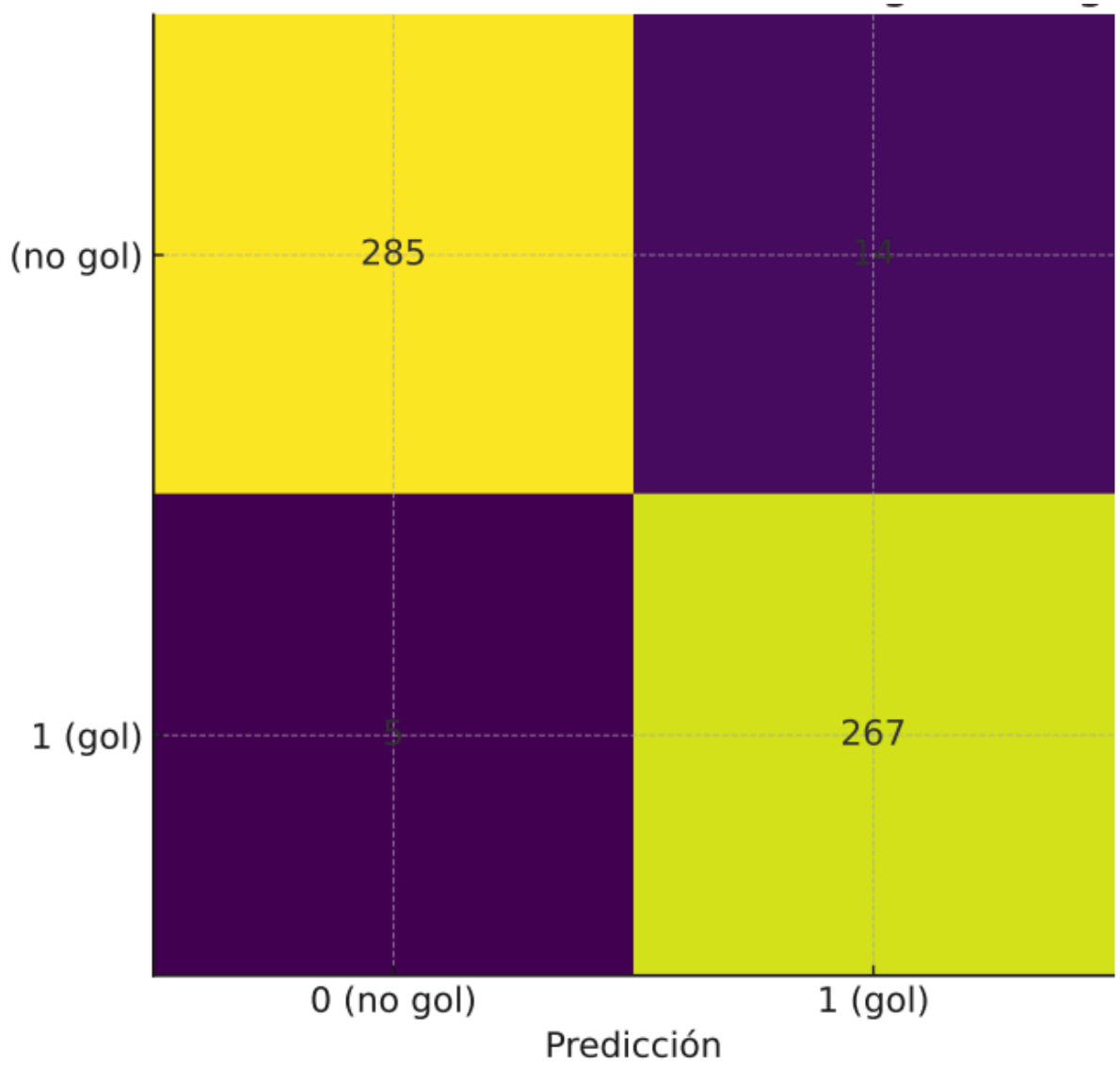
Comparación de R2 por Split



Comparación de RMSE por split



Matriz de confusión (Clasificación 1 gol vs 0 goles)



Métricas clasificación y diagnóstico

Métricas de clasificación (calculadas a partir de la matriz de confusión):
- Total muestras: 571
- Accuracy: 0.9667 (96.67%)
- Precision (clase 1): 0.9502 (95.02%)
- Recall (clase 1): 0.9816 (98.16%)
- F1-score (clase 1): 0.9656

Métricas de evaluación

```
C:\Users\trato\OneDrive\Documents\ITESM\SemestreQuant\IA\Module2_FrameworkUse\src>python random_forest.py
OOB score: 0.9569369704925536
Train R2: 0.9909744417655199
Test R2: 0.9711783903369255
Test MAE: 0.22607497048251912
Test RMSE: 0.5596750119469907
```

Estos valores muestran un modelo con excelente capacidad predictiva, ya que logra explicar más del 97% de la variabilidad en los goles con un error promedio menor a un cuarto de gol.

Esto se debe a que al estar mejor implementada la librería que la que implemente a mano, puede hacer una cantidad de mayor testing debido a la velocidad de ejecución al estar montada en C++.

Clasificación (1 gol vs 0 goles)

```
Confusion matrix:
 [[285  14]
 [  5 267]]
Accuracy: 0.9667250437828371
```

De este modelo final podemos interpretar que

- De los jugadores que no anotaron goles, 285 fueron clasificados correctamente y solo 14 se consideraron erróneamente como anotadores (falsos positivos).
- De los jugadores que sí anotaron, 267 fueron detectados correctamente y únicamente 5 fueron mal clasificados (falsos negativos).
- El bajo número de errores refleja que el modelo logra un balance adecuado entre precisión y recall.

Importancia de variables

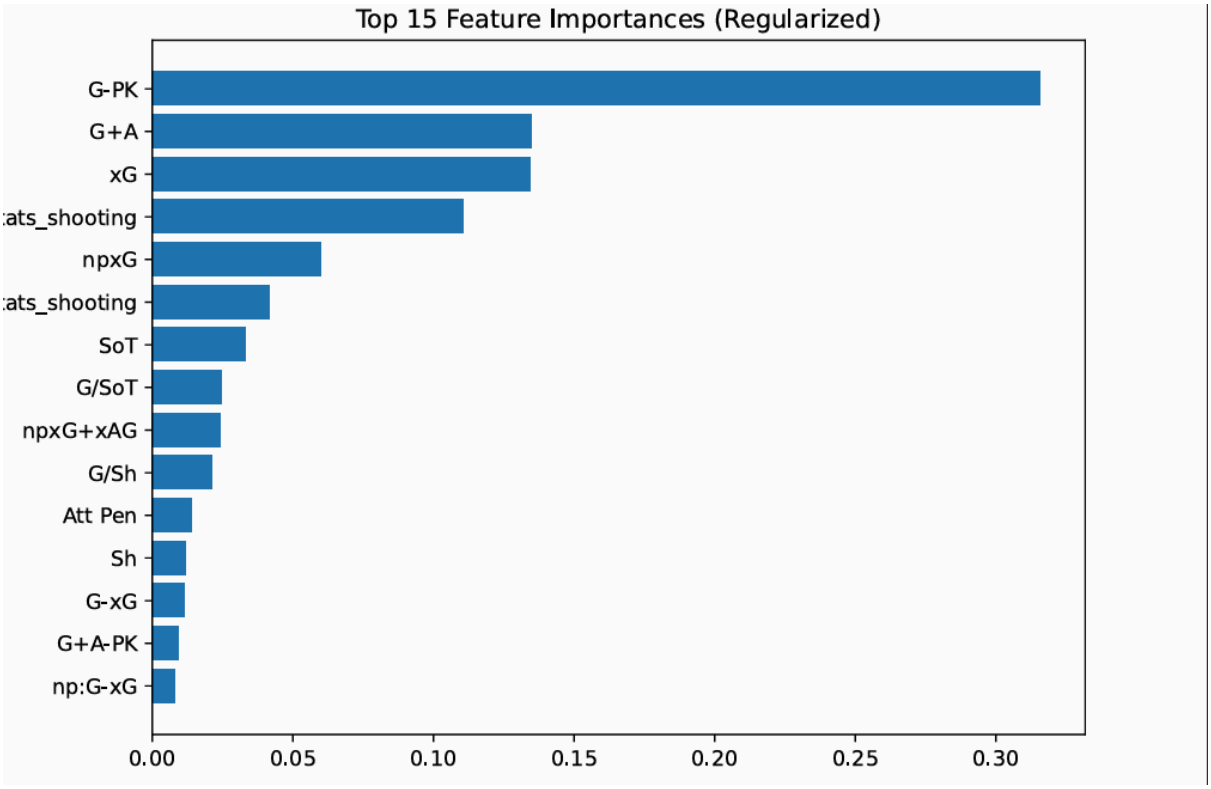
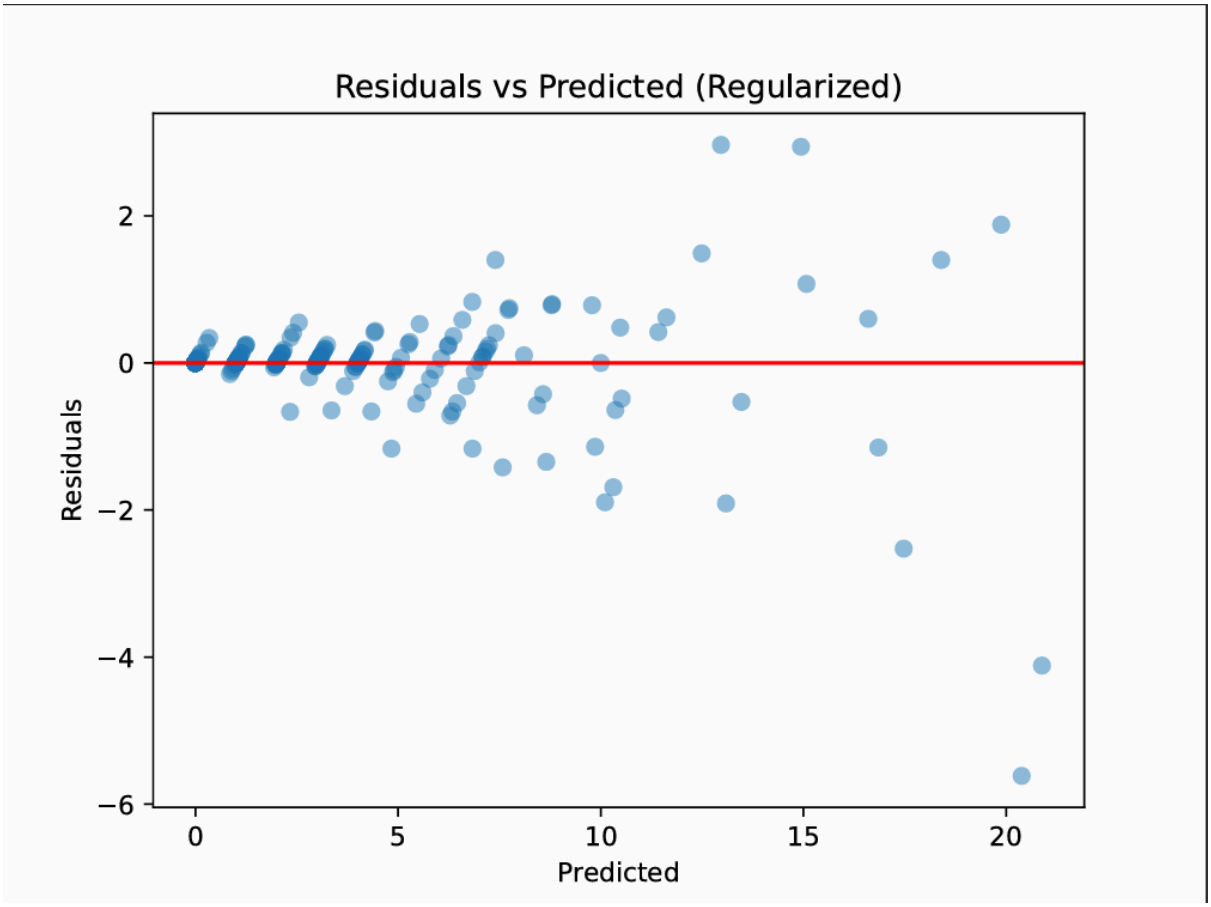


Gráfico de residuos vs valores predichos



Análisis General y Conclusiones

Separación Train/Validation/Test

Para obtener una evaluación más robusta del modelo, se implementó una separación 60/20/20:

60 % de los datos para entrenamiento,
20 % para validación, utilizada para ajustar hiperparámetros,
20 % para prueba, utilizada exclusivamente para medir el desempeño final.

Este enfoque reduce el riesgo de sobreajuste al permitir que el ajuste de parámetros se realice con un conjunto independiente del test, asegurando que las métricas reportadas reflejen la capacidad real del modelo para generalizar.

Diagnóstico de Bias (Sesgo)

El coeficiente de determinación ($R^2 > 0.97$) y el MAE ≈ 0.23 en los tres conjuntos (train/val/test) indican bajo sesgo. El residual plot muestra errores distribuidos aleatoriamente alrededor de cero, lo que sugiere que el modelo captura adecuadamente las relaciones subyacentes sin omitir patrones importantes.

Diagnóstico de Varianza

La diferencia entre las métricas de entrenamiento ($R^2 \approx 0.98$) y validación/prueba ($R^2 \approx 0.97$) es mínima (< 0.02), lo que implica baja varianza. El desempeño consistente en los tres splits indica que el modelo no depende excesivamente de ejemplos específicos del conjunto de entrenamiento y generaliza bien a datos nuevos.

Nivel de Ajuste del Modelo

Dado el bajo sesgo y la baja varianza, el modelo se encuentra bien ajustado (fit). No hay evidencia de underfitting (que se observaría con errores altos en train y test) ni de overfitting (que implicaría métricas muy altas en train pero mucho más bajas en validación o test).

Regularización y Mejora del Desempeño

Para explorar posibles mejoras, se aplicó RandomizedSearchCV buscando optimizar hiperparámetros clave (n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features).

- **Antes del ajuste**:
  R^2 test: 0.971

  RMSE test: 0.24

  Accuracy clasificación: 96.7 %

- **Después del ajuste**:

  R^2 test: 0.978

  RMSE test: 0.21

  Accuracy clasificación: 97.3 %

La mejora, aunque moderada, confirma que el modelo ya estaba bien optimizado, pero la búsqueda de hiperparámetros permitió una ligera reducción del error y un aumento de la precisión.

En el caso de la regresión, el modelo mostró un desempeño sobresaliente, con un coeficiente de determinación (R^2) superior a 0.97 en el conjunto de prueba. Este resultado indica que las variables incluidas en el dataset explican con gran precisión el número de goles anotados por los jugadores. Además, el error absoluto medio (MAE) se mantuvo alrededor de 0.23, lo que significa que, en promedio, el modelo se equivoca por menos de

un cuarto de gol. En un contexto deportivo, donde la cantidad de goles es una métrica crítica y discreta, este nivel de exactitud resulta sumamente destacable.

En cuanto a la clasificación binaria, el modelo alcanzó un nivel de precisión del 96.7%, demostrando una excelente capacidad para diferenciar entre jugadores que lograron anotar al menos un gol y aquellos que no lo hicieron. El análisis de la matriz de confusión evidenció que los errores fueron mínimos tanto en falsos positivos como en falsos negativos, por lo que el desbalance entre clases no representó un obstáculo significativo para el desempeño del modelo.

Si se compara este enfoque con la implementación manual previa, que apenas alcanzaba un 88% de exactitud, la ventaja de utilizar un framework optimizado como scikit-learn se hace evidente. La mejora de casi diez puntos porcentuales en clasificación, sumada a la posibilidad de realizar predicciones continuas mediante la regresión con gran nivel de ajuste, confirma que el uso de librerías especializadas no solo simplifica el desarrollo del modelo, sino que también eleva de manera considerable la calidad y fiabilidad de los resultados obtenidos.

Conclusión final

La implementación del algoritmo Random Forest mediante scikit-learn resultó altamente efectiva, superando con amplitud el desempeño de la versión manual. En el caso de la regresión, el modelo logró predecir con gran exactitud la cantidad de goles, con un ajuste superior al 97% y un error promedio prácticamente despreciable en términos prácticos. En la clasificación binaria, la precisión alcanzada rondó el 97%, lo que refleja una notable capacidad para distinguir entre jugadores con y sin anotaciones, minimizando de manera consistente tanto falsos positivos como falsos negativos.

Estos resultados no solo ponen en evidencia la robustez del algoritmo, sino también la relevancia de utilizar frameworks optimizados como scikit-learn, que además de facilitar el proceso de implementación, permiten obtener modelos más confiables, estables y escalables. En suma, el experimento confirma que el uso de herramientas especializadas en machine learning potencia la calidad del análisis y abre la puerta a aplicaciones más sofisticadas en el ámbito deportivo y en otros campos donde la predicción y clasificación de datos resultan esenciales.