Ganar dinero apostando al favorito. ¿Verdad o mito?

Francisco José Díez Corrales

Trabajo Fin de Máster - KSCHOOL

1. Introducción

Este trabajo analizará si el equipo que empieza el partido como favorito (según la casa de apuestas) acabará ganando el encuentro. Uno de los motivos de esta investigación es que muchas personas apuestan a estos partidos, debido a que se sienten más seguros al apostarlos, pensando que ganarán fácilmente.

El principal objetivo perseguido por este trabajo, es saber si tendremos beneficio o no a lo largo del tiempo si solo apostamos a los equipos favoritos.

Esta investigación simulará que realizamos apuestas simples, es decir, que apostaremos ficticiamente a los eventos deportivos individualmente sin mezclarlos.

La casa de apuestas nos ofrece distintos mercados a los que podemos apostar, los mercados son sucesos que pueden ocurrir en los eventos deportivos. El más básico es el "1X2", similar a la quiniela, nuestro estudio se centrará en este mercado.

Por último saber que la cuota es la probabilidad que nos da la casa de apuestas de que ganemos, al apostar al mercado "1X2" tendremos tres opciones, siendo la de cuota más pequeña la que mayor probabilidad tenga que ocurra. Funciona de la siguiente manera, si la cuota es 1,21, al apostar 100 euros multiplicaremos 1,21 por 100 y el beneficio bruto será 121 € y 21 € de beneficio neto.

Para el trabajo usaremos los datos históricos de los últimos años de tres deportes: fútbol (europeo y algunas ligas americanas), baloncesto (NBA) y tenis (WTA & ATP).

2. Datos

Todos los datos utilizados se encuentran dentro de la carpeta "Data". Estos datos han sido obtenidos de las siguientes fuentes:

- http://www.football-data.co.uk/
- http://www.tennis-data.co.uk/
- https://www.indatabet.com/

Los archivos contenían demasiada información por lo que nos hemos quedado solamente con las siguientes columnas:

Archivo	**Columnas**	**Descripción**
Football	Div	División
	Date	Fecha del partido
	HomeTeam	Equipo Local
	AwayTeam	Equipo visitante
	FTHG	Goles del equipo local
	FTAG	Goles del equipo visitante
	FTR	Ganador del partido
	IWH	Cuota del equipo local
	IWA	Cuota del equipo visitante

Football others	Country	País
	League	Liga
	Season	Temporada
	Date	Fecha del partido
	Time	Hora del Partido
	Home	Equipo Local
	Away	Equipo visitante
	HG	Goles del equipo local
	AG	Goles del equipo visitante
	Res	Ganador del partido
	PH	Cuota del equipo local
	PD	Cuota del empate
	PA	Cuota del equipo visitante
Basketball	Date	Fecha del partido
	Country	País
	League	Liga
	Seasons	Temporada
	Home	Equipo Local
	Away	Equipo visitante
	H	Puntos del equipo local
	A	Puntos del equipo visitante
	Winner	Ganador del partido
	H365	Cuota del equipo local
	A365	Cuota del equipo visitante
Tennis men & women	WTA/ATP	Categoría (ATP – Masculina, WTA – Femenina)
	Location	Ciudad donde se juega el torneo
	Tournament	Torneo
	Date	Fecha
	Tier / Series	Tipo de torneo
	Court	Lugar donde se juega el partido
	Surface	Tipo de pista
	Round	Ronda de la competición
	Best of	Modo de juego
	Winner	Ganador del partido
	Loser	Perdedor del partido
	B365W	Cuota del jugador que ha ganado
	B365L	Cuota del jugador que ha perdido

3. Metodología

Los notebooks deben ejecutarse en orden debido a que acciones realizadas en los primeros notebooks serán utilizadas en los siguientes.

1. Procesamiento de los datos

Lenguaje de programación: Python.

Objetivo del notebook: depurar y estructurar los datos que tenemos.

Librerías utilizadas:

- Pandas
- Numpy
- Os

Dentro del notebook crearemos rutas y definiremos las columnas que vamos a utilizar.

Como tendremos varios archivos con formato Excel y dentro de cada archivo hay varias hojas, juntaremos todas en una, creando un nuevo archivo que guardaremos en la ruta que hemos creado antes.

Al dataframe creado le dejaremos solamente las columnas que hemos definido anteriormente. Además en una nueva columna estableceremos cual es el "equipo favorito".

Para finalizar guardaremos los dataframes en archivos Excel para utilizarlos en los siguientes notabooks.

2. Hucha del favorito

Lenguaje de programación: Python.

Objetivo del notebook: creación de huchas ficticias para analizar si obtenemos beneficio de apostar siempre el favorito.

Librerías utilizadas:

- Pandas
- Numpy
- Os

Empezaremos creando una función que convierta la cuota en una probabilidad, para así hacer un análisis más sencillo, esta probabilidad será la que tiene de ganar el favorito. La función es la siguiente:

```
def prob(a):

return (1/a)*100
```

Realizaremos tres huchas, la primera será la más sencilla, la formaremos apostando siempre 10€ al equipo favorito. Para la segunda, apostaremos el resultado de multiplicar la cuota con su probabilidad.

Ambas huchas tienen resultados negativo, entonces crearemos un indicador para que nos diga que apostar a cada cuota. Para ello agruparemos por cuotas y obtendremos el porcentaje de favoritos que ganan en cada cuota.

Gracias a los indicadores reducimos las perdidas pero no conseguimos obtener beneficio.

Los resultados de las tres huchas para cada uno de los deportes son los siguientes:

	Sport	Hucha10	HuchaB	HuchaC
0	Basketball	-4328.40	-3039.198699	-1993.515818
1	Football	-99739.79	-50062.410331	-23660.392385
2	Tennis	-59220.81	-39093.534687	-21805.912572

El resultado global es negativo, pero en diversas ligas de fútbol obtenemos beneficios si solo apostamos al favorito. Son las siguientes:

	Tournament	Hucha10	HuchaB	HuchaC
41	USA	500.5	266.062265	158.748376
30	Russia	494.2	173.700524	76.140673
23	Japan	196.6	50.386316	43.205369
0	Argentina	-267.1	-63.285473	40.952720
3	Brazil	32.2	12.521216	32.008377
39	Switzerland	65.7	13.971395	13.970378
17	Finland	36.7	16.843731	7.774449
38	Sweden	28.9	15.365157	7.388849
8	Denmark	43.7	17.748560	6.702883
26	Norway	41.3	15.378291	3.843242

Para terminar el notebook calcularemos los datos estadísticos descriptivos de nuestros datos: media, desviación típica, mínimo, máximo, percentiles y moda.

3. Visualización

Lenguaje de programación: R.

Objetivo del notebook: crear gráficos para analizar los resultados de los anteriores notebooks.

Librerías utilizadas:

- ggplot2
- scales

Las visualizaciones creadas en este punto irán desde gráficos sencillos a más complejos, creando un total de 10 gráficos que nos ayudarán a comprender el análisis de los datos.

Guardaremos todos en un nuevo directorio con formato PDF.

4. Modelos

Lenguaje de programación: R.

Objetivo del notebook: creación de modelos para predecir si gana el equipo favorito.

Librerías utilizadas:

- plyr
- caTools
- ROCR
- e1071"
- glmnet
- dummies

Los modelos creados en este notebook serán aplicados solamente para el fútbol en las principales ligas europeas.

Las variables creadas son:

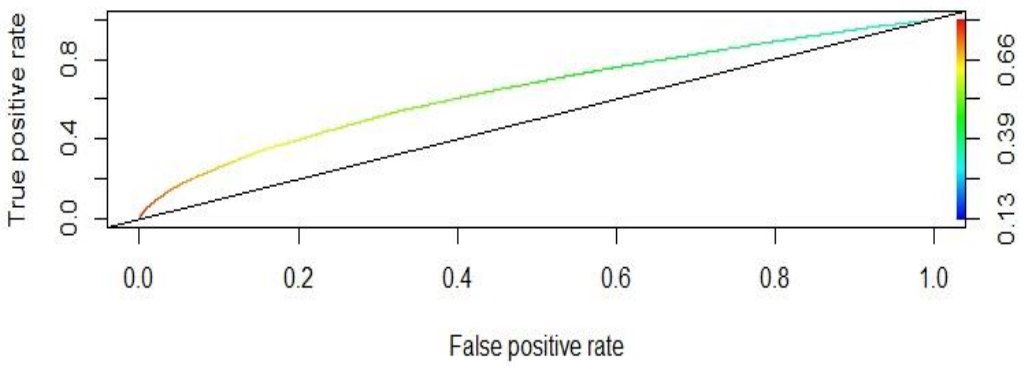
I. Div(Divisiones): España, Francia, Italia, Alemania, Portugal e Inglaterra.
II. LocalVisitante: 1 será local, 0 visitante.
III. Mes: que se juega el partido. Siendo 1 enero y 12 diciembre.
IV. JuegoEuropa: nos dice si el equipo participó en una competición europea durante la temporada. Siendo 1 que sí participó y 0 que no participó.
V. AñoMundialOEurocopa: está variable nos indicará si al terminar la temporada hay competición de selecciones, es decir, Eurocopa o Mundial. 1 será que si hubo competición y 0 que no hubo.
VI. Probabilidad: será la cuota del equipo antes de empezar el partido.
VII. GanaFavorito: si el equipo favorito ganó el partido tendrá 1 y sino 0.

Los modelos utilizados serán “Logit” y “Probit, en los dos modelos el test, entrenamiento y la capacidad del modelo son muy similares.

Además, aplicaremos variables dummies a nuestro dataframe, para intentar mejorar los resultados de los anteriores modelos, pero los resultados son bastante semejantes.

El accuracy en todos los modelos será aproximadamente 0,64.

Para poder medir si nuestro modelo es eficaz, utilizamos la curva ROC, y vemos que con nuestros modelos mejoramos al predecir si gana el favorito.



5. Machine Learning

Lenguaje de programación: Python.

Objetivo del notebook: a través de métodos de machine learning predecir si gana el equipo favorito.

Librerías utilizadas:

- pandas
- matplotlib.pyplot
- numpy
- os
- sklearn.model_selection & sklearn.metrics
- sklearn.neighbors import KNeighborsClassifier
- sklearn.tree import DecisionTreeClassifier
- sklearn.svm import SVC
- sklearn.linear_model import LogisticRegression
- sklearn.ensemble import RandomForestClassifier
- sklearn.naive_bayes import GaussianNB

Para poder emplear algoritmos de machine learning crearemos variables dummies en nuestro dataframe, para posteriormente definir X e Y.

Los algoritmos de machine learning utilizados son los siguientes:

K NEAREST NEIGHBORS

Es un método de clasificación supervisada, es usado como método de clasificación de elementos basado en un entrenamiento mediante ejemplos cercanos en el espacio de los elementos. *(Wikipedia)*

Los resultados obtenidos de este algoritmo son:

- Accuracy: 0.578
- Precision-recall score: 0.63

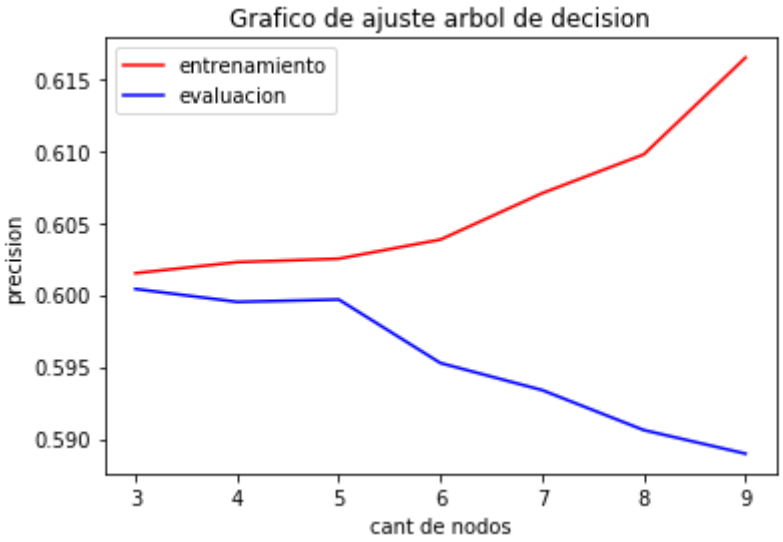
DECISION TREE

Es un método de clasificación supervisada, con un conjunto de datos se crean diagramas de construcciones lógicas, parecidos a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema. *(Wikipedia)*

Con este algoritmo obtenemos los siguientes resultados:

- Accuracy: 0.713
- Precision-recall score: 0.83

Debido a que obtenemos mejores resultados que en otros algoritmos comprobaremos si nuestro modelo tiene overfitting y si lo tiene lo ajustaremos.



Debido a que se produce overfitting, tenemos que buscar cuales son los mejores parámetros para nuestro algoritmo. Encontraremos que la máxima profundidad será tres y entonces nuestra accuracy será de 0,6.

SUPPORT VECTOR MACHINES

Es un método de clasificación supervisada, dado un conjunto de puntos, en el que cada uno de ellos pertenece a una de dos posibles categorías, un algoritmo basado en SVM construye un modelo capaz de predecir si un punto nuevo, cuya categoría desconocemos, pertenece a una categoría o a la otra. *(Wikipedia)*

El accuracy de este algoritmo es 0.599.

OTROS ALGORITMOS

Para finalizar el notebook haremos una comparativa rápida de otros algoritmos:
Logistic Regression, Random Forest Classifier y GaussianNB. Los resultados de estos algoritmos fueron:

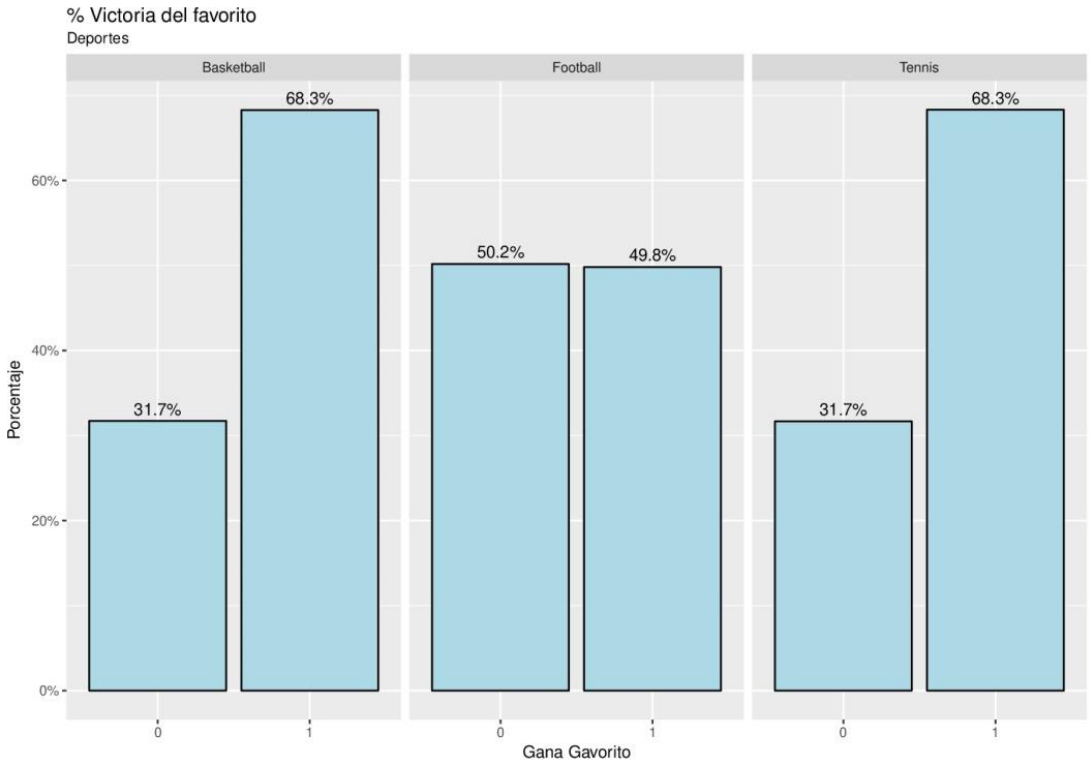
ACCURACY	RECALL
LR: 0.600297	LR: 0.648696
RF: 0.544994	RF: 0.550591
NB: 0.557132	NB: 0.852452

4. Conclusiones

El mundo del deporte profesional es difícil de analizar debido a que cualquier equipo podría ganar a otro aunque aparentemente uno es de mayor nivel que otro.

Como indica el título del trabajo, podemos afirmar que ganar dinero apostando al favorito es un mito, a lo largo del tiempo solamente tendremos perdidas.

Nuestro estudio se ha centrado, sobre todo en el fútbol, en el siguiente gráfico podemos ver que en el fútbol solamente el favorito gana el 49,8 % de las veces.



Los modelos que hemos elaborado en el trabajo nos ayudarán a predecir si gana el favorito, pero si queremos tener más precisión deberíamos tener más variables, con ellas podremos obtener mejores resultados.

5. Manual frontend

APLICACIÓN WEB

Se ha desarrollado con shiny una aplicación para que el usuario pueda predecir si gana el favorito. La aplicación está diseñada para PC.

Está disponible dentro de la carpeta Shiny y a través del siguiente enlace:

https://franciscojdiezc.shinyapps.io/bettingfav/

Se compone de dos pestañas, en la primera el usuario podrá seleccionar las características del equipo favorito, después de seleccionarlas tendrá que dar a predecir para obtener la probabilidad de que gane.

El usuario podrá guardar tus predicciones si pulsa primero en el botón “guardar”, cuando utiliza este botón, en la segunda pestaña se acumularán todas las predicciones y después podrá descargarlas en formato CSV.

TABLEAU PUBLIC

Para terminar se han creado tres dashboards, uno para cada deporte para que el usuario pueda explorar el profundidad los datos. Se pueden acceder a ellos a través de la carpeta Tableau o en los siguientes enlaces:

Fútbol → https://public.tableau.com/profile/franciscojdiezc#!/vizhome/football_22/Dashboard

Baloncesto → https://public.tableau.com/profile/franciscojdiezc#!/vizhome/tennis_5/Dashboard1

Tenis → https://public.tableau.com/profile/franciscojdiezc#!/vizhome/basketball_4/Dashboard1

Dentro de los cuadros de mando podemos ver los siguientes indicadores:

- Hucha semanal: es un indicador de todas las semanas del año que nos muestra, que si siempre apostamos al favorito cuando dinero acumularíamos en nuestra hucha.
- Gana Favorito semana a semana: nos refleja el porcentaje de victoria del favorito en cada semana del año.
- Goles/Puntos semana a semana: en fútbol y baloncesto respectivamente nos dirá la medía de goles y puntos conseguidos en esa semanal del año.
- Verdad o mito: con este KPI podremos ver, después de filtrar, bajo los criterios seleccionados si apostamos al favorito el porcentaje de acierto tendríamos.

El usuario podrá seleccionar los datos mediante diversos filtros: características del deporte, equipo/jugador favorito, fecha para comprobar como varían los indicadores de los cuadros de mando.