



ugr

Universidad
de Granada

TRABAJO DE FIN DE GRADO
DOBLE GRADO EN INGENIERÍA INFORMÁTICA Y
MATEMÁTICAS

Análisis de redes causales en deportes de equipo

Autora

Elena Merelo Molina

Directores

Juan Julián Merelo Guervós, Úrsula Torres Parejo

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE
TELECOMUNICACIÓN
FACULTAD DE CIENCIAS

Granada, septiembre de 2022

Análisis de redes causales en deportes de equipo

Elena Merelo Molina

Palabras clave: *software libre, redes bayesianas, redes causales, entropía, entropía de Tsallis, inferencia estadística, teorema de Bayes*

Resumen

En la parte informática, se creará una biblioteca para poder analizar de forma causal redes de pases de un equipo tanto a lo largo de un partido como a lo largo de una competición, y poder visualizar tanto el desempeño en el campo como su evolución, vinculándolo al desempeño de jugadoras específicas, pudiendo calificarlas de esta forma más allá de las puras medidas reticulares.

Los deportes de equipo, especialmente el fútbol, han sido analizados repetidamente desde el punto de vista de las redes complejas, buscando correlación entre meso y macroestructuras reticulares y el rendimiento del equipo. Sin embargo, no se ha hecho ningún análisis con redes causales, buscando relaciones causa-efecto en las redes de pases y, una vez más, su influencia en los resultados que se obtienen a lo largo de una temporada o su capacidad para encontrar descriptores macro del equipo a lo largo de una competición determinada, dependiendo o no del rival que haya enfrente.

Causal networks analysis in team sports

Elena Merelo Molina

Keywords: *open source, bayesian networks, causal networks, entropy, Tsallis entropy, inference, statistics, Bayes theorem*

Abstract

A library will be created to be able to causally analyze passes networks of a team both throughout a match and throughout a competition, and to be able to visualize both the performance on the field and its evolution, linking it to the performance of specific players, thus being able to qualify them in this way beyond the pure reticular measures.

Team sports, especially football, have been repeatedly analyzed from the point of view of of complex networks, looking for correlations between reticular meso and macrostructures and the performance of the team. However, no analysis has been done with causal networks, looking for cause-effect relationships in passing networks and, once again, their influence on the results obtained throughout a season or its ability to find team macro descriptors throughout a given competition, depending or not on the rival.

D. **Juan Julián Merelo Guervós**, Profesor del departamento de ATC, y D. **Úrsula Torres Parejo**, Profesora del Departamento de Estadística e Investigación Operativa

Informe:

Que el presente trabajo, titulado *Análisis de redes causales en deportes de equipo*, ha sido realizado bajo nuestra supervisión por **Elena Merelo Molina**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a 31 de septiembre de 2022.

El/la director(a)/es:

(Juan Julián Merelo Guervós, Úrsula Torres Parejo)

Agradecimientos

A mis padres, por creer en mí cuando yo no lo hago, animarme a continuar las innumerables veces que tiro la toalla y ser tan fundamentales en mi día a día, darle un toque especial y hacerlo mucho mejor todo. Por estar siempre ahí, haberme dado y aportado tantísimo en mi vida. A mis hermanas, por apoyarme y aceptarme, y hacer los días más entretenidos. A la tienda Alehop, porque sin el ventilador que les compramos habría sido mucho más pesado trabajar durante esta horrible ola de calor continua que llamamos verano en Granada. Al Mercadona, por vender chocolate y queso tan rico. A panadería Geni (kiosko de la plaza Mariana Pineda), por hacer que año tras año me levante con ilusión al saber que luego me tomaré una tostada del mejor pan de Granada, y porque su amabilidad, simpatía y consideración hacen que esperar las colas que suele tener sea más leve. A las personas que han aparecido aquí y allí, o me han acompañado y ayudado desde el principio en esta carrera de fondo, iluminando el camino por este túnel larguísimo que es el doble grado, y que más veces de las que no se sentía más cueva o pozo que otra cosa.

Índice general

1. Introducción	15
1.1. Motivación	15
1.2. Objetivos del trabajo	16
2. Descripción del problema	17
2.1. Metodología - Desarrollo ágil	17
2.2. Herramientas usadas	19
2.2.1. Plataforma de desarrollo	19
2.2.2. IDE	20
2.2.3. Lenguaje de programación	21
2.2.4. Sistema de composición de documentos	22
2.3. Clientes	23
2.4. Historias de usuario	23
2.5. Hitos o Productos Mínimamente Viables	24
2.5.1. <i>Milestone</i> 1 - Infraestructura	24
2.5.2. <i>Milestone</i> 2 - Planteamiento	25
2.5.3. <i>Milestone</i> 3 - Parte matemática	25
2.5.4. <i>Milestone</i> 4 - Implementación	26
2.6. Definiciones y terminología	26
2.6.1. Redes bayesianas	26
2.6.2. Redes bayesianas híbridas	27
2.6.3. Redes complejas	28
2.6.4. Sistemas de calificación	29
2.6.5. Entropía	30
2.6.6. Entropía de Tsallis	31
3. Estado del arte	33
4. Planificación	37
4.1. Temporización	37

4.2. Seguimiento del desarrollo	37
5. Desarrollo teórico	39
5.1. Teoría de la probabilidad	39
5.1.1. Conceptos básicos	39
5.2. Redes bayesianas	42
5.2.1. Introducción	42
5.2.2. Factorización de redes bayesianas	44
5.2.3. Criterios gráficos de independencia	45
5.2.4. Algoritmos de inferencia para redes bayesianas	45
5.2.5. Algoritmos de aprendizaje para redes bayesianas	46
5.3. Entropía conjunta	47
6. Implementación y resultados	49
6.1. Resultados	49
6.2. Elección de fuentes de datos	51
6.3. Paquetes de R	51
6.3.1. Igraph	51
6.3.2. Devtools	51
6.3.3. Tidyverse	51
6.4. Diseño experimental	52
6.5. Costes	53
7. Conclusiones y trabajos futuros	63

Índice de figuras

2.1. Cómo se ve nuestra terminal	20
2.2. Ejemplo de grafo dirigido acíclico (DAG). Generado mediante un script en R	28
6.1. Red de pases total de Inglaterra simplificado en la EURO 2022	53
6.2. Entropía por jugadoras de Inglaterra en la EURO 2022	55
6.3. Entropía por jugadoras de Inglaterra en la EURO 2022	56
6.4. Red de pases total simplificado de Noruega en la EURO 2022	57
6.5. Entropía por jugadoras de Noruega en la EURO 2022	58
6.6. Red de pases total de Noruega en la EURO 2022	59
6.7. Entropía de Inglaterra	60
6.8. Entropía de Noruega	61

Índice de tablas

6.1. Costes el proyecto en el escenario “ingeniera junior”	54
6.2. Costes el proyecto en el escenario “analista de datos junior”	54

Capítulo 1

Introducción

En este capítulo describiremos la motivación del proyecto y los objetivos que se quieren alcanzar durante el desarrollo.

1.1. Motivación

El fútbol es un deporte de equipo *complicado*. Suponiendo que jueguen veintidós personas, como parte del equipo técnico habrá que decidir quiénes jugarán, dónde se posicionará cada uno, cuándo y qué cambios se harán, si se aboga más por una táctica defensiva u ofensiva, así como en qué hacer incidencia durante los entrenamientos, entre otros. Son bastantes las variables que entran, literalmente, en juego.

Surgen, pues, de manera natural preguntas en torno a la práctica de este deporte alrededor del cual gira todo un mundo, y la importancia de las decisiones que se van tomando tanto dentro como fuera del campo: Si a un equipo le va mal en un campeonato, se echa al entrenador, pero ¿de quién es realmente "la culpa" [36]? ¿Qué cambios se pueden hacer? ¿Qué táctica es mejor? ¿Cómo se influyen los jugadores entre ellos, o cuando juegan con otro equipo? ¿Es ello determinante en los resultados que consiguen a lo largo de una temporada?

Principalmente, el presente trabajo pretende ayudar a un entrenador que, durante un partido, quiera saber el efecto de una alineación, técnica o formación, y tras el mismo saber qué mejorar, en qué trabajar más, qué ha causado un gol o pérdida de la posesión, o a un analista de datos que, antes de un partido, estudia al equipo rival o al propio, a nivel individual o colectivo: el número de contactos de cada jugador con el balón, el porcentaje de acierto tanto tirando a gol como pasando a un compañero, la dirección, el número de ataques realizados por sector del campo, las conexiones entre jugadores y así. Todo ello con énfasis en las redes de pases, viendo qué información podemos obtener de ellas y su importancia. Nos fijaremos en competiciones femeninas, y eventos "extremos" o "inesperados" en las mismas.

1.2. Objetivos del trabajo

Teniendo en cuenta lo anterior, los objetivos que nos planteamos son:

1. Entender cómo se pueden aplicar técnicas estadísticas en deportes de equipo, y adaptar resultados en ese campo a los tipos de datos que existen aquí.
2. Crear una herramienta para que personas del equipo técnico de un equipo de fútbol, como analistas de datos, analistas técnicos, analistas de rendimiento físico o entrenadores, puedan tomar decisiones en base al estudio que se haga antes, durante o después de un partido.

Este proyecto es software libre, y está liberado con la licencia [\[32\]](#).

Capítulo 2

Descripción del problema

El presente trabajo intentará responder entonces a: ¿La entropía mejora las posibilidades de marcar gol? Si cambia la entropía del equipo, ¿podemos determinar la causa? ¿Ha sido por un jugador específico, la alineación o más bien la influencia del equipo contrario?. Igualmente, ¿es posible ver la entropía reflejada en un tipo de visualización de las redes de pases?. Estudiaremos hasta qué punto es determinante la entropía a nivel de jugador o equipo, empleando técnicas estadísticas para obtener respuestas a lo planteado y siguiendo la forma de trabajar que pasamos a describir.

2.1. Metodología - Desarrollo ágil

La metodología nos dicta dos cosas esenciales:

1. ¿Qué tengo que hacer ahora?
2. ¿Lo que he hecho está bien y era lo que tenía que hacer?

Para el primer punto, hacemos uso de *issues* y *milestones* en Github, como veremos más adelante. Para lo segundo, hay que tener en cuenta que todos los *issues* son problemas, e idealmente deben de decir explícitamente (o implícitamente si está claro) cómo se resuelve el problema.

El desarrollo ágil tiene su origen en el manifiesto ágil [7], el cual fue una revolución frente al modelo en cascada, en la forma de desarrollar software, y planteaba que tienen más valor las personas que los procesos o herramientas, un producto funcionando que mucha documentación, colaboración con los clientes que contratos, y **flexibilidad** frente a seguir un plan. Desde el año 2001 que se redactó ha cambiado mucho la tecnología, y por ejemplo nosotros pondremos mucho peso en la documentación, mas la idea sigue intacta: lo importante son los usuarios, adaptarse a sus deseos. En general pues, con *ágil* nos referimos a una forma de pensar que se aplica a todo un ciclo de desarrollo del software centrado

en el cliente y que consiste en continuas mejoras de productos mínimamente viables [54]. En este apartado entenderemos mejor y dejaremos claro lo que esto significa e implica.

Lo esencial de esta forma de proceder es que su objetivo es **resolver problemas** y satisfacer al cliente, no hacer aplicaciones: importa el por qué, antes que el qué o el cómo. Estos últimos resultarán de ese primer análisis, seguidos de una empatización, que consiste en contactar con clientes, leer prensa y demás para enfocar el problema; posteriormente ideación, de la que saldrán los objetivos e hitos o historias de usuario; y por último diseño, que será aterrizar todo lo anterior.

Consecuentemente, el punto de partida es pensar la motivación o problema que queremos resolver. ¿Por qué queremos hacer este TFG? ¿A quién ayuda? ¿Quién lo usaría? ¿Qué solución proponemos? Los problemas deben estar antes que nada; es complicado comprar ingredientes si no sabes qué receta vas a preparar.

Luego, de la motivación saldrán los objetivos que nos planteamos. Estos habrán de indicar qué es lo que se quiere conseguir, incluyendo qué tipo de medios tenemos disponibles. Lo más importante es que han de estar en el dominio del problema, tienen que ser específicos, medibles y alcanzables [53]. De estos objetivos saldrán una serie de productos mínimamente viables.

Seguidamente, las **historias de usuario** sirven para centrarnos en los problemas que queremos solucionar y los objetivos a alcanzar. Están relacionadas con la lógica de negocio del proyecto y siempre son un beneficio para los posibles usuarios del proyecto.

Una regla del pulgar para las historias de usuario: Siempre tienen que expresar un deseo y un beneficio para el usuario. Si ponemos "ojalá quéz te lo imaginas en la boca del usuario y suena creíble, es que es una historia de usuario. Si no, es un *issue* o tarea que necesitamos que el usuario haga para que cumpla sus deseos. El problema reside en poner lo que nosotros queremos que haga el usuario para conseguir algo, y no lo que el usuario quiere. Más adelante expondremos las historias de usuario planteadas.

Por otro lado, las **issues** plantean un problema. Siempre están enmarcadas en una **milestone**, y tienen que tener un criterio de aceptación para ser cerradas. Hacer tareas lo más atómicas posibles ayuda, porque se hace un **pull request**, se termina una tarea, y se avanza más fácil y suavemente.

Todo el código se incorpora mediante *pull requests*, eventos que ocurren cuando un contribuidor está preparado para iniciar el proceso de mezclar el nuevo código, normalmente desarrollado en una **rama**, con el repositorio del proyecto principal. Facilita así la revisión por parte de otra persona del equipo, y el asegurarnos de que en producción siempre hay algo que funciona y está testado.

De manera adicional, un **milestone** describe un producto mínimamente viable, y el estado en el que tiene que estar el repositorio al terminar, además de los criterios que se daban seguir para validarlo.

Dentro de la metodología descrita, nos enmarcaremos en el **design thinking**,

un proceso iterativo y no lineal que consta de una fase inicial consistente en empatizar con el cliente, para saber qué necesita, seguida de la definición del problema, pensar en una solución al mismo, el desarrollo de un prototipo, y finalmente se testea todo.

2.2. Herramientas usadas

En esta sección profundizaremos en las herramientas empleadas para desarrollar el trabajo y seguir la metodología ya descrita, además de explicitar por qué las hemos escogido.

2.2.1. Plataforma de desarrollo

Necesitábamos una plataforma que nos permitiera guardar el proyecto en algún sitio aparte de nuestro ordenador, para que en caso de problemas con el mismo pudiéramos continuar el desarrollo y no perderlo todo, a la vez que consultar versiones anteriores de lo escrito, dado que en algunas ocasiones se borran cosas que luego quieres incluir, o simplemente surge la necesidad de consultar el estado del trabajo en otro momento, para restablecerlo incluso. Es esencial tener un visión global de la construcción paso a paso de lo desarrollado, así como ver quién ha aportado qué, permitir que se sugieran cambios o añadan comentarios de una manera fácil, sin tener que andar pasando *zips* para arriba y para abajo, y no entrar en pánico cada vez que se me quedaba el ordenador calado por el programa de R. Teniendo en consideración el potencial tamaño de un proyecto como el que nos ocupa, es importante también que a la plataforma “no le importe”, siga funcionando rápido y no limite de ninguna manera lo que se tiene pensado hacer, sino más bien lo haga sencillo, intuitivo, *disfrutable*. Sin tener que instalar *gigas* de paquetes ni nada, tan solo sea una parte más y se funda con el proyecto. Por supuesto, que sea gratuita es un buen añadido, al igual que el que tenga una gran comunidad detrás. Todo esto nos llevó a escoger [Github](#) (y seguro que nos dejamos algo; GitHub ofrece increíbles posibilidades).

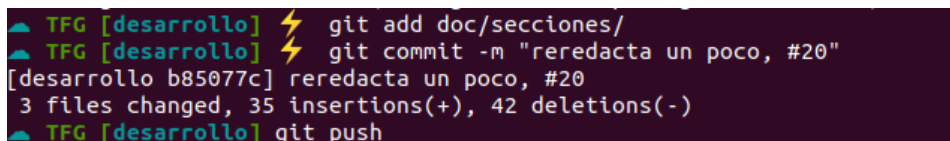
GitHub es una aplicación con web (también API o marco de aplicaciones, entre otras) muy completa para que desarrolladores y programadores puedan trabajar colaborativamente en repositorios. Su punto fuerte es el sistema de control de versiones, que permite llevar cuenta detallada de los cambios realizados por cada persona, discutirlos, revisarlos y proponer modificaciones, así como separar el producto final de las funcionalidades que se vayan añadiendo y sobre las que cada equipo o persona esté trabajando, mediante *ramas*. Adicionalmente, proporciona herramientas como las [Github Actions](#), las cuales permiten la automatización de los flujos de trabajo, incluyendo integración y despliegue continuos. Nosotros por ejemplo lo empleamos para chequear la ortografía cada vez que subimos algo al repositorio, o construir el *pdf* de la memoria cuando se cambie algún archivo *.tex*, pero por supuesto hay muchísimas más funcionalidades posibles. Uno de los objetivos principales y subyacentes de este trabajo de

fin de grado es aprender a gestionar un proyecto siguiendo las mejores prácticas, y GitHub o Git lo piden como requisito algunas empresas, y en general es ampliamente usado.

Se puede consultar nuestro repositorio en [el siguiente enlace](#), y navegar por las [issues creadas](#), [pull requests mergeados](#) o [actions usadas](#). Para el desarrollo de nuestro trabajo es esencial Github, y lo hemos escogido por ser de los más usados y establecidos, con más de 45 millones de usuarios, y porque frente a [gitlab](#), contiene justo lo que necesitamos. Gitlab es para proyectos más grandes y profesionales.

Para tener conciencia situacional del estado del repositorio de Github, instalamos [oh my zsh](#), un *framework* libre para administrar la configuración del *shell* Zsh. Permite el acceso y fácil instalación de plantillas para la terminal, que hacen más fácil saber en qué rama se está trabajando, qué cambios se han realizado, si están ya añadidos o no, y en general casa genial con *Github* y sus funcionalidades. Tiene detrás una comunidad muy grande de contribuidores, incluye numerosos *plugins* que hacen el desarrollo de *software* más fácil y bonito, y la posibilidad de personalizar la terminal, con temas ya creados o propios.

Añadir que Zsh es un *shell* que se presenta como alternativa a *bash*, el cual viene por defecto en *Ubuntu*. Las ventajas de las que más nos hemos aprovechado y por las que lo escogimos parten de su mayor configurabilidad, el que te corrija errores de escritura y complete palabras, y esa necesidad que ya hemos mencionado de tener un buen conocimiento de la situación y el desarrollo.



```
TFG [desarrollo] ⚡ git add doc/secciones/
TFG [desarrollo] ⚡ git commit -m "reredacta un poco, #20"
[desarrollo b85077c] reredacta un poco, #20
3 files changed, 35 insertions(+), 42 deletions(-)
TFG [desarrollo] git push
```

Figura 2.1: Cómo se ve nuestra terminal

2.2.2. IDE

Como editor de código precisábamos de uno que diera soporte a varios lenguajes de programación, fuera bien en Ubuntu, se pudiera coordinar con Github sin problemas, hiciera la escritura de código algo más fácil con funcionalidades como autocompletación de palabras, resaltado de código y corchetes, autoindentación, herramientas de compilación o *debugging*, seguimiento de definiciones, atajos de teclado o un mercado de plugins y temas para facilitar lo anterior, y más en general hacer sencillo el desarrollo. Por ello su enorme potencial y comunidad detrás, escogimos [Visual Studio Code](#). Combina pues la simplicidad de un editor de código fuente con herramientas de desarrollo potentes, como completación de código o *debugging*. Y lo mejor, es software libre, y se integra perfectamente con git y Github, además de permitirte construir el proyecto y compilarlo, todo

sin cambiar de pantalla ni rompederos de cabeza; es intuitivo y hasta diríamos inteligente.

Como opciones, destacamos [Apache Netbeans](#), [Eclipse](#), [RStudio](#), [Vim](#) o [Emacs](#). De todos ellos, es el que mejor puntuación y *reviews* tiene en la web. Lo elegimos antes que Netbeans dado que este no se puede integrar con Github, el desarrollo es más personalizado y permite revisión de código o seguimiento de errores; se ajusta más que Netbeans para lo que queremos hacer y es más fácil de usar, con mayor número de funcionalidades. VS Code es más eficiente que Vim, y gana con diferencia, al ser un IDE completo y no solo un editor de textos, en el que habría que hacerlo casi todo a mano. Por supuesto, hay muchos puristas de Vim o Emacs, que ya han pasado el periodo de adaptación y saben usarlo genial, lo tienen chetadísimo, pero las ventajas y características de VS Code hacen que nos quedemos con él, aparte de la ya reiterada estupenda integración con Github de la que estos últimos carecen. Por último, hemos incluido RStudio debido a que la parte informática será en ese lenguaje, y para ello es maravilloso, pero desde VS Code también se puede hacer, así que mejor tenerlo todo centralizado.

2.2.3. Lenguaje de programación

Cuando pensamos en estadística + lenguaje de programación, lo primero que nos viene a la cabeza es R. Pero por supuesto hay más opciones en el mercado, que pasamos a comparar. Sobre todo, escogimos [R](#) puesto que necesitábamos un lenguaje que no nos pusiera límites ni trabas en cuanto a la aplicación de la teoría matemática, esto es, fuera uno con la estadística computacional y análisis de datos, nos permitiera realizar gráficos, y constara de numerosos paquetes profusamente documentados y desarrollados por expertos, junto con una gran comunidad detrás. Que fuera gratis y software libre son requisitos que damos por sabido, además de que llevara muchos años asentado como herramienta de desarrollo, y en continua expansión y crecimiento.

Como opciones teníamos [Python](#), [Mathematica](#) y [julia](#), entre otros. Ambos Python y R son muy usados en ciencia y análisis de datos o *machine learning*, si bien R en general se dice que es preferible para la parte estadística de un proyecto al ir mejor los paquetes, y para visualizar grafos, a pesar de ser algo más lento. [Aquí](#) se puede consultar una de las comparativas que miramos a la hora de decidirnos. Hemos de admitir también que en la asignatura de Estadística Computacional nos familiarizamos bastante bien con el lenguaje, y nos gustó su potencial y forma de funcionar o traducir las matemáticas a algo que el ordenador entienda. Python obviamente sirve para muchísimo, y nos habría servido igual de bien para hacer el trabajo, pero es un poco como ir con un ferrari por una ciudad, a 30km/h; nosotros necesitábamos únicamente librerías bien trabajadas para experimentación con datos, exploración, simplificación de problemas matemáticos y visualización, y para ello R va que chuta. Mathematica la descartamos casi al instante, al no ser software libre. Por último, Julia ha ganado muchísimo tirón estos años, y se dice que es más rápida que R (podemos

ver una comparativa [aquí](#)). Existen librerías para entropía o redes bayesianas, si bien nos quedamos con R al estar su uso más extendido, con vistas sobre todo a que, ante el surgimiento de problemas, podamos consultar una documentación sin huecos, y haya una extensiva batería de preguntas en StackOverflow que nos sirvan para aprender de nuestros errores.

2.2.4. Sistema de composición de documentos

Nuevamente, cuando piensas en escritura de matemáticas, lo primero que viene a la cabeza y no se cuestiona demasiado es LaTeX. [LaTeX](#) es un sistema estándar para la composición de documentos científicos y técnicos, software libre y gratuito. Lo escogimos ante la necesidad de un programa que nos permitiera escribir la memoria, con soporte para lenguaje matemático o código informático, sin tener que andar ajustando tamaños de letras, imágenes o índices. Queríamos centrarnos en el contenido, y no tanto en el formato. En un principio, puede parecer más fácil usar Word o LibreOffice, pero para escribir proyectos más grandes que no comprendan únicamente texto, tras una pequeña fase de aprendizaje, LaTeX gana con ventaja, al disponer también de plantillas, una documentación excelente, y años y años de consolidación en la comunidad científica. Otra opción que ha aparecido de forma relativamente reciente es [Quarto](#), un sistema de publicación de documentos científicos y técnicos software libre construido con [Pandoc](#), escribiéndolos como si fuera una *notebook* de Python o simple [Markdown](#). Lo que sin duda nos echó más para atrás es la falta de plantillas. Llama la atención y nos reafirmamos en lo productivo y genial que es GitHub al ver que todos estos lenguajes de programación que vamos mencionando han sido desarrollados y podemos encontrar los repositorios en él; por lo que este no es excepción, y a la hora de valorar si usarlo encontramos una [issue](#) abierta en noviembre de 2021 sobre la ausencia de *templates* de Quarto para tesis y artículos académicos, que finalmente fue solucionada en abril de este año. Al ser tan reciente, preferimos LaTeX, de manera que si surgen dudas o problemas haya una literatura más extensa que consultar. Si nos sorprendió que ya existieran extensiones para trabajar con él en VS Code, pero no fue suficiente como para “ganar” a LaTeX en cuanto a ventajas.

LaTeX da pues al usuario un control extremadamente bueno sobre el formato de los documentos. Una vez que se domina, puede ser mucho más fácil trabajar con él que con un procesador de texto convencional. Tuve problemas (que fui reflejando en [mi Github](#)) a la hora de usarlo ya que no lo había empleado hasta este momento, por lo que tenía que entender el funcionamiento de la plantilla que uso, [cómo escribir las fórmulas y símbolos matemáticos](#), [la definición de entornos](#), [cómo hacer citas y referencias](#), [cómo se incluyen imágenes](#), cómo incluir y usar paquetes, o incluso cómo poner títulos, letra en itálica, enlaces. Hay numerosos tutoriales en internet, pero lo que más consulté ha sido [la documentación de Overleaf](#), aunque tenga otro editor de LaTeX. Ciertamente, la curva de aprendizaje no es muy grande, y una vez entiendes la lógica detrás, se escribe

relativamente rápido. En cualquier caso, es mejor que ir buscando la fórmula en el cuadro de mandos de los editores de texto tradicionales, y al ser ficheros de texto los edito en VisualCode Studio y puedo subirlos al repositorio de GitHub junto con el resto del proyecto.

2.3. Clientes

Basándonos en la [metodología basada en personas](#), hemos llegado a los siguientes usuarios ¹:

- Analista táctico: estudia los equipos y cómo se desenvuelven en los partidos.
- *Scouter*: en algunos equipos grandes existe esta figura que asiste a encuentros de otras ligas, otras divisiones, o incluso el filial, para identificar qué jugadores pueden entrar en los planes de adquisición o de ascenso a equipos de categorías superiores del club.
- Analistas de datos: tienen que generar la estrategia de recogida de datos del club y preparar aplicaciones e informes para ayudar en la toma de decisiones del mismo.
- Persona que apuesta: recopila información sobre los equipos y jugadores, para poder apostar en base a una alineación, o una vez se sabe quién va a jugar en un partido.
- Entrenador: es el encargado de decidir a quién sacar durante un partido, los cambios, dónde poner a quién.
- Presidente del equipo de fútbol: tiene que decidir antes del comienzo de cada temporada cuánto dinero ha de invertir en nuevos jugadores.
- Periodista deportivo: debe escribir un artículo sobre un partido con gráficos y datos estadísticos, para lo que debe conocer medianamente a los jugadores y equipos, junto con su desempeño a lo largo de una liga o temporada.

2.4. Historias de usuario

A partir de los clientes, y como parte de la metodología, definimos en [las siguientes historias de usuario](#):

- Como analista táctico, quiero obtener un análisis del propio equipo y de los rivales.

¹Para más información sobre las figuras que aparecen en el análisis de datos deportivos, consultar [Objetivo Analista](#)

- Como scouter, quiero tener elementos cuantitativos de juicio, a partir de las observaciones, para tomar decisiones que beneficien el desempeño deportivo del club.
- Como analista de datos, deseo conocer qué metodología de obtención, tratamiento y presentación de datos es la que me permite elaborarlos con más claridad para presentarlos a los que toman las decisiones en el club.
- Como persona que apuesta, querré inferir el resultado de un partido y quién marcará, una vez se publiquen los jugadores.
- Como entrenador, me gustaría poder usar la herramienta desarrollada, y que sea intuitiva.
- Como presidente del equipo de fútbol, quiero poder calcular los ingresos probables por las ventas de jugadores no deseados, el gasto neto relativo de otros equipos y el posible impacto negativo en el desempeño del equipo de hacer demasiados cambios de personal de una vez.
- Como periodista deportivo, desearía tener una herramienta más en mi arsenal, y usarla para obtener gráficos, estadísticas, y conclusiones acerca de un partido, temporada, equipo o jugador.
- Como tutores, nos gustaría que el proyecto siga un desarrollo ágil.
- Como matemática, quiero que la teoría sobre redes causales se aplique bien al caso concreto, y se conecte sin problema con la parte informática.

2.5. Hitos o Productos Mínimamente Viables

Pasamos a describir los prototipos del producto que queremos desarrollar.

2.5.1. *Milestone 1* - Infraestructura

El objetivo de esta *milestone* es configurar la infraestructura del proyecto. Para ello, tendremos que:

- Borrar los archivos de la plantilla que no sean necesarios.
- Definir los primeros *milestones* e issues.
- Configurar un corrector ortográfico.
- Documentar la configuración inicial.
- Formular los objetivos principales del trabajo.

De esta manera, tendremos hecho el esqueleto sobre el que seguir construyendo. El objetivo principal y producto esperado de este *milestone* es pues tener claro el problema a resolver y los objetivos de este trabajo.

Al final de esta *milestone*, se pretende tener el repositorio configurado de forma que tengamos las bases para crear un producto de calidad (sin faltas ortográficas, por ejemplo).

2.5.2. *Milestone 2 - Planteamiento*

El objetivo de esta *milestone* es dejar escritas:

- Prefacio
- Introducción y objetivos del trabajo
- Descripción/Motivación del problema
- Metodología y justificación de las herramientas usadas
- Clientes, historias de usuario y *milestones*
- Estado del arte
- Planificación y costes

Habiendo terminado así la parte de ideación, planteamiento del problema, y pasar a estar preparado para solucionarlo.

2.5.3. *Milestone 3 - Parte matemática*

El objetivo de esta *milestone* es dejar clara la teoría sobre la que basaremos la implementación, así como la recopilación y procesamiento de datos, eliminando los que no sirvan, para obtener conclusiones. Habrá que dar nociones de fundamentos de redes, probabilidad, probabilidad condicionada, redes probabilísticas, técnicas empleadas para hallar probabilidad, al igual que de redes causales, modelos gráficos e identificación de efectos causales o algoritmos empleados.

Como producto esperado se tendrá la base teórica terminada y bien explicada, para ello tendremos que:

- Añadir conceptos básicos de teoría de la probabilidad
- Incluir fundamentos de redes bayesianas
- Ver criterios gráficos de independencia
- Explicar algoritmos de inferencia para redes bayesianas
- Profundizar en algoritmos de aprendizaje para redes bayesianas

2.5.4. *Milestone* 4 - Implementación

El objetivo de esta *milestone* es aplicar lo aprendido y obtenido en la anterior, analizando los datos de partidos de fútbol, generando visualizaciones de las redes de pases y calculando la entropía.

Como producto esperado se tendrá unos programas en R funcionando y grafos a partir de los cuales podamos obtener conclusiones acerca del desempeño de un equipo a lo largo de una competición, dejando pinceladas de trabajos futuros.

2.6. Definiciones y terminología

El objetivo de esta sección es introducir los términos que serán usados en el estado del arte, donde veremos todavía mejor cómo se conectan con nuestro problema.

Enunciamos en primer lugar lo que son los modelos gráficos probabilísticos, ya que permiten realizar inferencia en dominios complejos dotados de incertidumbre.

Definición 1. Modelos gráficos probabilísticos Un modelo gráfico probabilístico es una representación de la distribución de probabilidad conjunta sobre un dominio, que consta de dos partes: una componente cualitativa en forma de grafo que codifica afirmaciones de independencia condicional sobre el dominio que se está estudiando, y una componente cuantitativa que consiste en una colección de distribuciones de probabilidad locales que cumplen la propiedades de independencia especificadas en el modelo [1].

Un tipo de modelo gráfico probabilístico que usaremos y estudiaremos en profundidad en los próximos capítulos son las redes bayesianas.

2.6.1. Redes bayesianas

Definición 2 (Redes bayesianas). Una red bayesiana [44] $\mathfrak{B} = \{\mathcal{G}, \mathbb{P}\}$ está definida por:

- Un grafo dirigido acíclico $\mathcal{G} = (V, E)$ donde V es un conjunto de nodos y E es un conjunto de aristas.
- Un espacio de probabilidad (Ω, \mathbb{P}) .
- Un conjunto de variables aleatorias $\mathbf{V} = V[i], i = 1, \dots, N$ con N lista de variables aleatorias, asociadas con los nodos del grafo (Ω, \mathbb{P}) de tal manera que $\mathbb{P}(V[1], \dots, V[N]) = \prod_{i=1}^n \mathbb{P}(V[i] | pa(V[i]))$, donde $pa(V[i])$ es el conjunto de los nodos padre de $V[i]$ en \mathcal{G} .

En otras palabras, una red bayesiana es un grafo donde los nodos representan variables aleatorias discretas o continuas y los bordes o aristas representan las influencias entre ellas. Las aristas representan pues las causalidades, que pueden

ser determinísticas o probabilísticas. Para un borde uniendo el hecho A con el B , con $P(B|A)$ representamos la relación de probabilidad de un nodo, conocidos sus padres. Para los nodos sin padres o nodos raíz, se asignará una probabilidad previa.

Vamos a ver un poco más específicamente qué queremos decir cuando hablamos de grafos.

2.6.1.1. Grafos

Definición 3. Un grafo es un par $\mathcal{G} = (V, E)$, donde V es un conjunto finito de vértices distinguibles y $E \subseteq V \times V$ es un conjunto de aristas. Un par ordenado $(u, v) \in E$ denota un borde dirigido del vértice u al vértice v , y se dice que u es padre de v y v hijo de u . Al conjunto de padres e hijos de un vértice v los denotaremos por $pa(v)$ y $ch(v)$, respectivamente. Los bordes dirigidos se representarán con flechas, y los bordes no dirigidos con líneas.

Un camino $\langle v_1, \dots, v_n \rangle$ es una secuencia de vértices distinguibles tales que $u \rightarrow v$, $v \rightarrow u$ o $u = v$ para cada $i = 1, \dots, N-1$. El camino es dirigido si $v_i \rightarrow v_{i+1}$ para cada $i = 1, \dots, N-1$; v_i es por tanto un ancestro de v_j y v_j es un descendiente de v_i para cada $j > i$.

Un ciclo es un camino, $\langle u, \dots, v \rangle$ de longitud mayor que dos (excepto el caso $v_1 = v_N$); un grafo dirigido sin ciclos dirigidos es el ya mencionado DAG.

Generalmente, las redes bayesianas se utilizan como un marco eficiente para la toma de decisiones con conocimiento incierto, y miden la estructura de dependencia condicional de un conjunto de variables aleatorias, tomando como base el teorema de Bayes:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad (2.1)$$

donde $P(A)$ es la probabilidad previa, $P(B)$ son las observaciones y la probabilidad posterior viene dada por $P(A|B)$ [66].

En 5 profundizaremos más sobre este tipo de redes.

En aplicaciones prácticas, es común encontrarse con escenarios que involucran variables discretas y continuas simultáneamente. Este tipo de problemas se pueden modelar utilizando las llamadas redes bayesianas híbridas.

2.6.2. Redes bayesianas híbridas

Definición 4 (Redes bayesianas híbridas). Una red bayesiana híbrida [43] es un tipo de red bayesiana que permite modelar incertidumbre sobre variables discretas y continuas.

Han recibido una atención creciente durante los últimos años, debido a que amplían la aplicabilidad de las redes bayesianas en general. Sin embargo, esta

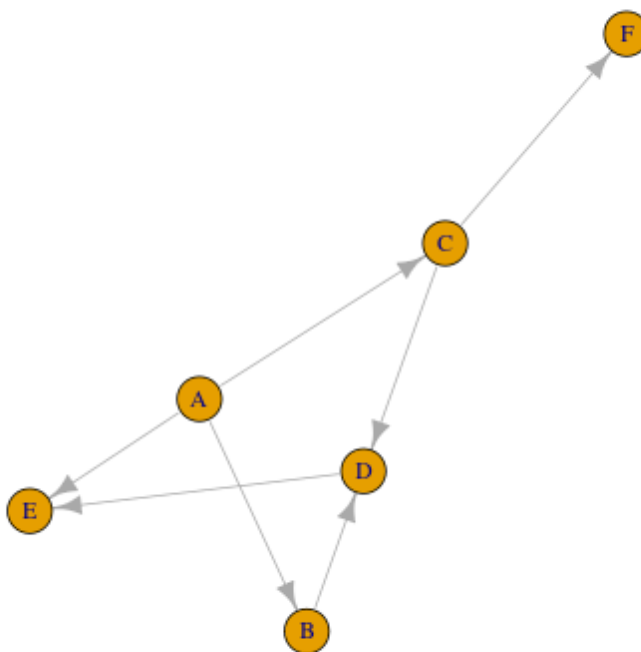


Figura 2.2: Ejemplo de grafo dirigido acíclico (DAG). Generado mediante [un script en R](#).

característica adicional también tiene un coste: la inferencia en este tipo de modelos es computacionalmente más desafiante [1].

Tradicionalmente se han aplicado más las redes complejas que las bayesianas al análisis del fútbol [5]. Este tipo de redes nos interesan ya que su estudio está inspirado en análisis empíricos de redes reales, y permiten entender el funcionamiento de los equipos de fútbol, ya sea a nivel individual o colectivo. Usaremos variables definidas sobre la red compleja para "alimentar" la red bayesiana [6].

2.6.3. Redes complejas

Definición 5 (Redes complejas). Una [red compleja](#) es un grafo con características topológicas que no se dan en redes simples como [celosías](#) o [grafos aleatorios](#), pero a menudo ocurren en redes que representan sistemas reales.

Estos sistemas son llamados complejos debido a que no es posible predecir directamente su comportamiento colectivo a partir de sus componentes individuales. Una de las principales razones por las que las redes complejas son populares es su flexibilidad y generalidad para representar prácticamente cualquier estructura natural, incluidas las que experimentan cambios dinámicos de topología [52].

La teoría de redes complejas se desarrolla en base a la teoría de grafos y la física estadística; en esta, todo sistema complejo puede abstraerse como una red. Los nodos de la red se pueden considerar como los elementos en el sistema, y las relaciones entre cada elemento como conexiones. Hay muchos indicadores estructurales en la teoría de redes complejas que pueden cuantificar la importancia de los nodos desde el nodo mismo o el nivel de relación de red, usando [centralidad de grado](#), [centralidad de vectores propios](#) y [coeficiente de agrupamiento](#) [3].

Otra manera de medir la importancia de cada nodo es mediante sistemas de calificación.

2.6.4. Sistemas de calificación

Determinar la habilidad relativa entre adversarios es uno de los elementos más importantes en el análisis de fútbol junto con la predicción de resultados de partidos; las posiciones en una liga tienen numerosos inconvenientes que hacen que no sean fiables para la predicción. Por ejemplo, una liga de fútbol sufre una variación alta al inicio de la temporada, y baja al final. Asimismo, es posible que los equipos que compiten durante una temporada no compartan un número equivalente de partidos jugados debido a aplazamientos y por lo tanto, la clasificación de la liga será errónea durante unas cuantas semanas. De hecho, la clasificación de una liga está sesgada hasta que se juega el último partido de la temporada, porque para que la clasificación sea justa, cada equipo tiene que jugar contra el resto, en casa y fuera. Incluso al final de una temporada, el *ranking* representa el rendimiento general durante ese período, pero no logra demostrar cómo varió la habilidad de un equipo globalmente, al ignorar los partidos de Copa u otras competiciones (como la Champions League), o no compara equipos en diferentes divisiones/ligas. **En resumen, una tabla de clasificación no es un buen indicador de la situación actual de un equipo. Un sistema de calificación proporciona medidas relativas de superioridad entre adversarios y supera todos los problemas anteriores** [23].

Cuando se trata de fútbol, para generar calificaciones que capturen con precisión la capacidad actual de un equipo, tenemos que considerar al menos:

- La ventaja que supone jugar en casa [37]
- El hecho de que los resultados más recientes son más importantes que los menos recientes al estimar las habilidades actuales.

- El hecho de que una victoria es más importante para un equipo que aumentar la diferencia de goles.

Definición 6. La calificación general de un equipo es la calificación promedio entre las actuaciones en casa y fuera, y esto se define como

$$R_{\tau} = \frac{R_{\tau H} + R_{\tau A}}{2}$$

donde R_{τ} es la calificación para un equipo τ , $R_{\tau H}$ es la calificación para el equipo τ cuando juega en casa, y $R_{\tau A}$ es la calificación de un equipo τ cuando juegan fuera.

Nosotros calificaremos a los equipos según la entropía obtenida de las redes de pases entre los jugadores, la cual caracteriza su grado de heterogeneidad (uniformidad) u homogeneidad (variabilidad). Esto es, cuanto más cercana sea la distribución de pases a la uniformidad, menos certeza tendremos de quién obtendrá el balón [56].

2.6.5. Entropía

La entropía es el concepto clave para extraer características universales de un sistema a partir de sus detalles. Aparece en muchos contextos como termodinámica, mecánica estadística o teoría de la información, como una medida de diferentes propiedades: energía que no puede producir trabajo, nivel promedio de información, sorpresa, desorden, incertidumbre, aleatoriedad o complejidad [4], [16].

Definición 7 (Entropía). Para una variable aleatoria discreta X definimos su entropía como:

$$H(X) := - \sum_x P(x) \log[P(x)]$$

bits, donde X toma valores en \mathcal{X} , $P: \mathcal{X} \rightarrow [0,1]$, $P(x)$ es la probabilidad de que X esté en el estado x . La entropía conjunta de las variables $X_1 \dots X_N$ se define por

$$H(X_1, \dots, X_N) = - \sum_{x_1} \dots \sum_{x_N} P(x_1, \dots, x_N) \log[P(x_1, \dots, x_N)]$$

El concepto de entropía (o entropía de Shannon) se lo debemos al matemático e ingeniero eléctrico americano [Claude Shannon](#), que lo introdujo en 1948 en el artículo "Una teoría matemática de la comunicación" [65].

La [entropía de Shannon](#) nos da una manera de estimar el número medio mínimo de *bits* que se necesitan para codificar una cadena de símbolos, basándonos en la frecuencia de estos.

Con la entropía de Tsallis de los pases entre los jugadores, se puede medir la organización asociada al comportamiento de un equipo de fútbol [56].

2.6.6. Entropía de Tsallis

En 1988 el físico [Constantino Tsallis](#) introdujo una nueva definición para entropía que describe las características estadísticas de sistemas complejos. Fue diseñada para analizar sistemas donde existen correlaciones entre sus microestados [27].

Definición 8 (Entropía de Tsallis). Dado un conjunto discreto de probabilidades p_i con la condición $\sum_i p_i = 1$, y con q cualquier número real, se define la entropía de Tsallis como

$$S_q(p_i) = \frac{k}{q-1} \left(1 - \sum_{i=1}^N p_i^q \right)$$

donde q se denomina índice entrópico, N son los microestados y k es una constante positiva.

La fórmula se reduce a la de la entropía de Shannon cuando $q = 1$. En último lugar, notar que si se combinan dos sistemas idénticos, la entropía de Tsallis del sistema combinado no es igual a la suma de la entropía de sus subsistemas [51]. Esto quiere decirnos que podemos trabajar con la entropía de alguna variable calculada a partir de los jugadores.

Capítulo 3

Estado del arte

La historia del análisis de fútbol comienza en 1950, cuando Charles Reep [61], teniente coronel de la *Royal Air Force* británica y contable, empezó a registrar sistemáticamente los eventos que tenían lugar a lo largo de un partido de Swindon Town. Con ello, no pretendía únicamente tener detalle de lo que había ocurrido, sino también saber por qué pasaba y qué podían aprender los equipos de los datos, para mejorar su juego. Tras décadas, llegó a la conclusión de que el mayor número de goles resultaba de posesiones de tres pases o menos, por lo que aludía que los equipos debían simplificar sus tácticas para llegar a portería más rápido.

El problema de Reep no fue su recolección de datos, sino su análisis; podría haberse preguntado sobre la tasa de goles en posesiones de varias distancias, o cómo se desarrollaron exactamente esas posesiones de tres pases.

Otro pionero en el uso de datos en el deporte fue [Abdus Salam Qureishi](#), filántropo e informático en Silicon Valley, hacía *datascouting* en la reconstrucción de los [Dallas Cowboys](#), un equipo profesional de fútbol americano de los años 60 [17].

En los años 70, [Valeri Lobanovski](#) fue un jugador y entrenador de fútbol ucraniano que formalizó el cuerpo técnico moderno y la captura de eventos en el fútbol del equipo [Dinamo de Kiev](#). Declaraba que "Hoy en día los jugadores no pueden quejarse, saben que mañana después del partido habrá colgada una hoja con las cifras que describen en detalle su juego" [41].

En 1996, [Opta](#) empezó a acumular datos de partidos para la *Premier League* inglesa. Ello incluía todas las estadísticas a las que un aficionado al fútbol está acostumbrado: número de pases, número de regateos, distancia recorrida, por ejemplo. Este podría considerarse el punto de comienzo del análisis de fútbol moderno. Desde 2019, se llama [Stats Perform](#), y es todo un referente a la hora de análisis de datos de deportes, incorporando inteligencia artificial.

En 2003, Michael Lewis lanzó el libro [Moneyball](#), el cual tuvo una gran influencia también fuera de Estados Unidos. Trata sobre cómo un equipo de béisbol de bajo presupuesto se convirtió en uno de los mejores, usando estadísticas pa-

ra reclutar a jugadores con habilidades hasta entonces minusvaloradas, como el [porcentaje de veces que un bateador llega a una base](#), o la [productividad de bateo](#) [39]. El lanzamiento del libro fomentó que se generaran [preguntas](#) en torno al empleo de datos en deportes de equipo.

A su vez, en 2005 hubo muchos intentos de relacionar redes complejas y fútbol [46], e incluso en el 2004 se lanzó un desafío en la lista [redes](#) [] para predecir el resultado de la Eurocopa usando redes sociales. En cualquier caso, tanto el estallido del análisis de redes sociales como el libro contribuyeron a la expansión de este campo.

En los últimos años ha habido una revolución en el análisis de la organización y rendimiento de los equipos de fútbol y sus jugadores, por lo que es posible tener acceso a todos los eventos que ocurren en el campo, tales como pases, disparos o goles, todo ello con las coordenadas exactas de tiempo y posición, y el jugador responsable de cada evento. Por otro lado, también es posible llevar un registro de las posiciones de todos los jugadores en el campo, junto con el balón, lo que permite determinar la posición, velocidad y aceleración de cada jugador, y dando información muy útil sobre su desempeño físico y táctico.

No obstante, los avances más determinantes e importantes tienen que ver con la posibilidad de aplicar o definir nuevos métodos y herramientas. En ese sentido, se pueden entender los roles de los jugadores como un todo, no solo como componentes aisladas sin interacciones entre ellos. Consecuentemente, es posible contruir redes de pases compuestas por nodos, correspondientes a los jugadores, y enlaces, que representan los pases entre ellos. Esta organización está además lejos de ser aleatoria. El análisis de la evolución de las redes de pases ha mostrado que sus propiedades cambian continuamente a lo largo de un partido y que eventos importantes tales como goles pueden afectar la organización de la red [50].

Asimismo, durante la última década las redes bayesianas 2 se han popularizado en el campo de la inteligencia artificial, y hay numerosos estudios que buscan predecir los resultados de partidos de fútbol empleándolas, para lo que construyen diferentes modelos [58]. En [62] consiguen una precisión predictiva del 75.09%. [20] introduce el uso de calificaciones dinámicas 2.6.4 y redes bayesianas híbridas 4 para hacer una predicción del resultado de un partido entre dos equipos a y b a partir de datos históricos de partidos en los que no participan ni a ni b . Ello nos muestra el increíble potencial que tienen estas redes también, y no solo las más tradicionales redes complejas 5 [13].

A diferencia de los sistemas simples, muy pocas metodologías tratan la evaluación de la confiabilidad de sistemas complejos, especialmente los configurados como redes, donde es difícil tomar en consideración los diferentes vínculos y factores que pueden afectar la disponibilidad y confiabilidad de tales sistemas. En este contexto, las redes bayesianas permiten el modelado de sistemas configurados como red y el cálculo de probabilidades marginales de los nodos del sistema utilizando probabilidades previas y condicionales [29].

Adicionalmente, hay estudios [21] que ponen un gran énfasis en aplicar co-

nocimiento causal al proceso de desarrollo del modelo, basándose en los datos que son necesarios para la predicción. De esa manera, consiguen predicciones precisas del cambiante rendimiento de equipos de fútbol. El modelo permite predecir, antes de que empiece una temporada, los puntos totales en la liga que se espera un equipo acumule a lo largo de la temporada, lo que supone un cambio de perspectiva con respecto a los artículos mencionados con anterioridad, y en su metodología construye y trabaja sobre la literatura anterior, extendiéndola.

Por otro lado, hay vertientes que hacen análisis de fútbol desde la inferencia estadística, desde el *machine learning*, o ambos [9]. En este *paper*, muestran que los equipos están caracterizados por dónde en el campo realizan pases, y se pueden identificar por la manera en que pasan el balón. Usando mapas de calor de las localizaciones de los pases, consiguen una precisión del 87% en una clasificación de veinte equipos. Emplean además la localización de los pases a lo largo de una posesión del balón para predecir tiros a puerta. Finalmente, usan los pesos del modelo de predicción para categorizar a los jugadores según el valor de sus pases. Nos muestra una vez más la diversidad que existe en el estudio y análisis del fútbol; no hay límites en cuanto a lo que se puede considerar y obtener.

[56] afirma que para optimizar su rendimiento, un equipo debe mantener un equilibrio entre orden u organización, el cual propicia la cooperación entre sus miembros, y desorden o desorganización, que confunde al oponente y favorece el mantener un cierto grado de libertad. Para medir, usan entropía de Tsallis 8 a partir de los pases entre los jugadores, y concluyen que la posición de un equipo al final de la temporada está correlada con la entropía del mismo.

Más específicamente, saber quién pasa el balón a quién significa reducir la entropía 7 de los pases y maximizar la comunicación y cooperación entre los jugadores del equipo. A su vez, esto implicaría limitar sus grados de libertad al mover la pelota y sorprender al oponente, lo que expondría al equipo a contraataques. El precio pues de maximizar la certeza de los pases puede ser un comportamiento predecible y hacer al equipo vulnerable. Lo más importante de este *paper* es que reivindica que el rendimiento de un equipo de fútbol puede ser modelado usando la entropía de sus pases de balón.

Y no solo eso, [50] cuantifica la entropía espacial y temporal de equipos de fútbol y sus jugadores, también mediante interacciones basadas en pases. Sus resultados muestran que la entropía espacial cambia de acuerdo con la posición de los jugadores en el campo, y la organización de las redes de pases cambian a lo largo de un partido.

[25] fue uno de los primeros en investigar la entropía entre los jugadores de un equipo de fútbol. Analizan el desempeño y estilo de juego de la selección española en los mundiales del 2010 desde una perspectiva temporal, con lo que medidas globales como el número de pases consecutivos o el número de pases por minuto reflejan el éxito de un equipo en imponer su forma de juego. Así, España consigue tener mayor posesión del balón y pases completados, lo que es una buena estrategia defensiva al privar al otro equipo de influenciar en el juego.

Esto en último lugar determina el resultado final del partido, si bien es discutible la necesidad de tener tanto tiempo el balón, sin muchas veces avanzar a puerta. En el último partido muestran que la precisión disminuye y el juego es menos elaborado, al introducirse más emoción y nervios.

Pero, ¿cómo medimos la calidad de un equipo? No hay muchos artículos aplicando redes causales a partidos de fútbol. [15] muestra que las estrategias ofensivas son más influyentes que las defensivas.

Capítulo 4

Planificación

En este capítulo describiremos cómo nos hemos organizado y repartido el trabajo.

4.1. Temporización

En los meses de julio y agosto, hemos dedicado nueve horas diarias al trabajo, lo que implica 558 horas, más aquellas *issues* y *milestones* en las que trabajamos en [junio](#), a la cual dedicamos tres horas, en [abril](#) y [marzo](#) diez horas conjuntamente, lo que nos deja con 571 horas empleadas.

4.2. Seguimiento del desarrollo

Se puede consultar en [nuestro GitHub](#), más específicamente en [el historial de commits](#).

Capítulo 5

Desarrollo teórico

En este capítulo estableceremos la base matemática del proyecto. Principalmente nos basaremos en los libros *Análisis de datos avanzado desde un punto de vista elemental* [64], *Learning Bayesian Networks* [55] y *Redes probabilísticas para aficionados - Una guía para la construcción y análisis de redes bayesianas y diagramas de influencia* [42], junto con el capítulo *Visión general de la representación y descubrimiento de relaciones causales usando redes bayesianas* [24] y la tesis *Modelos de clasificación de documentos basados en redes bayesianas* [63].

5.1. Teoría de la probabilidad

5.1.1. Conceptos básicos

En esta subsección revisaremos algunos conceptos de teoría de la probabilidad básicos, en el caso discreto.

5.1.1.1. Funciones y espacios de probabilidad

Un *espacio muestral* es un conjunto contable¹ $\Omega = \{x_1, x_2, \dots\}$ que representa los posibles resultados de un experimento.

Una función $P: 2^\Omega \rightarrow \mathbb{R}$, donde 2^Ω es el conjunto de todos los subconjuntos (eventos) de Ω , se denomina *función de probabilidad* si satisface los axiomas de probabilidad de Kolmogorov:

1. $0 \leq P(A) \leq 1, \forall A \subseteq \Omega$.
2. $P(\Omega) = 1$.

¹Un conjunto es *contable* si tiene el mismo número de elementos (cardinalidad) que algún subconjunto de los números naturales, \mathbb{N} .

3. Si E_1, E_2, \dots son tales que $E_i \cap E_j = \emptyset, \forall i \neq j$, entonces

$$P(\cup_i E_i) = \sum_i P(E_i).$$

Si P es una función probabilística, el par (Ω, P) se llama *espacio de probabilidad*.

5.1.1.2. Probabilidad condicionada y teorema de Bayes

Sean $A, B \subseteq \Omega$ tales que $P(B) \neq 0$. Entonces, la *probabilidad condicionada* de A , dado B y notada como $P(A|B)$ viene dada por:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Diremos que dos eventos A, B , donde $P(A) \neq 0$ y $P(B) \neq 0$ son *independientes* si $P(A|B) = P(A)$, y son *condicionalmente independientes* si dado C , $P(A|B \cap C) = P(A|C)$.

Si E_1, \dots, E_n es un conjunto de eventos mutuamente disjuntos tales que $\cup_i E_i = \Omega$, esto es, son una *partición* del espacio muestral, y $P(E_i) > 0, \forall i = 1, \dots, n$, entonces para cualquier evento A :

$$P(A) = \sum_{i=1}^n P(A \cap E_i) = \sum_{i=1}^n P(A|E_i)P(E_i). \quad (5.1)$$

Esta propiedad se llama *regla de probabilidad total*, y puede ser demostrada con teoría básica de conjuntos.

5.1.1.3. Variables aleatorias

Una *variable aleatoria* (discreta) es una función que mapea eventos a valores de un conjunto contable, donde cada valor tiene una probabilidad mayor o igual a cero. Representaremos las variables aleatorias con letras mayúsculas, y los valores a los que los eventos son mapeados, con minúsculas.

Una variable aleatoria induce una nueva función de probabilidad sobre sus valores. Usaremos la notación $X = x$ para representar el conjunto de eventos de Ω mapeados por X a x . Si definimos $P_X(\{x\}) := P(X = x)$, entonces P_X es una función de probabilidad. De ahora en adelante escribiremos simplemente $P(X = x)$, que se denominará la *distribución de probabilidad* de X (o solo *distribución* de X). Si sabemos que X es una variable aleatoria, podemos denotar su distribución como $P(X)$ sin ambigüedad.

En ocasiones, dada una variable aleatoria X y siendo x uno de sus valores, escribiremos para ser más breves $p(x)$ en vez de $P(X = x)$. Esta cantidad es la *probabilidad de x* .

Asimismo, para una variable aleatoria A , denominaremos al conjunto de valores que A puede tener como "rango de A ", Ω_A .

Dadas X, Y variables aleatorias definidas en el mismo espacio muestral, y dos valores x de X e y de Y , podemos medir la probabilidad del evento intersección $P(X = x, Y = y)$, la cual se denomina *distribución de probabilidad conjunta* de X e Y . Podemos definir la distribución de probabilidad conjunta de un conjunto arbitrario de variables aleatorias $\{X_1, X_2, \dots, X_n\}$ como $P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$.

Pasamos a introducir la operación conocida como *marginalización*. Dada una distribución de probabilidad conjunta $P(X = x, Y = y)$, podemos obtener la distribución $P(X = x)$ (llamada *distribución marginal de probabilidad de X^2*), sumando sobre todos los valores de y en el rango de Y :

$$P(X = x) = \sum_y P(X = x, Y = y).$$

Esta es otra expresión de la ley de probabilidad total 5.1.

En último lugar, una vez entendida la analogía entre variable aleatoria y eventos, podemos definir la *distribución de probabilidad condicionada* de una variable aleatoria X dado Y :

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}.$$

Este caso se puede extender a un conjunto de dos o más variables. Un resultado que permite calcular la distribución conjunta de un conjunto de variables aleatorias usando únicamente probabilidades condicionadas es la *regla de la cadena*:

$$P(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i | X_{i-1} = x_{i-1}, \dots, X_1 = x_1),$$

la cual puede probarse fácilmente como aplicación directa del teorema de Bayes 2.1.

5.1.1.4. Independencia condicionada y observaciones de variables aleatorias

Dos de las nociones "clásicas" en teoría de la probabilidad (y ya mencionadas en la sección 5.1.1.2) son el concepto de *independencia* y *independencia condicionada*.

Desde el punto de vista de variables aleatorias, podemos reescribir la definición de *eventos independientes*; diremos que dos eventos $X = x$ e $Y = y$ (siendo x, y dos valores de las variables aleatorias X, Y) son independientes si

$$p(x, y) = p(x)p(y).$$

²El término "marginal" se usa cuando una distribución se obtiene por marginalización de otra distribución conjunta, pero de hecho es también una distribución "convencional".

Además, X e Y son *independientes* sí y solo sí

$$P(X = x, Y = y) = P(X = x)P(Y = y),$$

para todos los valores x, y de X e Y , respectivamente.

Para problemas con al menos tres variables aleatorias, X, Y, Z , puede observarse *independencia condicionada*; diremos que, dada Z , X e Y son *independientes condicionadamente* \leftrightarrow todo par de valores x (de X) e y (de Y) es independiente condicionadamente para cada z (de Z), tal que $p(z) > 0$, esto es:

$$\forall x, y, z, \quad p(z) > 0 \Rightarrow p(x, y|z) = p(x|z)p(y|z).$$

Este concepto no está solo limitado a tres variables, y se puede extender a conjuntos grandes.

Diremos que una variable es *observada* si toma uno de los valores de su rango. Si queremos testear la *independencia con respecto a una variable observada*, necesitamos únicamente testear la independencia con respecto a ese valor asignado. Es de notar que observar una variable puede cambiar las relaciones de independencia entre un conjunto de variables.

Escribiremos $A \perp B|C$ cuando una variable A es independiente de otra, B , tras la observación de C . Si A es independiente de B , sin observaciones adicionales, escribiremos $A \perp B$.

Dados $\mathbf{W}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}$ conjuntos de variables aleatorias, las relaciones de independencia tienen las siguientes propiedades:

- Simetría $(\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}) \Rightarrow (\mathbf{Y} \perp \mathbf{X}|\mathbf{Z})$.
- Descomposición $(\mathbf{X} \perp \mathbf{Y}, \mathbf{W}|\mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y}|\mathbf{Z})$.

5.2. Redes bayesianas

5.2.1. Introducción

Las redes bayesianas nos ayudan a modelar y entender las muchas variables que informan nuestro proceso de toma de decisiones. Las decisiones más complejas están normalmente basadas en una multitud de factores o variables. Por ejemplo, para el presidente de un equipo de fútbol 2.4, podemos mapear la decisión que tiene que tomar y las diferentes variables usando una red bayesiana, esto es, un modelo gráfico que captura la relación entre variables que están bajo supuestos de causalidad o influencia [22].

Como ya hemos comentado en 2.6.1, **una red bayesiana es un diagrama que usa flechas o arcos dirigidos para mostrar cómo distintos factores, representados por nodos elípticos, se influyen los unos a los otros.** Cada nodo viene con una tabla de probabilidades condicionadas, la cual refleja las posibilidades de que tenga lugar distintos desenlaces, provenientes de las

influencias que le afectan directamente. Una vez la estructura del grafo y dicha tabla han sido definidas, hay algoritmos estándar que calculan los estados de las variables desconocidas basándose en los estados de las variables conocidas en el modelo [2], [30], [8].

Una de las razones por las que las redes bayesianas son tan potentes es que pueden realizar inferencias tanto predictivas como diagnósticas. Por ejemplo, podemos por un lado predecir la posición en la liga de un equipo para un valor dado de rendimiento, y por otro ingresar un estado de posición en la liga como observación para examinar qué nivel de desempeño del equipo podría explicarla. Estos algoritmos estándar son llamados algoritmos de "propagación bayesiana" [14], [35], [19] porque se basan en el teorema de Bayes; en ellos, la probabilidad de una variable desconocida se actualiza después de que se obtenga evidencia relevante para esa variable [34].

Las clases de causalidad que producen redes bayesianas son:

- **Cadena causal:** describe variables que tienen un efecto dominó las unas sobre las otras. Por ejemplo, *cambios en la calidad de los jugadores* tiene impacto sobre el *desempeño del equipo* que a su vez influencia la *posición en la liga*. Esto quiere decir que la *posición en la liga* es independiente de los *cambios en la calidad de los jugadores* una vez conocemos el *desempeño del equipo*.
cambios en la calidad de los jugadores → *Desempeño del equipo* → *Posición en la liga* Se dirá que son *nodos en secuencia*.
- **Efecto común:** ocurre cuando dos variables diferentes, tales como *fichajes* y *jugadores vendidos*, tienen influencia sobre una tercera variable tal como *gasto neto en transferencias*. Esto significa que *jugadores vendidos* depende de *fichajes* una vez que conocemos el *gasto neto en transferencias*.
Fichajes → *Gasto neto en transferencias* ← *Jugadores vendidos* Se dirá que son *nodos convergentes*.
- **Causa común:** tiene lugar cuando dos variables distintas, tales como *posición en la liga* y *asistencia*, se ven influenciados por la misma variable, tal como *desempeño del equipo*. Ello significa que *asistencia* es independiente de *posición en la liga* una vez conocemos *desempeño del equipo*.
Posición en la liga ← *Desempeño del equipo* → *Asistencia* Se dirá que son *nodos divergentes*.

Probablemente el aspecto más importante de las redes bayesianas es que son representaciones directas del mundo, no de procesos de razonamiento. Las flechas en el DAG (ver por ejemplo 2.2) representan conexiones causales reales, y no el flujo de información durante un razonamiento, como en [sistemas basados en reglas](#) o [redes neuronales](#) [40]. Procesos de razonamiento pueden operar en redes bayesianas mediante la propagación de información en cualquier dirección.

La mayor parte de los modelos probabilísticos, incluidas las redes bayesianas generales, describen una distribución sobre eventos que pueden haber sido observados, pero no dicen nada sobre qué pasará si una cierta intervención tiene lugar. Una *red causal* es una red bayesiana con la propiedad añadida de que los padres de cada nodo son sus causas directas [40]. Las redes causales se definen pues como redes bayesianas en las cuales el modelo de probabilidad viene dado por la eliminación de enlaces de nodos del padre; en general (y a no ser que indiquemos lo contrario), cuando hablemos de redes bayesianas nos referiremos a redes bayesianas causales. Aparte de que las redes bayesianas son usadas para descubrir estructuras causales en datos estadísticos no procesados, como iremos viendo [12].³

5.2.2. Factorización de redes bayesianas

Nos interesa reducir el número de parámetros que encontraremos en las redes bayesianas, de manera que sean menos costosas de analizar, y la predicción sea mejor. En esta subsección estudiamos maneras de hacerlo.

Dado un conjunto de variables aleatorias $\mathbf{U} = \{X_1, \dots, X_n\}$, un *modelo de dependencia* M es un conjunto de relaciones de independencia suficiente para determinar, para cualesquiera $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3 \subset \mathbf{U}$ con $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ disjuntos, y $\mathbf{A}_1, \mathbf{A}_2$ no vacíos, si la relación de independencia $\mathbf{A}_1 \perp \mathbf{A}_2 | \mathbf{A}_3$ es verdadera o falsa.

Si tenemos un conjunto de variables aleatorias cada una en un nodo de un DAG \mathcal{G} , la *condición de Markov* nos dice que una variable aleatoria de un grafo es independiente de sus no-descendientes, dados sus padres [38].

Una red bayesiana causal satisface la *condición de Markov causal*: dadas las causas directas, el fenómeno asociado a un nodo es independiente de los que no tienen efecto sobre él. Esta asunción permite que la distribución conjunta de las variables en una red causal sea factorizada como [48]:

$$P(X_1, \dots, X_N) = \prod_{i=1}^N P(X_i | pa(X_i))$$

Una distribución de probabilidad P en un DAG D constituye una *red bayesiana* $\leftrightarrow D$ es un l-map mínimo de P . Esto nos dice que una red bayesiana sobre un conjunto de variables aleatorias puede verse como un conjunto de relaciones de independencia almacenadas usando un DAG [31].

Teorema 5.1. (Factorización de una red bayesiana) *La distribución conjunta de una red bayesiana $P(X_1, \dots, X_n)$ se factoriza de la siguiente manera:*

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | pa(X_i))$$

donde $pa(X_i)$ es el conjunto de padres de la variable en el grafo.

³Para obtener una idea más clara de la distinción redes bayesianas / redes bayesianas causales, consultar [el siguiente enlace](#).

El recíproco también es cierto:

Teorema 5.2. *Sea P una distribución en un grafo tal que $P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | Pa(X_i))$. Entonces, el grafo es un I-map de P .*

Las redes bayesianas cumplen el conjunto de condiciones de Markov, si bien suele aplicarse (y son equivalentes) más el concepto de d -separación, un criterio gráfico de independencia, por su buen rendimiento computacional [60].

5.2.3. Criterios gráficos de independencia

Las redes bayesianas nos dan un conjunto de *criterios gráficos* para comprobar si, dado otro conjunto, dos conjuntos de variables aleatorias son independientes.

Diremos que X está d -separado (o simplemente *separado*) de Y dados Z_1, \dots, Z_k si todos los caminos entre X e Y recorriendo los arcos de la red en ambas direcciones, están *bloqueados* por cualquier nodo de Z_1, \dots, Z_k . Las dos maneras en que un nodo puede bloquear un camino son:

- Hay observaciones de nodos en secuencia en el camino.
- Hay un nodo convergente sin observar, cono todos sus descendientes no observados

Estamos pues en condiciones de ver el resultado principal de d -separación e independencia:

Teorema 5.3. *Sean A y B variables de una red bayesiana separadas por C , entonces $A \perp B | C$.*

Alternativamente, podemos definir un I-map usando d -separación: un DAG sobre un conjunto de variables aleatorias es un I-map del modelo de dependencia M si la d -separación sobre dos conjuntos de variables implica que son independientes.

Consecuentemente, una red bayesiana de nodos X_i , $i = 1, \dots, n$ es útil para encontrar probabilidades $p(x_1, \dots, x_n)$, dado que el conjunto de variables satisface la lista de independencias representada por el grafo. Esta expresión gráfica de la distribución ha ayudado a desarrollar algoritmos para aplicar algunas herramientas de probabilidad como el Teorema de Bayes, o la regla de marginalización. Asimismo, también se puede inferir la distribución de probabilidad a partir de datos, asumiendo que hay una red bayesiana representando el conjunto de independencias entre las variables [11].

5.2.4. Algoritmos de inferencia para redes bayesianas

La inferencia en redes bayesianas es el método por el cual, dado un conocimiento previo o *evidencia*, podemos calcular las probabilidades de que ocurran

ciertos resultados [60]; dado un conjunto de variables aleatorias $\mathbf{X} = \{X_1, \dots, X_n\}$, $\mathbf{E} \subset \mathbf{X}$, con las variables previamente conocidas $\mathbf{E} = \{E_1, \dots, E_m\}$, y un conjunto de valores $\mathbf{e} = \{e_1, \dots, e_m\}$ tales que $e_i \in \Omega_{E_i}$, un *algoritmo de inferencia* o *algoritmo de propagación* es un método que halla el valor de probabilidad

$$p(x_i|\mathbf{e}), \quad x_i \in \Omega_{X_i}, \quad \forall X_i \in \mathbf{X} \setminus \mathbf{E}.$$

donde el conjunto de variables observadas puede ser vacío.

Si el valor de probabilidad $p(x_i|\mathbf{e})$ se calcula de una manera exacta, se denomina *algoritmo exacto de propagación*, mientras que si hacemos una aproximación $\hat{p}(x_i|\mathbf{e})$ de $p(x_i|\mathbf{e})$, se denominará *aproximado* [14].

El problema de inferencia exacta para el caso general de redes bayesianas es NP-complejo [24], y aunque el de inferencia aproximada también lo sea [26], estos métodos requieren menos carga computacional, y funcionan relativamente bien en casos en los que no es posible hacer inferencia exacta.

Los algoritmos de inferencia exacta [28, 60] están basados en los conceptos de teoría de la probabilidad explicados con anterioridad. Entre ellos tenemos el *algoritmo de eliminación de variables* o el *algoritmo de árbol de unión* [45].

Algunos ejemplos de algoritmos de inferencia aproximada son los *métodos de Monte Carlo* [14], entre los cuales encontramos el *algoritmo de muestreo de importancia* [49].

5.2.5. Algoritmos de aprendizaje para redes bayesianas

En este apartado, veremos algunos algoritmos que se pueden usar para obtener redes bayesianas.

5.2.5.1. Conceptos

La tarea de aprender una red bayesiana consiste en, dado un conjunto de muestras de la distribución conjunta, recuperar la estructura del grafo y el conjunto de distribuciones de probabilidad [8].

Formalmente, si tenemos un conjunto $(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)})$ de muestras del conjunto de variables $\mathbf{X} = \{X_1, \dots, X_n\}$, nuestro objetivo es encontrar la red bayesiana que mejor representa los datos. Los pasos para encontrar una solución óptima son [14]:

1. Aprendizaje de la estructura de la red.
2. Aprendizaje del conjunto de parámetros.

El segundo ha sido muy estudiado, y tiene una solución exacta [55]. Para el primero, podemos:

1. Detectar el conjunto de independencias entre las variables
2. Buscar una red que represente correctamente las muestras que tenemos.

1 detecta independencias haciendo tests y añadiéndolas a una lista, para dar una red bayesiana que representa a la mayoría o todas las independencias obtenidas de esa manera [2], mientras que 2 (llamado “buscar y puntuar”) busca en el espacio de todas las posibles estructuras de redes bayesianas, asignando a cada una un valor real o puntuación con una cierta función o métrica, que mide lo adecuada que es la red para esos datos [35].

En ambos casos, el espacio de búsqueda, esto es, el conjunto de todos los DAG posibles entre las n variables, es de tamaño *hiper exponencial*; es un hecho que la búsqueda de la estructura óptima del grafo no se puede hacer con un ordenador [18]. Para encontrar una buena estructura de red se usan algoritmos de búsqueda aproximada. Entre ellos, aquellos que emplean heurísticas o metaheurística suelen encontrar buenas soluciones.

5.3. Entropía conjunta

Pasamos a extender en este apartado el ya introducido concepto de entropía 7, para mejor entenderlo y aplicarlo a nuestro problema.

Sea una variable aleatoria unidimensional X con valores posibles en $\mathcal{X} = x_1, \dots, x_N$ y una función de probabilidad $P(X)$. Un evento con una probabilidad de ocurrencia $P(x)$, $x \in \mathcal{X}$ tiene un contenido informativo $I[P(x)] = -\log P(x)$. El valor esperado de I es la *entropía de Shannon*,

$$S(X) = \sum_{x \in \mathcal{X}} [-\log P(x)] P(x)$$

expresión que mide la incertidumbre en $P(X)$ [47].

La entropía conjunta mide la forma de la función asociada a un conjunto de variables aleatorias [33]. La entropía de Shannon conjunta de las variables aleatorias (X, Y) es

$$S(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} -P(x, y) \log P(x, y)$$

donde $P(X, Y)$ denota la función de probabilidad conjunta. Si X e Y son independientes, su entropía conjunta es la suma de las entropías individuales, $S(X, Y) = S(X) + S(Y)$ [47].

Capítulo 6

Implementación y resultados

En este capítulo presentaremos los resultados experimentales, y qué nos llevó a ellos, siguiendo la metodología ya descrita en 2.1.

6.1. Resultados

Como caso de estudio, nos centraremos en el equipo femenino de Noruega, dejando para trabajos futuros el resto de análisis y preguntas que nos podemos plantear.

Más específicamente, usaremos como fuente la [EURO Femenina de la UEFA 2022](#), en la cual Inglaterra resultó ganadora, y Noruega perdió [un partido](#) contra ellas por ocho goles a cero (es de notar que seis fueron tan solo en la primera parte), constituyendo uno de los márgenes más grandes de victoria (en las fases finales) de la historia del campeonato, además de uno de los partidos (nuevamente, de los finales) en los que más goles se han marcado. Nuestro objetivo será estudiar por qué ocurrió esto, si podemos verlo reflejado de alguna manera en las redes de pases o entropía de ambos equipos. Añadir que luego tampoco pasaron a cuartos de final, al perder contra Austria por [un gol](#). Será interesante comparar más a fondo el desempeño de Noruega a lo largo de este campeonato, del cual (para nuestra agradable sorpresa) se [liberaron los datos](#) el cuatro de agosto, por [StatsBomb](#), una empresa de análisis de fútbol y visualización de datos con sede en el Reino Unido.

Esa derrota fue sorprendente teniendo en cuenta que, si consultamos el [“cuadro de honor de la EURO femenina”](#), vemos que previamente Noruega ha ganado dos veces esta competición, contra una vez que ya ganó Inglaterra; han estado seis veces en la final versus tres que ha estado Inglaterra, y nueve veces en semifinales, contra seis de Inglaterra. Si consultamos las [estadísticas por equipo y jugadoras](#), sin duda Inglaterra y Alemania son las que se imponen, apareciendo Noruega únicamente en un quinto puesto en cuanto a posesión del balón (52.7%, contra un máximo de 63.8% por parte de España), un sexto puesto por una pre-

cisión de pases del 79.7% (contra máximo de un 86.8%, también de España), y otro quinto puesto por catorce paradas de la portera, en comparación con el primer puesto por veintitrés salvadas de Holanda.

Nuestra pregunta es pues ¿qué les pasó a las noruegas este año? El entrenador (ahora ex-entrenador) Martin Sjögren [fue muy criticado](#) por no hacer cambios de formación o jugadores en esos primeros 47 minutos en los que volaron goles, y al ser ese partido de la EURO 2022 contra Inglaterra la mayor derrota de Noruega en la historia; anteriormente habían perdido 0-4 contra Alemania en la Eurocopa de 2009, 0-5 contra China en el mundial de 1999, o [un 0-7](#) contra Holanda el año pasado. [Se culpa](#) a la falta de estructura defensiva y presión ofensiva; alegan que empezaron bien, mas una vez Inglaterra marcó el primer gol en el minuto 12 no pudieron pararles, y fueron cuesta abajo. Veremos si esto se refleja en los datos obtenidos y por qué salió tan mal, considerando en último lugar que Noruega tiene [jugadoras de los mejores equipos del mundo](#).

No obstante, desde que en Inglaterra tienen a la entrenadora Sarina Wiegman, no han perdido los últimos dieciséis partidos. ¿Tan grande es la influencia del entrenador? [En este artículo](#) ponen en evidencia los problemas que se han ido arrastrando desde que Sjögren es entrenador, con un rendimiento lejos del esperado (teniendo en cuenta la buena prestación de las jugadoras cuando juegan en los equipos en que están fichadas) en las últimas competiciones. Conceptualizar y medir cuánto y en qué se diferencia un todo de la suma de sus partes es un desafío no trivial. Una posible dirección para abordar esto consiste en cuantificar hasta qué punto la entropía del todo no es aditiva [57]. Se discute que como equipo cometieron errores a la hora de posicionarse, dejaron libre la zona de defensa y jugadoras que podrían haber sido decisivas se quedaron aisladas, con lo que abrieron las puertas a las inglesas a dominar el juego.

[En el partido contra Austria](#), que debían ganar, no hubo un solo disparo a puerta antes del minuto 88, por lo que siguieron siendo una sombra de sí mismas, pese a que, en sus palabras, usaron los días posteriores para reconstruir su orgullo y autoestima, y recomponerse de tal golpe bajo. No consiguieron estar a la altura, dejar de ser una sombra de ellas mismas, y hubieron de pasar 63 minutos antes de que el entrenador hiciera un cambio (el gol de Austria fue en el minuto 37). Se cometieron también errores aquí y allí, pero está lejos de ser el partido contra Inglaterra. ¿Podremos ver esta diferencia reflejada en las redes de pases o entropía del equipo? Principalmente, ¿es problema de la formación (en [la nueva entrenadora Hege Riise](#) afirma que habría sido mejor jugar 4-3-3, en vez de 4-2-3-1), del entrenador, de las jugadoras que no supieron coordinarse, o simplemente no están a la altura de Inglaterra? Con este trabajo intentaremos encontrar respuestas.

6.2. Elección de fuentes de datos

Escogemos las de [StatsBomb](#), al contener datos de los partidos y equipos que nos interesaban. Otras opciones disponibles en el mercado son [StatsPerform](#), para el cual hay que pagar, [FbRef](#), el cual dispone de datos libres, pero para descargarlos y tratar con ellos es más aparatoso y manual, aparte de que no tiene los datos temporales ni de los pases con tanto detalle y extensión como Statsbomb.

6.3. Paquetes de R

En este apartado describiremos los paquetes de R de los que hemos hecho uso, justificando nuestra elección.

6.3.1. Igraph

La librería y paquete [igraph](#) de R lo hemos escogido por su facilidad a la hora de calcular la entropía; está destinada al análisis de redes y consta por tanto de [funciones sencillas](#) para calcular medidas como *closeness*, *network density*, *centrality*, *betweenness*, *centralization*, *robustness*, *efficiency*, *effectiveness and diversity* sobre grafos, mediante la creación de “[graph dataframes](#)”, que viene genial dado que podemos usarlos con [plot](#) y te traza la red, y si aparte quieres usarlos como dataframes para hacer otros cálculos, también se puede. Un ejemplo de uso podemos verlo [aquí](#), programa en el cual nos inspiramos para trazar nuestra red de pases.

6.3.2. Devtools

Si bien la mayoría de los paquetes de R están en [CRAN](#), hay muchos que vamos a necesitar que se encuentran en Github. [devtools](#) da la posibilidad de descargarlos directamente.

6.3.2.1. Statsbomb

[Este paquete](#) lo empleamos, una vez decidido que íbamos a hacer uso de sus datos, para poder quedarnos con las columnas que nos interesaban y demás operaciones, esto es, *parsear* la información.

6.3.3. Tidyverse

[Tidyverse](#) contiene paquetes como [dplyr](#) que son útiles para manipular, explorar y visualizar datos, y prácticamente es un estándar en estadística y ciencia de datos. [Sus ventajas](#) incluyen funciones consistentes y uso a lo largo de todo el *workflow*, lo que se ve reflejado en una mayor productividad, y no nos hizo dudar

al escogerlo. [Más ventajas](#) incluye que carga una *suite* complete y extensiva de las últimas herramientas de tratamiento de datos, haciendo más fácil el desarrollo al despreocuparte así de tener que andar con librerías y paquetes de aquí y allí, que son intuitivamente distintos a la hora de pasar parámetros o tratar las tablas.

6.3.3.1. Ggplot2

Está contenido dentro de *tidyverse*, y lo usamos [ggplot](#) por lo sencillo que hace crear gráficos de caja, de dispersión, temporales o de radar. De hecho, fue una de las razones por las que elegimos usar R para este proyecto; es una librería extremadamente trabajada y documentada, que nos representa datos de un [dataframe](#) complejos de manera personalizada, sin mucho problema; justo lo que necesitábamos, teniendo en cuenta el gran tamaño de las tablas que manejamos. Además es el paquete de R más popular para visualizar datos.

6.3.3.2. Dplyr

[Este paquete](#) hace extremadamente fácil reducir la cantidad de datos que tenemos, gracias a funciones como *filter*, *slice*, *arrange*, *select*, *rename*, *mutate*, *relocate* o *summarise*. [Aquí](#) podemos consultar más a fondo.

6.4. Diseño experimental

En este apartado veremos los experimentos realizados, y lo obtenido.

En 6.1 hemos representado cada jugador con un peso correspondiente a la suma de sus interacciones, y al calcular la entropía de los pases a lo largo de la competición, podemos observar lo compacta que es la red, lo bien conectadas que están todas, y cómo en el centro (y por tanto las que tocan más balón y hacen más juego) se encuentran aquellas que han jugado más minutos a lo largo de la competición, son las que han marcado más goles, o juegan como centrocampistas, seguidas por un alto número de defensas, lo que ciertamente llama la atención, dado que normalmente suelen ser las centrocampistas las que mueven más la pelota, y nos indica que defendieron bien. En definitiva, de la red podemos ver que todas participan y se coordinan de manera óptima.

En 6.2 estudiamos cómo es consistente con lo ya comentado, y hay poca variabilidad dentro de la entropía, conteniendo únicamente extremos mínimos correspondientes a la portera y máximos correspondientes a las defensas. Nos reafirma en que la entropía refleja bien el comportamiento en el campo, la posición de las jugadoras y cuántos minutos jugaron.

En 6.3 vemos las redes de peso sin simplificar las aristas, lo que nos permite apreciar lo densa que es, pero a su vez están bien repartidas, y no apelonadas como en Noruega ??.

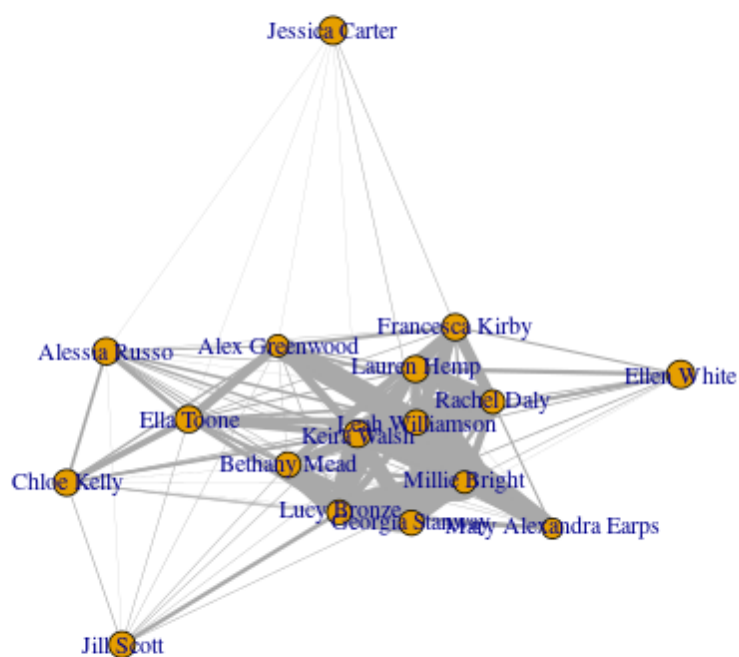


Figura 6.1: Red de pases total de Inglaterra simplificado en la EURO 2022

En 6.4 y 6.5 vemos que no es tanto una estructura casi romboidal, sino circular; están más juntas y sin una forma reticular clara.

En 6.7 y 6.8 vemos cómo Noruega se mantuvo con una entropía bastante alta en general (lo que podemos ver reflejado en 6.6), mientras que la de Inglaterra es una curva decreciente, por lo que tienen orden y desorden, y consiguen un buen equilibrio.

6.5. Costes

En cuanto a coste de amortización anuales, he usado mi portátil Asus Vivobook 14, comprado hace casi tres años por 799€, lo que se corresponde a 150€. El coste de desarrollo, teniendo en cuenta que hemos empleado 571 horas, entre 40 que son las horas que se echan por semana en un trabajo a tiempo

Concepto	Coste unitario	Unidades	Total
Amortización portátil	150€	1	150€
Ventilador	12€	1	14€
Costes laborales	1750€	3.5	6125€
Total			6289€

Tabla 6.1: Costes el proyecto en el escenario “ingeniera junior”

Concepto	Coste unitario	Unidades	Total
Amortización portátil	150€	1	150€
Ventilador	12€	1	14€
Costes laborales	2000€	3.5	7000€
Total			7164€

Tabla 6.2: Costes el proyecto en el escenario “analista de datos junior”

completo, nos quedamos con 14 semanas, esto es, unos tres meses y medio. Si consultamos [LinkedIn](#), las ofertas de trabajo para desarrolladores *junior* están en torno a los 21.000€ al año, lo que viene a ser 1750€ por mes, añadiendo los 14€ que costaron el ventilador de Alehop mencionado en los agradecimientos. Esto se muestra en la tabla 6.1

Sin embargo, el trabajo realizado aquí corresponde más bien al de un analista junior de datos deportivos. Según [Payscale](#), el sueldo medio está en torno a los 24000€. En este escenario los costes serían los indicados en la tabla 6.2.

El coste de explotación, despliegue y producción se corresponderán a servicios de desarrollo a medida, adaptación, implantación, cursos, entre otros, pero este tipo de trabajos se suelen hacer en nómina o bien vendiendo los informes; el coste del informe trataría de ponerse de acorde con las horas empleadas y la amortización del equipo durante el tiempo necesario. Una vez llevado a cabo todo el análisis inicial y la exploración de los métodos, la elaboración de un nuevo informe de este tipo se estima que duraría unos 15 días.

La biblioteca tendremos que mirarla cada cierto tiempo, actualizar dependencias, responder a los issues, para lo que echaremos 40 horas al mes, que serán costes de producción que tendremos que factorizar en el coste del producto.

Lauren Hemp	0.8735
Leah Williamson	0.8205
Bethany Mead	0.8304
Rachel Daly	0.7756
Keira Walsh	0.8783
Millie Bright	0.7238
Lucy Bronze	0.8
Mary Alexandra Earps	0.6649
Georgia Stanway	0.813
Francesca Kirby	0.8765
Ellen White	0.9138
Ella Toone	0.8881
Alessia Russo	0.8826
Chloe Kelly	0.8641
Alex Greenwood	0.7233
Jill Scott	0.8581
Jessica Carter	0.9172

Figura 6.2: Entropía por jugadoras de Inglaterra en la EURO 2022

Lauren Hemp	0.8735
Leah Williamson	0.8205
Bethany Mead	0.8304
Rachel Daly	0.7756
Keira Walsh	0.8783
Millie Bright	0.7238
Lucy Bronze	0.8
Mary Alexandra Earps	0.6649
Georgia Stanway	0.813
Francesca Kirby	0.8765
Ellen White	0.9138
Ella Toone	0.8881
Alessia Russo	0.8826
Chloe Kelly	0.8641
Alex Greenwood	0.7233
Jill Scott	0.8581
Jessica Carter	0.9172

Figura 6.3: Entropía por jugadoras de Inglaterra en la EURO 2022

Julie Blakstad	0.8482
Maria Thorisdottir	0.7752
Ingrid Syrstad Engen	0.897
Maren Nævdal Mjelde	0.8021
Frida Maanum	0.8637
Guro Reiten	0.8261
Ada Hegerberg	0.8937
Anja Sønstevoid	0.8837
Caroline Graham Hansen	0.9117
Amalie Vevle Eikeland	0.8969
Guro Pettersen	0.8711
Tuva Hansen	0.9038
Vilde Bøe Risa	0.9185
Karina Sævik	0.8917
Guro Bergsvand	0.779
Anna Langås Jøsendal	0.9591
Celin Bizet Ildhusøy	0.9474
Sophie Roman Haug	0.9183

Figura 6.5: Entropía por jugadoras de Noruega en la EURO 2022

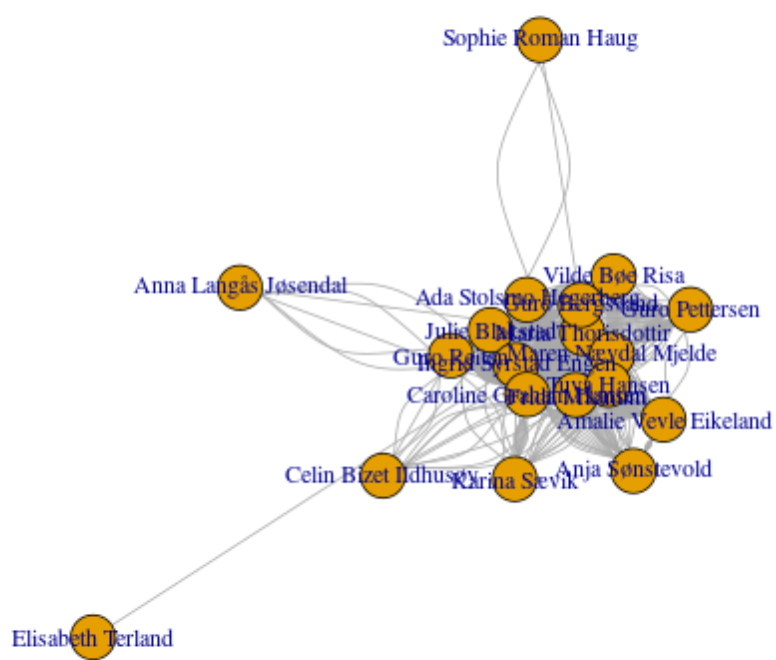


Figura 6.6: Red de pases total de Noruega en la EURO 2022

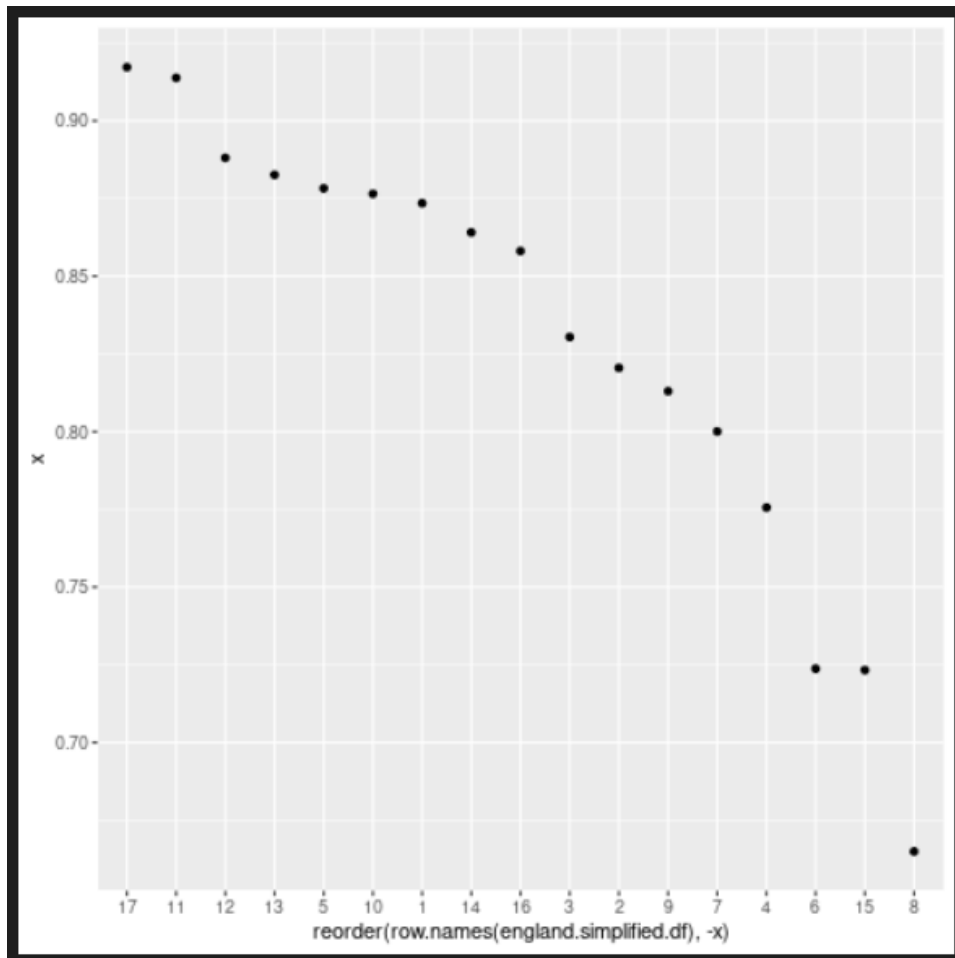


Figura 6.7: Entropía de Inglaterra

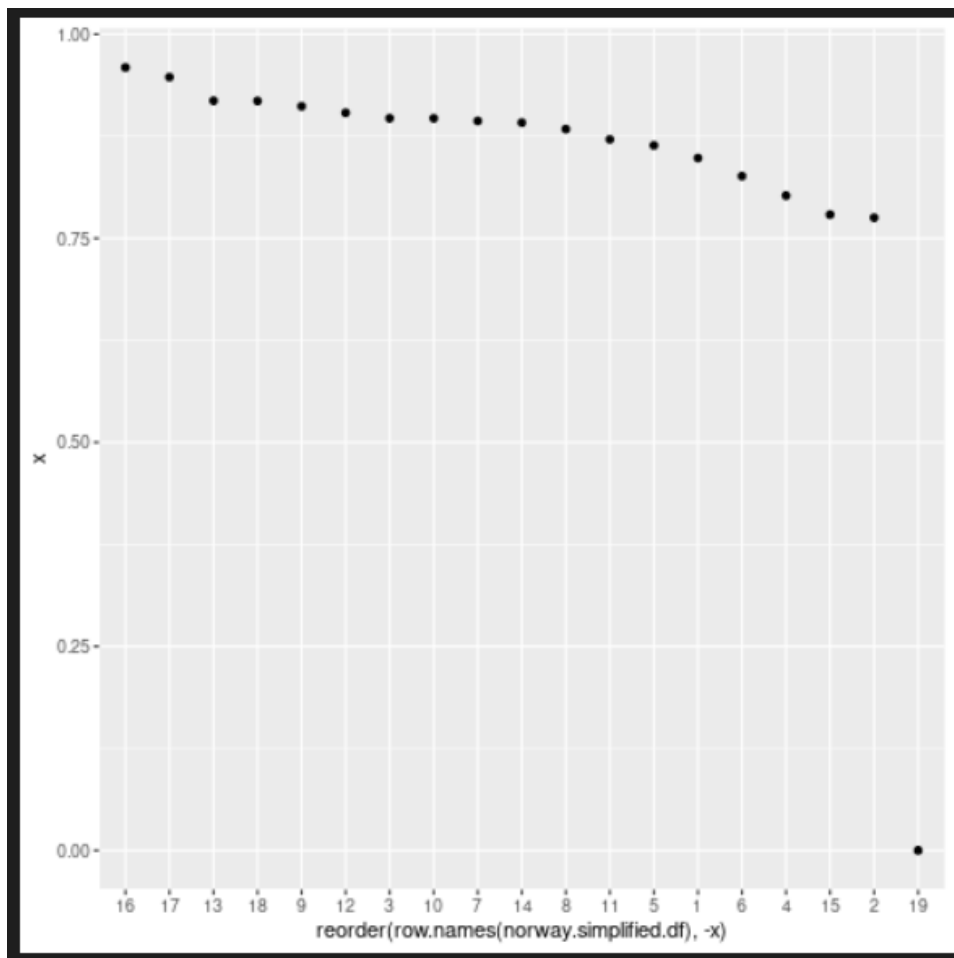


Figura 6.8: Entropía de Noruega

Capítulo 7

Conclusiones y trabajos futuros

En este trabajo hemos intentado modelar la manera en que los equipos consiguen, en mayor o menor medida, alcanzar un equilibrio entre organización y desorganización. Hemos visto reflejado el hecho de que una mayor entropía implica una posición más baja en un campeonato, y en trabajos futuros nos gustaría estudiarlo más en profundidad. En este apartado veremos específicamente las conclusiones a las que hemos llegado a partir de este estudio preliminar, y los caminos que se pueden seguir a partir de aquí.

Hemos de tener en mente que por un lado, un equipo ha de intentar minimizar su entropía para poder maximizar la comunicación entre sus jugadores, a la vez que por el otro, deben maximizar su entropía para maximizar sus grados de libertad y evitar que el rival identifique un patrón ordenado de comportamiento. [56]

El desempeño de un equipo de fútbol está claramente influenciado por el equipo contra el que juegan, así que para trabajos futuros dejamos lo de analizar, visualizar y estudiar el partido en el que se enfrentaron Inglaterra y Noruega. [59] teorizó sobre la naturaleza adaptativa de un jugador, pero no midió el desempeño de los jugadores en términos de entropía. Nosotros hemos estudiado el rendimiento de un equipo de fútbol en base a la entropía de los pases a lo largo de una competición, y queda ver cómo esta medida se adapta al equipo rival, o comparar cómo varía de partido a partido.

Otra cuestión importante que no podemos dejar de lado es que la correlación entre el buen desempeño de un equipo y la entropía que se obtiene de sus redes de pases puede tener su base en la posesión del balón del mismo; cuánto más tiempo lo tengan, más pases harán, luego más se pueden organizar y son los que dominan el juego, pero puede salir una entropía más alta, que no es que indique más caos o desorden, sino que hay más grados de libertad, y no tiene por qué significar que vayan a tener una posición más baja en la competición. Sería interesante pues ampliar este estudio que hemos hecho incluyendo los datos correspondientes a la posesión del balón de cada equipo en la tabla de distribuciones condicionadas de

la red bayesiana, o como causalidad. En todo caso, introducir introducir redes bayesianas de las maneras que estudiábamos en el apartado de análisis puede ser muy esclarecedor y ampliar lo que se obtiene, y lo dejamos para próximos trabajos.

Bibliografía

- [1] Salmerón A., Rumi R., Langseth H., Nielsen T.D., and Madsen A.L. A review of inference algorithms for hybrid Bayesian networks. *Journal of Artificial Intelligence Research*, 62:799–828, 2018.
- [2] Silvia Acid, Luis M. de Campos, Juan M. Fernandez-Luna, Susana Rodriguez, Jose Maria Rodriguez, and Jose Luis Salcedo. A comparison of learning algorithms for Bayesian networks: a case study based on data from an emergency medical service. *Elsevier Health*, 30:215–232, 2004.
- [3] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.
- [4] Jose M. Amigo, Samuel G. Balogh, and Sergio Hernandez. A brief review of generalized entropies. *Entropy*, 20(813), 2018.
- [5] E. Arriaza-Ardiles, J.M. Martín-González, M.D. Zuniga, J. Sánchez-Flores, Y. de Saa, and J.M. García-Manso. Applying graphs and complex networks to football metric interpretation. *Human Movement Science*, 57:236–243, 2018.
- [6] Yiping Bai, Yuxuan Xing, and Jiansong Wu. Integrating knowledge graph, complex network and Bayesian network for data-driven risk assessment. *Chemical Engineering Transactions*, 90:31–36, May 2022.
- [7] Kent Beck, Mike Beedle, Arie van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, Jon Kern, Brian Marick, Robert C. Martin, Steve Mellor, Ken Schwaber, Jeff Sutherland, and Dave Thomas. Principios del manifiesto Ágil.
- [8] Stefano Beretta, Mauro Castelli, Ivo Gonçalves, Roberto Henriques, and Daniele Ramazzotti. Learning the structure of Bayesian networks: A quantitative assessment of the effect of different algorithmic schemes. *Complexity*, 2018, 2018.

-
- [9] Joel Brooks, Matthew Kerr, and John Guttag. Using machine learning to draw inferences from pass location data in soccer. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(5):338–349, 2016.
- [10] Javier Bundio and Conde Matías. Exploraciones en fútbol y redes sociales. análisis del desempeño deportivo durante la eurocopa 2004 a partir del análisis de redes sociales. *Redes. Revista hispana para el análisis de redes sociales*, 13, ene. 2008.
- [11] W. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8(2):195–210, 1996.
- [12] Glymour C., Zhang K., and Spirtes P. Review of causal discovery methods based on graphical models. *frontiers*, 2019.
- [13] Guido Caldarelli. *Scale-free networks: complex webs in nature and technology*. Oxford University Press, 2007.
- [14] Andrés Cano, Serafín Moral, and Antonio Salmerón. *Algorithms for Approximate Probability Propagation in Bayesian Networks*, pages 77–99. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.
- [15] Pedro Cerqueira, Luiz Nakamura, Rodrigo Pescim, and Roseli Leandro. Investigating the underlying causal network on european football teams. *Journal of data science: JDS*, 15, 04 2017.
- [16] Saptarshi Chakraborty, Debolina Paul, and Swagatam Das. t-entropy: A new measure of uncertainty with some applications. *CoRR*, abs/2105.00316, 2021.
- [17] Victor Chazan-Pantzalis. Sports analytics algorithms for performance prediction. *IEEE*, 2020.
- [18] David Maxwell Chickering. *Learning Bayesian Networks is NP-Complete*, pages 121–130. Springer New York, New York, NY, 1996.
- [19] C. G. Chua and A. T. C. Goh. A hybrid Bayesian back-propagation neural network approach to multivariate modelling. *International Journal for Numerical and Analytical Methods in Geomechanics*, 27(8):651–667, 2003.
- [20] A. C. Constantinou. Dolores: a model that predicts football match outcomes from all over the world. *SpringerLink*, pages 49–75, 2018.
- [21] Anthony Constantinou and Norman Fenton. Towards smart-data: Improving predictive accuracy in long-term football team performance. *Knowledge-Based Systems*, 124:93–104, 2017.

-
- [22] Anthony C. Constantinou and Norman Fenton. Things to know about Bayesian networks: Decisions under uncertainty, part 2. *Significance*, 15(2):19–23, 2018.
- [23] Anthony C. Constantinou and Norman Elliott Fenton. Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, 9(1):37–50, 2013.
- [24] Gregory F. Cooper. The computational complexity of probabilistic inference using bayesian belief networks. *Artificial Intelligence*, 42(2):393–405, 1990.
- [25] Carlos Cotta, Antonio Mora, Juan Julián Merelo, and Cecilia Merelo-Molina. A network analysis of the 2010 FIFA world cup champion team play. *Journal of Systems Science and Complexity*, 2013.
- [26] Paul Dagum and Michael Luby. Approximating probabilistic inference in bayesian belief networks is np-hard. *Artificial Intelligence*, 60(1):141–153, 1993.
- [27] Amir Hossein Darooneh, Ghassem Naeimi, Ali Mehri, and Parvin Sadeghi. Tsallis entropy, escort probability and the incomplete information theory. *Entropy*, 12(12):2497–2503, 2010.
- [28] Castillo E., Gutierrez J.M., and Hadi A.S. *Expert Systems and Probabilistic Network Models*. Springer New York, 1997.
- [29] Carlos Echegoyen, Alexander Mendiburu, Roberto Santana, and Jose A. Lozano. Estimation of Bayesian networks algorithms in a class of complex networks. In *IEEE Congress on Evolutionary Computation*, pages 1–8, 2010.
- [30] N Fenton, M Neil, and D Marquez. Using Bayesian networks to predict software defects and reliability. *Proc ImechE*, 222, 2008.
- [31] Juan Manuel Fernández Luna. *Modelos de recuperación de información basados en redes de creencia*, 2013.
- [32] Free Software Foundation. GNU General Public License. <http://www.gnu.org/licenses/gpl.html>.
- [33] Korn G.A. and Korn T.M. *Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review*. McGraw-Hill, 1968.
- [34] Zoubin Ghahramani and Matthew Beal. Propagation algorithms for variational Bayesian learning. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

-
- [35] Zoubin Ghahramani and Matthew J. Beal. Propagation algorithms for variational Bayesian learning. *Advances in Neural Information Processing Systems*, 2001.
- [36] Francisco González-Gómez, Andrés J. Picazo-Tadeo, and Miguel A. García-Rubio. The impact of a mid-season change of manager on sporting performance. *Sport, business and management: An International journal*, 1(1):28–42, 2011.
- [37] Chris Goumas. Modelling home advantage for individual teams in uefa champions league football. *Journal of Sport and Health Science*, 6(3):321–326, 2017.
- [38] Venkat N. Gudivada, Dhana Rao, and Vijay V. Raghavan. Chapter 9 - big data driven natural language processing research and applications. In Venu Govindaraju, Vijay V. Raghavan, and C.R. Rao, editors, *Big Data Analytics*, volume 33 of *Handbook of Statistics*, pages 203–238. Elsevier, 2015.
- [39] Jahn K. Hakes and Raymond D. Sauer. An economic evaluation of the moneyball hypothesis. *Journal of Economic Perspectives*, 20(3):173–186, September 2006.
- [40] Pearl J. and Rusell S. Bayesian networks. *eScholarship*, 2000.
- [41] David Kilpatrick. Inverting the pyramid: A history of football tactics. *Journal of Sport History*, 38(3):530–531, 2011.
- [42] Uffe B. Kjaerulff and Anders L. Madsen. *Probabilistic Networks for Practitioners — A Guide to Construction and Analysis of Bayesian Networks and Influence Diagrams*. Springer New York, NY, 2006.
- [43] P. V. Kumar, B. Naveenand Kumar. Learning parameters for hybrid Bayesian network. *Springer*, pages 255–260, 2021.
- [44] Radouane Laggoune, Ait Mokhtar El Hassene, and Alaa Chateauneuf. Bayesian networks for the evaluation of complex system’s availability. *CEUR Workshop Proceedings*, 1256, 09 2014.
- [45] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.
- [46] Jegoo Lee, Stephen P Borgatti, Jose Luis Molina, and Juan J Merelo Guervos. Who passes to whom: Analysis of optimal network structure in soccer matches. In *XXV International Sunbelt Social Network Conference*, 2005.
- [47] A. M. Lopes and J. A. Tenreiro Machado. Entropy analysis of soccer dynamics. *Entropy (Basel, Switzerland)*, 21(187), 2019.

-
- [48] Ashique Rupam Mahmood. Structure learning of causal bayesian networks: A survey, 2011.
- [49] Irene Martínez, Carmelo Rodríguez, and Antonio Salmerón. Dynamic importance sampling in bayesian networks using factorisation of probability trees. pages 187–194, 01 2006.
- [50] Johann H. Martínez, David Garrido, José L Herrera-Diestra, Javier Busquets, Ricardo Sevilla-Escoboza, and Javier M. Buldú. Spatial and temporal entropies in the Spanish football league: A network science perspective. *Entropy*, 22(2), 2020.
- [51] Tomasz Maszczyk and Wlodzislaw Duch. Comparison of shannon, renyi and tsallis entropy used in decision trees. *Springer*, 2008.
- [52] Angélica Sousa da Mata. Complex networks: a mini-review. *Brazilian Journal of Physics*, 2020.
- [53] J. J. Merelo. Cómo formular buenos objetivos en un trabajo, 2019.
- [54] J. J. Merelo. Agile (data) science: a (draft) manifesto. *arXiv*, 2022.
- [55] Richard Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 01 2003.
- [56] Y. Neuman, N. Israeli, D. Vilenchik, and Y. Cohen. The adaptive behavior of a soccer team: an entropy-based analysis. *Entropy*, 2018.
- [57] Yair Neuman, Denis Noble, and Yochai Cohen. Is the whole different from the sum of its parts? a proposed procedure for measuring divergence from additivity. *International Journal of General Systems*, 47(7), 2018.
- [58] F. Owramipour, P. Eskandarian, and F. S. Mozneb. Football result prediction with Bayesian network in Spanish league-Barcelona team. *International Journal of Computer Theory and Engineering*, 2013.
- [59] P. O’Donoghue. Interacting performances theory. *Int. J. Perform. Anal. Sport*, 9(24-46), 2009.
- [60] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.
- [61] Richard Pollard. Charles Reep (1904-2002): pioneer of notational and performance analysis in football. *Journal of Sports Sciences*, 20(10):853–855, 2002.
- [62] Nazim Razali, Aida Mustapha, Faiz Ahmad Yatim, and Ruhaya Ab Aziz. Predicting football matches results using Bayesian networks for English Premier League (EPL). *IOP Conference Series: Materials Science and Engineering*, 226:012099, 2017.

- [63] Alfonso E. Romero. *Document Classification Models Based On Bayesian Networks*. PhD thesis, April 2010.
- [64] Cosma Rohilla Shalizi. *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press, 2021.
- [65] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 1948.
- [66] Xin-She Yang. 2 - mathematical foundations. In Xin-She Yang, editor, *Introduction to Algorithms for Data Mining and Machine Learning*, pages 19–43. Academic Press, 2019.