Artículo Original Pág. 61 - 75 ISSN 2415-2323

ILUMATE VOLUMEN 12

PREDICCIÓN DE LA CLASIFICACIÓN FINAL EN LA LIGA ESPAÑOLA DE FÚTBOL

Prediction of the final classification in the Spanish soccer league

Hugo Herrera Lunario
hugolunario@gmail.com
Instituto de Investigaciones en Ciencia y Tecnología,
Universidad La Salle-Bolivia

Ramiro Paucara Ochoa
ramiropaucara@gmail.com
Instituto de Investigaciones en Ciencia y Tecnología,
Universidad La Salle-Bolivia

Resumen

En este trabajo se propone el uso de herramientas de machine learning, en particular del algoritmo de regresión logística para intentar predecir la clasificación final del siguiente campeonato de la Liga española de fútbol. El método es experimental de tipo exploratorio, los datos que se utilizaron se armaron en base a datos disponibles en el sitio web de Kaggle, se hizo uso de 3 conjuntos de datos en los cuales se tiene información de los jugadores, el historial de partidos desde 1996 hasta 2020 y el conjunto de equipos que participaran en el siguiente campeonato con su ranking actual de la FIFA. Las principales variables para predecir los resultados fueron el ranking FIFA de cada equipo, el talento de sus jugadores y el historial de

enfrentamientos. Los experimentos realizados sobre el conjunto de datos muestran cuantitativamente que el modelo desarrollado es apropiado para predecir una clasificación de los equipos de la Liga española en el siguiente año.

Palabras claves

Machine learning, regresión logística, Kaggle, ranking.

Abstract

In this work, the use of machine learning tools is proposed, in particular the logistic regression algorithm to try to predict the final classification of the next championship of the Spanish Soccer League. The method is experimental of an exploratory type, the data used was assembled based on data available on the Kaggle website, 3 data sets were used in which there is information on the players, the history of matches since 1996 to 2020 and the set of teams that will participate in the next championship with their current FIFA rankings. The main variables to predict the results were the FIFA ranking of each team, the talent of its players and the history of confrontations. The experiments carried out on the data set show quantitatively that the model developed is appropriate to predict a classification of the teams of the Spanish league in the following year.

Key words

Machine learning, logistic regression, Kaggle, ranking.

Introducción

La liga de fútbol española comúnmente conocida como La Liga es la primera liga nacional de fútbol de España, siendo una de las ligas deportivas profesionales más populares del mundo.

Actualmente, está formado por 20 equipos distribuidos de forma bastante uniforme por todo el país, pero principalmente de las regiones más desarrolladas: Madrid, Barcelona y País Vasco. Los cuatro mejores equipos están clasificados para la Liga de Campeones, mientras que los tres equipos peor ubicados (posiciones 18-20) son relegados a la segunda división.

Las opciones de aplicación de los algoritmos de aprendizaje automático en el fútbol son diversas ya que comprenden el sector de la industria de las apuestas deportivas; el análisis del desempeño de los jugadores que conforman el equipo; la formación táctica del equipo y plan de juego; la cantidad de goles anotados por encuentro hasta el posible ganador antes de que comience el encuentro. (ARIAS, 2019).

Se utilizará la metodología de aprendizaje automático comprendido por los pasos de construcción del conjunto de datos, categorización de los datos, posteriormente un preprocesamiento de la información y limpieza de los datos, luego se realizará la extracción de características acompañadas de validación de datos que permitan la construcción del modelo de predicción que finalmente tendrá como salida una categorización en una tabla de clasificación. (ARIAS, 2019).

El resultado de este experimento basado en aprendizaje automático mediante la técnica de la regresión logística tiene como fin apoyar la toma de decisiones de los equipos del fútbol profesional para generar un fútbol más competitivo, por otra parte, puede ser de utilidad para las empresas patrocinadoras para ver la eficiencia partido tras partido de los equipos de fútbol para promover sus productos y apoyar de forma económica a los equipos del fútbol profesional.

Referentes conceptuales.

Métodos.

Regresión Logística Binaria. La regresión logística es un método estadístico

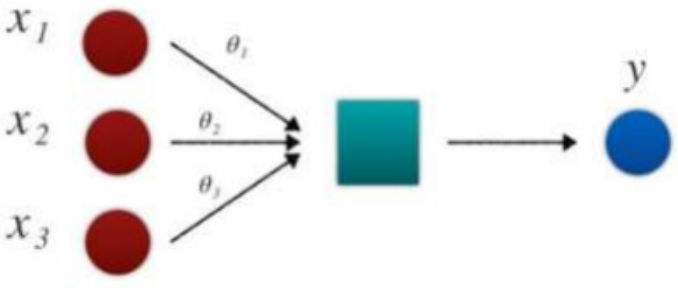
para analizar un conjunto de datos en el que hay una o más variables independientes que determinan un resultado. El resultado se mide con una variable dicotómica (en la que solo hay dos resultados posibles). Se utiliza para predecir un resultado binario dado un conjunto de variables independientes. Para representar el resultado binario o categórico, se utilizan variables ficticias. También la regresión logística se puede expresar como un caso especial de regresión lineal, es decir, cuando la variable de resultado es categórica, donde se está utilizando el registro de probabilidades como variable dependiente. En palabras simples, predice la probabilidad de ocurrencia de un evento al ajustar los datos a una función. (Hernández, 2019)

Este modelo logístico binario según Ponce (2020) se usa para estimar la probabilidad de una respuesta binaria basada en una o más variables (características) predictoras (o independientes). Permite decir que la presencia de un factor de riesgo aumenta la probabilidad de un resultado dado en un porcentaje específico.

Como todos los análisis de regresión, la regresión logística es un análisis predictivo.
La regresión logística se utiliza para describir datos y para explicar la relación entre una variable binaria dependiente y una o más variables independientes nominales, ordinales, de intervalo o de relación. (Gandhi, 2018).

En la siguiente figura se observa un modelo de regresión logística.

Fig. 1. Representación gráfica de un modelo de regresión logística.



Fuente: (ARIAS, 2019)

Esta función de probabilidad es la ' Función sigmoidal' como se observa en la siguiente figura:

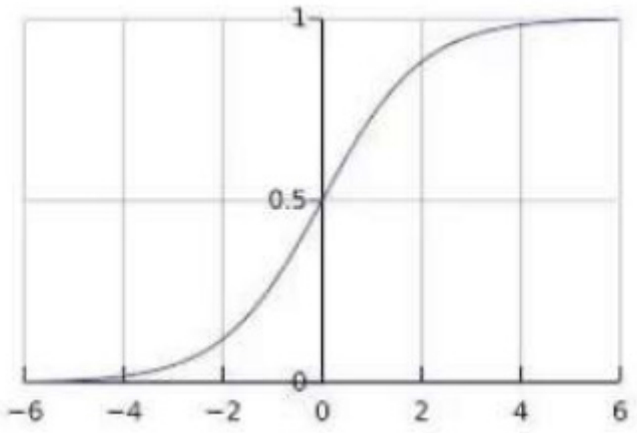
Fig. 2. Función sigmoidal

$$\frac{1}{1 + e^{(-z)}}$$

Fuente: (ARIAS, 2019)

En la siguiente figura se muestra gráficamente la función, se vería de la siguiente forma:

Fig. 3. Función sigmoidal



Fuente: (ARIAS, 2019)

Regresión Logística Multinomial. La regresión logística multinomial es un método de clasificación que generaliza la regresión logística a problemas a problemas multiclase, es decir, con más de dos posibles resultados discretos. Es un modelo que se utiliza para predecir las probabilidades de los diferentes resultados posibles de una variable dependiente distribuida categóricamente, dado un conjunto de variables independientes. La regresión logística multinomial es una solución particular para los problemas de clasificación que utilizan una combinación lineal de las características observadas y algunos parámetros específicos del problema para estimar la probabilidad de cada valor particular de la variable dependiente. (Catalan, 2021).

Al igual que en otras formas de regresión lineal, la regresión logística multinomial utiliza una función predictiva lineal $f(k, i)$ para predecir la probabilidad de que la observación i tenga un resultado i, de la siguiente forma:

Fig. 4. Función de predicción regresión logística multinomial

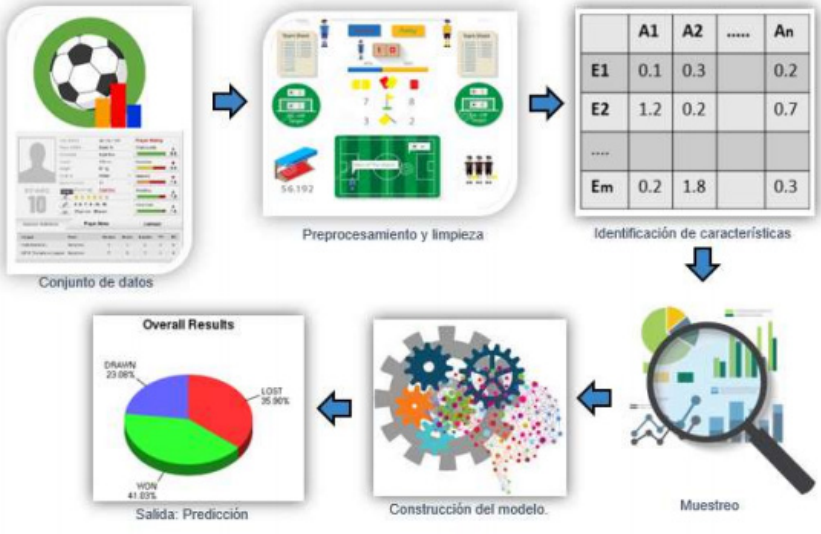
$$f(k,i) = \boldsymbol{\beta}_k \cdot \mathbf{x}_i$$

Fuente: (Catalan, 2021)

donde Xi es el vector de variables explicativas que describen la observación i, βk es un vector de ponderaciones correspondientes al resultado k, y la puntuación (Xi, k) es la puntuación asociada con la asignación de la observación i a la categoría k. En la teoría de la elección discreta, donde las observaciones representan personas y los resultados representan elecciones, la puntuación se considera la utilidad asociada con la persona i que elige el resultado k. El resultado predicho es el que tiene la puntuación más alta.

Para realizar el proceso de predicción de la clasificación final de la liga española de fútbol, es necesario realizar una serie de etapas en donde cada una se complementa directamente con la anterior, de esta forma se describen 6 pasos como se puede observar en la siguiente figura.

Fig. 5. Metodología aplicada predicción de clasificación



Fuente: (ARIAS, 2019)

Conjunto de datos

Se cuenta con información en tres conjuntos de datos:

- Equipos: Este dataset contiene los datos de los jugadores en cada equipo. (Kishan, 2020).
- La Liga_Matches_1995-2020: Este dataset contiene el historial de partidos en que se enfrentaron los equipos que participan de la liga, desde 1996 hasta el 2020. (Kishan, 2020).
- Copa_20_21: Este dataset contiene la composición del total de equipos que participan de la Liga con su ranking FIFA actual. (Tadhg, 2018).

Preprocesamiento y limpieza de datos.

En esta etapa se realiza la selección de datos más importantes para el proceso

de clasificación con el objetivo de conservar solo las características o los atributos más relevantes. (Dominguez, 2019).

Se eliminaron varios campos irrelevantes para esta predicción, como ser: sueldo de jugadores, fecha de vencimiento de contratos, nuero de camiseta, etc. También se relleno datos faltantes con el promedio correspondiente de cada columna.

Identificación de características

Los métodos de selección de características ayudan a reducir las dimensiones sin perder mucho la información total. También ayuda a dar sentido a las características y su importancia. (AED - Asociación Española de Directivos, 2019).

Usamos las características de si un equipo juega en casa o fuera de casa, estas características se transforman en una codificación en una variable discreta para su procesamiento posterior.

Muestreo

En esta etapa se va a clasificar el conjunto de datos en dos de manera que una sea para realizar el entrenamiento y la segunda sea para realizar las pruebas.

Puede haber muchos más datos seleccionados disponibles de los que necesita para trabajar. Más datos pueden resultar en tiempos de ejecución mucho más largos para los algoritmos y mayores requisitos computacionales y de memoria. Se tomó una muestra representativa más pequeña de los datos seleccionados que pueden ser mucho más rápidos para explorar y crear prototipos de soluciones antes de considerar el conjunto de datos completo.

Construcción del modelo de predicción.

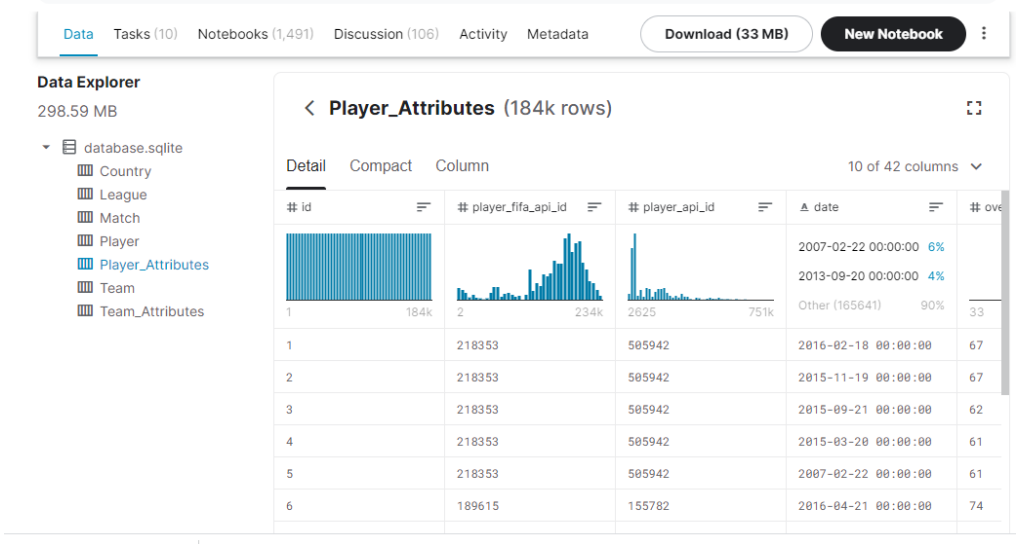
En esta etapa, se aplica los clasificadores de aprendizaje automático necesarios para realizar predicción de la clasificación final de la liga española de fútbol.

Salida.
En esta etapa se va a obtener el resultado que genera el modelo predictivo donde se indica el resultado del modelo desarrollado.

Resultados y discusión

Para construir el conjunto de datos, fue necesario indagar varias fuentes las cuales debían contar con la información relacionada a los partidos de fútbol de la liga española, por otra parte, la información debía ser confiable en su totalidad, finalmente se optó por seleccionar la fuente de datos de la página kaggle.

Fig. 6. Obtencion de informacion del sitio web Kaggle.com



Fuente: Elaboracion Propia

Para el conjunto de datos obtenido de la fase anterior, fue necesario construir la variable overall en el conjunto de datos de los partidos, uno del equipo local, otro para el equipo visitante y un campo adicional de la diferencia entre overall del equipo local y del visitante, en base a la información de los jugadores de cada equipo como se observa en la siguiente ilustración.

Fig. 7. Codigo del modelo

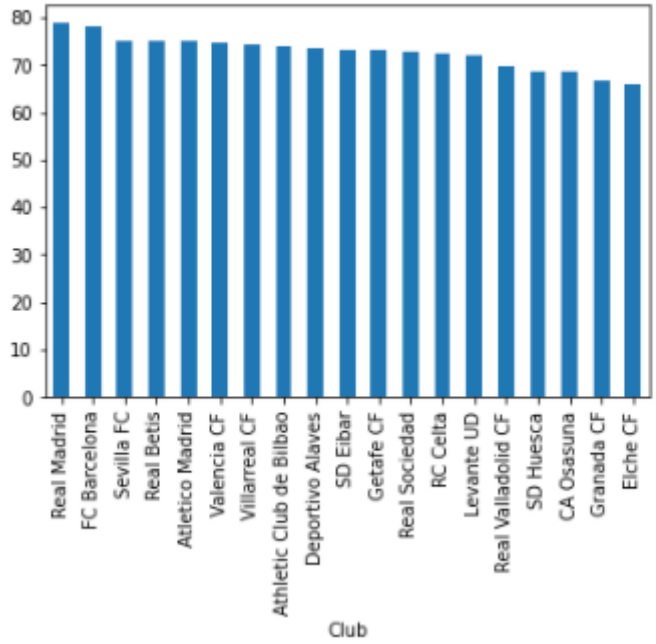
```
# Agregando el pontencial del equipo al dataset de partidos
potencial = equipos.groupby('Club').mean()['Overall']
partidos = partidos.merge(potencial,
                          left_on=['HomeTeam'],
                          right_on=['Club'])
partidos = partidos.merge(potencial,
                          left_on=['AwayTeam'],
                          right_on=['Club'],
                          suffixes=('_local', '_visitante'))

# Agregar diferencia de potencial entre los equipos
partidos['dif_potencial'] = partidos['Overall_local'] - partidos['Overall_visitante']
```

Fuente: Elaboracion Propia

El siguiente análisis, corresponde a la calidad de los jugadores por equipo, como se muestra en la siguiente ilustración, como se puede ver los grandes candidatos de acuerdo a la calidad de sus jugadores son Real Madrid, Barcelona, Sevilla.

Fig. 8. Potencial por equipo

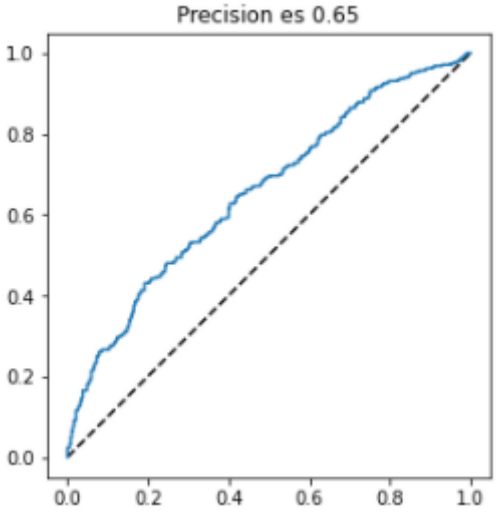
```
plot = potencial.sort_values( ascending=False).plot(kind='bar')
```



Fuente: Elaboracion Propia.

Al ejecutar el algoritmo, se muestra en la ilustración siguiente la precisión del modelo:

Fig. 9. Precision del modelo



Fuente: Elaboracion Propia

Finalmente se asigna una probabilidad de victoria del local o visitante en cada partido según los campos predictores de cada partido, posteriormente se suman los puntos por cada victoria o empate de cada equipo y al final se realiza una sumatoria de los puntos al final de todos los partidos todos contra todos, y el resultado es el siguiente:

Fig. 10. Clasificacion final

	equipo	puntos
0	Real Madrid	54
1	FC Barcelona	51
2	Sevilla FC	48
3	Atletico Madrid	45
4	Valencia CF	36
5	Villarreal CF	33
6	CA Osasuna	33
7	Athletic Club de Bilbao	30
8	Elche CF	30
9	Deportivo Alaves	27
10	Granada CF	24
11	Real Betis	21
12	SD Huesca	21
13	Getafe CF	15
14	Real Valladolid CF	15
15	Levante UD	12
16	RC Celta	9
17	Real Sociedad	6
18	SD Eibar	3

Fuente: Elaboracion Propia.

Conclusiones:

En el transcurso del desarrollo del presente trabajo se identificaron inconvenientes en base a la información del conjunto de datos, la precisión de predicción no subió lo esperado, pues ya que los atributos del conjunto de datos son limitados y no cuentan con características relevantes que en efecto puedan impactar en quien será el ganador de un partido de fútbol, como, por

ejemplo, datos de tipo climatológico, factores emocionales, puntuación del árbitro, factores sociales, etc.

En el presente trabajo de investigación, como se pudo observar la precisión del modelo construido es cercano al 65%, es decir, no existe un criterio claro y definido para predecir la clasificación final de un campeonato en función de la probabilidad que se obtuvo como resultado.

Referencias:

- AED - Asociación Española de Directivos. (2019). *Machine Learning, Inteligencia Artificial y Big Data. Lo que todo directivo debe saber .* Madrid: Accenture.
- ARIAS, E. (2019). *Repositorio Universidad Catolica de Colombia.* Obtenido de https://repository.ucatolica.edu.co/
- Catalan, A. (2021). Predicción y clasificación con Data Science. *Revista de Marina*.
- Dominguez, K. (2019). Inteligencia artificial. *Pro Magazine*, 42-43.
- Gandhi. (2018). *towardsdatascience.* Obtenido de http://www.towardsdatascience.com
- Hernández, A. (2019). *TRES MODELOS PREDICTIVOS DEPOTENCIAL ARQUEOLÓGICO.* Obtenido de https://es.scribd.com/document/411566036/Modelos-Predictivos-de-Potencial-Arqueologico
- Kishan, K. (2020). *kaggle.* Obtenido de https://www.kaggle.com/kishan305/la-liga-results-19952020/version/4
- Ponce, H. (2020). Buscando la integridad academica de la inteligencia artificial. *Integridad academica*, 6-8.
- Tadhg, F. (2018). *kaggle*. Obtenido de https://www.kaggle.com/tadhgfitzgerald/fifa-international-soccer-mens-ranking-1993now

Artículo recibido: 09-09-2020
Artículo Aceptado: 09-11-2020