Modelos predictivos aplicados al análisis del salario de futbolistas profesionales

Javier Moreno Morón

Tutores: Pablo Mesejo Santiago, Óscar Cordón García

Grado en Ingeniería Informática

Universidad de Granada, España

30 de Junio de 2024

Índice

Índice

Definición del problema

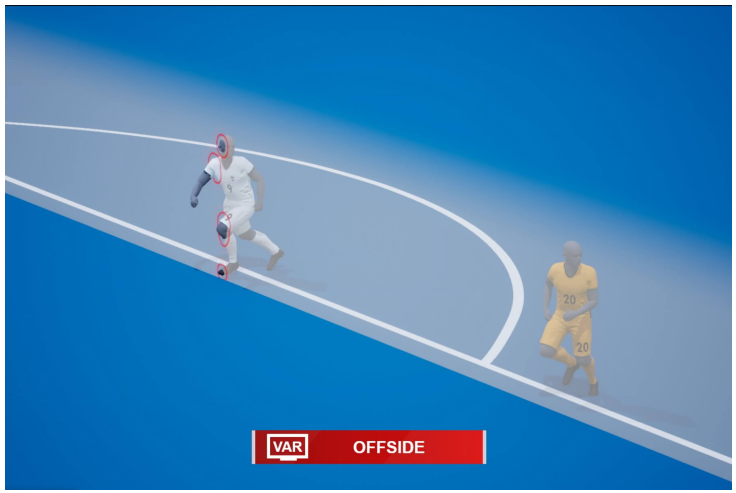
Buscamos diseñar un **sistema inteligente explicable** para **estimar** el **salario**.

Cuatro actores que necesitan conocer el salario de un futbolista:

- El **jugador** que es el **beneficiario del salario**.
- El **representante** que se encarga de **negociar el sueldo** del futbolista.
- El **club** que es quien **paga el salario** al jugador.
- Las **organizaciones** que se encargan de **controlar que los clubes no incumplen las reglas financieras**.

Motivación

- **Auge** del uso de la **Inteligencia Artificial** (IA) en **aplicaciones del mundo del fútbol**.
- Uso de **métodos subjetivos** a la hora de **estimar el salario**.
- **Conocer** que **factores** son más **influyentes en el salario**.



Ejemplo del uso de la IA para la detección del fuera de juego[1].

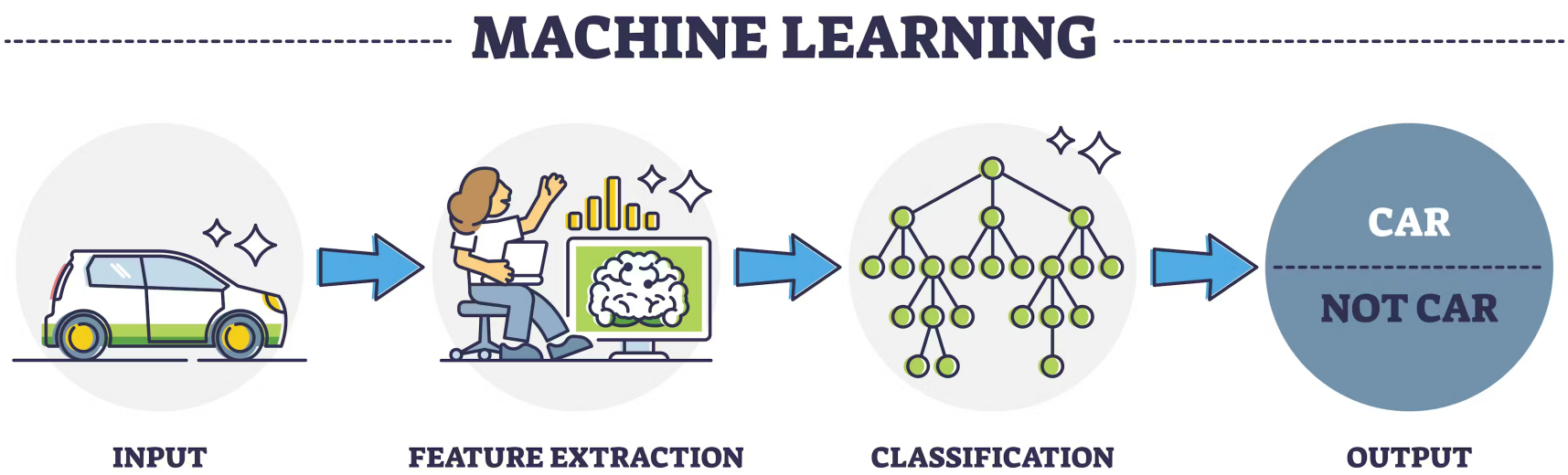
[1] **The Athletic** (Nov 29, 2023). *Premier League continuing to monitor semi-automated offside technology amid behind-the-scenes testing*.

Objetivos

1. Revisar el **estado del arte**.
2. **Diseñar** y **preprocesar** el conjunto de **datos**.
3. **Seleccionar** e **implementar** el conjunto de **hipótesis**.
4. **Seleccionar** y **comparar** con otros estudios el **mejor modelo**.
5. **Analizar** el **estimador** para obtener la **explicación** de las **predicciones**.

Índice

Machine Learning



Esquema del funcionamiento del ML[2].

[2] **Turing**. *Deep Learning vs Machine Learning: The Ultimate Battle*.

Métricas de error

Coeficiente de determinación o R^2

$$R^2(y, y') = 1 - \frac{\sum_{i=1}^{n}(y_i - y'_i)^2}{\sum_{i=1}^{n}(y_i - Y')^2}.$$

Donde $Y' = \frac{1}{n}\sum_{i=1}^{n} y_i$

Error cuadrático medio (MSE)

$$MSE(y, y') = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - y'_i)^2$$

Error absoluto medio (MAE)

$$MAE(y, y') = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - y'_i|$$

Selección de características

Tipos de **métodos existentes**:

- Enfoque de **filtro**: Relaciona **características** y **variable a predecir**.
- Enfoque de **envoltura**: Relaciona **estimador** y **características**.
- Enfoque **integrado**: El **estimador** realiza la selección **internamente**.

Estado del arte

Li et al. (2022)[3]:

- **Explicabilidad** de las predicciones.
- Modelo basado en **RF**.
- Utiliza **datos reales**.
- R^2 de **0.606**.

Behravan and Razavi (2021)[4]:

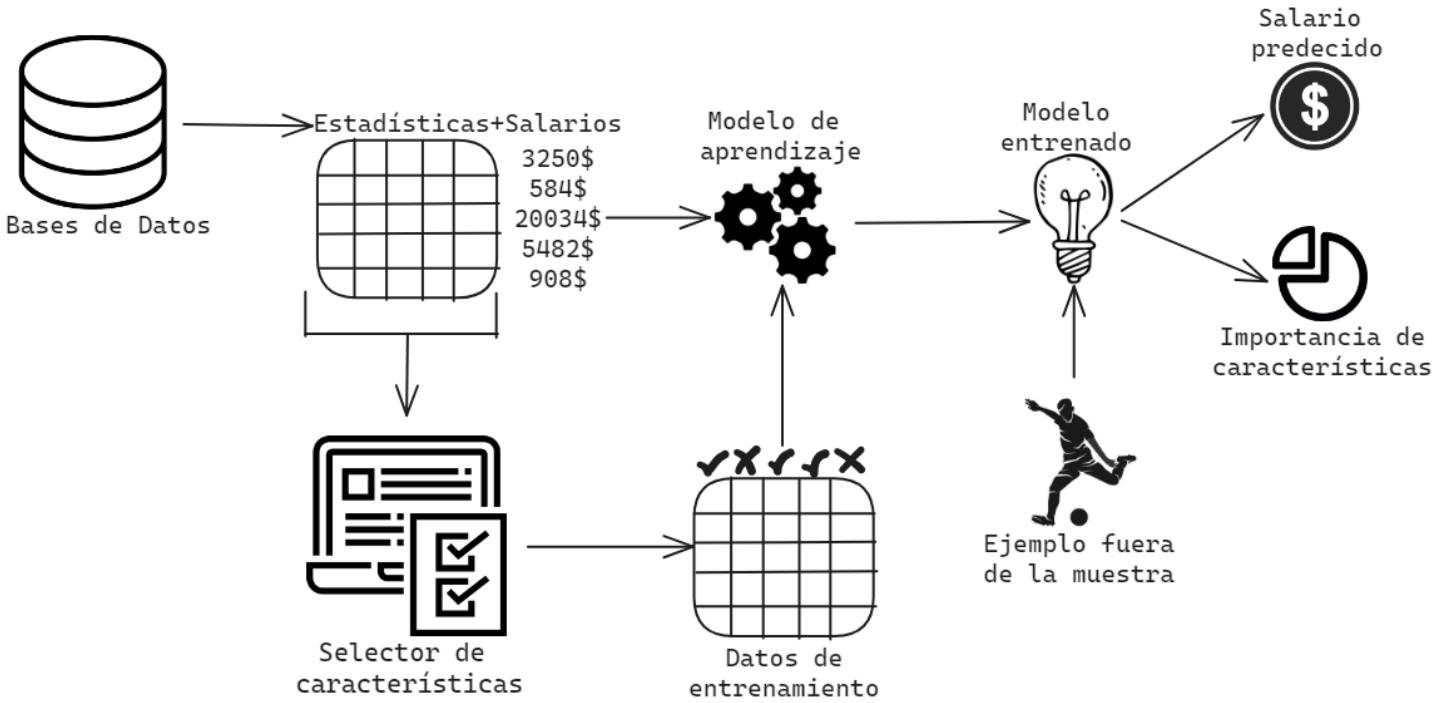
- Uso de técnicas de **selección de características**.
- Modelo basado en **SVR** más **PSO**.
- **Base de datos** de un **videojuego**.
- R^2 de **0.74**, **mejores resultados** en este campo.

[3] C. Li, S. Kampakis, and P. Treleaven (2022). "Machine learning modeling to evaluate the value of football players". In: *arXiv preprint arXiv:2207.11361*

[4] I. Behravan and S. M. Razavi (2021). "A novel machine learning method for estimating football players' value in the transfer market". In: *Soft Computing* 25.3, pp. 2499–2511

Índice

Metodología



Esquema del procedimiento llevado a cabo para resolver el problema propuesto.

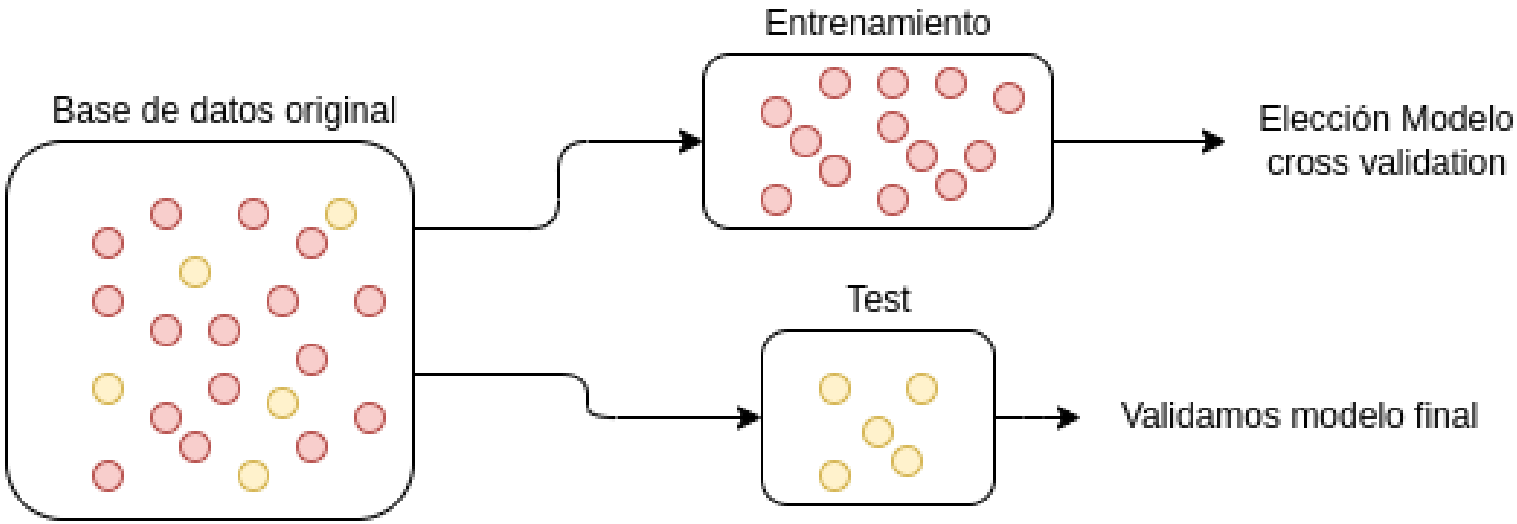
Conjunto de datos

- **10786 ejemplos** con **76 características**.
- Pertenecientes a las **5 grandes ligas europeas**.
- Datos de las temporadas **2017-2018 a la 2022-2023**.
- Obtenidos mediante técnicas de **web scrapping**.
- **Variable a predecir** en **euros** y ajustada a la **tasa de inflación de 2023**.

Partición de los datos

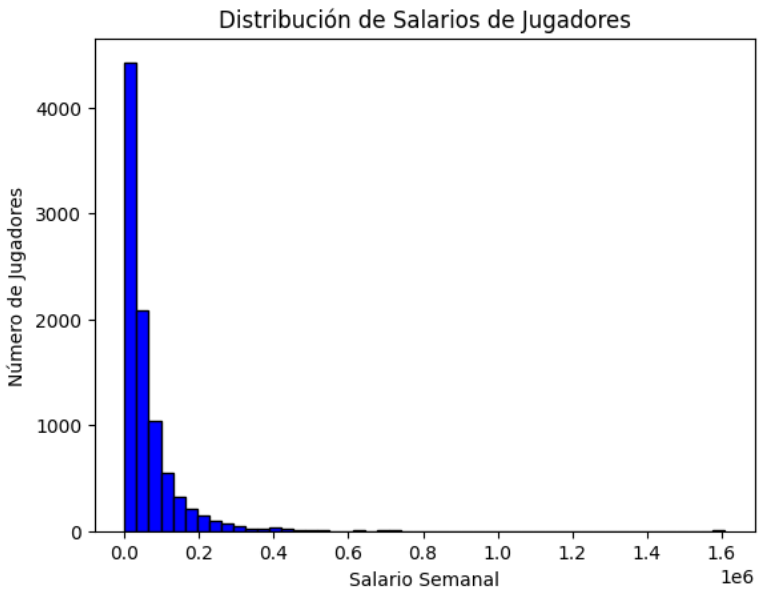
Se dividen los datos en entrenamiento y test:

- **9168** ejemplos para **entrenamiento** (**85%**).
- **1618** ejemplos para **test** (**15%**).
- Sobre el conjunto de **entrenamiento** usaremos **5 fold cross-validation** como protocolo de validación.

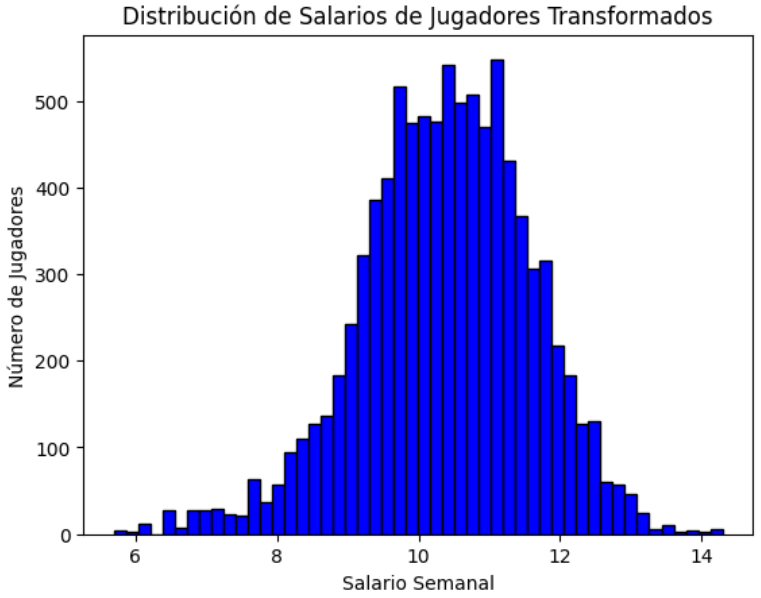


Preprocesamiento de los datos

Escalado logarítmico del salario.



Distribución de salarios antes de escalar.



Distribución de salarios después de escalar.

Preprocesamiento de los datos

Normalización de las características:

Utilizamos **normalización min-max** para que el **rango de las variables** se encuentre en **[0-1]**. La función para normalizar es la siguiente:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Métodos seleccionados

Conjunto de hipótesis:

- **Regresión Lineal** (RL): Modelo sencillo, lineal.
- **K-Nearest Neighbor** (k-NN): Modelo sencillo, no lineal.
- **Random Forest** (RF) y **Gradient Boosting** (GB): Modelos basado en árboles[5].
- **Perceptrón Multicapa** (MLP): Red neuronal.

Métodos de selección de características utilizados:

- Información Mutua para Selección de Características (MIFS).
- Coeficiente r de *Pearson* y *F-value*.
- Eliminación Recursiva de Características (RFE).
- Selector de Características Las Vegas (LVF/LVW).

[5] L. Grinsztajn, E. Oyallon, and G. Varoquaux (2022). "Why do tree-based models still outperform deep learning on typical tabular data?" In: *Advances in neural information processing systems* 35, pp. 507–520

Aplicación *web*

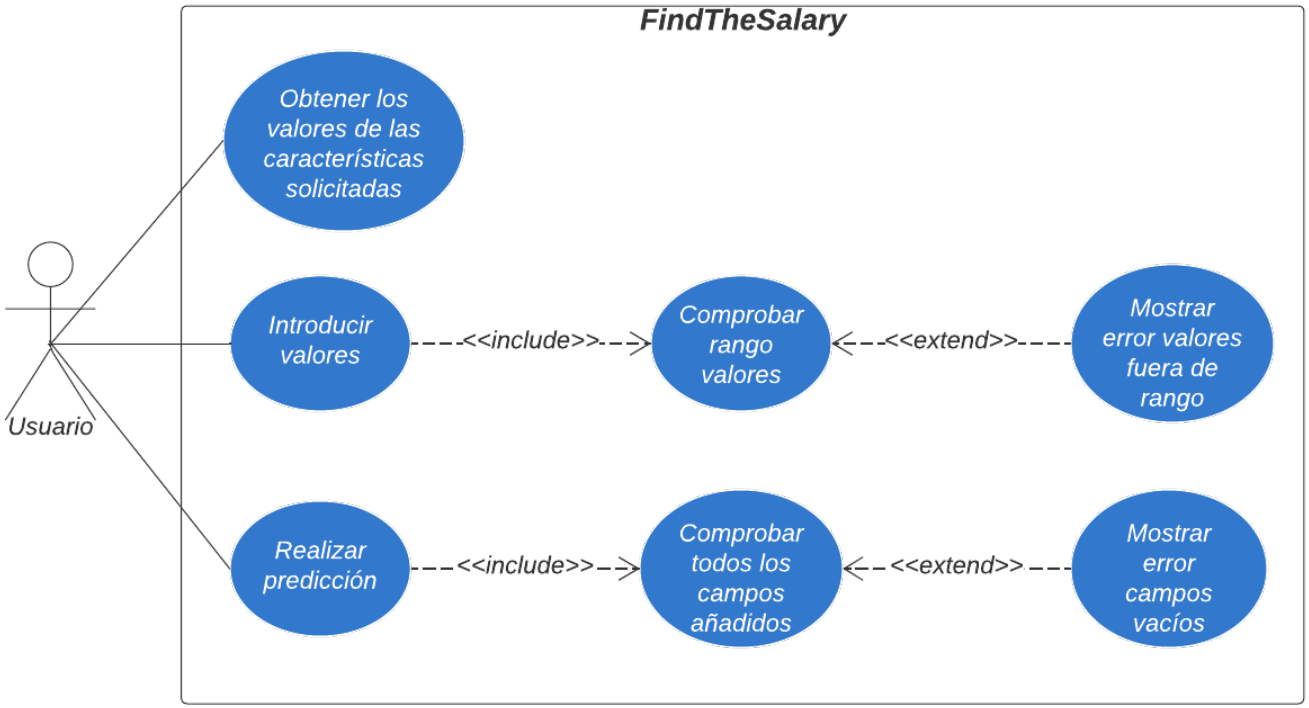


Diagrama de casos de uso de la *app*.

Índice

Cronología de los experimentos

- Experimentos utilizando técnicas de **reducción de características** para **seleccionar** los **mejores modelos**.
- Experimentos con los **mejores modelos reduciendo complejidad del dataset**.
- Experimentos con los **mejores modelos** aplicando técnicas de **selección de características**.
- **Análisis** de los resultados y **selección** del **mejor modelo**.
- **Integración** del **estimador** desarrollado dentro de la **aplicación**.

Experimentos iniciales. Reducción de características

Coeficiente de correlación de Pearson:

- **71 características** continuas **iniciales**.
- **Eliminamos** variables con $> 99\%$ de correlación.
- **70 características finales** tras aplicar *Pearson*.

Análisis de Componentes Principales (PCA):

- **70 características** continuas **iniciales**.
- Buscamos **quedarnos** con al menos un 99% de la varianza explicada total.
- **38 características finales** tras aplicar PCA.

Experimentos iniciales. Resultados obtenidos

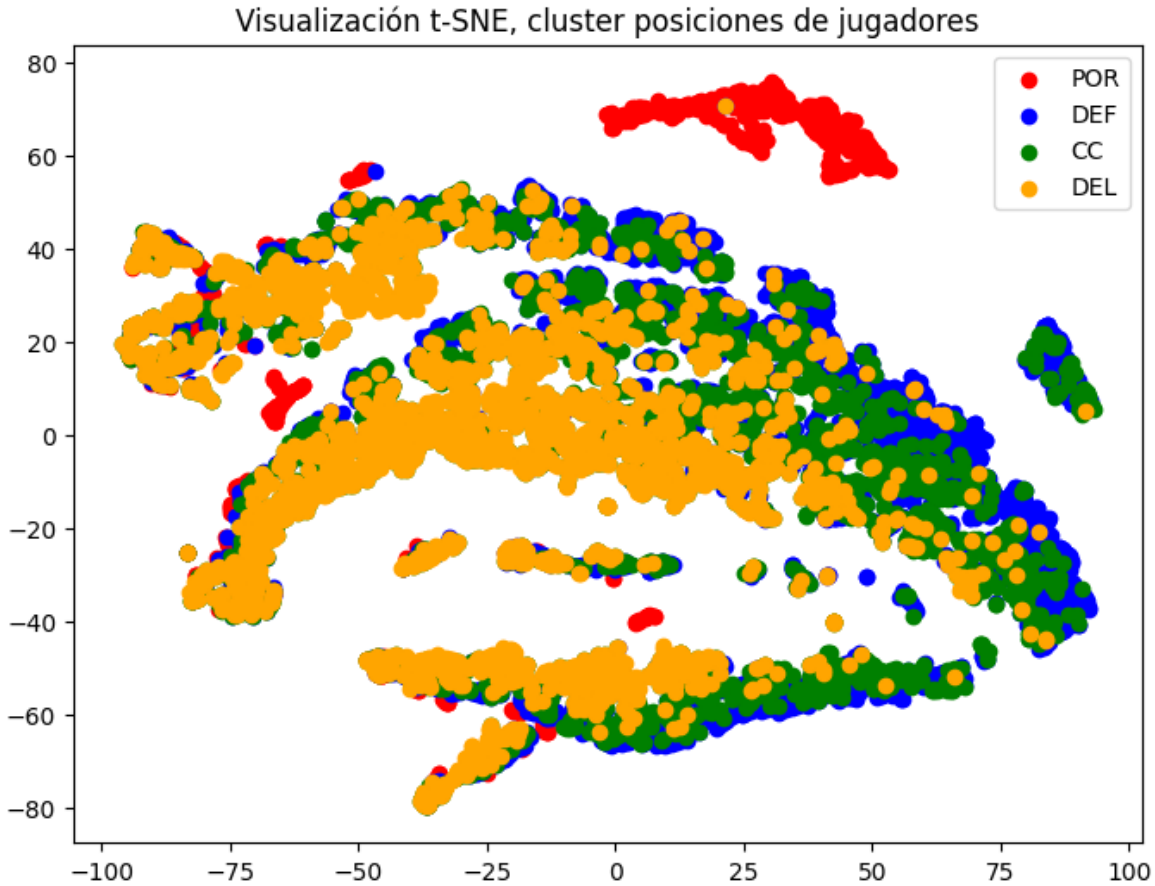
Métricas error	k-NN	RL	RF	GB	MLP	D-m[6]
MSE_{train}	**0.0**	0.5884	0.1509	0.209	0.4155	1.4602
MAE_{train}	**0.0**	0.5948	0.303	0.3293	0.4906	0.9532
$R^2{}_{train}$	**1.0**	0.597	0.8967	0.8569	0.7154	0.0
MSE_{val}	0.7289	0.5908	0.5227	0.4822	**0.464**	1.4603
MAE_{val}	0.6763	0.5958	0.5517	0.5262	**0.5206**	0.9533
$R^2{}_{val}$	0.5008	0.5955	0.642	0.6698	**0.6823**	0.0

Mejores modelos: **Gradient Boosting** y el **Perceptrón Multicapa**.

[6] **Dummy-mean**, un estimador que siempre **predice el salario medio**.

Pruebas sin porteros

¿Por qué sin porteros?



Pruebas sin porteros

- Eliminamos los ejemplos y variables de los porteros, quedándonos con **9872 instancias** y **64 características**.
- Utilizamos solo los **mejores modelos** (**GB** y **MLP**).
- Aplicamos de nuevo **Pearson** y **PCA**, quedándonos con **37 características** finales.

Métricas error	**GB**	**MLP**
MSE_{train}	**0.3547**	0.4091
MAE_{train}	**0.4443**	0.4855
$R^2{}_{train}$	**0.76**	0.7232
MSE_{val}	0.4923	**0.4554**
MAE_{val}	0.5335	**0.5142**
$R^2{}_{val}$	0.6668	**0.6918**

Pruebas considerando una única posición

- Nos quedamos solo con los **centrocampistas**, quedándonos con **4839 instancias** y **61 características**.
- Utilizamos solo los **mejores modelos** (**GB** y **MLP**).
- Aplicamos de nuevo **Pearson** y **PCA**, quedándonos con **35 características** finales.

Métricas error	**GB**	**MLP**
MSE_{train}	**0.3084**	0.4297
MAE_{train}	**0.4107**	0.4995
$R^2{}_{train}$	**0.8016**	0.7236
MSE_{val}	0.5092	**0.4731**
MAE_{val}	0.5425	**0.5256**
$R^2{}_{val}$	0.6711	**0.6943**

Experimentos de selección de características

- Ya **no** se utilizan técnicas de **reducción de características**.
- Una vez obtenido el nuevo *dataset*, usamos un modelo **básico** de **GB** para **medir la calidad del dataset seleccionado**.
- Para cada método de selección de características, seleccionamos quedarnos con **10, 20, 30, 40 y 50 variables** y posteriormente **elegimos** la que **mejores resultados** consigue.
- Una vez tenemos el mejor conjunto de cada modelo, los **comparamos entre ellos**.

Experimentos de selección de características. Mejores resultados

Métodos:	**MIFS**	**Coef. r Pearson**	**F-value**	**RFE**	**LVW**
MSE	0.4683	0.4676	0.4678	**0.4637**	0.4722
MAE	0.5164	0.5159	0.516	**0.5129**	0.5168
R^2	0.681	0.6815	0.6814	**0.6842**	0.6783
Nº variables	30	30	40	**20**	28

La mejor selección de características la consigue la **Eliminación Recursiva de Características**.

Experimentos finales

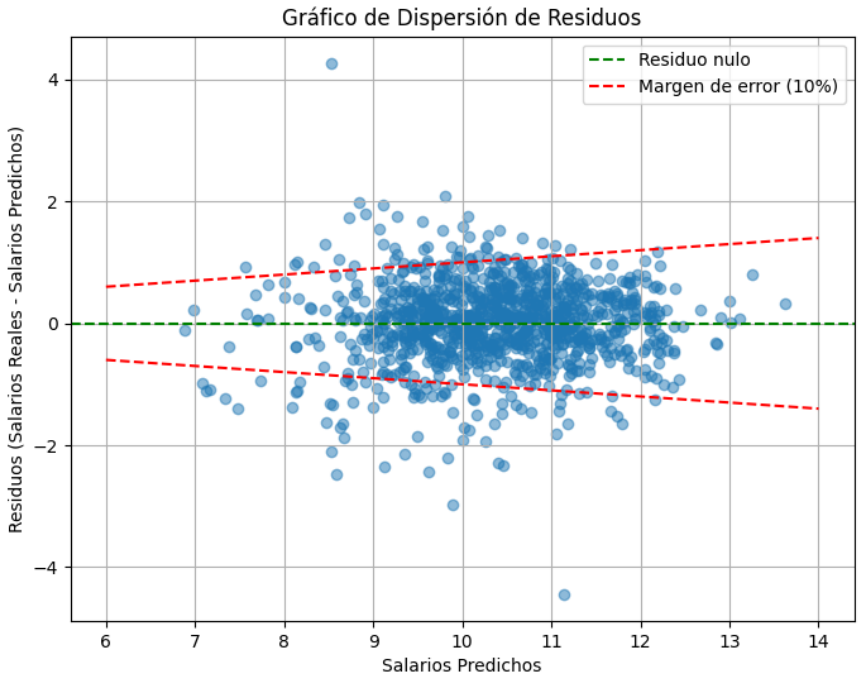
- Utilizamos el **mejor conjunto** obtenido en la selección de características.
- Entrenamos y comparamos los **mejores modelos** (**GB** y **MLP**).

Métricas error	**GB**	**MLP**
MSE_{train}	**0.228**	0.461
MAE_{train}	**0.35**	0.518
$R^2{}_{train}$	**0.845**	0.686
MSE_{val}	**0.421**	0.471
MAE_{val}	**0.487**	0.524
$R^2{}_{val}$	**0.713**	0.679

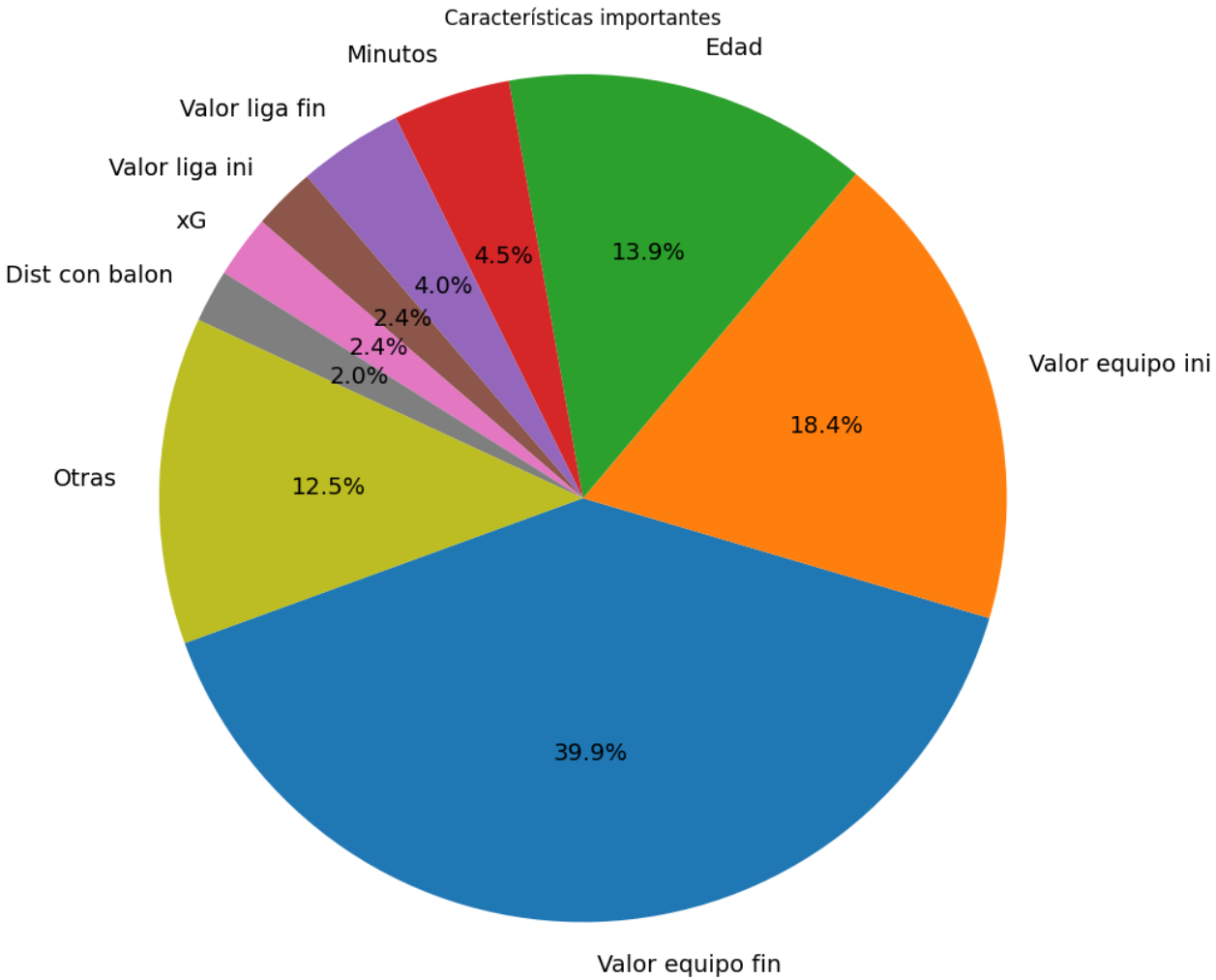
Seleccionamos **Gradient Boosting** como el mejor modelo.

Resultados para el conjunto de test

- **Reentrenamos** el modelo seleccionado utilizando ahora **todo el conjunto de entrenamiento** y **sin aplicar CV**.
- Resultados: **MSE**: 0.4573, **MAE**: 0.4978, $\mathbf{R^2}$: 0.7034
- **88.79%** de ejemplos estimados correctamente dentro de un **margen de error** del **10%**.



Explicabilidad de las predicciones



Aplicación web

https://findthesalary.streamlit.app/

Índice

Conclusiones

- **Todos los objetivos cumplidos**:
  1. Revisión del **estado del arte**.
  2. **Diseño** y **preprocesamiento** del conjunto de **datos**.
  3. **Selección** e **implementación** del conjunto de **hipótesis**.
  4. **Selección** y **comparación** con otros estudios del **mejor modelo**.
  5. **Análisis** del **estimador** para obtener la **explicación** de las **predicciones**.
- Además, se han logrado **desarrollar una aplicación** para probar los resultados obtenidos en la investigación.
- $\mathbf{R}^2$ solo **0.04** puntos por **debajo** del mejor estudio en este campo.

Trabajos futuros

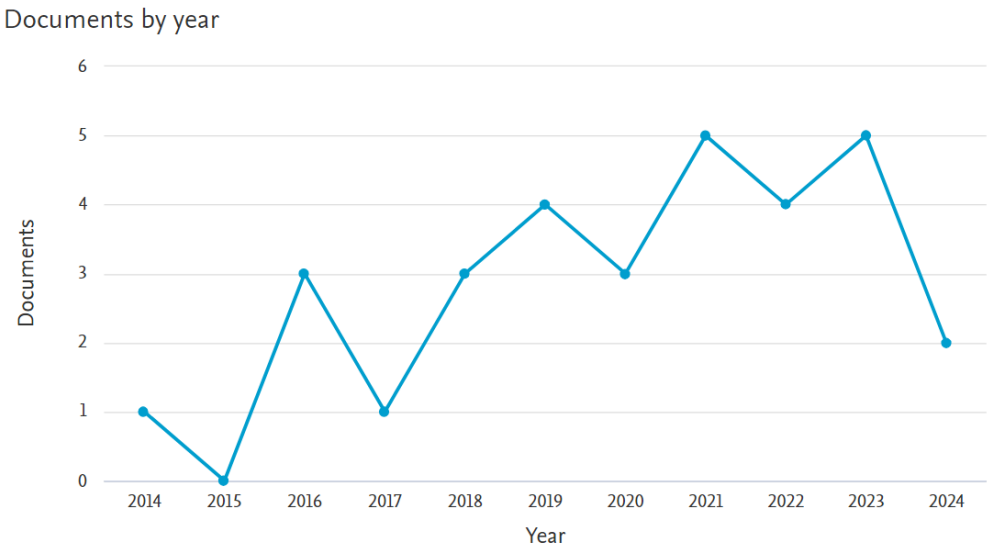
- Métodos más sofisticados de **web scrapping**.
- **Heurísticas más complejas** para la **selección de características**, como Búsqueda Local o algoritmos genéticos.
- Realizar las **predicciones** teniendo en cuenta las **estadísticas de anteriores temporadas**.
- Facilitar al usuario la **introducción de los datos** en la **app web**.
- Otras mejoras en la *app* como **mejoras estéticas**.
- Repositorio del proyecto: **https://github.com/JMMelcrack/code**

Preguntas

Búsqueda Scopus

Aplicando técnicas de IA:

- **Cantidad reducida** de artículos.
- Tendencia **ligeramente ascendente**.
- Temática **muy reciente**.

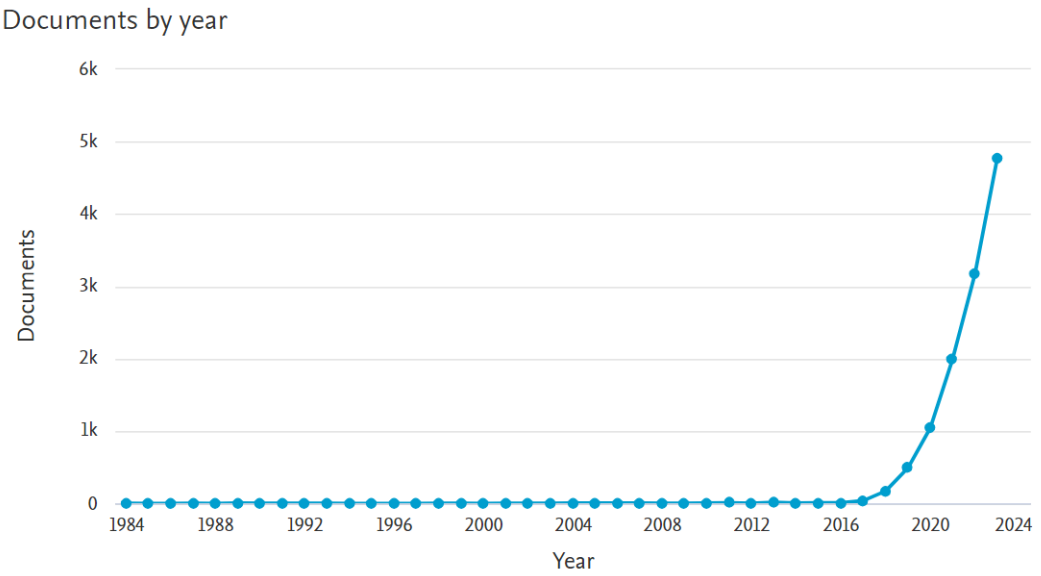


Query: TITLE-ABS-KEY ((((deep AND learning) OR (machine AND learning) OR (artificial AND intelligence) OR (data AND mining)) AND (wage OR salary OR (market AND value)) AND (estimation OR prediction) AND (sport OR football OR soocer))). Fecha: 02/05/2024.

Búsqueda Scopus

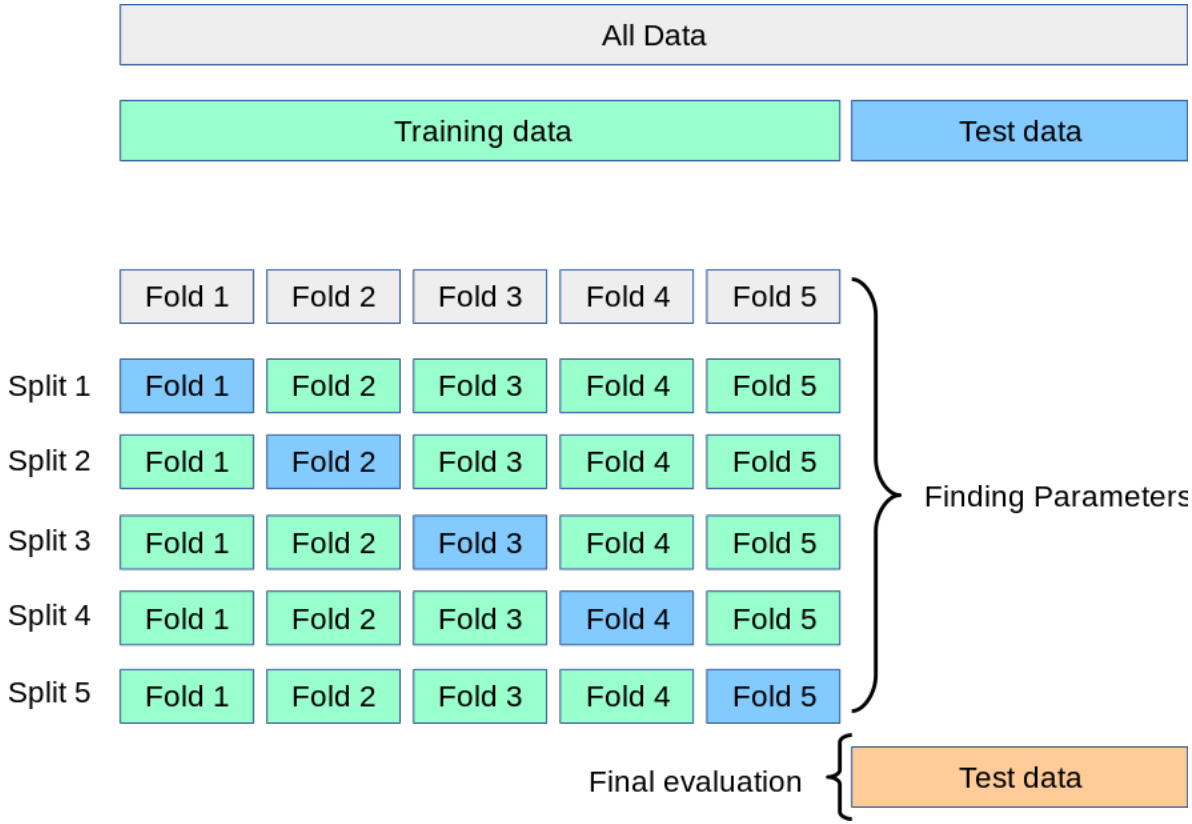
Explicabilidad de las predicciones:

- **Gran cantidad** de artículos.
- **Tendencia exponencial** en los últimos años.



Query: TITLE-ABS-KEY ((explainable AND ((deep AND learning) OR (machine AND learning) OR (soft AND computing) OR (artificial AND intelligence) OR (data AND mining)))). Rango de la búsquedad: Hasta 2023. Fecha: 05/05/2024.

Protocolo de validación



Esquema de *cross validation* (CV) con $k = 5$[7].

[7] **Scikit Learn**. *Cross-validation: evaluating estimator performance*.