LA PREDICTIBILIDAD EN EL FUTBOL

PLANTEAMIENTO

Como usuario esporádico de **apuestas deportivas** siempre me he preguntado hasta que grado un resultado de fútbol puede ser predecible y si podemos usar esas predicciones para **optimizar el riesgo y las ganancias de este tipo de apuestas**. La idea será responder a estas tres preguntas:

- ¿ Con que grado de confianza podemos predecir un resultado (1 X 2)?
- ¿ Podemos identificar que apuestas están sobre o infra-pagadas?
- ¿ Podemos identificar ciertos nichos donde la predictibilidad sea mayor ?

STATE OF ART

Existen páginas donde se indica la probabilidad de que un partido acabe en victoria local, empate o victoria visitante pero no se indica como estas probabilidades son calculadas, ni ofrece diferentes modelos de predicción dependiendo del partido en cuestión u otras variables **ni señalan que partidos son más ventajosos en términos de ganancia potencial**. Mediante este análisis se pretende cubrir todos estos puntos.

ENTORNO

Será necesario ejecutar todos los notebooks o el notebook que contiene todo el código desde el siguientes **repositorio de GitHub.**

Será necesario tener acceso a **JupyterLab** (en uno de los notebooks se procederá a la instalación de dos paquetes necesarios: datapackage and tensorflow) **[NOTEBOOK 2]**

Será necesario una **cuenta de Tableau** para acceder a al dashboard para el usuario.

Acceso a las APIs **betsapi2** (el usuario y password para acceder a ellas ya están incluidas en el notebooks).

LAS FUENTES

Necesitaremos por un lado un histórico con los resultados y las cuotas de las 10 últimas temporadas más los de la temporada en curso. Además necesitaremos identificar los partidos de la siguiente jornada con sus cuotas asociadas.

Para ello utilizaremos las siguientes fuentes:

- Temporadas 09-10 17-18
- Temporadas 19 20 y 20 21
- Partidos de la jornada siguiente y las cuotas asociadas

LOS DATOS

Será necesario la **creación de una base de datos histórica** con las fuentes anteriores proveen una gran cantidad de datos de los cuales nos quedaremos con los siguientes:

- **Div:** League Division
- **Date**: Match Date (dd/mm/yy)
- **Time**: Match Time
- **HomeTeam**: Home Team
- **AwayTeam**: Away Team
- **FTHG**: Full Time Home Team Goals
- **FTAG**: Full Time Away Team Goals
- **FTR**: Full Time Result (H Home Win, D Draw, A Away Win)
- **HS**: Home Team Shots
- **AS**: Away Team Shots
- **HST:** Home Team Shots on Target
- **AST**: Away Team Shots on Target
- **B365H**: Bet365 home win odds
- **B365D**: Bet365 draw odds
- **B365A**: Bet365 away win odds

Para los partidos de la jornada solo necesitaremos los siguientes:

- **Div:** League Division
- **Date**: Match Date (dd/mm/yy)
- **Time**: Match Time
- **HomeTeam**: Home Team
- **AwayTeam**: Away Team
- **B365H**: Bet365 home win odds
- **B365D**: Bet365 draw odds
- **B365A**: Bet365 away win odds

METODOLOGÍA

A) IDENTIFICAR PRÓXIMA JORNADA : PARTIDOS Y CUOTAS [NOTEBOOK 3]

Por una lado es necesario identificar los partidos de la jornada que viene a través de la siguiente API y por otro lado las cuotas asociadas a esos partidos desde la siguiente API. Ambos tienen un identificador único del partido que utilizaremos para crear una tabla única:

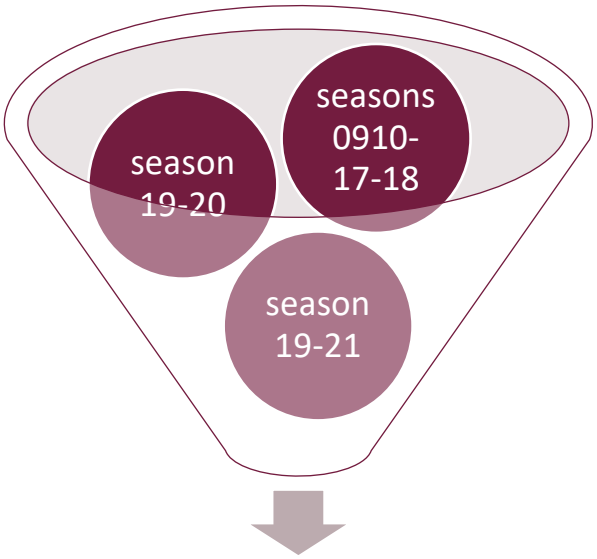
id	home_name	away_name	id	1	X	2
102540251	Levante	Cadiz	102540251	1.727	4.100	4.200
102540243	Celta Vigo	Real Betis	102540243	3.200	3.800	2.050
102540245	Eibar	Barcelona	102540245	6.000	5.000	1.444
102540249	Huesca	Valencia	102540249	1.615	4.200	5.000

id	home_name	away_name	1	X	2
102540251	Levante	Cadiz	1.727	4.100	4.200
102540243	Celta Vigo	Real Betis	3.200	3.800	2.050
102540245	Eibar	Barcelona	6.000	5.000	1.444
102540249	Huesca	Valencia	1.615	4.200	5.000

B) CREAR BASE DE DATOS HISTÓRICA [NOTEBOOK 4]

B.1) Importando los datos

La información está contenida en diferentes bloques por lo que será necesario agregarla por bloques:



Historical dataset

	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HS	AS	HST	AST	B365H	B365D	B365A	season
0	SP1	2018-08-17	Betis	Levante	0	3	A	22	6	8	4	1.66	4	5	1819
1	SP1	2018-08-17	Girona	Valladolid	0	0	D	13	2	1	1	1.75	3.6	5	1819
2	SP1	2018-08-18	Barcelona	Alaves	3	0	H	25	3	9	0	1.11	10	21	1819
3	SP1	2018-08-18	Celta	Espanol	1	1	D	12	14	2	5	1.85	3.5	4.5	1819
4	SP1	2018-08-18	Villarreal	Sociedad	1	2	A	16	8	7	4	2.04	3.4	3.8	1819

El dataset estará compuesto por ~ 4500 partidos y en el proceso añadiremos la variable 'season' para identificar a que temporada pertenece cada partido.

B.2) Creación de una base de datos por equipo y temporada

Las **variables disponibles** se antojan **muy limitadas** para poder pensar en tener predicciones mínimamente fiables.

Por ello es conveniente **disponer del número de jornada,de los puntos y de la clasificación de cada equipo antes de comenzar cada jornada**.

Es por ello será necesario calcular estos campos por cada uno de los equipos en cada una de las jornadas, calcular si jugaba en casa o fuera y dependiendo del resultado asignar los puntos.

season	Team	Date	Home/Away	match	Div	FTHG	FTAG	FTR	HS	AS	HST	AST	B365H	B365D	B365A	points
0910	Almeria	2009-08-30	H	1	SP1	0	0	D	20	7	5	1	2.1	3.3	3.5	1
0910	Almeria	2009-09-13	A	2	SP1	1	0	H	16	7	4	0	2.38	3.25	3	0
0910	Almeria	2009-09-20	H	3	SP1	1	0	H	11	23	3	11	2.5	3.25	2.8	3
0910	Almeria	2009-09-23	A	4	SP1	2	2	D	24	12	9	7	1.44	4.33	7	1
0910	Almeria	2009-09-27	H	5	SP1	2	2	D	13	13	4	6	2.25	3.25	3.2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2021	Villarreal	2021-04-25	H	33	SP1	1	2	A	10	15	4	5	4.5	4.2	1.66	0
2021	Villarreal	2021-05-02	H	34	SP1	1	0	H	4	15	3	3	2.1	2.9	4.1	3
2021	Villarreal	2021-05-09	H	35	SP1	2	4	A	16	11	7	6	1.95	3.6	3.8	0
2021	Villarreal	2021-05-13	A	36	SP1	0	2	A	11	9	2	3	3.8	3.5	1.95	3
2021	Villarreal	2021-05-16	H	37	SP1	4	0	H	12	17	5	5	2.37	3.4	2.87	3

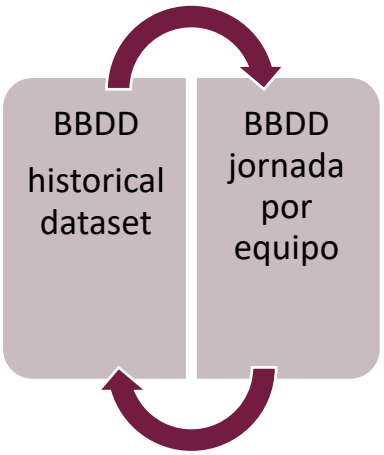
Posteriormente realizaremos los cálculos de los disparos, los goles y los puntos (a favor y en contra) acumulados hasta la jornada anterior de cada uno de los equipos en cada una de las jornadas.

goals_for_acc	goals_against_acc	shots_for_acc	shots_against_acc	shots_target_for_acc	shots_target_against_acc	points_acc
0	0	0	0	0	0	0
0	0	20	7	5	1	1
0	1	27	23	5	5	1
1	1	38	46	8	16	4
3	3	50	70	15	25	5
...	...	...	...	...	...	...
49	36	349	327	139	103	49
50	38	359	342	143	108	49
51	38	363	357	146	111	52
53	42	379	368	153	117	52
55	42	388	379	156	119	55

Una vez tenemos los puntos acumulados de cada equipo hasta la jornada inmediatamente pasamos a calcular la clasificación (inversa para que todas las variables tengan una relación positiva y facilite la interpretabilidad).

season	Date	match	Team	Div	points_acc	ranking
0910	2009-08-30	1	Almeria	SP1	0	1
0910	2009-09-13	2	Almeria	SP1	1	9
0910	2009-09-20	3	Almeria	SP1	1	3
0910	2009-09-23	4	Almeria	SP1	4	11

B.3) Incorporando los datos calculados a la bbdd de los partidos.



Cruzamos la base de datos de los partidos con la creado por equipo y jornada. De esta formada tendremos para cada uno de los equipos los goles, puntos, disparos a favor y en contra acumulados y clasificación.

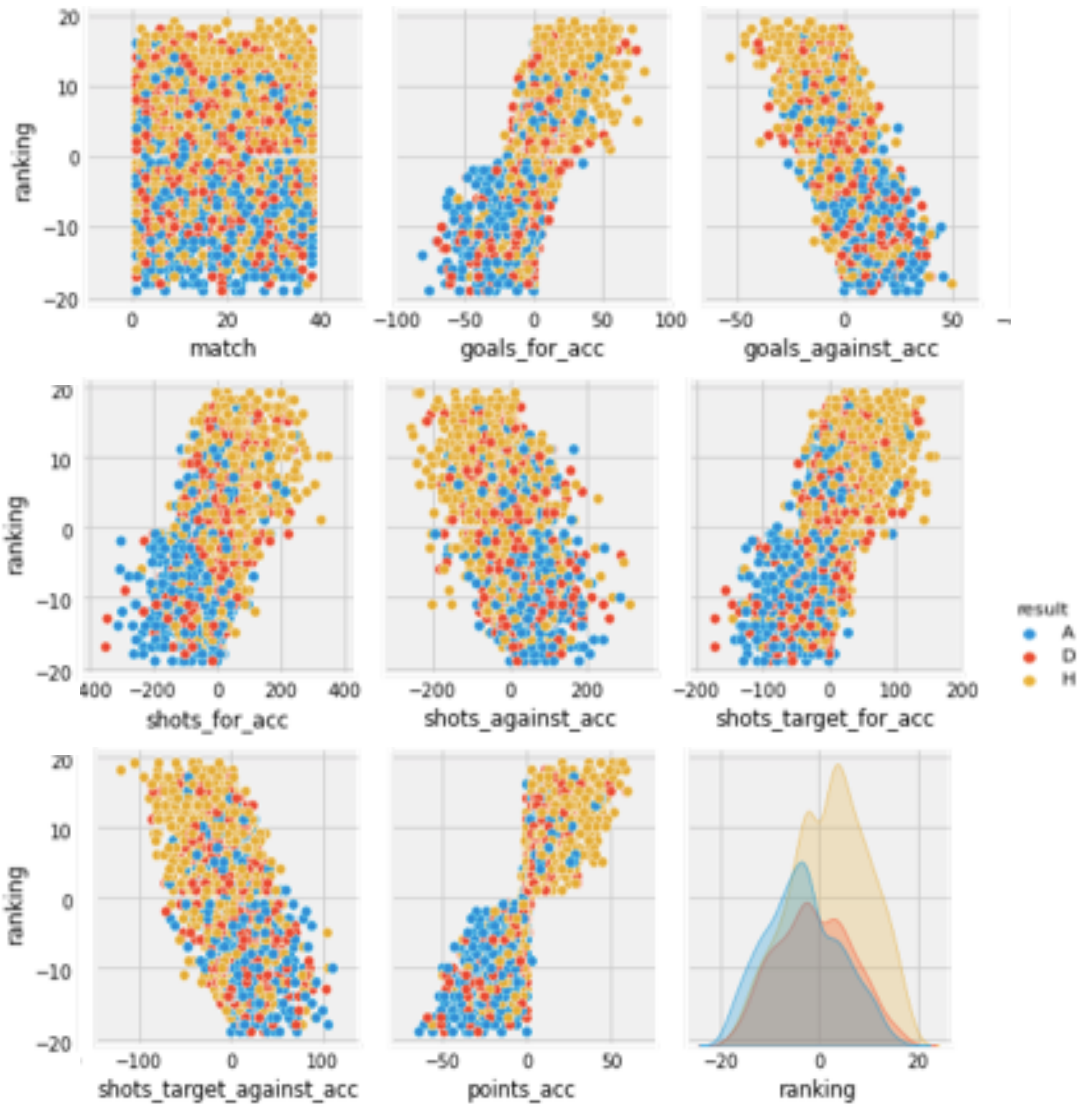
Para que los modelos de machine learning puedan interpretar cada partido será necesario calcular la diferencia entre equipos de cada una de las variables de cada equipo.

season	match	HomeTeam	AwayTeam	goals_for_acc	goals_against_acc	shots_for_acc	shots_against_acc
2021	37	Cadiz	Elche	4	-1	52	-47
2021	37	Getafe	Levante	-17	-11	-22	-140
2021	37	Sociedad	Valladolid	22	-14	66	-135
2021	37	Valencia	Eibar	18	5	-40	171
2021	37	Villarreal	Sevilla	3	13	-45	52

season	match	HomeTeam	AwayTeam	shots_target_for_acc	shots_target_against_acc	points_acc	ranking	result
2021	37	Cadiz	Elche	15	-15	13	7	A
2021	37	Getafe	Levante	-24	-52	-6	-3	H
2021	37	Sociedad	Valladolid	35	-49	25	13	H
2021	37	Valencia	Eibar	7	31	9	6	H
2021	37	Villarreal	Sevilla	19	2	-19	-2	H

Ejemplo: El Getafe llegó a la jornada 37 con 17 goles menos que el Levante, 3 posiciones por debajo al tener 6 puntos menos que el Levante.

C) EXPLORACION DEL DATASET CREADO [NOTEBOOK 5]



No es sorprendente que si el ranking es positivo (el equipo de casa clasifica mejor que el equipo de fuera) y variables como puntos, goles o disparos acumulados también lo son las victorias locales son más comunes.

Los empates no siguen un patrón claro por lo que será presumiblemente el resultado más complicado de predecir.

La jornada del partido no influye significativamente en el patrón de victorias locales, visitantes o empates.

D) PREPARANDO EL TRAIN/TEST DATASETS [NOTEBOOK 6]

D.1) Preprocesando datos

La idea es aplicar modelos de ML tanto al dataset como a su versión 'escalada' para determinar posteriormente con cual podemos trabajar de una forma más optima.

Para ello aplicaremos un LabelEncoder a los nombres de los equipos (Home Team & AwayTeam) y a los resultados (H-D-A).

Modificaremos la naturaleza de algunas de las variables:

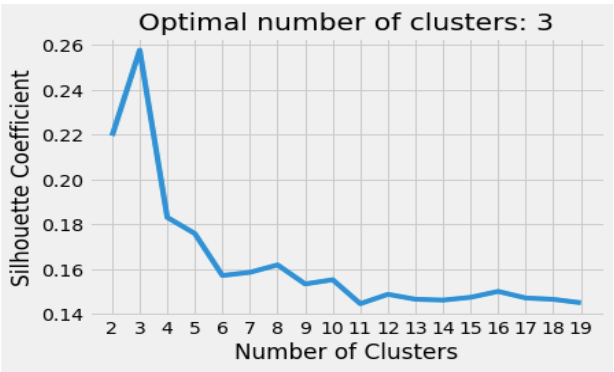
- categories = ['Div','season','HomeTeam','AwayTeam','result']
- floats = ['B365H','B365D','B365A']
- integers = ['season']

D.2) Dividiendo el dataset para poder entrenarlo

Antes de dividir el dataset definiremos la parte de las 'features' obviando la parte de las cuotas (pues implícitamente son las probabilidad asignadas por B365 a ese partido) y definiremos el 'target' como el resultado.

Una vez definida las features y el target dividiremos el dataset estableciendo la parte de test en un 10% (~ 450 partidos). Posteriormente aplicaremos a la parte de train un StandarScaler.

D.3) K-Means clustering

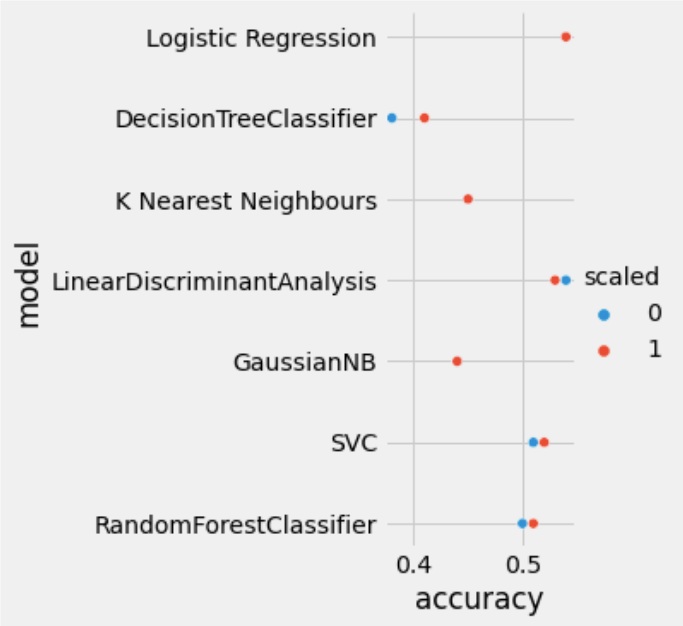


Previendo que puede haber partidos 'parecidos' entre sí que pueden ser agrupados en distintos grupos y que estos pueden tener distintos niveles de predictibilidad aplicamos un K-Means al Train Dataset Scaled.

Número de Clusters : 3

E) APLICANDO DIFERENTES ML MODELOS [NOTEBOOK 7]

E.1) Primera aproximación con varios modelos

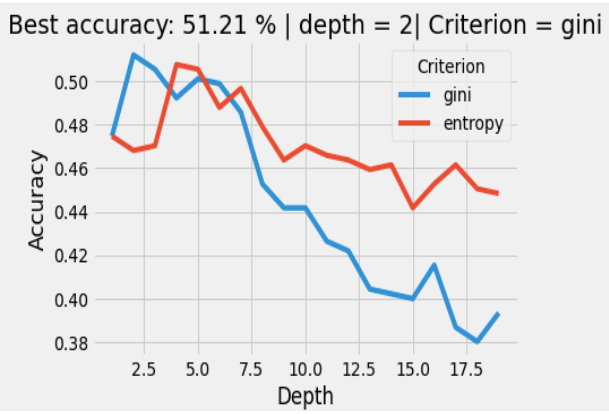


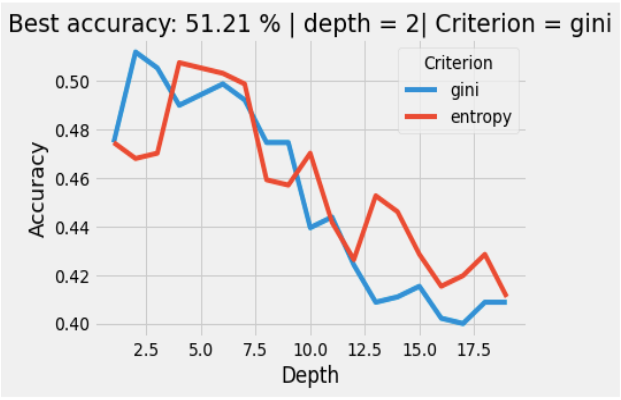
- Probamos 5 modelos sobre el dataset de train normal y sobre el que hemos aplicado un StandardScaler y obtenemos en el mejor de los casos un accuracy del 50%.

- Las diferencias entre el dataset escalado y el normal no presenta diferencias significativas.

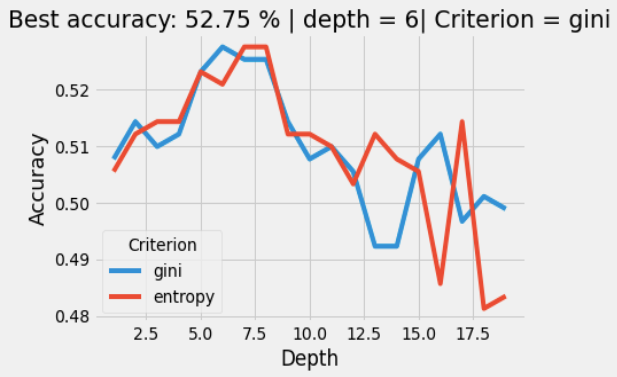
E.2) Búsqueda de los parámetros óptimos para cada par de datasets

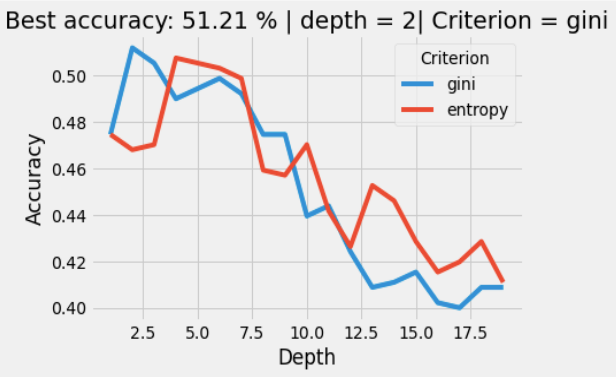
Decision Tree Classifier



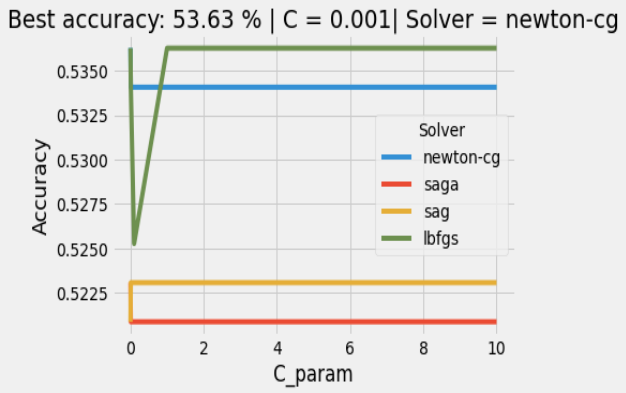


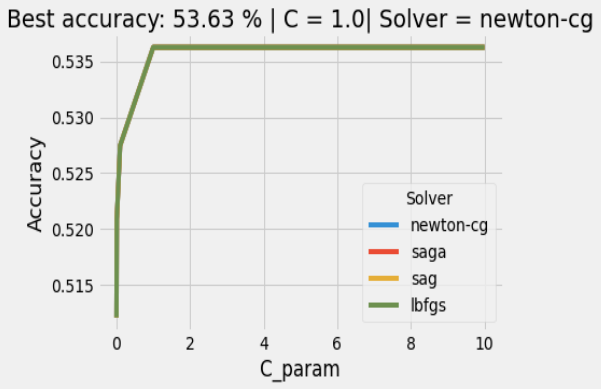
Random Forest Classifier



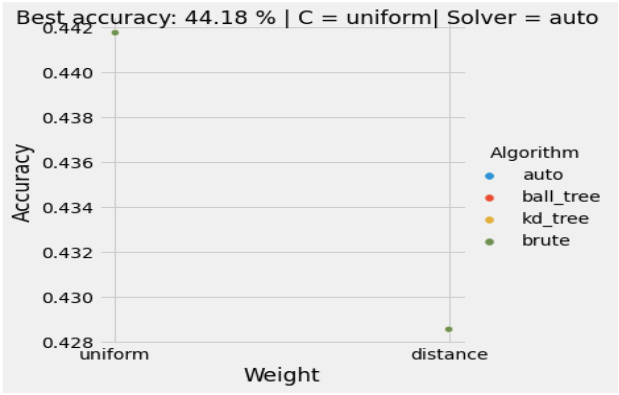


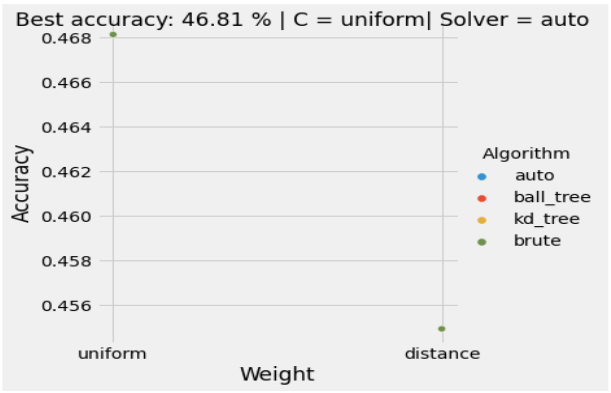
Logistic Regression



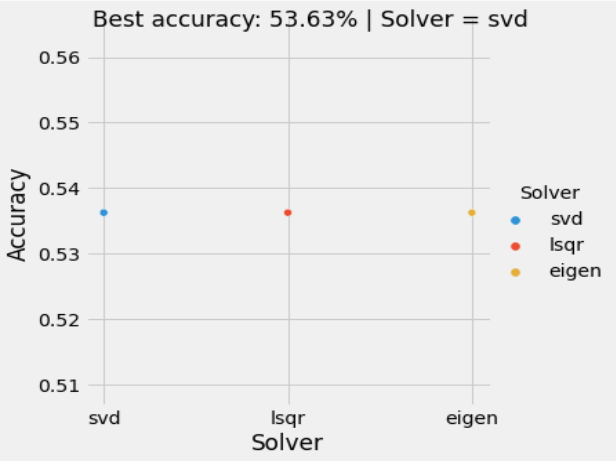


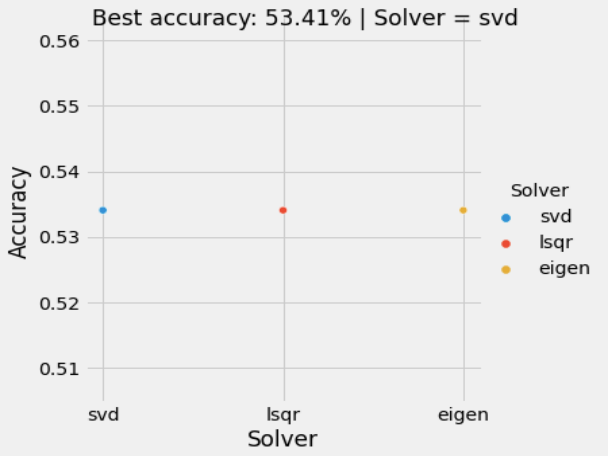
K-Neighbors Classifier



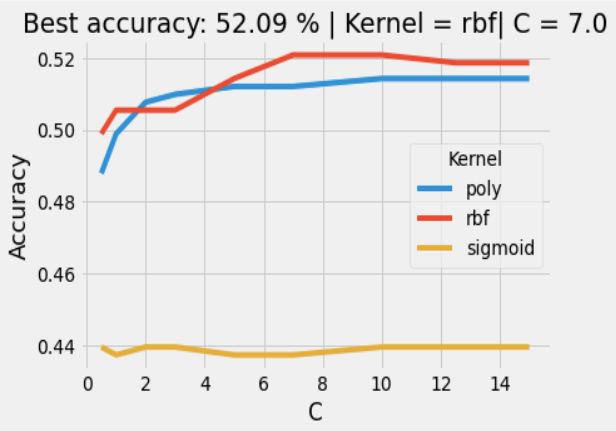


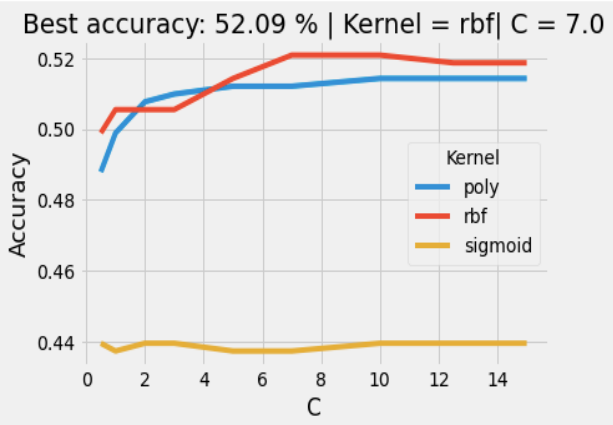
Linear Discriminant Analysis





SVC





E.3) Aplicación de los modelos

Aplicación para cada modelo su versión con la parametrización que optimiza el accuracy tanto al 'X_test' para al dataset completo.

De esta forma tendremos para cada partido las predicciones de cada uno de los modelos así como las probabilidades de que se cada uno de los resultados (H,D,A).

RESULTADOS

F) GENERACIÓN DEL DATASET FINAL Y CONCLUSIONES [NOTEBOOK 8]

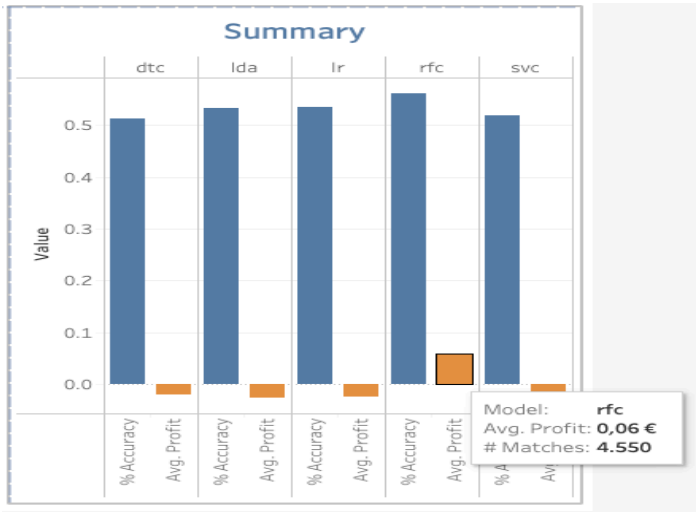
Generamos una base de datos completa con todos los partidos, con sus variables calculadas asociadas, con sus predicciones y probabilidades la cual servirá como base para un análisis de conclusiones contenidos en el siguiente fichero de Tableau(link pendiente de linkar)

season	HomeTeam	AwayTeam	match_home	match_away	B365H	B365D	B365A	FTR	cluster
1819	Barcelona	Alaves	1	1	1.11	10	21	H	2
1516	Betis	Barcelona	36	36	21	9	1.13	A	0
2021	Eibar	Alaves	34	34	2,5	3,1	3	H	2

season	HomeTeam	AwayTeam	Prob_Home%	Prob_Draw%	Prob_Away%	Prediction	hit_result	paid_quota	Prize	Profit	Difficulty	Model
1819	Barcelona	Alaves	42,34	28,16	29,5	H	1	1.11	1.11	0.11	1	dtc
1516	Betis	Barcelona	10,81	21,34	67,85	A	1	1.13	1.13	0.13	1	rfc
2021	Eibar	Alaves	54,36	25,22	20,42	H	1	2,5	2,5	1,5	2	lr

Es importante señalar que un accuracy óptimo será aquel que provea beneficios. Un accuracy alto en apuestas con bajo retorno podría no ser interesante mientras que un accuracy más bajo podría ser interesante si el retorno del gasto en apuestas es positivo.

Resultados Globales



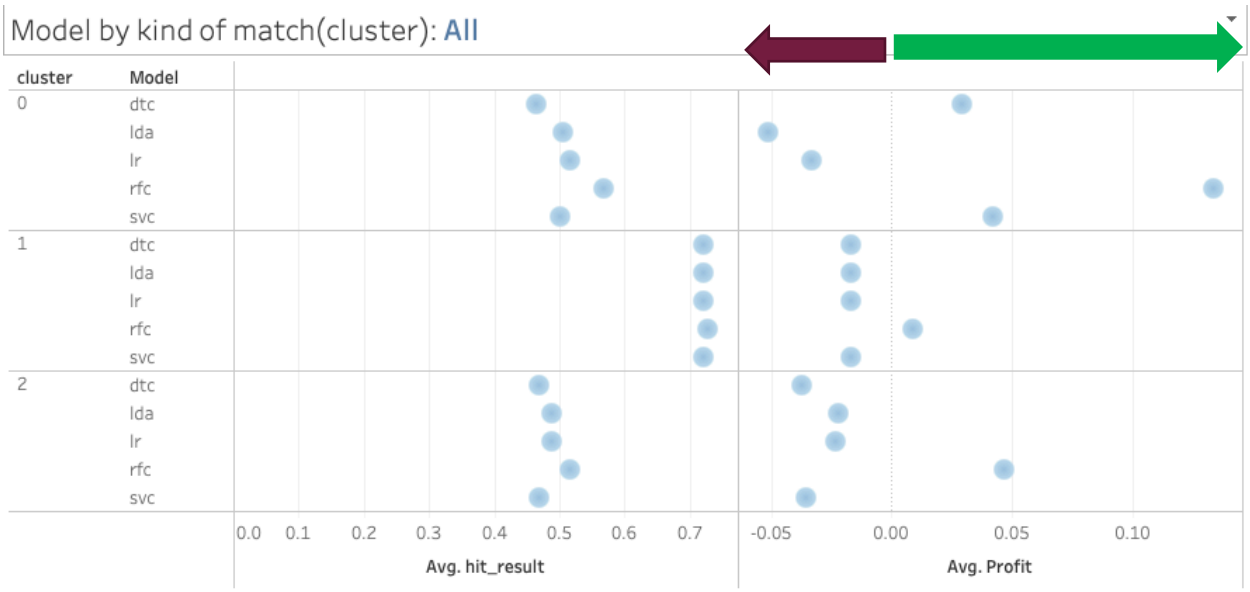
Todos los modelos rondan el 50% de accuracy pero sólo el **Random Forest Classifier** da un beneficio medio de 0.06€ por euro jugado.

Resultados por tipo de partido (Cluster)

Cuando comencé a crear el dataset decidí aplicar un KMeans para saber en cuantos subgrupos podía dividir los partidos para posteriormente analizar si alguno de estos subgrupos era más interesante que otros en términos de accuracy/ganancia.

Los partidos del cluster 1 son muy interesantes en términos de predictibilidad, ya que tienen un accuracy del 70%, sin embargo los del cluster 0 tienden a ser más beneficios en términos de ganancias que 3 de los 5 modelos escogidos dan beneficios medios positivos (dtc, rfc, svc).

El cluster 2 contiene partido tan complicados de predecir como los del cluster 0 sin embargo otorgan beneficios medios mucho menores por lo que por su relación predictibilidad-beneficio medio resultan ser los menos interesantes.



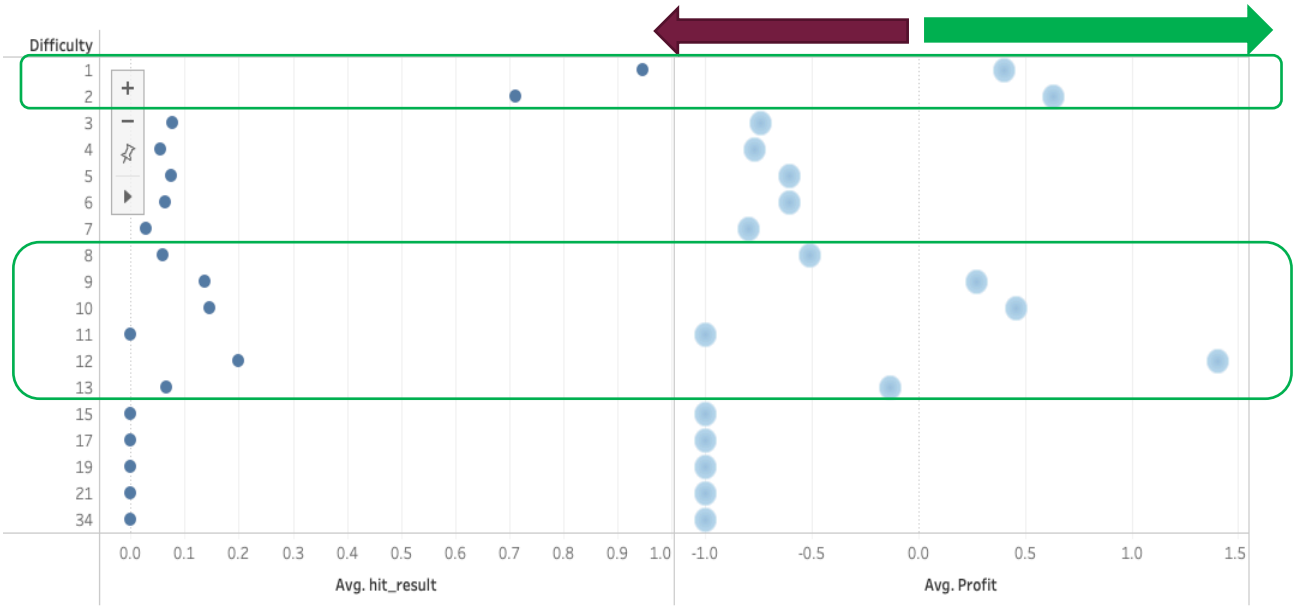
Resultados por cuota pagada (Dificultad)

Otro análisis interesante sería saber si es más interesante apostar a cuotas bajas (más probables) o a cuotas altas (menos probables). **¿Es mejor apostar , en términos generales, una actitud conservadora o una actitud de riesgo?**

Para responder a esta pregunta, manteniendo la simplicidad, he definido de la siguiente manera los niveles de dificultad:

Cuotas que finalmente fueron las acertadas:

- De 1.01 a 1.99 → 1
- De 2.00 a 2.99 → 2
- ….
- De 34.00 a 34.00 → 34

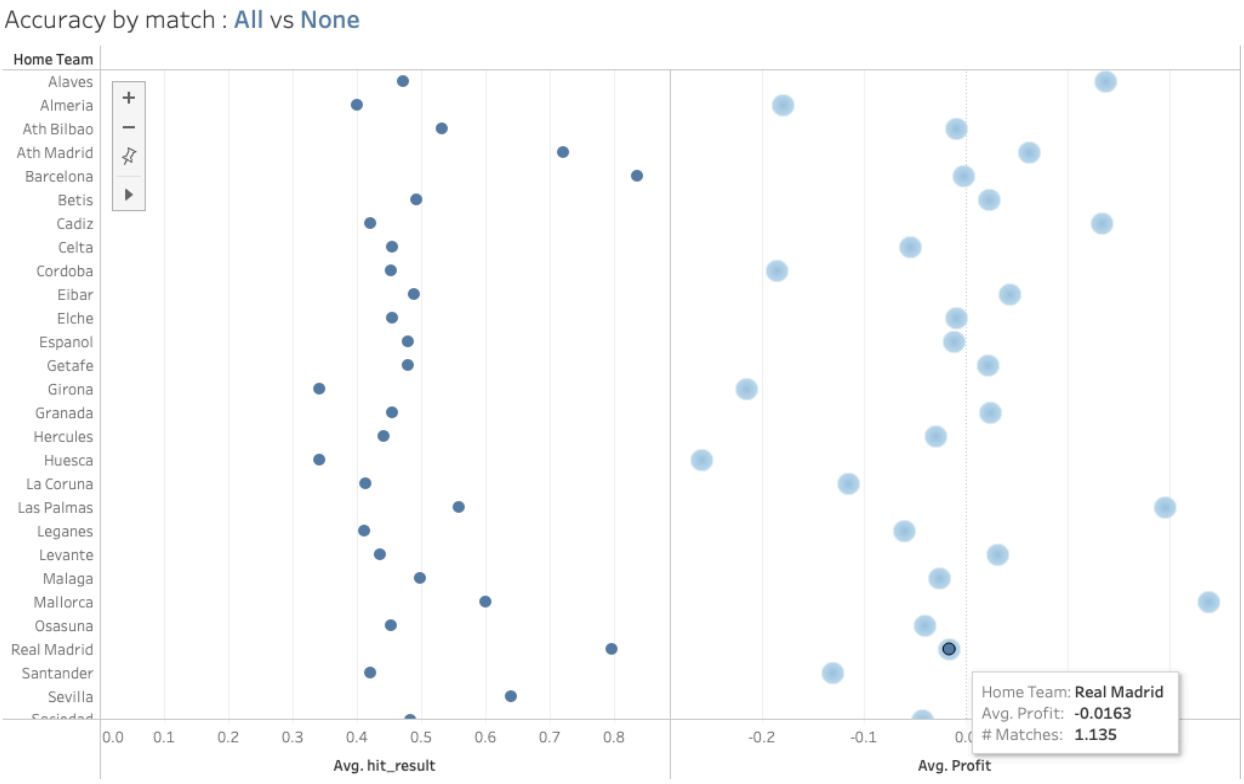


Atendiendo a los resultados resulta más interesante hacer apuestas menores de 2.99€ y en caso de arriesgar es más interesante hacerlo en los tramos de 8 a 13€ por euro apostado ya que aunque el número de partidos que son acertados con estas cuotas es pequeño su beneficio medio es más interesante que los que tienen cuotas de 3€ a 7€ o de 15€ a 34€.

Resultados por Equipo Local o Visitante

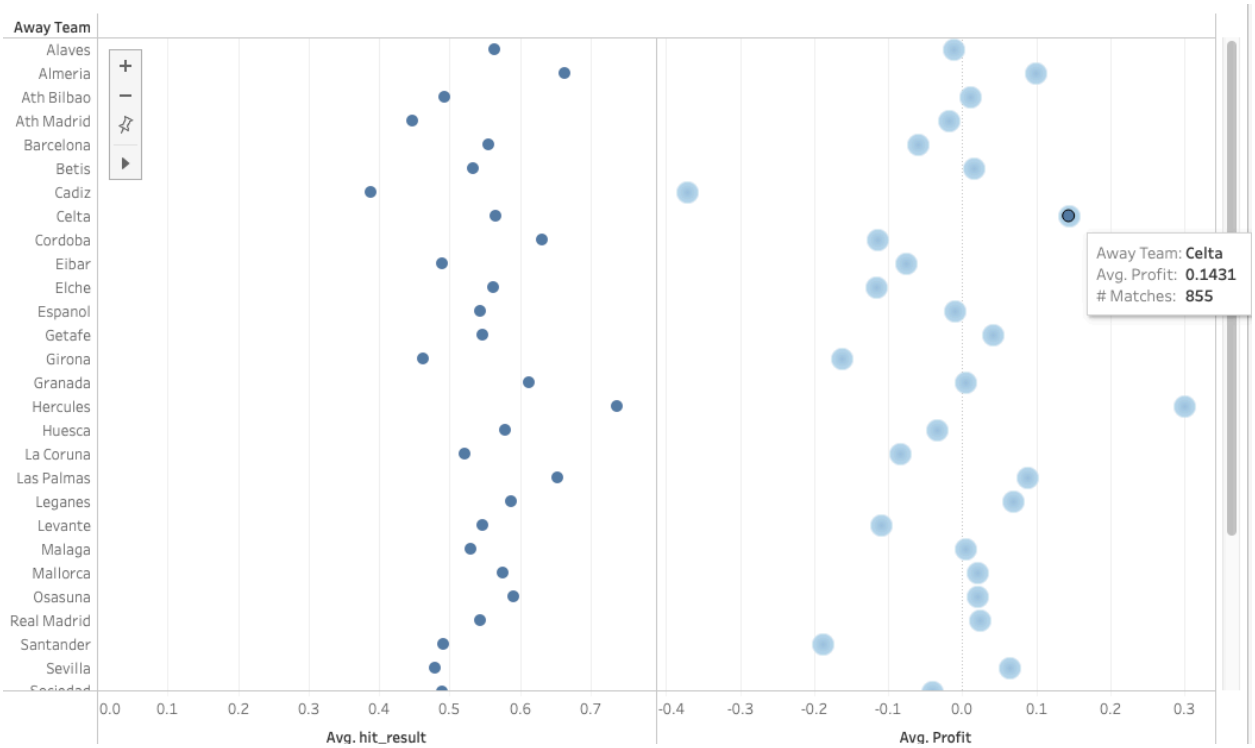
Aún sin datos para afirmarlo categóricamente es intuitivo afirmar que las mayorías de apuestas deportivas implican a equipos como el Madrid, el Barca o el Atlético. Pero ¿Son las opciones más interesantes?

Como podemos ver con el ejemplo del Real Madrid a pesar de tener una capacidad de predicción media de alrededor del 80% cuando juega en casa, el beneficio medio por euro jugado es negativo por lo que otros equipos como el Alavés o el Cádiz resultan apuestas más interesantes cuando juegan en casa.



¿ Y fuera de casa?

Mismo patrón. Apostar al Celta como visitante es mucho más interesante que hacerlo por los equipos grandes ya que sus partidos pueden ser predichos correctamente en el 56% de las ocasiones lo que lleva a un beneficio medio por euro jugado de 15 céntimos.



Todos estos análisis están muy bien a nivel global, pero llega la jornada que viene y queremos saber a que partido apostar.

¿Como facilitamos que toda esta información este disponible para cualquier usuario?

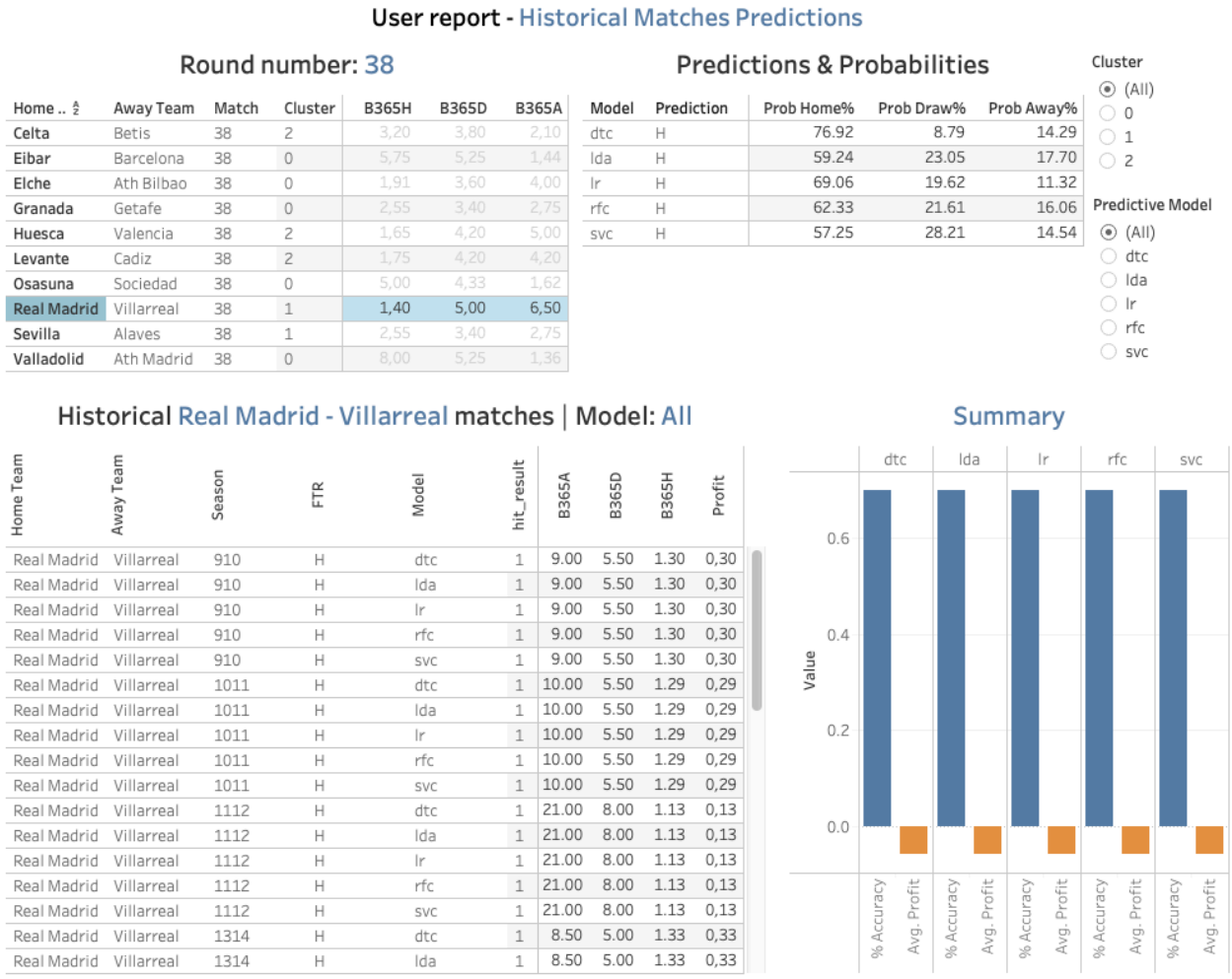
A través del reporte que relaciona todos estos resultados con la jornada siguiente que puedes encontrar aquí (INTRODUCIR HYPERLINK!!!)

G) GENERACIÓN DEL INFORME PARA EL USUARIO [NOTEBOOK 9]

La idea del reporte es dar toda la información contenida hasta ahora aplicada sobre los 10 partidos de la jornada de liga siguiente [al haber acabado la liga utilizaremos a modo de muestra la última jornada jugada].

Para ello cruzaremos los partidos, junto con sus cuotas, obtenidas en el notebook 2 y le aplicaremos todos los procesos (cálculo de variables, K-Means clustering y modelos de ML).

De esta forma podremos ir viendo partido a partido cual podría ser más Interesante. ¿Apostamos por el Real Madrid?



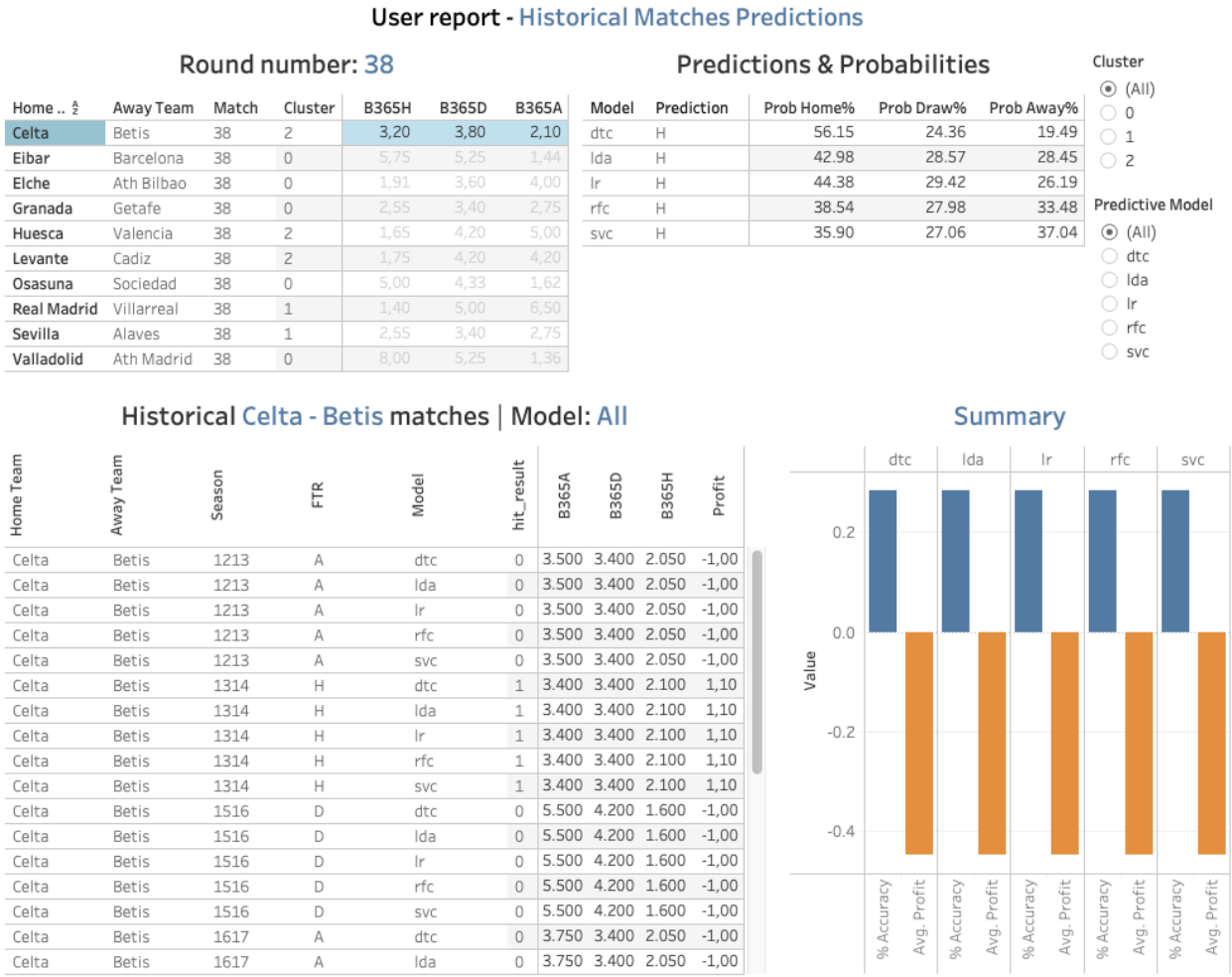
Home ..	Away Team	Match	Cluster	B365H	B365D	B365A
Celta	Betis	38	2	3,20	3,80	2,10
Eibar	Barcelona	38	0	5,75	5,25	1,44
Elche	Ath Bilbao	38	0	1,91	3,60	4,00
Granada	Getafe	38	0	2,55	3,40	2,75
Huesca	Valencia	38	2	1,65	4,20	5,00
Levante	Cadiz	38	2	1,75	4,20	4,20
Osasuna	Sociedad	38	0	5,00	4,33	1,62
Real Madrid	Villarreal	38	1	1,40	5,00	6,50
Sevilla	Alaves	38	1	2,55	3,40	2,75
Valladolid	Ath Madrid	38	0	8,00	5,25	1,36

Model	Prediction	Prob Home%	Prob Draw%	Prob Away%
dtc	H	76.92	8.79	14.29
lda	H	59.24	23.05	17.70
lr	H	69.06	19.62	11.32
rfc	H	62.33	21.61	16.06
svc	H	57.25	28.21	14.54

Home Team	Away Team	Season	FTR	Model	hit_result	B365A	B365D	B365H	Profit
Real Madrid	Villarreal	910	H	dtc	1	9.00	5.50	1.30	0,30
Real Madrid	Villarreal	910	H	lda	1	9.00	5.50	1.30	0,30
Real Madrid	Villarreal	910	H	lr	1	9.00	5.50	1.30	0,30
Real Madrid	Villarreal	910	H	rfc	1	9.00	5.50	1.30	0,30
Real Madrid	Villarreal	910	H	svc	1	9.00	5.50	1.30	0,30
Real Madrid	Villarreal	1011	H	dtc	1	10.00	5.50	1.29	0,29
Real Madrid	Villarreal	1011	H	lda	1	10.00	5.50	1.29	0,29
Real Madrid	Villarreal	1011	H	lr	1	10.00	5.50	1.29	0,29
Real Madrid	Villarreal	1011	H	rfc	1	10.00	5.50	1.29	0,29
Real Madrid	Villarreal	1011	H	svc	1	10.00	5.50	1.29	0,29
Real Madrid	Villarreal	1112	H	dtc	1	21.00	8.00	1.13	0,13
Real Madrid	Villarreal	1112	H	lda	1	21.00	8.00	1.13	0,13
Real Madrid	Villarreal	1112	H	lr	1	21.00	8.00	1.13	0,13
Real Madrid	Villarreal	1112	H	rfc	1	21.00	8.00	1.13	0,13
Real Madrid	Villarreal	1112	H	svc	1	21.00	8.00	1.13	0,13
Real Madrid	Villarreal	1314	H	dtc	1	8.50	5.00	1.33	0,33
Real Madrid	Villarreal	1314	H	lda	1	8.50	5.00	1.33	0,33

¿Que tal Celta – Betis?

Parecería un partido a evitar ya que la capacidad de predicción de este partido ha sido muy baja durante sus duelos históricos. Ningún modelo llega a un 30% de acierto y se pierden de media más de 40 céntimos por euro jugado. Lo complicado de acertar este partido no está proporcionalmente ajustado por cuotas mejor pagadas/más justas.



Home ..	Away Team	Match	Cluster	B365H	B365D	B365A
Celta	Betis	38	2	3,20	3,80	2,10
Eibar	Barcelona	38	0	5,75	5,25	1,44
Elche	Ath Bilbao	38	0	1,91	3,60	4,00
Granada	Getafe	38	0	2,55	3,40	2,75
Huesca	Valencia	38	2	1,65	4,20	5,00
Levante	Cadiz	38	2	1,75	4,20	4,20
Osasuna	Sociedad	38	0	5,00	4,33	1,62
Real Madrid	Villarreal	38	1	1,40	5,00	6,50
Sevilla	Alaves	38	1	2,55	3,40	2,75
Valladolid	Ath Madrid	38	0	8,00	5,25	1,36

Model	Prediction	Prob Home%	Prob Draw%	Prob Away%
dtc	H	56.15	24.36	19.49
lda	H	42.98	28.57	28.45
lr	H	44.38	29.42	26.19
rfc	H	38.54	27.98	33.48
svc	H	35.90	27.06	37.04

Home Team	Away Team	Season	FTR	Model	hit_result	B365A	B365D	B365H	Profit
Celta	Betis	1213	A	dtc	0	3.500	3.400	2.050	-1,00
Celta	Betis	1213	A	lda	0	3.500	3.400	2.050	-1,00
Celta	Betis	1213	A	lr	0	3.500	3.400	2.050	-1,00
Celta	Betis	1213	A	rfc	0	3.500	3.400	2.050	-1,00
Celta	Betis	1213	A	svc	0	3.500	3.400	2.050	-1,00
Celta	Betis	1314	H	dtc	1	3.400	3.400	2.100	1,10
Celta	Betis	1314	H	lda	1	3.400	3.400	2.100	1,10
Celta	Betis	1314	H	lr	1	3.400	3.400	2.100	1,10
Celta	Betis	1314	H	rfc	1	3.400	3.400	2.100	1,10
Celta	Betis	1314	H	svc	1	3.400	3.400	2.100	1,10
Celta	Betis	1516	D	dtc	0	5.500	4.200	1.600	-1,00
Celta	Betis	1516	D	lda	0	5.500	4.200	1.600	-1,00
Celta	Betis	1516	D	lr	0	5.500	4.200	1.600	-1,00
Celta	Betis	1516	D	rfc	0	5.500	4.200	1.600	-1,00
Celta	Betis	1516	D	svc	0	5.500	4.200	1.600	-1,00
Celta	Betis	1617	A	dtc	0	3.750	3.400	2.050	-1,00
Celta	Betis	1617	A	lda	0	3.750	3.400	2.050	-1,00