



ANALÍTICA DE DATOS FUTBOLÍSTICOS



INTELIGENCIA DE NEGOCIO

**Sandra Gamero Ortega
Antonio Manuel Claro González
Aitor Muñoa Carrasco
Miguel Ángel Sanz Guijarro**

Índice

1. Introducción	4
2. Plan	5
2.1 Objetivos	5
2.2 Plan de trabajo	5
2.2.1 Extracción de datos	5
2.2.2 Preprocesamiento	6
2.2.3 Analítica de datos	6
2.2.4 Cuadro de mandos	14
2.2.4.1 Reporte dinámico	14
2.2.4.2 Analítica covid	17
2.2.4.3 EXTRA	18
2.3 Análisis de la viabilidad	18
2.4 Riesgos	18
3. Análisis	19
3.1 Establecimiento de los requisitos del sistema	19
4. Diseño	20
4.1. Capacidad de memoria de la organización.	20
4.2. Capacidad de integración de información.	23
4.2.1 Eliminar columnas que no necesitamos	23
4.2.2 Solucionar valores inexistentes (missing values)	23
4.2.3 Renombramiento de columnas	23
4.2.4 Cambiar tipos de datos	23
4.3. Capacidad de crear conocimiento.	24
-Regresión lineal basada en las estadísticas del propio partido:	24
-Regresión Random forest basada en las estadísticas del propio partido:	24
-Regresión potenciado de gradiente basada en las estadísticas del propio partido:	25
-Regresión potenciado de gradiente basada en serie temporal:	26
4.4. Capacidad de presentación.	26
-Reporte dinámico	26
-Analítica covid	27
5. Implementación	28
5.1. Pasos para la instalación y configuración de los datos.	28
5.2. Software necesario para la extracción de los datos	34
5.3. Software necesario para el análisis de los datos.	34
5.4. Software necesario para crear los reportes.	34
5.5. Otro software necesario	34
6. Despliegue	34
Primera Rama	37
Segunda Rama	39
7. Conclusiones	39

Extracción de datos	39
Análisis en KNIME	39
PowerBI:	41
Reporte dinámico:	41
Analítica covid:	41
9. Manual de instalación	46
KNIME	46
POWER BI	48

1. Introducción

El proyecto que llevaremos a cabo en la asignatura de Inteligencia de Negocio estará relacionado con la analítica de datos futbolísticos. En este caso trabajaremos con datos de partidos de La Liga, es decir partidos de distintas temporadas (serie temporal) de la primera división de fútbol española.

Hoy en día tenemos muchas aplicaciones que nos permiten ver datos de fútbol de forma objetiva, pero hay pocas que lo traten de forma analítica. Nuestra intención es poder utilizar los datos previos para realizar predicciones y poder detectar ciertos factores que hayan sido determinantes a la hora del resultado.

Nos hemos definido 3 objetivos para este proyecto:

- Predicción de estadísticas/resultado dado 2 equipos: Pondremos cuál será el equipo local y el visitante y obtendremos la predicción del resultado y/o estadísticas para ese partido.
- Reporte dinámico: Se realizará un reporte dinámico donde seleccionando un equipo se podrán visualizar sus estadísticas generales, además de poder verlas desglosadas por temporadas o de forma general.
- Estudio sobre la afectación del Covid-19 a los resultados de los partidos: Siempre se ha dicho que la afición ha jugado un papel fundamental para ciertos equipos al jugar en casa. A raíz del Covid-19 se han jugado muchos partidos a puerta cerrada y se puede estudiar si la dinámica de resultados ha cambiado o no, además de ver si afecta en otras estadísticas.

En cuanto a las tecnologías usadas, los datos serán sacados de una API de deportes, "<https://www.api-football.com>". Se realizará un script en Python para hacer llamadas a la API y montar el dataset para el posterior estudio.

La parte predictiva, se realizará usando Knime. El reporte dinámico se realizará con PowerBI. En cuanto a la afectación del Covid-19, podremos realizar también las gráficas con PowerBI.

2. Plan

2.1 Objetivos

Tras una reunión poniendo puntos en común hemos decidido que vamos a tener un total de tres objetivos:

- **Predicción de estadísticas/resultado dados 2 equipos:** la idea de este objetivo es, dado un partido (equipo local y visitante), obtener una predicción del resultado del partido mediante los diferentes algoritmos de predicción que hemos aplicado.
- **Reporte dinámico:** Crearemos un reporte dinámico donde podremos visualizar las estadísticas de un equipo en una temporada concreta. Podremos seleccionar un equipo y mostrar sus estadísticas generales y a su vez podremos verlas desglosadas por temporada o de manera general.
- **Estudio sobre la afectación del Covid-19:** Siempre se ha dicho que la afición ha jugado un papel fundamental para ciertos equipos al jugar en casa. A raíz del Covid-19 se han jugado muchos partidos a puerta cerrada y se puede estudiar si la dinámica de resultados ha cambiado o no, además de ver si afecta en otras estadísticas.

2.2 Plan de trabajo

El plan de trabajo que hemos desarrollado consta de 4 fases principalmente, las cuales iremos detallando a continuación.

2.2.1 Extracción de datos

Hemos extraído los datos de la siguiente [página](#), la cual hemos estado viendo y tiene los datos que necesitamos para el proyecto. Para la extracción hemos empleado la API que nos ofrece la página, con la cual poco a poco hemos ido sacando los datos que necesitamos.

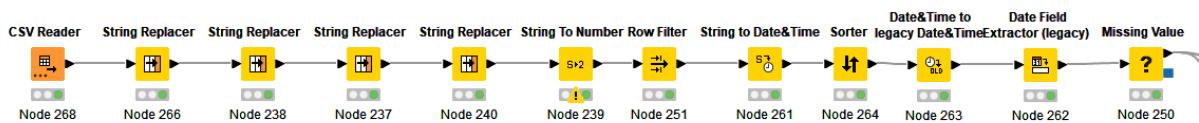
Para la extracción hemos empleado una serie de scripts programados en python:

- **get_seasons.py:** extrae los datos de las temporadas en un rango de años deseados. Guardamos esas temporadas en formato json.
- **get_matches_ids.py:** extrae los ids de los partidos de todas las temporadas extraídas.
- **get_matches.py:** toma los ids de todos los partidos, detecta cuáles han sido ya extraídos y va haciendo peticiones a la api para extraer los datos de cada partido, y va guardando cada partido en formato json.
- **parse_json_to_csv.py:** con los archivos .json anteriores, se queda con los datos más importantes o representativos de las estadísticas y los guarda en un .csv, para que sea más simple a la hora de la carga en los posteriores programas.

Además se ha utilizado un archivo config.json para mantener la coherencia y parametrizar los scripts.

2.2.2 Preprocesamiento

Para comenzar con el preprocesamiento cargaremos los datos mediante un .csv. Una vez realizada esta carga, se procede a transformar y a procesar todos esos datos mediante diferentes nodos de KNIME vistos en la asignatura y también algunos nuevos. A continuación se muestra lo realizado:



Básicamente, se han quitado los % para poder tomarlo como datos numéricos, se han convertido esas columnas de porcentaje a número y se han filtrado las filas para coger los datos con posterioridad al 2014, ya que es cuando empiezan a guardarse las estadísticas de los partidos.

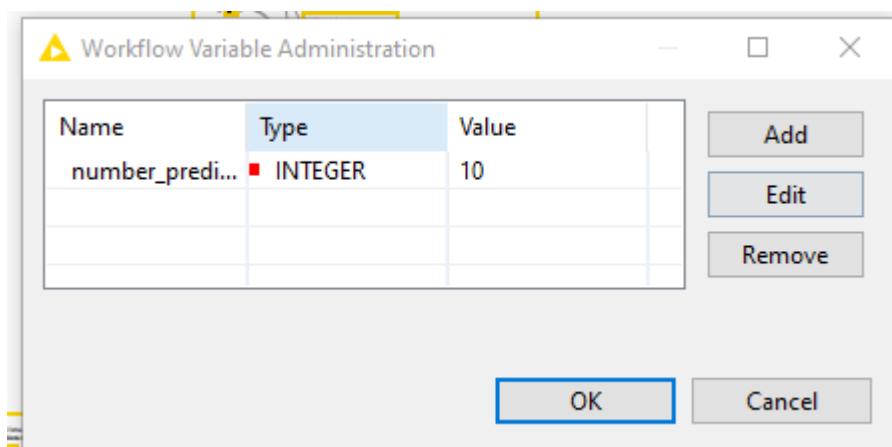
Por último, se ha transformado la fecha para que sea más entendible para Knime y se han extraído sus campos, y se ha ordenado por fecha.

En el caso de los missing, nos hemos fijado en los datos que si había missing en la mayoría de los casos se correspondía con un 0 (número de tarjetas rojas, amarillas, tiros a puerta, etc.), por lo que se ha optado por sustituir por 0 aquellos missing pertenecientes a columnas numéricas.

2.2.3 Analítica de datos

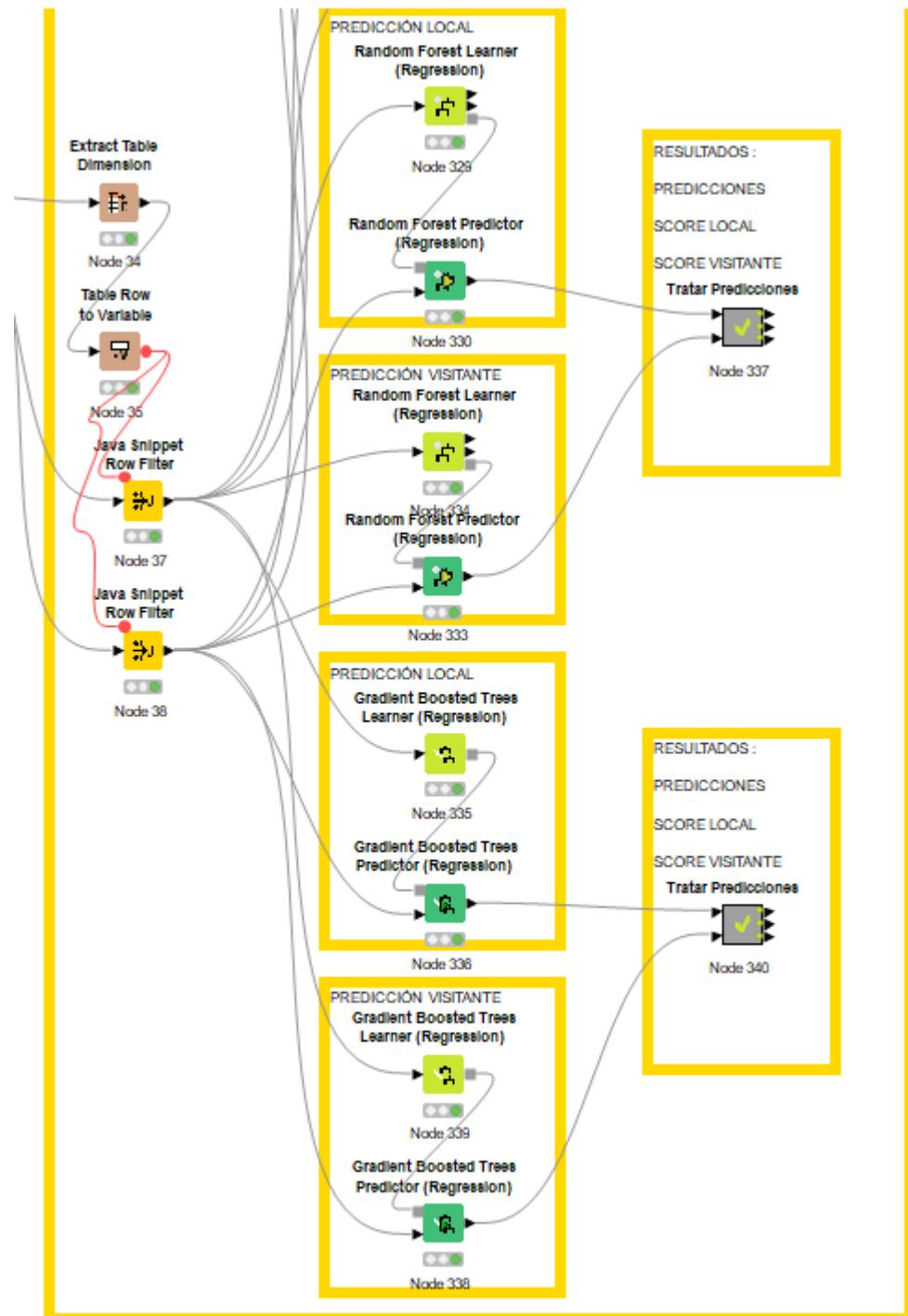
La fase de la analítica se lleva a cabo mediante diferentes nodos de KNIME, vistos a lo largo de la asignatura. Hemos creado 2 ramas para esta parte, en cada una usamos la regresión de forma distinta, aunque cabe mencionar que el preprocesamiento es el mismo para ambas ramas.

En ambos casos se ha querido predecir el resultado de los partidos de la última jornada, que está compuesta por 10 partidos.



El número de partidos a predecir es parametrizable en las variables de flujo del workflow. Siempre tomará los últimos N partidos para hacer la predicción.

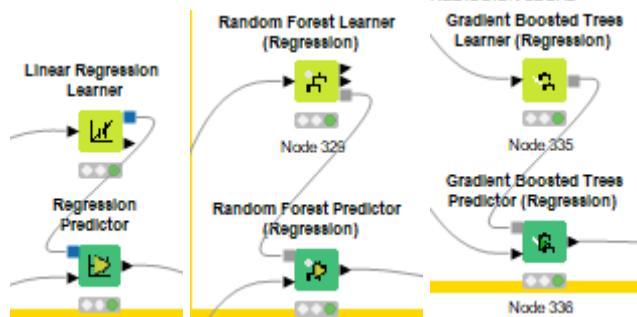
La primera rama emplea regresión basada en estadísticas del propio partido.



Utilizamos java snippets para separar en training y test para que el test abarque solo 10 filas.

Además realizamos 2 predicciones, uno para los goles del equipo local y otro para los goles del equipo visitante.

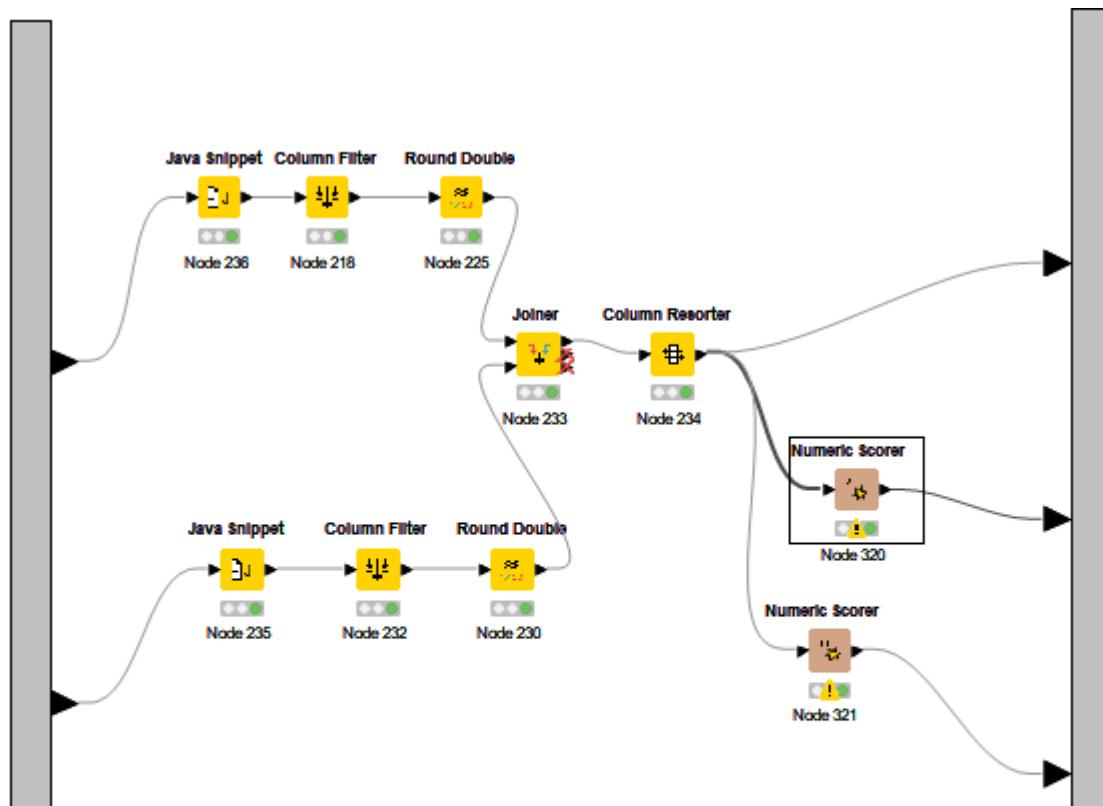
Hemos probado a predecir de 3 formas distintas:



Regresión lineal, regresión random forest y regresión de gradiente.

Más adelante comentaremos los resultados y cuál creemos que es la más positiva.

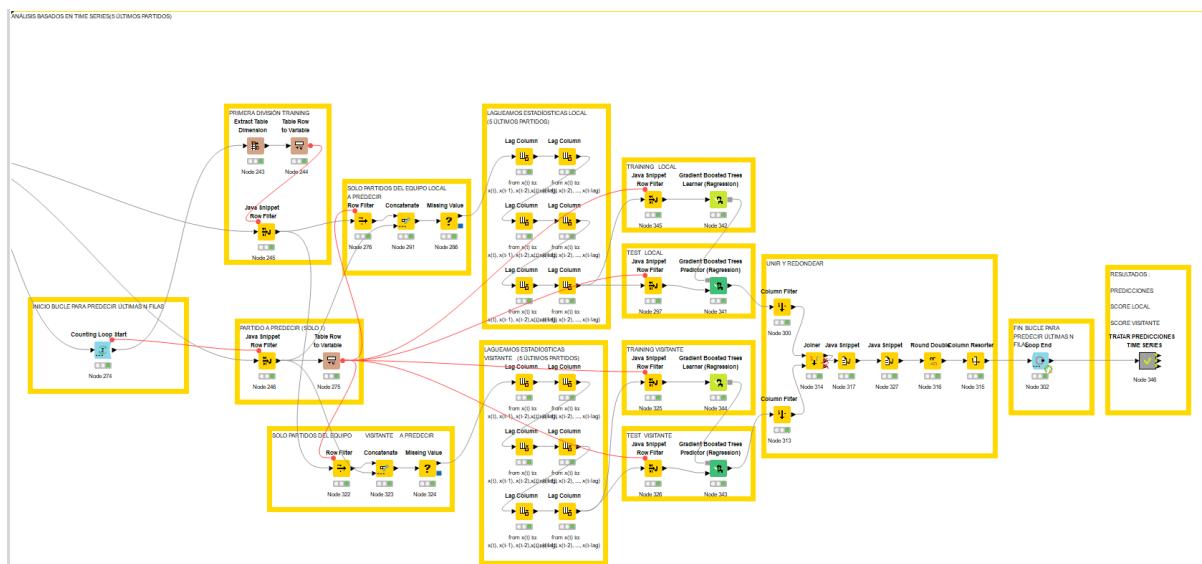
Por último, tenemos un metanodo para tratar las predicciones



Al ser los goles un campo entero, redondeamos los resultados que nos han salido y reordenamos las columnas.

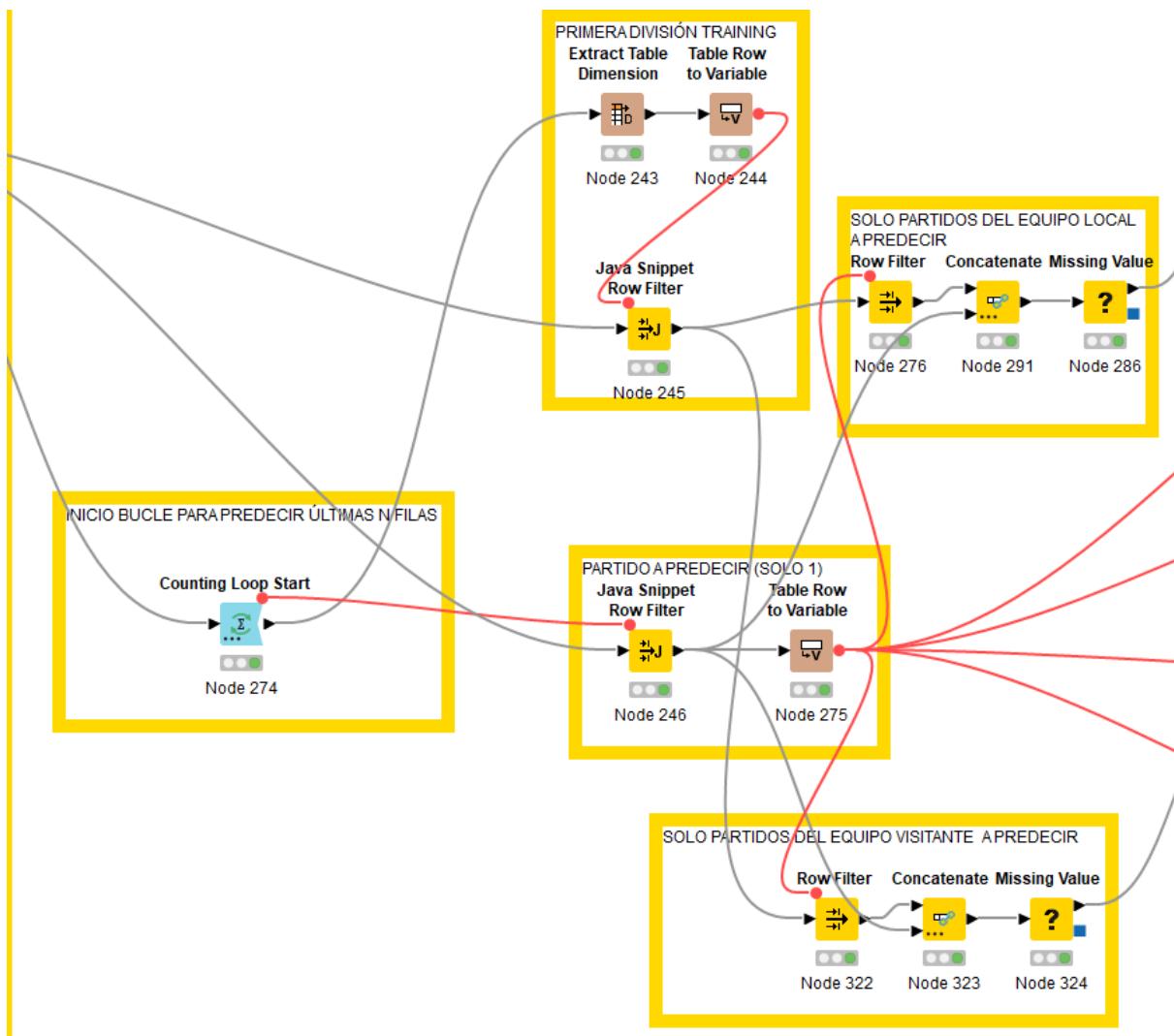
La predicción para los goles del equipo local se basan en sus estadísticas y el nombre de su equipo y el equipo visitante. Lo mismo ocurre para el visitante, pero a la inversa.

Por otro lado, la segunda rama emplea regresión basada en series temporales, el cual nos sirve para predecir el siguiente partido sin tener los datos del mismo.

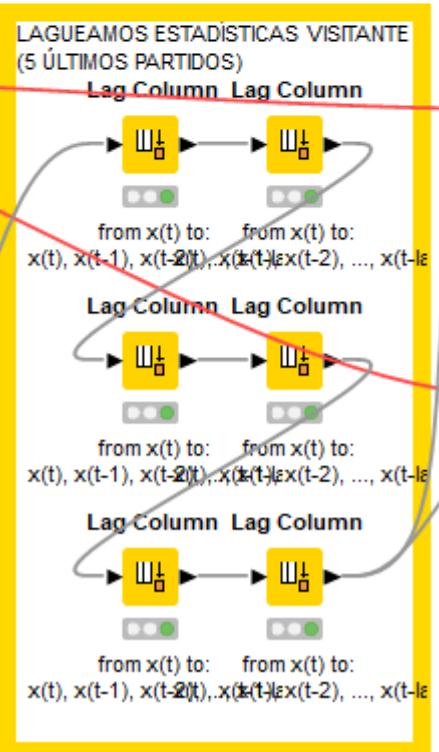
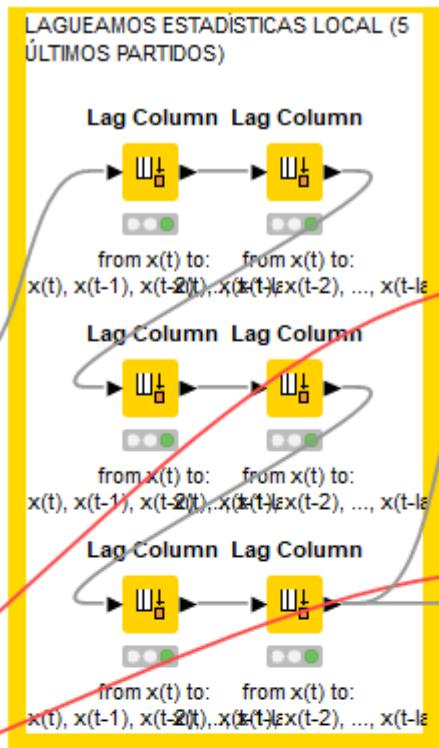


Este flujo es algo más complejo. Para predecir un partido, nos basaremos únicamente en los partidos que ha jugado ese equipo como local y el otro equipo para visitante.

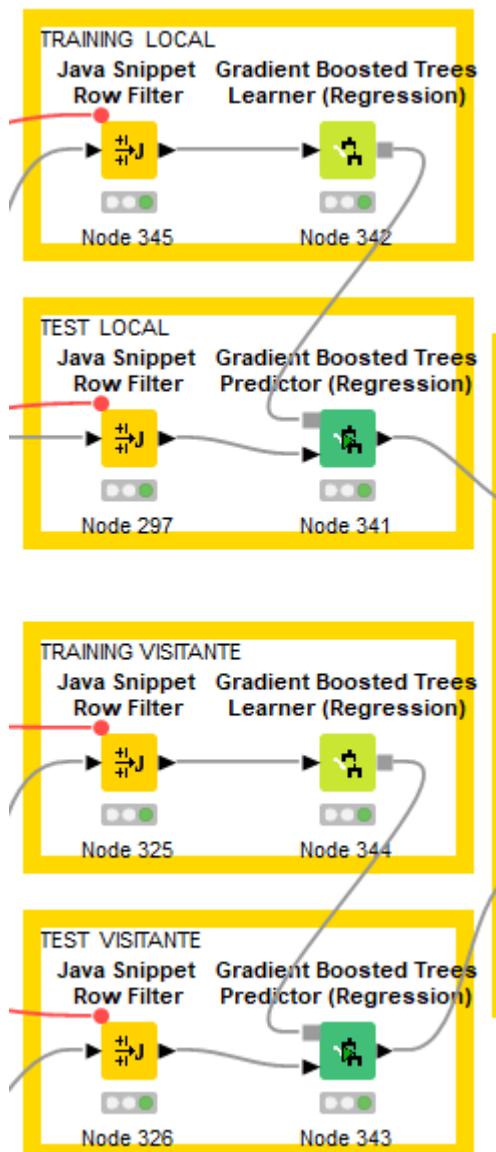
Para esto, hacemos un bucle que cuenta hasta 10, con java snippets sepáramos el dataset en 2 (todos menos los 10 últimos y el último de la iteración actual), filtramos del primer grupo aquellos que tengan al equipo local como el equipo local del partido que queremos predecir y el segundo grupo, nos quedamos con aquellos partidos que tengan como equipo visitante el equipo visitante del partido que queremos predecir.



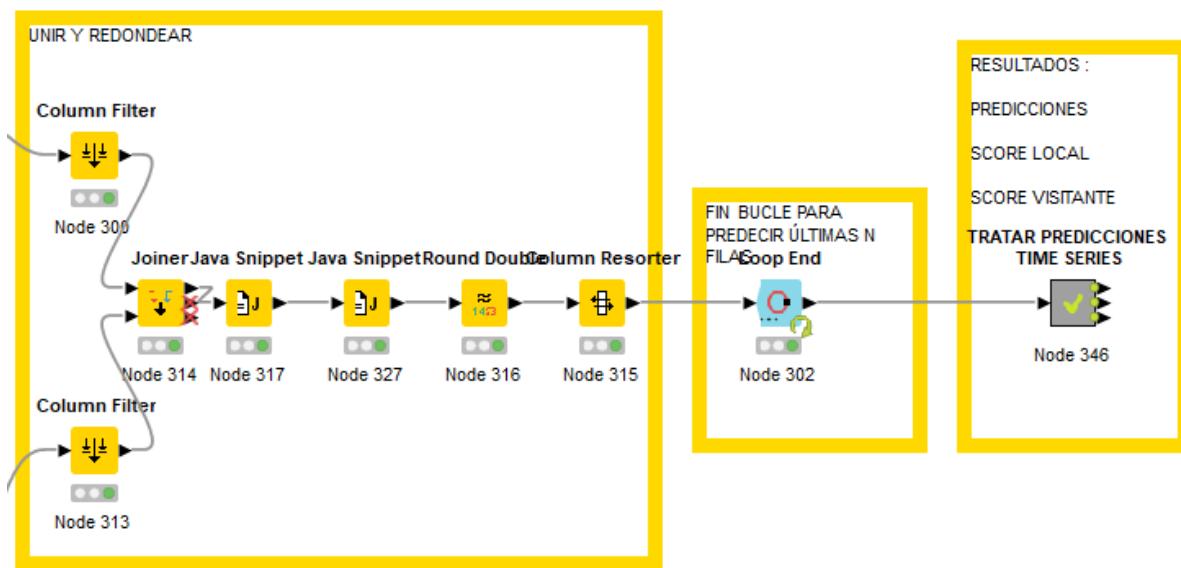
A continuación, lagueamos las columnas con las que haremos las predicciones tanto para la parte local como para la visitante. En este caso, se tendrán en cuenta los últimos 5 partidos.



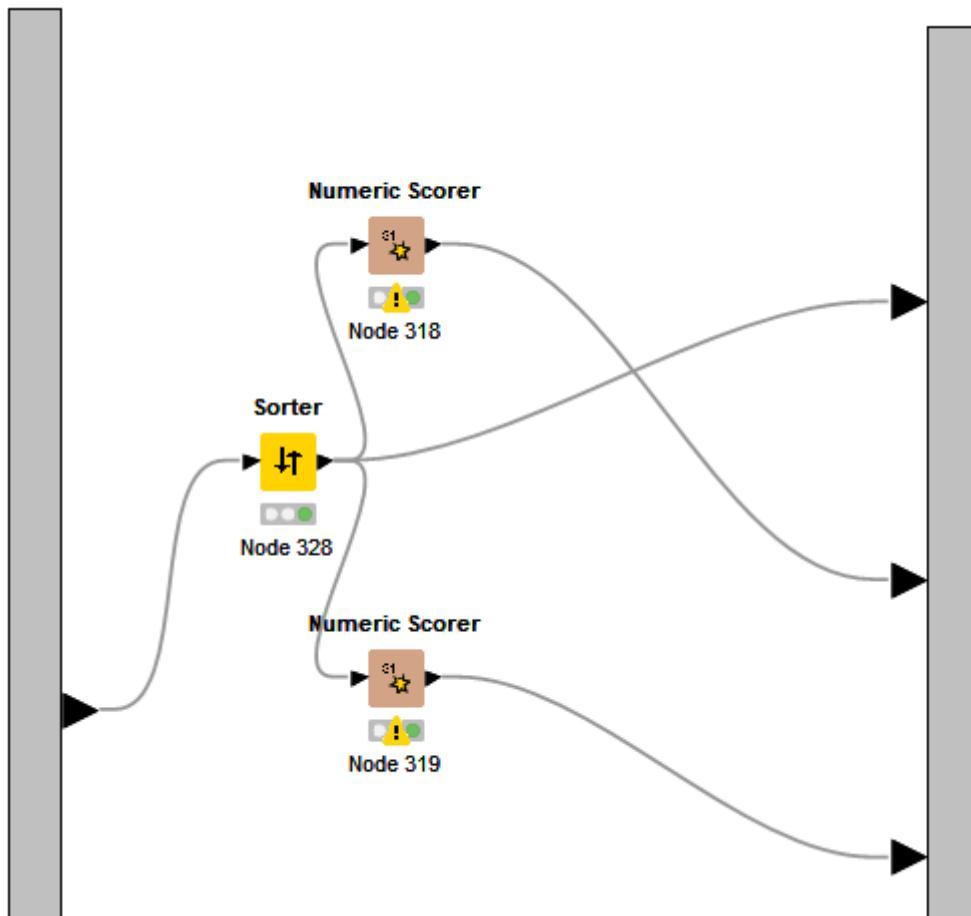
Tras esto, hacemos las predicciones para el equipo local y visitante con la regresión basada en potenciación de gradiente.



Ahora, unimos, ordenamos y redondeamos las columnas, las agregamos al final del bucle y obtenemos los resultados finales y los score.



Por último, el metanodo final lo único que hace es sacar los score y la tabla.



2.2.4 Cuadro de mandos

Esta es la última fase, donde nos encargamos de representar diferentes gráficos.

Esta fase consiste principalmente en representar los datos de manera visual para así crear un impacto en la persona que esté leyendo el informe final.

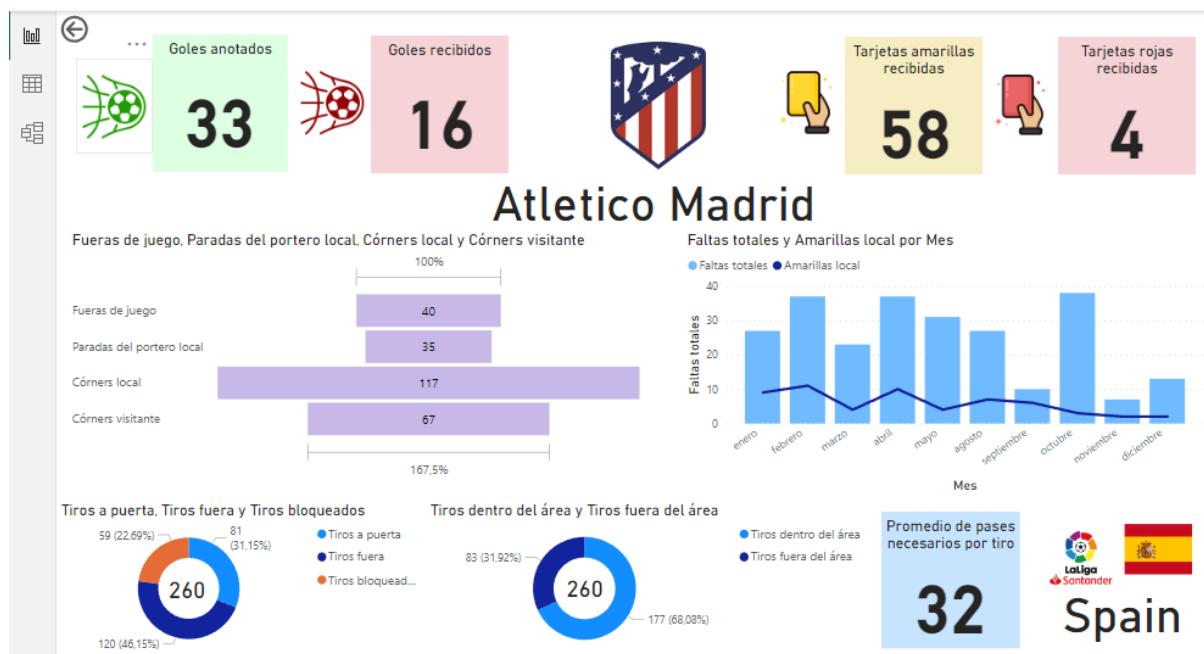
Para poder emplear el cuadro de mandos hemos empleado los datos obtenidos en la fase de extracción.

2.2.4.1 Reporte dinámico

Hemos realizado un reporte dinámico para poder ver las estadísticas de un equipo en una temporada concreta.

Al cargar los datos, hemos realizado conversiones de tipos para las columnas y categorización para algunas, como las url de las imágenes.

Este reporte dinámico lo tenemos dividido en 2 informes, las estadísticas cuando el equipo juega como local y las estadísticas cuando juega como visitante.



Esta imagen de arriba es el dashboard cuando el equipo 'Atlético de Madrid' juega como local en la temporada 2021. A continuación mostramos lo mismo como visitante.



Podemos filtrar por equipo y temporada en los filtros de la página.

Filtros

Buscar

Filtros de esta página

...

away_team_name

es Atletico Madrid

Buscar

<input type="checkbox"/> Alaves	19
<input type="checkbox"/> Athletic Club	19
<input checked="" type="checkbox"/> Atletico Madrid	19
<input type="checkbox"/> Barcelona	19
<input type="checkbox"/> Cadiz	19
<input type="checkbox"/> Celta Vigo	19
<input type="checkbox"/> Elche	19

Requerir selección única

season

es 2021

Buscar

<input type="checkbox"/> 2010	19
<input type="checkbox"/> 2011	19
<input type="checkbox"/> 2012	19
<input type="checkbox"/> 2013	19
<input type="checkbox"/> 2014	19
<input type="checkbox"/> 2015	19
<input type="checkbox"/> 2016	19
<input type="checkbox"/> 2017	19

Requerir selección única

Agregar campos de datos aquí

Filtros de todas las páginas

...

Agregar campos de datos aquí

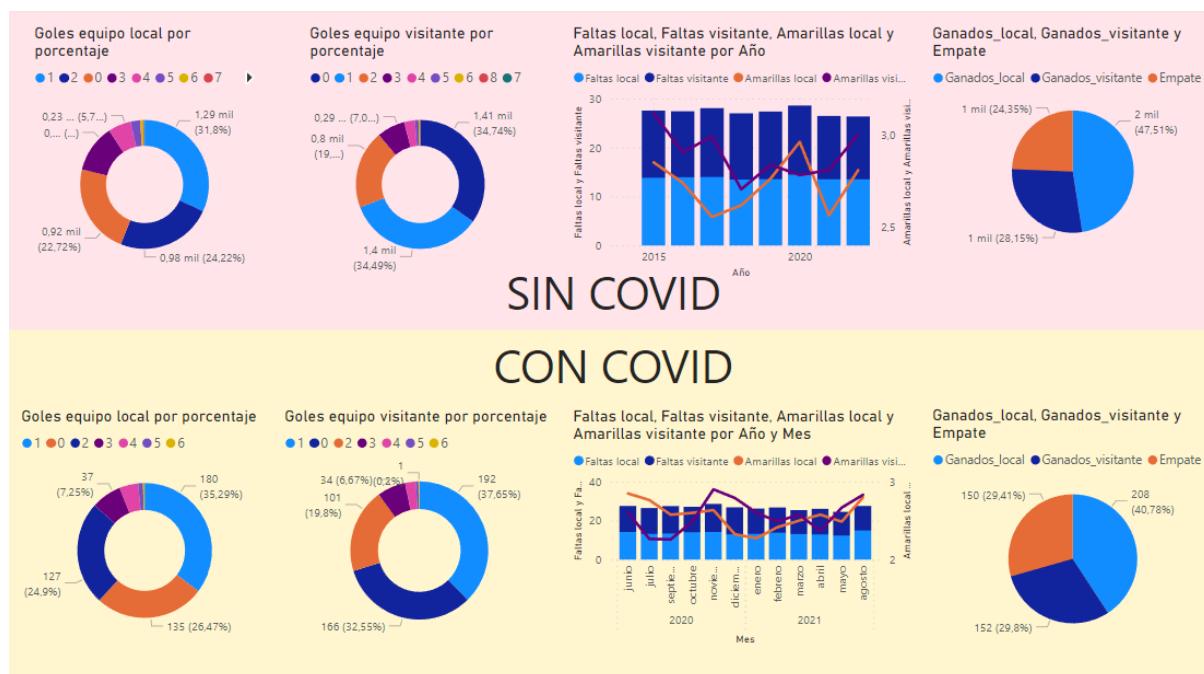
En este dashboard principalmente se muestran estadísticas básicas e importantes (goles anotados, recibidos, tarjetas amarillas y rojas recibidas), gráficos intermedios (desglose de tiros en tiros a puerta, fuera y bloqueados; desglose de tiros dentro del área y fuera), y otros

gráficos algo más complejos (embudo con otras estadísticas para su comparativa y faltas totales por mes vs tarjetas amarillas por mes). Además se ha sacado una métrica a través de DAX para sacar el promedio de pases que necesita el equipo para lograr un tiro a portería.

También se renderizan imágenes del escudo del equipo, de la liga y bandera del país al que pertenece, por lo que sería aplicable a otras ligas. También se incluyen el nombre del equipo y el nombre del país.

2.2.4.2 Analítica covid

Tenemos un informe con las gráficas sesgadas por fechas: partidos en los que no había covid y había público en los estadios vs partidos en los que había covid y no había público en los estadios.



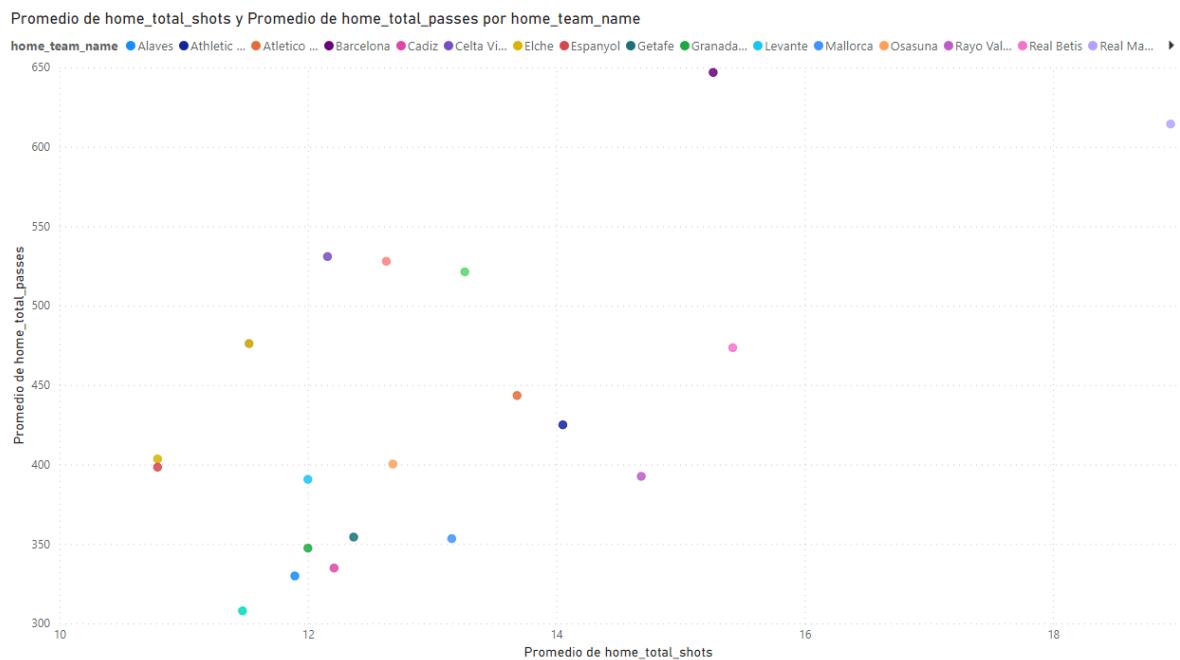
Aquí hemos desglosado los goles en número, tanto de local como visitante, para analizar si el público influye a la hora de marcar más o menos goles.

Además también hemos analizado el promedio de faltas de local y visitante, para ver si el público puede llegar a influir en la agresividad de los jugadores, así como el número de amarillas en promedio. En la época covid añadimos el desglose por mes.

Por último, mediante consultas DAX, hemos obtenido el número de partidos ganados por el equipo local, visitante y empelados. Hemos representado un gráfico circular para ver el % que representan sobre el total, para ver si existen diferencias.

2.2.4.3 EXTRA

Incluimos una gráfica extra que nos parecía interesante, en otra página del informe.



Aquí podemos ver reflejado el promedio de pases por partido vs el promedio de tiros por partido desglosado por equipo. Esto nos da mucha información visual sobre la manera de jugar de un equipo (más tiros Real Madrid, más pases Barcelona, juego directo Betis y Rayo...).

2.3 Análisis de la viabilidad

El coste del proyecto no es muy elevado, analizando los distintos aspectos tenemos que:

- El coste en tiempo no ha sido muy alto, los scripts eran automáticos y se podían poner mientras se hacía otra cosa en segundo plano. El mayor tiempo empleado ha sido realizando el flujo de knime, el cuadro de mandos y la documentación.
- El coste económico no ha sido muy alto tampoco, hemos invertido 20€ para poder sacar los datos de la página web indicada con comodidad, ya que la versión gratuita tenía un límite de 100 peticiones diarias y nos ha dado ciertos problemas debido a baneos de cuentas y de ip.

2.4 Riesgos

En caso de no poder alcanzar los objetivos en cada uno de las distintas fases, se produciría una pérdida de datos provocando un menor nivel de exactitud en las predicciones que pretendemos obtener con este proyecto.

3. Análisis

3.1 Establecimiento de los requisitos del sistema

Sistema operativo Windows 10 o superior. Si no se utilizara powerBI, también sería posible linux.

Para la recogida de datos vamos a lanzar unos scripts usando el lenguaje python, además se necesita tener instalado las librerías ‘pandas’ y ‘requests’ para usar estos scripts. Los scripts son lanzados desde la carpeta principal del repositorio.

Usaremos KNIME para el preprocesamiento y el análisis de datos.

Para los cuadros de mando se usará POWER BI.

Hemos usado git para trabajar con control de versiones (https://github.com/antonioclaro99/INTELIGENCIA_NEGOCIO).

Los datos se recogen en archivos JSON pero los scripts antes mencionados los convierte a csv para su posterior manejo.

Es necesaria conexión a internet para la extracción de datos desde la api: <https://www.api-football.com>

4. Diseño

4.1. Capacidad de memoria de la organización.

Hemos usado un fichero csv que reúne datos de partidos de fútbol jugados entre la temporada 2010 y 2021. Cada fila corresponde a un partido.

A	B	C	D	E	F	G	H	I
	fixture_id	date	season	league_id	league_name	league_logo	country_name	country_logo
0	203337	2015-08-21T1	2015	140	La Liga	https://media Spain		2015
1	203338	2015-08-22T1	2015	140	La Liga	https://media Spain		2015

- fixture_id: Identificador asignado al partido por la api.
- date: Fecha en la que se jugó el partido.
- season: Temporada en la que se jugó.
- league_id: Identificador asignado a la liga en la que se jugó el partido.
- league_name: Nombre de la liga.
- league_logo: Logo de la liga.
- country_name: Nombre del país al que pertenece la liga.
- country_logo: Bandera del país en el que se jugó el partido.

J	K	L	M	N	O	P	Q
home_team_id	home_team_name	home_team_img	away_team_id	away_team_name	away_team_img	home_goals	away_goals
535	Malaga	https://media-3.ap	536	Sevilla	https://media-2.ap	0	0
540	Espanyol	https://media-3.ap	546	Getafe	https://media-2.ap	1	0

- home_team_id: Identificador asignado al equipo local por la api.
- home_team_name: Nombre del equipo local.
- home_team_img: Imagen del escudo del equipo local.
- away_team_id: Identificador asignado al equipo visitante por la api.
- away_team_name: Nombre del equipo visitante.
- away_team_img: Imagen del escudo del equipo visitante.
- home_goals: Goles marcados por el equipo local.
- away_goals: Goles marcados por el equipo visitante.

R	S	T	U	V
home_shots_on_goal	home_shots_off_goal	home_total_shots	home_blocked_shots	home_shots_insidebox
5.0	13.0	25.0	7.0	9.0
2.0	2.0	5.0	1.0	3.0

- home_shots_on_goal: Disparos a puerta realizados por el equipo local
- home_shots_off_goal: Disparos que van fuera realizados por el equipo local.
- home_total_shots: Disparos totales realizados por el equipo local.
- home_blocked_shots: Disparos bloqueados por los defensas del equipo local.
- home_shots_insidebox: Disparos dentro del área del equipo local.

home_shots_outsidebox	home_fouls	home_corner_kicks	home_offsides	home_ball_possession	home_yellow_cards
16.0		7.0	1.0	57%	3
2.0		4.0	5.0	33%	2

- home_shots_outsidebox: Disparos desde fuera del área del equipo local.
- home_fouls: Faltas realizadas por el equipo local.
- home_corner_kicks: Saques de córner del equipo local.
- home_offsides: Saques de banda del equipo local.
- home_ball_possession: Posesión de la pelota del equipo local en porcentaje.
- home_yellow_cards: Tarjetas amarillas recibidas por el equipo local.

AC	AD	AE	AF	AG
home_red_cards	home_goalkeeper_saves	home_total_passes	home_passes_accurate	home_passes_percentage
2.0	374.0	298.0		80%
2.0	263.0	174.0		66%

- home_red_cards: Tarjetas rojas recibidas por el equipo local.
- home_goalkeeper_saves: Paradas realizadas por el portero del equipo local.
- home_total_passes: Total de pases realizados por el equipo local.
- home_passes_accurate: Precisión del equipo local para realizar un pase con éxito.
- home_passes_percentage: Porcentaje de pases realizados con éxito por el equipo local.

AH	AI	AJ	AK	AL	AM
away_shots_on_goal	away_shots_off_goal	away_total_shots	away_blocked_shots	away_shots_insidebox	away_shots_outsidebox
2.0	7.0	10.0	1.0	8.0	2.0
3.0	7.0	13.0	3.0	5.0	8.0

- away_shots_on_goal: Disparos a puerta realizados por el equipo visitante.
- away_shots_off_goal: Disparos que van fuera realizados por el equipo visitante.
- away_total_shots: Disparos totales realizados por el equipo visitante.
- away_blocked_shots: Disparos bloqueados por los defensas del equipo visitante.
- away_shots_insidebox: Disparos dentro del área del equipo visitante.
- away_shots_outsidebox: Disparos desde fuera del área del equipo visitante.

AN	AO	AP	AQ	AR	AS
away_fouls	away_corner_kicks	away_offsides	away_ball_possession	away_yellow_cards	away_red_cards
2.0	3.0		43%	5	1.0
6.0	2.0		67%	5	1.0

- away_fouls: Faltas realizadas por el equipo visitante.
- away_corner_kicks: Saques de córner del equipo visitante.
- away_offsides: Saques de banda del equipo visitante.
- away_ball_possession: Posesión de la pelota del equipo visitante en porcentaje.

-away_yellow_cards: Tarjetas amarillas recibidas por el equipo visitante.

-away_red_cards: Tarjetas rojas recibidas por el equipo visitante.

AT	AU	AV	AW
away_goalkeeper_saves	away_total_passes	away_passes_accurate	away_passes_percentage
4.0	283.0	207.0	73%
1.0	519.0	434.0	84%

-away_goalkeeper_saves: Paradas realizadas por el portero del equipo visitante.

-away_total_passes: Total de pases realizados por el equipo visitante.

-away_passes_accurate: Precisión del equipo visitante para realizar un pase con éxito.

-away_passes_percentage: Porcentaje de pases realizados con éxito por el equipo visitante.

4.2. Capacidad de integración de información.

Para que los resultados obtenidos en nuestras soluciones sean válidos, es necesario que la información con la que tratamos sea, a su vez, válida. Por lo tanto, los datos que no necesitamos o que contengan algún fallo, necesitaremos corregirlos para poder utilizar de forma adecuada dichos datos.

4.2.1 Eliminar columnas que no necesitamos

Al fusionar conjuntos de datos o al tratar con ellos, hay columnas que no son relevantes en nuestras soluciones. Por lo tanto, descartamos dichas columnas. Al igual que algunos datos se repetirán, también será necesario eliminarlas.

4.2.2 Solucionar valores inexistentes (missing values)

Durante el proyecto trataremos con valores inexistentes. Se realizan dos tipos de soluciones que son descartar dichos valores o asignarles un valor(0). En función de la necesidad de la problemática se optará por una o por otra.

4.2.3 Renombramiento de columnas

Al venir los datos en json será necesario poner correctamente los nombres de las columnas. Aquellos nombres que se puedan mantener se mantendrán.

4.2.4 Cambiar tipos de datos

Durante el proyecto será necesario transformar algunos datos para que estos nos sean de utilidad en las soluciones. Por lo tanto, habrá diferentes tipos de transformaciones a lo largo del proyecto.

4.3. Capacidad de crear conocimiento.

A continuación podemos ver la solución correspondiente a la actividad 1 y cómo ha sido reflejada:

-Regresión lineal basada en las estadísticas del propio partido:

ID	home_team_name	away_team_name	home_goals	away_goals	Prediction (home_goals)	Prediction (away_goals)
41...	Real Madrid	Real Betis	0	0.0	1.0	0.0
41...	Rayo Vallecano	Levante	2	4.0	2.0	3.0
41...	Valencia	Celta Vigo	2	0.0	1.0	1.0
41...	Elche	Getafe	3	1.0	3.0	1.0
41...	Alaves	Cadiz	0	1.0	1.0	2.0
41...	Granada CF	Espanyol	0	0.0	2.0	1.0
41...	Olasana	Mallorca	0	2.0	0.0	2.0
41...	Barcelona	Villarreal	0	2.0	1.0	1.0
41...	Real Sociedad	Atletico Madrid	1	2.0	1.0	2.0
41...	Sevilla	Athletic Club	1	0.0	1.0	1.0

ID	Prediction (home_goals)	ID	Prediction (away_goals)
R^2	0.2660550458715598	R^2	0.6153846153846153
mean absolute error	0.5999999999999999	mean absolute error	0.6
mean squared error	0.7999999999999999	mean squared error	0.6
root mean squared error	0.8944271909999159	root mean squared e...	0.7745966692414834
mean signed difference	0.3999999999999997	mean signed differe...	0.2
mean absolute percentage error	NaN	mean absolute perc...	NaN
adjusted R^2	0.2660550458715598	adjusted R^2	0.6153846153846153

-Regresión Random forest basada en las estadísticas del propio partido:

ID	home_team_name	away_team_name	home_goals	away_goals	Prediction (home_goals)	Prediction (away_goals)	Prediction (home_goals) (Prediction Variance)	Prediction (away_goals) (Prediction Variance)
4174_4174	Real Madrid	Real Betis	0	0.0	2.0	1.0	2.0	1.0
4178_4178	Rayo Vallecano	Levante	2	4.0	2.0	3.0	2.0	2.0
4177_4177	Valencia	Celta Vigo	2	0.0	2.0	1.0	1.0	1.0
4179_4179	Elche	Getafe	3	1.0	3.0	1.0	2.0	1.0
4170_4170	Alaves	Cadiz	0	1.0	1.0	2.0	1.0	1.0
4172_4172	Granada CF	Espanyol	0	0.0	2.0	2.0	2.0	1.0
4173_4173	Olasana	Mallorca	0	2.0	2.0	3.0	2.0	2.0
4171_4171	Barcelona	Villarreal	0	2.0	2.0	1.0	1.0	1.0
4175_4175	Real Sociedad	Atletico Madrid	1	2.0	1.0	2.0	1.0	1.0
4176_4176	Sevilla	Athletic Club	1	0.0	1.0	1.0	1.0	1.0

ID	Prediction (home_goals)	ID	Prediction (away_goals)
R^2	-0.5596330275229355	R^2	0.2948717948717948
mean absolute error	0.9	mean absolute error	0.8999999999999999
mean squared error	1.7	mean squared error	1.1
root mean squared error	1.3038404810405297	root mean squared error	1.0488088481701516
mean signed difference	0.9	mean signed difference	0.5
mean absolute percentage error	NaN	mean absolute percentage error	NaN
adjusted R^2	-0.5596330275229355	adjusted R^2	0.2948717948717948

-Regresión potenciado de gradiente basada en las estadísticas del propio partido:

ID	home_team_name	away_team_name	home_goals	away_goals	Prediction (home_goals)	Prediction (away_goals)
4174_4174	Real Madrid	Real Betis	0	0.0	1.0	0.0
4178_4178	Rayo Vallecano	Levante	2	4.0	2.0	3.0
4177_4177	Valencia	Celta Vigo	2	0.0	2.0	1.0
4179_4179	Elche	Getafe	3	1.0	3.0	1.0
4170_4170	Alaves	Cadiz	0	1.0	0.0	1.0
4172_4172	Granada CF	Espanyol	0	0.0	2.0	1.0
4173_4173	Olasuna	Mallorca	0	2.0	1.0	1.0
4171_4171	Barcelona	Villarreal	0	2.0	2.0	1.0
4175_4175	Real Sociedad	Atletico Madrid	1	2.0	1.0	2.0
4176_4176	Sevilla	Athletic Club	1	0.0	1.0	0.0

ID	Prediction (home_goals)	ID	Prediction (away_goals)
R^2	0.08256880733944971	R^2	0.6794871794871795
mean absolute error	0.6	mean absolute error	0.5
mean squared error	1.0	mean squared error	0.5
root mean squared error	1.0	root mean squared error	0.7071067811865476
mean signed difference	0.6	mean signed difference	-0.1000000000000002
mean absolute percentage error	NaN	mean absolute percentage error	NaN
adjusted R^2	0.0825688073394496	adjusted R^2	0.6794871794871795

-Regresión potenciado de gradiente basada en serie temporal:

ID	home_team_name	away_team_name	home_goals	away_goals	Prediction (home_goals) (rounded)	Prediction (away_goals) (rounded)
4174_4174#9	Real Madrid	Real Betis	0	0	1.0	1.0
4178_4178#8	Rayo Vallecano	Levante	2	4	2.0	1.0
4177_4177#7	Valencia	Celta Vigo	2	0	2.0	1.0
4179_4179#6	Elche	Getafe	3	1	2.0	2.0
4170_4170#5	Alaves	Cadiz	0	1	2.0	1.0
4172_4172#4	Granada CF	Espanyol	0	0	2.0	2.0
4173_4173#3	Osasuna	Mallorca	0	2	1.0	2.0
4171_4171#2	Barcelona	Villarreal	0	2	3.0	1.0
4175_4175#1	Real Sociedad	Atletico Madrid	1	2	1.0	1.0
4176_4176#0	Sevilla	Athletic Club	1	0	1.0	0.0

ID	Prediction (home_goals)	ID	Prediction (away_goals) (
R^2	-0.8348623853211006	R^2	-0.15384615384615397
mean absolute error	1.0	mean absolute error	1.0
mean squared error	2.0	mean squared error	1.8
root mean squared error	1.4142135623730951	root mean squared error	1.3416407864998738
mean signed difference	0.7999999999999999	mean signed difference	0.0
mean absolute percentage error	NaN	mean absolute percentage error	NaN
adjusted R^2	-0.8348623853211008	adjusted R^2	-0.15384615384615397

4.4. Capacidad de presentación.

-Reporte dinámico

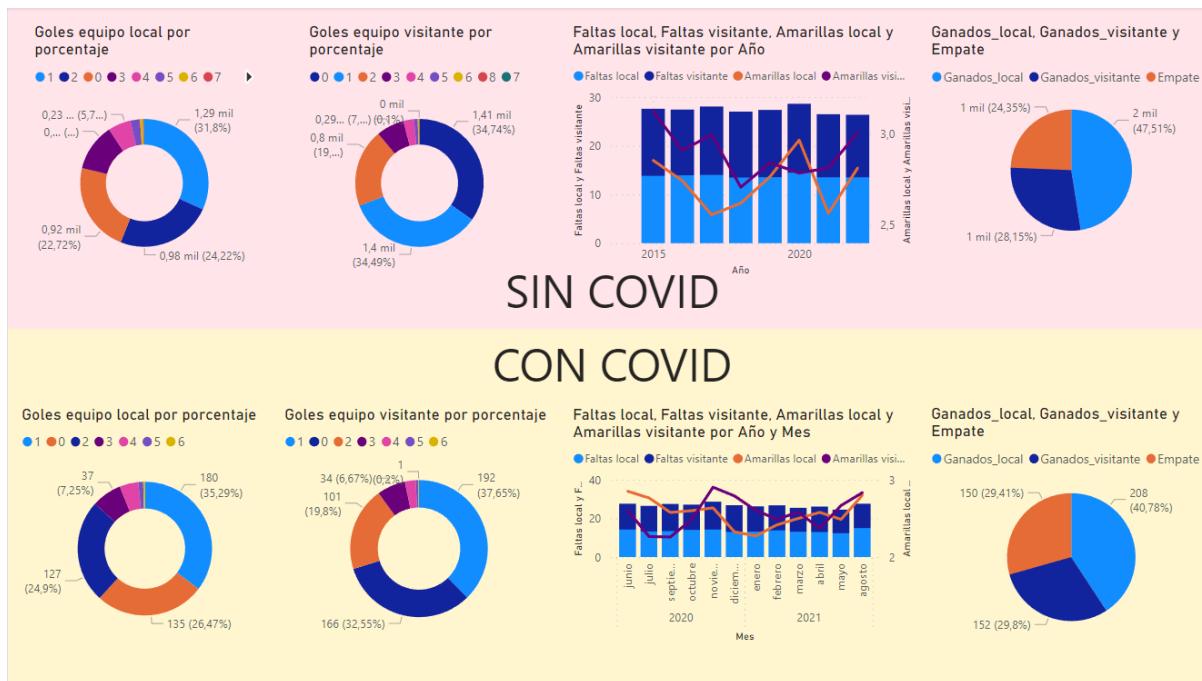
Nos permite ver de un rápido vistazo gráficas interesantes de un equipo como local o visitante de una temporada concreta.



Además, los colores nos ayudan a centrarnos en estadísticas positivas o negativas de manera visual.

-Analítica covid

Al separar el informe por la mitad, nos permite comparar las gráficas entre sí de manera muy sencilla. Para poder comparar las estadísticas en los períodos es importante utilizar porcentajes o promedios.



5. Implementación

5.1. Pasos para la instalación y configuración de los datos.

Para comenzar, debemos instalar el software de KNIME. Para ello, introducimos en el buscador “Descargar knime” y nos lleva a la página que se visualiza a continuación:



New to the KNIME family? Let us help you get started with a short series of introductory emails. These messages are possible and introduce you to resources that will maximize your success with the KNIME Analytics Platform.

First Name

Last Name

Seleccionamos directamente la opción de descargar y la versión que más nos convenga en función del equipo que se esté utilizando.

The KNIME Analytics Platform version is intended for end users and provides everything needed to immediately begin using KNIME as well as extend KNIME with extension packages developed by others.

Windows		
KNIME Analytics Platform for Windows (installer) <i>The installer adds an icon to the desktop and suggests suitable memory settings</i>	64 Bit (441.03 MB) 32 Bit (437.42 MB)	
KNIME Analytics Platform for Windows (self-extracting archive) <i>The self-extracting archive only creates a folder holding the KNIME installation</i>	64 Bit (444.58 MB) 32 Bit (441.15 MB)	
KNIME Analytics Platform for Windows (zip archive)	64 Bit (529.54 MB) 32 Bit (525.59 MB)	

Para finalizar, debemos pulsar en la opción de descargar.

 Open for Innovation
KNIME

Hub Blog Forum Events Careers Contact [Download](#) Q

SOFTWARE / SOLUTIONS / LEARNING / PARTNERS / COMMUNITY / ABOUT

Home > Downloads > Download KNIME Analytics Platform for Windows (installer) 64 Bit

You decided to download the installer for KNIME Analytics Platform for Windows (installer) 64 Bit (441.03 MB).

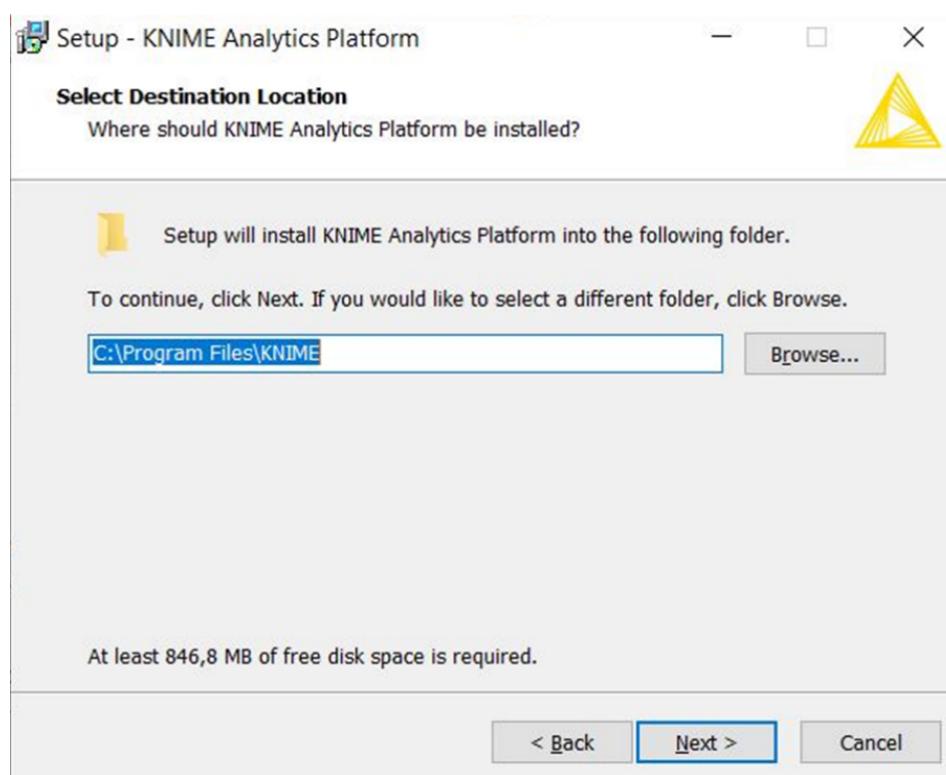
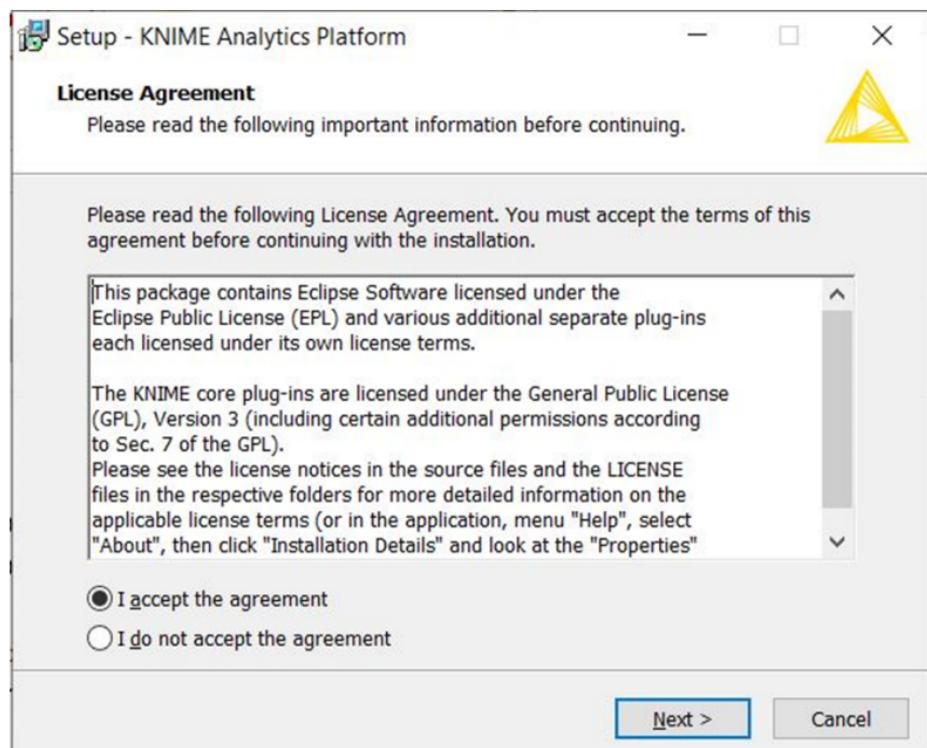
The installer adds an icon to the desktop and suggests suitable memory settings

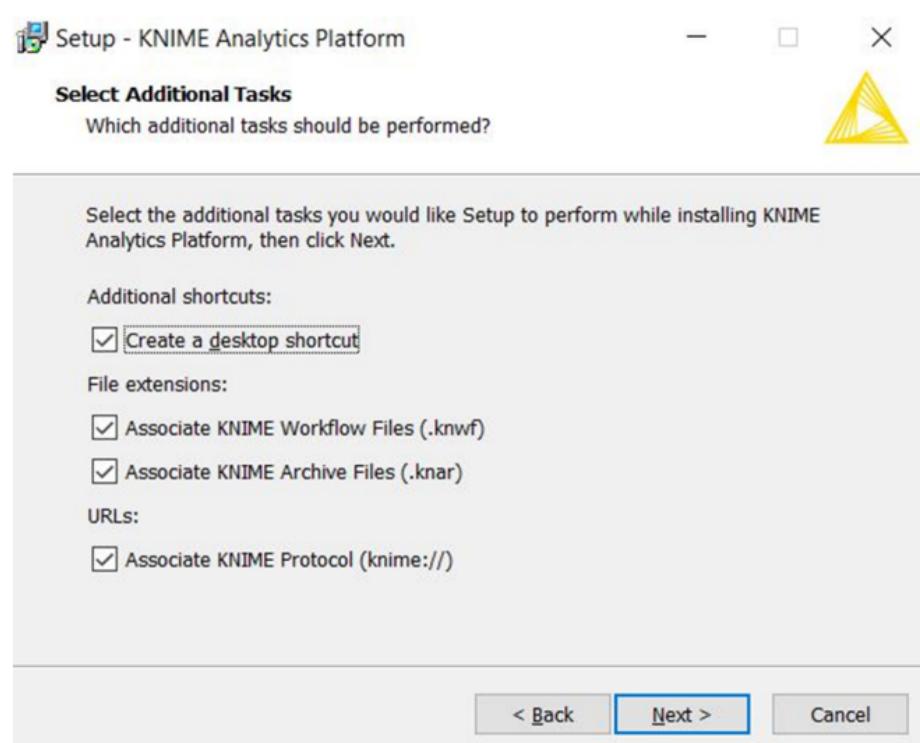
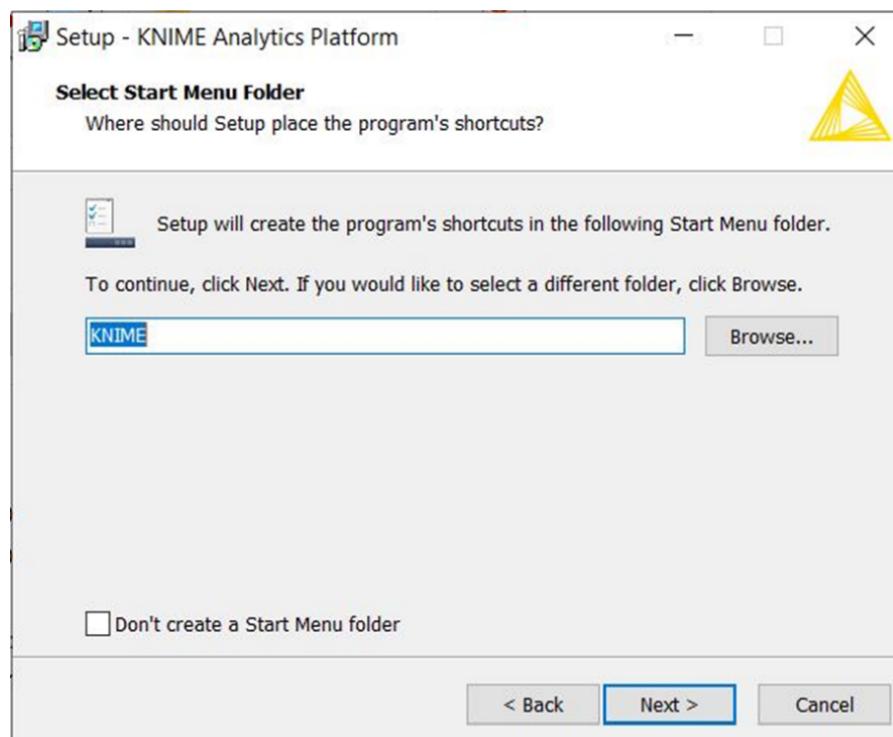
If you want to run the KNIME installer or self-extracting archive for Windows you might experience some difficulty because of the Microsoft SmartScreen filter which was introduced with Internet Explorer 9 and Windows 8. Find out how to solve the problem.

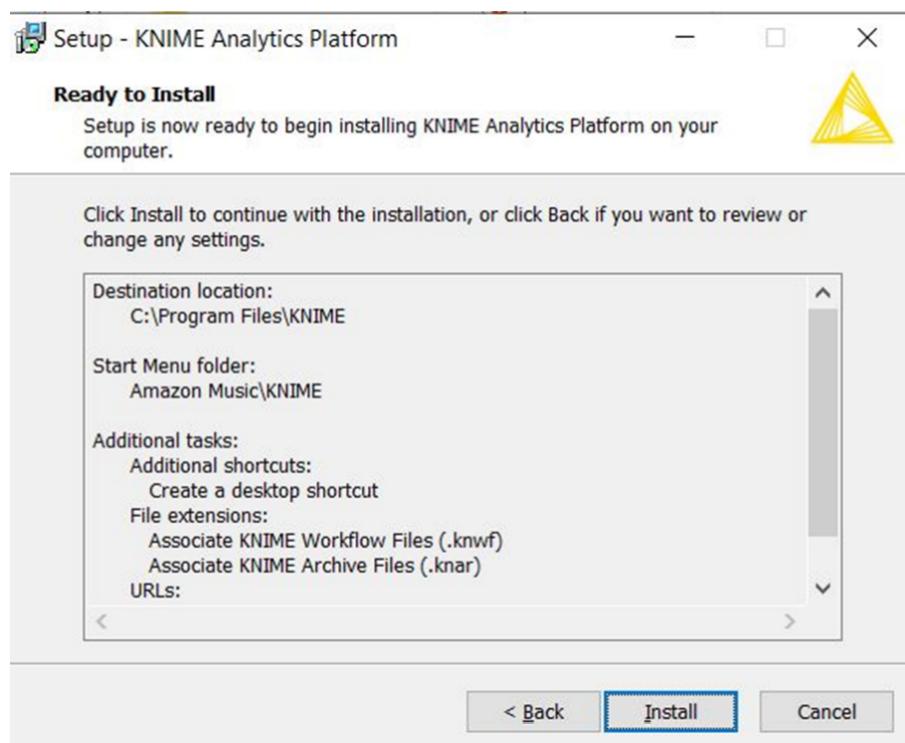
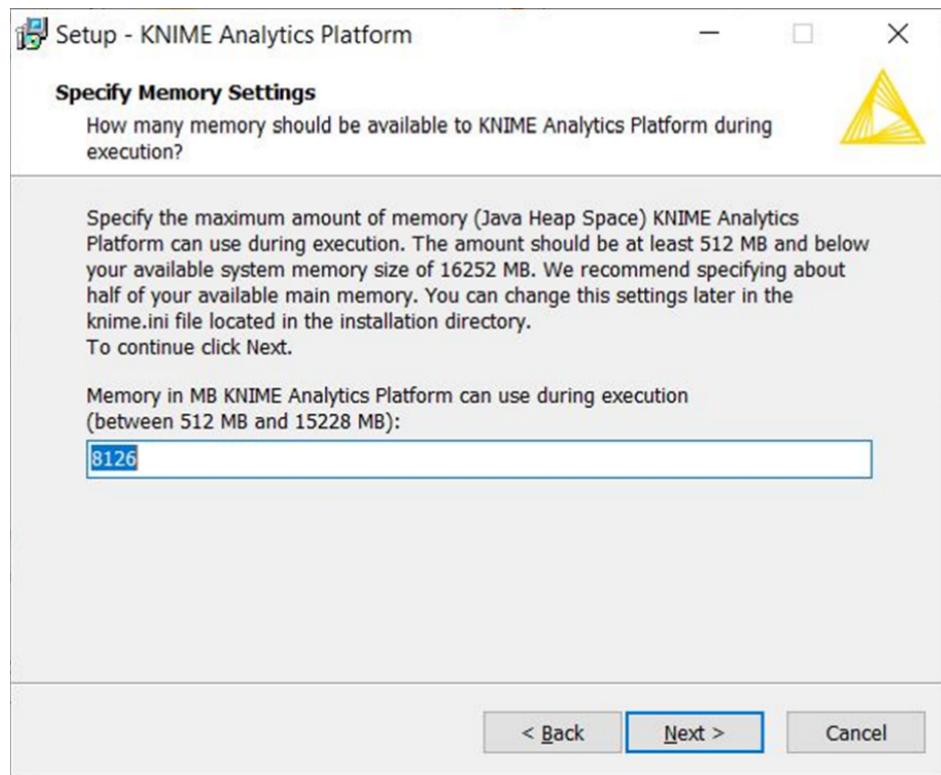
I have read and accept the [privacy policy](#) and the [terms and conditions](#) *

[④ Download](#)

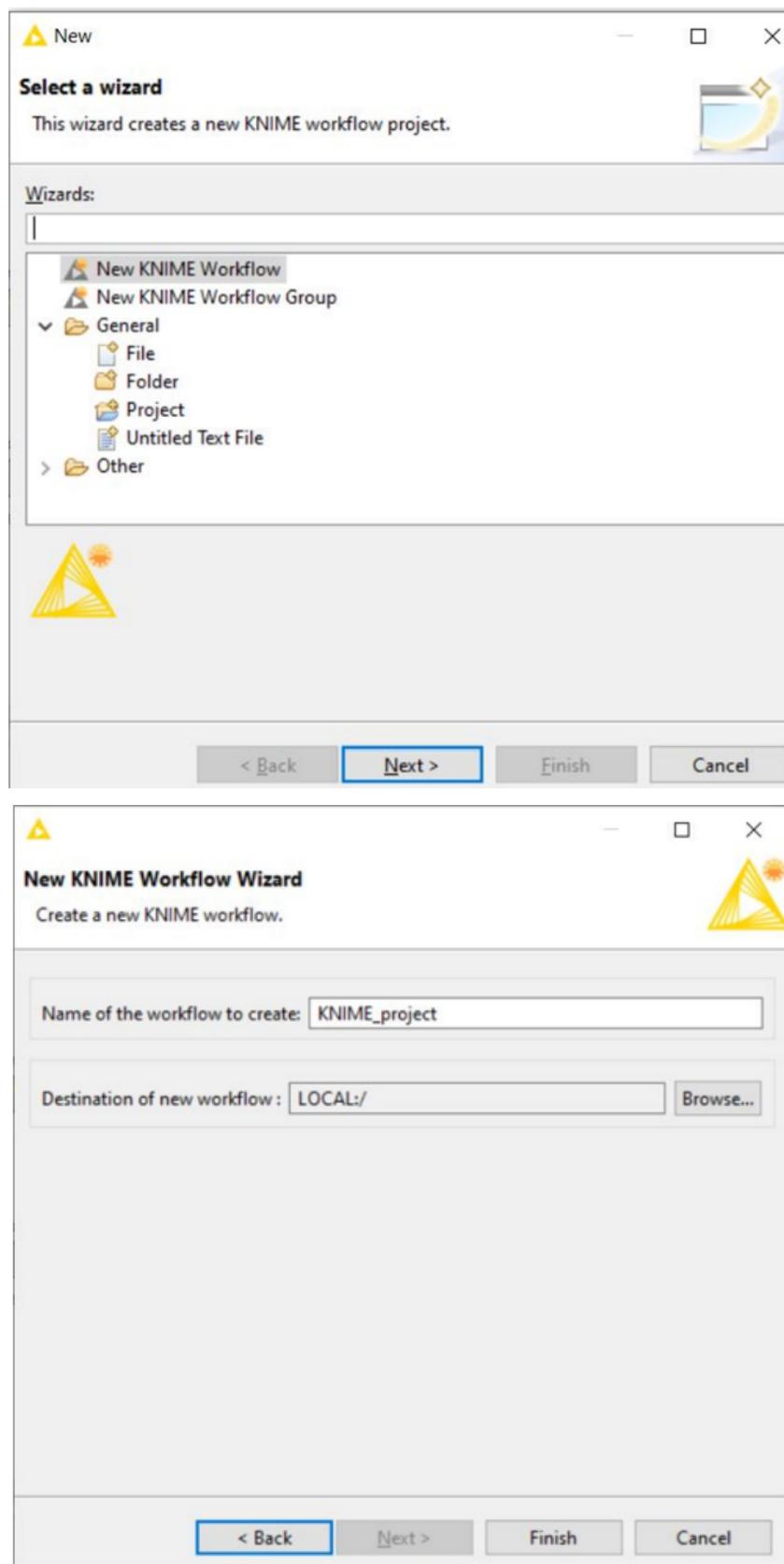
Para la instalación, debemos dejar las opciones por defecto a no ser que se quiera modificar, o bien la ruta de la instalación, o bien la cantidad de memoria RAM que se quiera aportar.







Después, abrimos el programa y seleccionamos la opción File y New.



5.2. Software necesario para la extracción de los datos

Python 3.8+, con dependencias pandas y pip. Sistema operativo Windows/Linux.

5.3. Software necesario para el análisis de los datos.

Knime. Sistema operativo Windows/Linux.

PowerBI. Sistema operativo Windows.

5.4. Software necesario para crear los reportes.

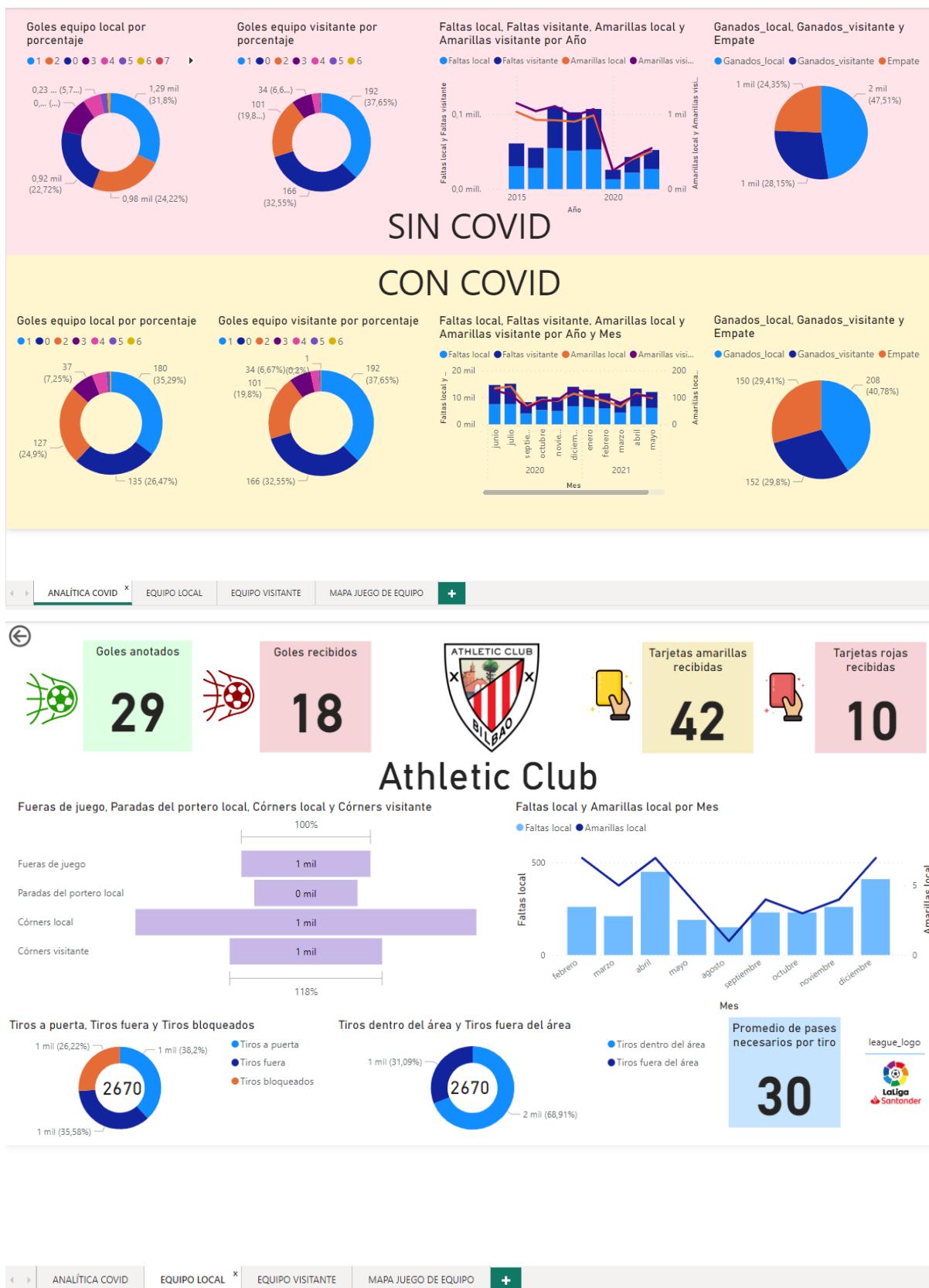
Los reportes se crean con Power BI.

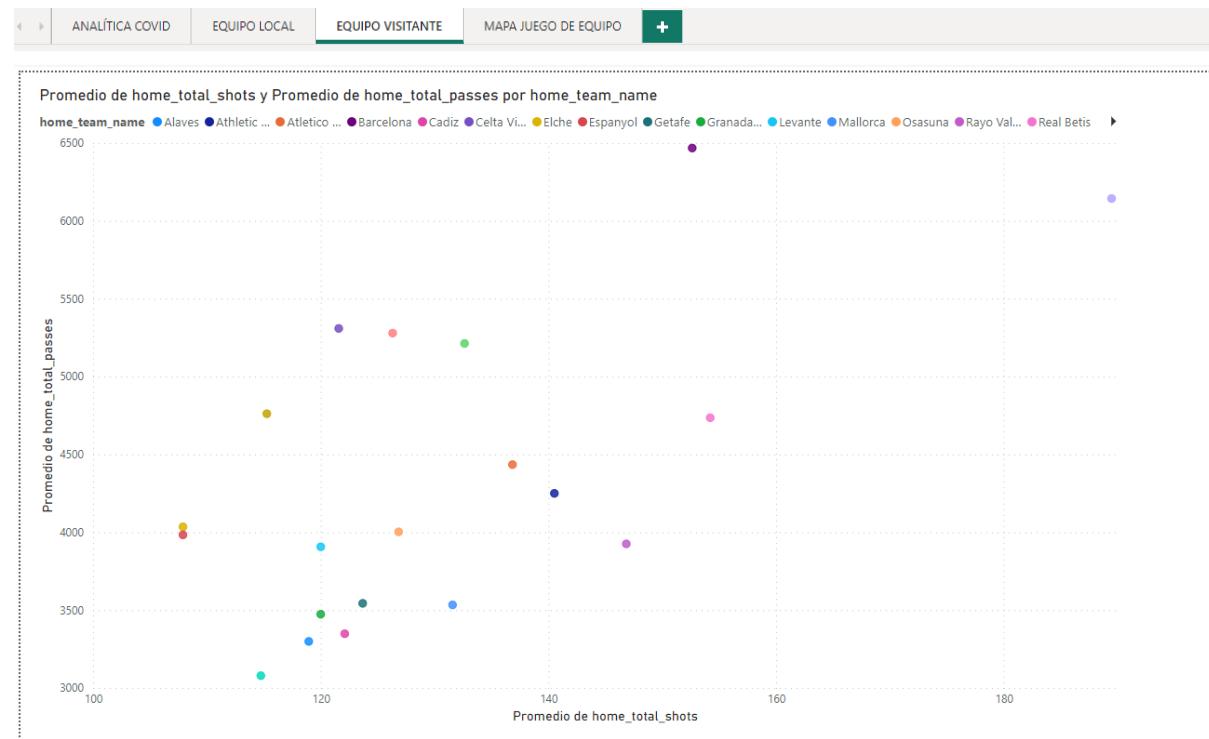
5.5. Otro software necesario

El seguimiento y la realización del trabajo se ha realizado a través de git.

6. Despliegue

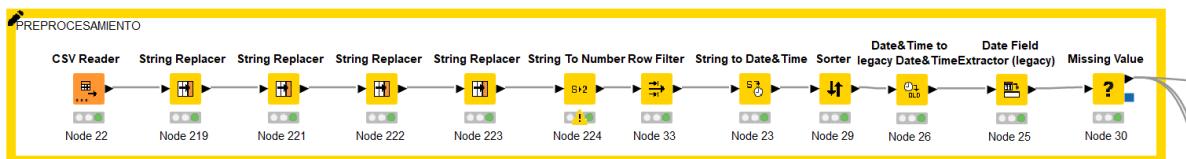
Como resultado del despliegue del cuadro de mandos de Power BI, encontramos lo que mostramos a continuación:

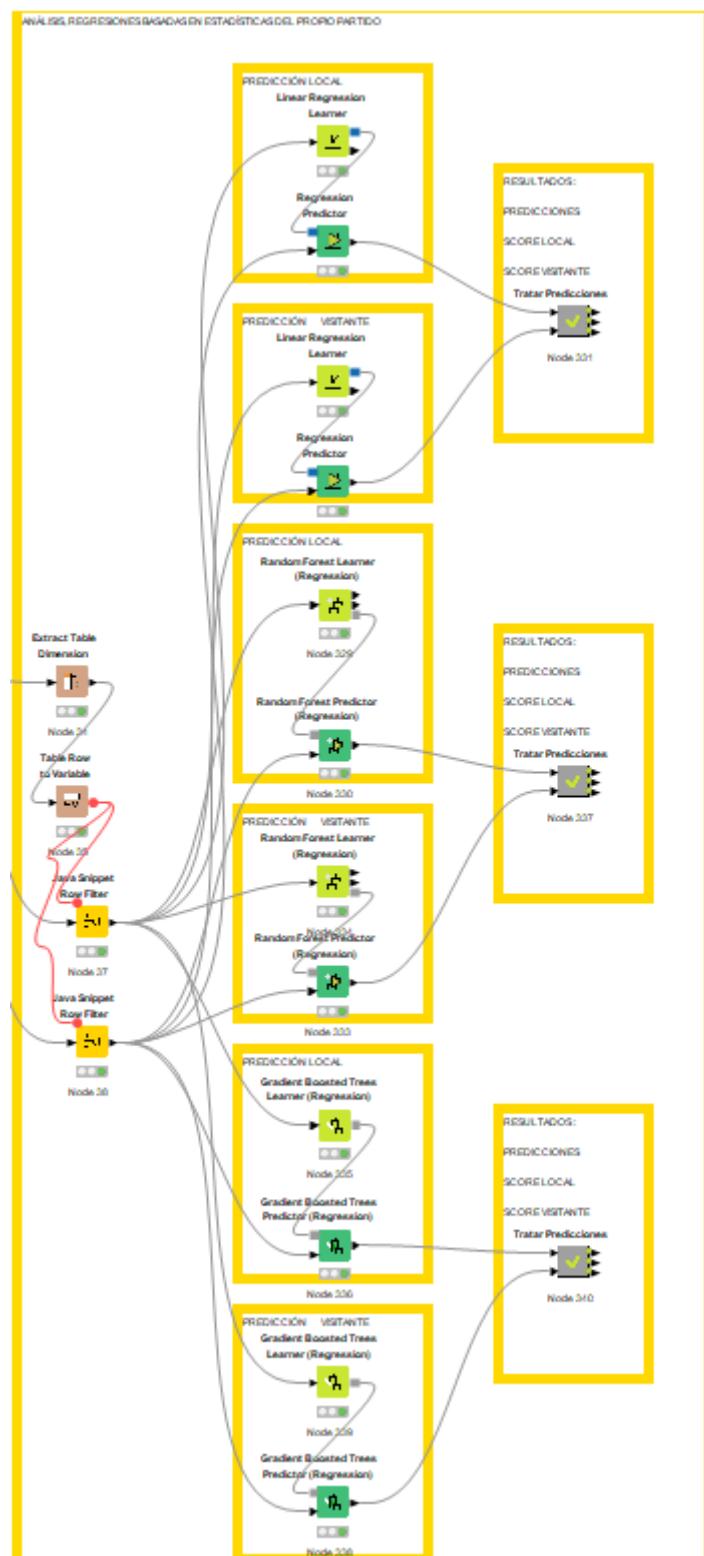




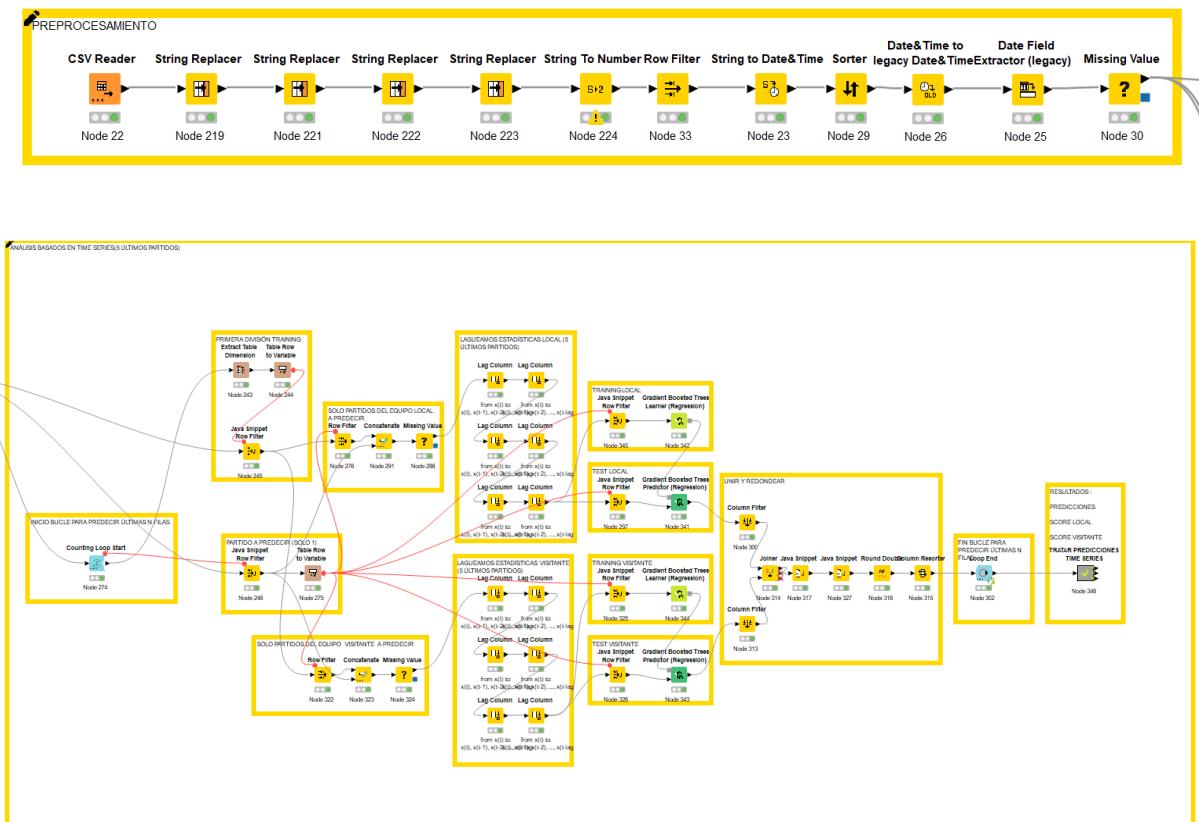
A su vez, como resultado del despliegue del KNIME, encontramos lo que mostramos a continuación:

Primera Rama





Segunda Rama



7. Conclusiones

Extracción de datos

Tras muchos intentos con la api gratuita y distintos baneos a cuentas e ips, decidimos pagar la suscripción durante 1 mes para poder extraer todos los datos, ya que teníamos el resto de flujos montado y era lo único que quedaba, extraer todos los datos.

Creemos que esta ha podido ser la parte más desesperante por los baneos continuos de la api.

Análisis en KNIME

En cuanto a la regresión en estadísticas del propio partido, creemos que es útil aplicarlo para poder aprender de por qué un partido ha quedado como ha quedado o para saber según las estadísticas los goles que debería haber marcado cada equipo.

En cuanto a la regresión, creemos en este caso que la más acertada fue la del gradiente potenciado debido a los siguientes factores:

ID	home_team_name	away_team_name	home_goals	away_goals	Prediction (home_goals)	Prediction (away_goals)
4174_4174	Real Madrid	Real Betis	0	0.0	1.0	0.0
4178_4178	Rayo Vallecano	Levante	2	4.0	2.0	3.0
4177_4177	Valencia	Celta Vigo	2	0.0	2.0	1.0
4179_4179	Elche	Getafe	3	1.0	3.0	1.0
4170_4170	Alaves	Cadiz	0	1.0	0.0	1.0
4172_4172	Granada CF	Espanyol	0	0.0	2.0	1.0
4173_4173	Olasuna	Mallorca	0	2.0	1.0	1.0
4171_4171	Barcelona	Villarreal	0	2.0	2.0	1.0
4175_4175	Real Sociedad	Atletico Madrid	1	2.0	1.0	2.0
4176_4176	Sevilla	Athletic Club	1	0.0	1.0	0.0

ID	Prediction (home_goals)	ID	Prediction (away_goals)
R^2	0.08256880733944971	R^2	0.6794871794871795
mean absolute error	0.6	mean absolute error	0.5
mean squared error	1.0	mean squared error	0.5
root mean squared error	1.0	root mean squared error	0.7071067811865476
mean signed difference	0.6	mean signed difference	-0.1000000000000002
mean absolute percentage error	NaN	mean absolute percentage error	NaN
adjusted R^2	0.0825688073394496	adjusted R^2	0.6794871794871795

Aunque el R2 sale muy cercano a 0 en los goles como local (esto quiere decir que no se establece una relación), puede estar condicionado por tener la mayoría 0 goles de local en este caso concreto de test. Si nos centramos en el MAE(mean absolute error), obtenemos que nos equivocamos de media 0.5 goles con respecto al resultado, lo que creemos que es un buen resultado para nuestro modelo. Con los otros algoritmos de aprendizaje obtenemos peores medidas en estas métricas.

Dado estos resultados, decidimos utilizar de nuevo este mismo algoritmo para la serie temporal.

ID	home_team_name	away_team_name	home_goals	away_goals	Prediction (home_goals) (rounded)	Prediction (away_goals) (rounded)
4174_4174#9	Real Madrid	Real Betis	0	0	1.0	1.0
4178_4178#8	Rayo Vallecano	Levante	2	4	2.0	1.0
4177_4177#7	Valencia	Celta Vigo	2	0	2.0	1.0
4179_4179#6	Elche	Getafe	3	1	2.0	2.0
4170_4170#5	Alaves	Cadiz	0	1	2.0	1.0
4172_4172#4	Granada CF	Espanyol	0	0	2.0	2.0
4173_4173#3	Olasuna	Mallorca	0	2	1.0	2.0
4171_4171#2	Barcelona	Villarreal	0	2	3.0	1.0
4175_4175#1	Real Sociedad	Atletico Madrid	1	2	1.0	1.0
4176_4176#0	Sevilla	Athletic Club	1	0	1.0	0.0

ID	Prediction (home_goals)	ID	Prediction (away_goals) (
R^2	-0.8348623853211006	R^2	-0.15384615384615397
mean absolute error	1.0	mean absolute error	1.0
mean squared error	2.0	mean squared error	1.8
root mean squared error	1.4142135623730951	root mean squared error	1.3416407864998738
mean signed difference	0.7999999999999999	mean signed difference	0.0
mean absolute percentage error	NaN	mean absolute percentage error	NaN
adjusted R^2	-0.8348623853211008	adjusted R^2	-0.15384615384615397

En este caso, obtenemos lo contrario, un buen R2 en el local, pero malo en el visitante. De todas formas, indagando en internet hemos visto que para series temporales no es recomendable usar R2 como métrica. En este caso, si nos volvemos a fijar en el MAE, vemos que nos equivocamos 1 gol de media en los goles tanto de local o visitante. Quizás este resultado sea mejorable, pero adivinar un resultado en un partido de fútbol es algo muy complicado, y estamos teniendo en cuenta pocos factores.

En líneas generales, hemos aprendido mucho y creemos que tratamos con datos muy complejos y estamos contentos con el resultado de las predicciones.

PowerBI:

Reporte dinámico:

En cuanto a los reportes dinámicos de powerbi, nos hubiera gustado hacer un solo cuadro de mandos donde se pudiera elegir si querías ver las estadísticas de local o visitante o ambas conjuntas, de un equipo en una temporada. Esto implicaba muchísimas modificaciones y consultas DAX, y no hemos tenido el tiempo suficiente para desarrollarlo.

Otro de los problemas en los que nos hemos visto envueltos y encontramos más difícil es que no sabemos cuándo parar de llenar un informe o qué información puede llegar a ser más útil para un usuario.

Finalmente creo que hemos conseguido el objetivo, aunque sea teniendo local y visitante por separado. Además hemos aprendido a usar consultas DAX para sacar por ejemplo el promedio de pases por tiro.

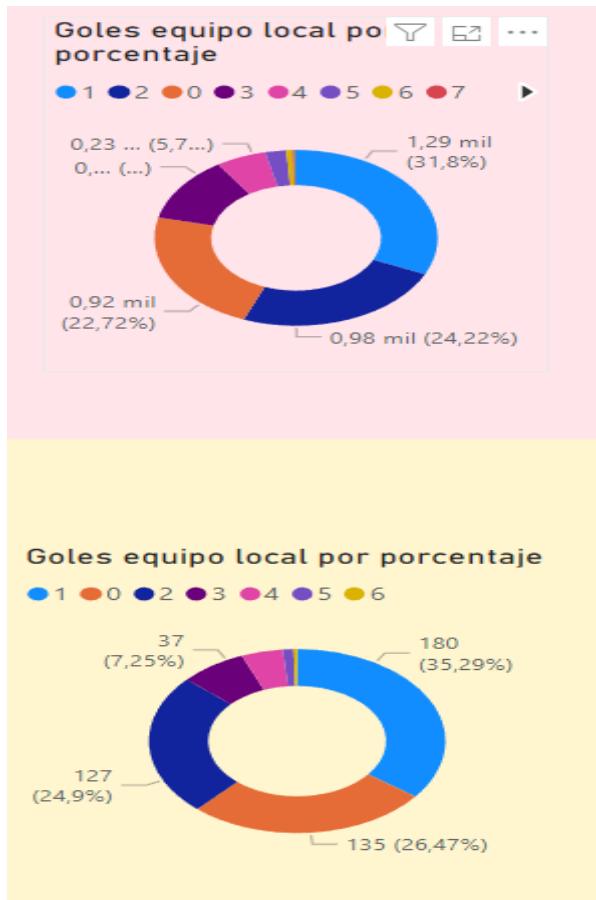
Analítica covid:

Al principio del trabajo, nos hacíamos una pregunta. ¿Influyó el que no hubiera público en la tendencia de los resultados de local y visitante?

Aunque no podemos afirmar que fuera solo por el público, tras realizar el análisis de las gráficas de este cuadro de mandos sí que vemos un cambio de tendencia en los resultados de local y visitante.

Hay que tener en cuenta varios factores además del público: más cantidad de partidos en menos tiempo, más lesiones de jugadores, entran los 5 cambios en vez de 3...

Pero si nos fijamos en las gráficas (parte superior sin covid, inferior con covid)...



Aquí observamos que el equipo local refuerza los porcentajes de partidos con 0, 1 y 2 goles, en la época covid. Es decir, vemos muchas menos goleadas del equipo local.



Lo mismo ocurre con los equipos visitantes, se refuerzan los porcentajes de partidos con 1 y 0 goles.

En cuanto al gráfico de faltas y tarjetas de local y visitante, en este caso no conseguimos observar muchas diferencias, por lo que no parece que el público sea un factor influyente para que un equipo haga más faltas o le saquen más tarjetas.

Por último, el gráfico más impactante y determinante es el siguiente:



Aquí podemos ver cómo se ve reducido en un 7% el porcentaje de victorias de los equipos como local durante el covid, subiendo de forma notable el porcentaje de empate y de una forma más leve la victoria visitante.

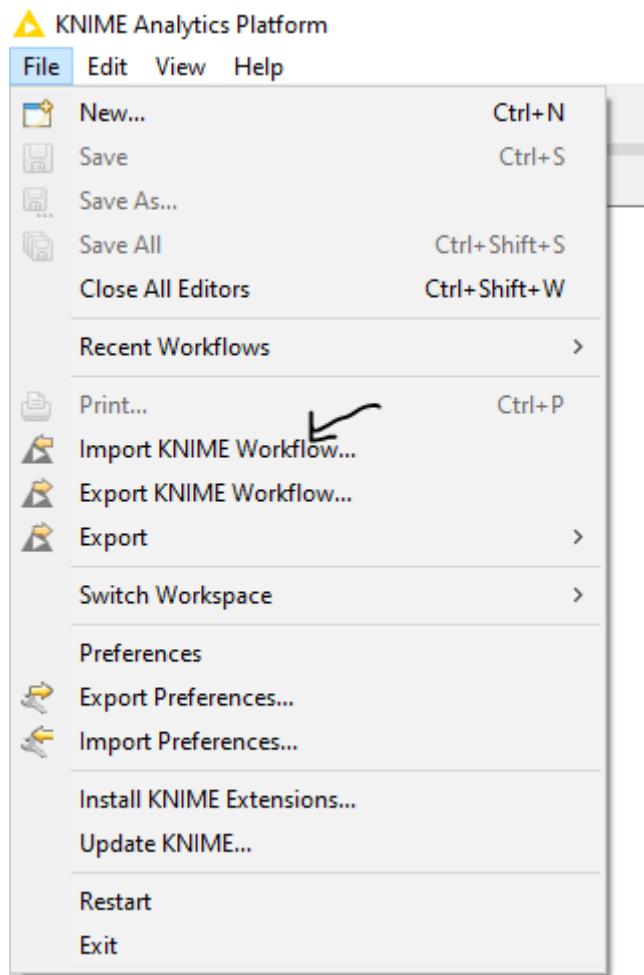
Con este gráfico podemos afirmar que el público tiene un papel diferencial para hacer que el equipo local gane su partido, y que los partidos a puerta cerrada benefician mucho más al equipo visitante.

8. //TODO ANTONIO

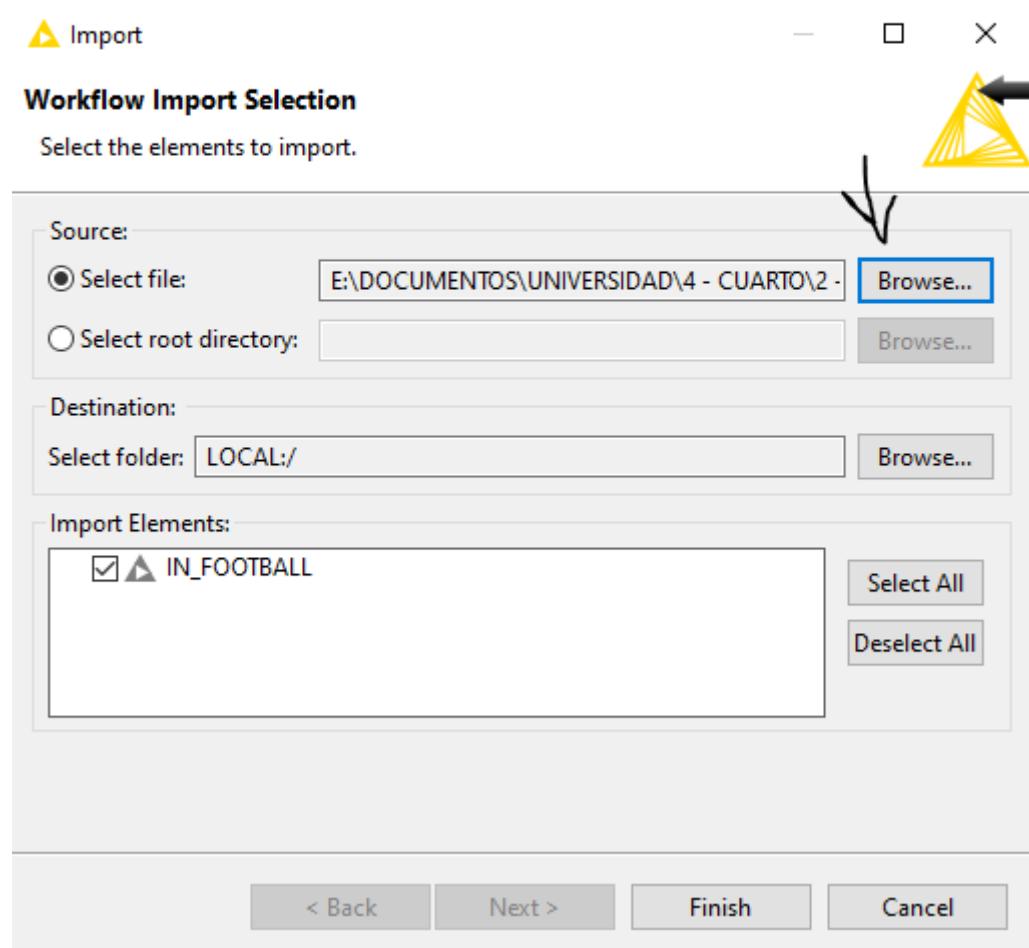
9. Manual de instalación

KNIME

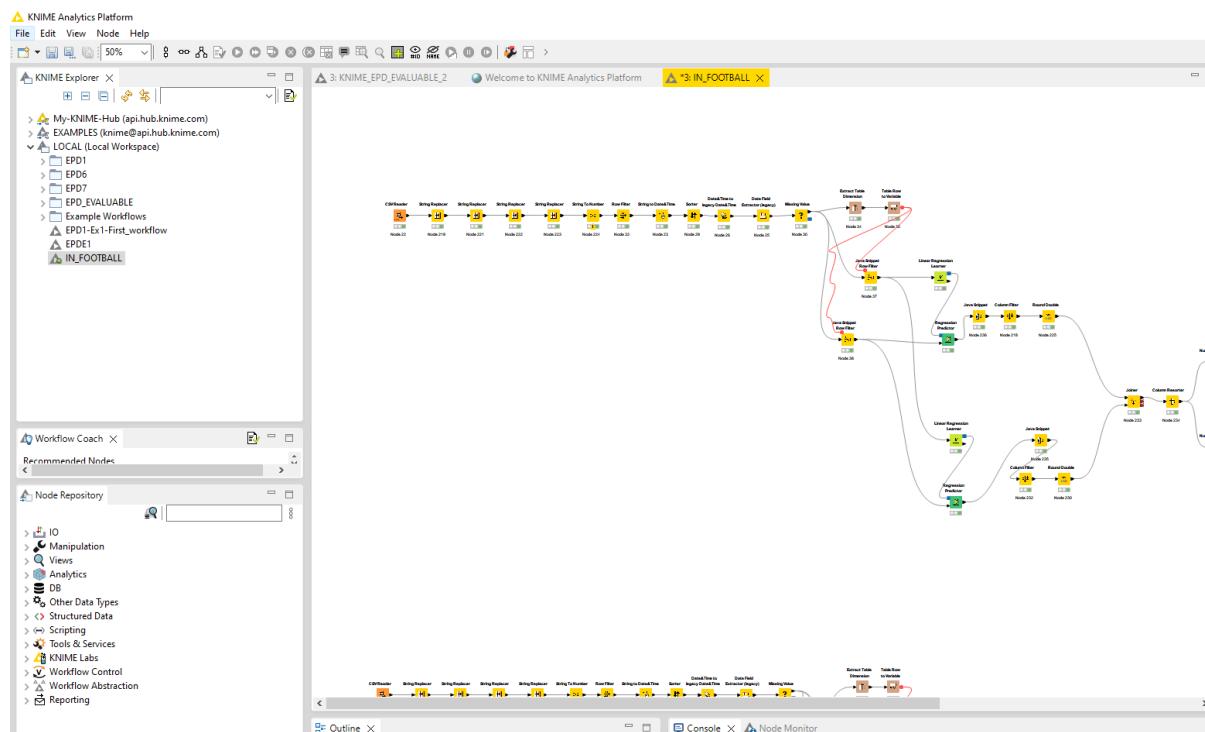
Comenzar por lo explicado en el punto (5.1). Una vez tengamos instalado ya el Knime procederemos a importar nuestro proyecto.



Seleccionamos Nuestro proyecto en browse.

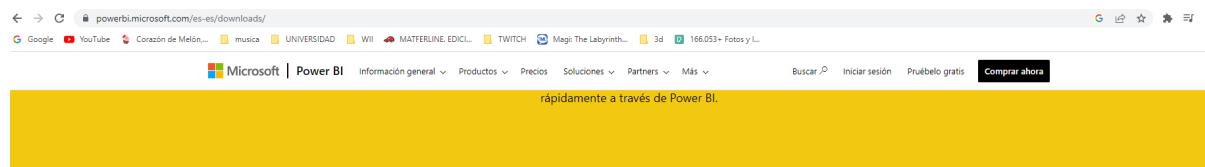


Una vez seleccionamos nuestro archivo, ya tendríamos nuestro proyecto importado.



POWER BI

Instalamos el programa Power BI.



Microsoft Power BI Desktop Con Power BI Desktop, puede explorar visualmente los datos con un lienzo de arrastrar y colocar de forma libre, una amplia gama de visualizaciones modernas de datos y una experiencia de creación de informes fácil de usar. Descargar > Opciones avanzadas de descarga >	Microsoft Power BI Mobile Acceda a los datos en cualquier lugar y en cualquier momento. Estas aplicaciones nativas proporcionan acceso directo, interactivo y móvil a la información empresarial importante. Get it from Microsoft Download on the App Store GET IT ON Google Play	Puerta de enlace de datos local de Microsoft Mantenga sus paneles y sus informes actualizados mediante la conexión a los orígenes de datos locales, sin necesidad de desplazar los datos. Descargar modo estándar > Descargar modo personal >
Informes locales con Power BI Report Server Implemente y distribuya informes interactivos de Power BI, así como informes paginados tradicionales, dentro de los límites del firewall de la organización. Descargar >	Generador de informes de Microsoft Power BI Cree informes paginados con pixeles perfectos para su impresión o distribución electrónica gracias a una experiencia conocida en la que confían miles de autores de informes. Descargar >	

Capturas de pantalla

Power BI Desktop
Microsoft Corporation

4,8 ★ 839
Promedio Clasificación
Empresa

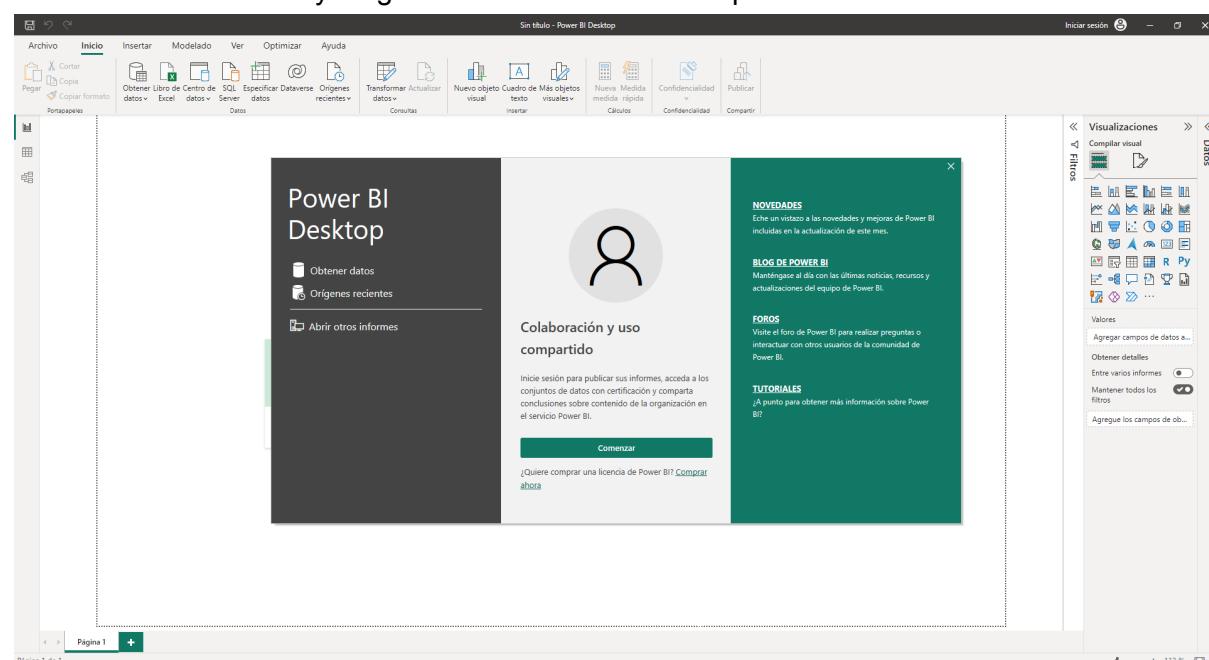
Obtener

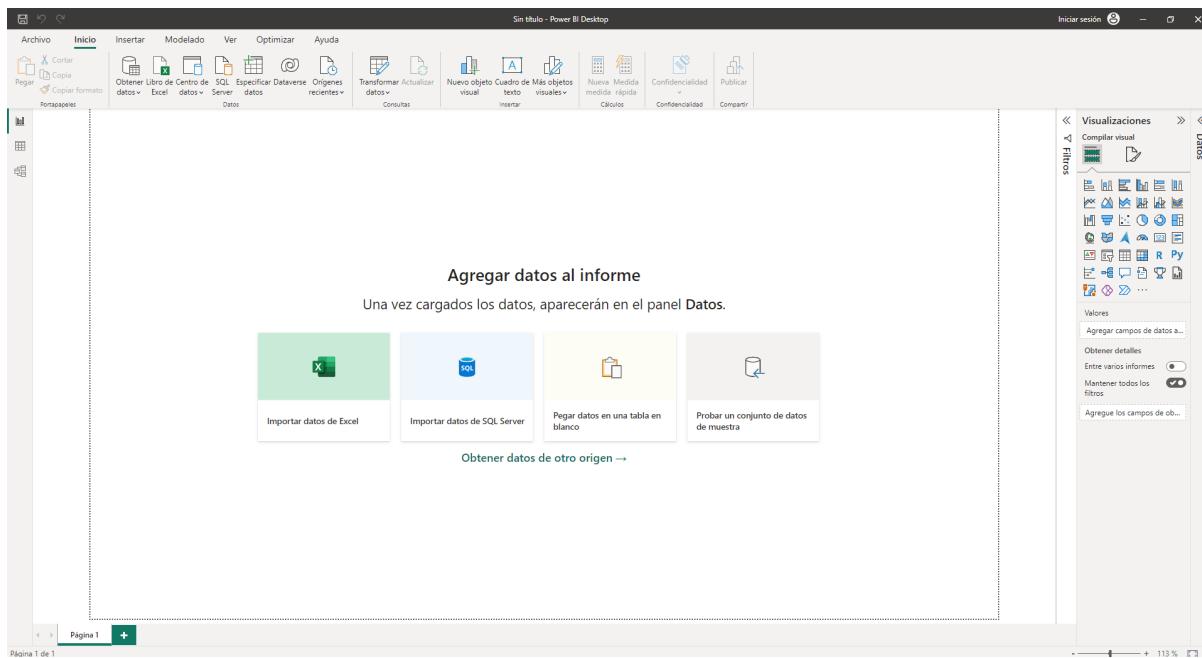
Descripción

Power BI Desktop pone el análisis visual a su alcance. Con esta herramienta de creación eficaz, puede crear visualizaciones de datos interactivos.

Conecte, combine, modele y visualice sus datos. Coloque los elementos visuales exactamente donde los quiere, analice y explore sus datos y nublínate contenido al servicio web de Power BI para compartirlo con su equipo.

Si clicamos en obtener y luego en abrir nos saldrá esta pantalla de a continuación.





Para exportar nuestro proyecto tiene qué pinchar en:

The first screenshot shows the "Abrir informe" (Open report) screen with a sidebar containing options: Nuevo, Abrir informe, Guardar, Guardar como, Obtener datos, Importar, Exportar, Publicar, Opciones y configuración, and Comenzar. The "Abrir informe" option is highlighted. A callout arrow points to the "Examinar informes" (Browse reports) button at the bottom of the sidebar. The second screenshot shows the "informe" (Report) screen with a message: "No ha abierto ningún informe recientemente. Seleccione Examinar para abrir un informe entre sus archivos." It includes two buttons: "Pegar datos en una tabla en blanco" (Paste data into a blank table) and "Probar un conjunto de datos de muestra" (Test a sample dataset). A link "Obtener datos de otro origen →" (Get data from another source →) is also present.

