UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

School of Informatics

Characterization of trajectories in football matches from anomalous diffusion models.

End of Degree Project

Bachelor's Degree in Data Science

AUTHOR: Castillo Esteve, Joan

Tutor: Conejero Casares, José Alberto

ACADEMIC YEAR: 2022/2023

Resum

L'anàlisi esportiu és un camp el qual ha anat creixent molt estos últims anys. De la mà del *Big Data* s'han començat a fer molts més estudis i amb més possibilitats. En el món del futbol la Ciència de Dades junt amb el machine learning ha canviat per complet molts dels anàlisis que existien i s'han creat uns altres de nous. En este projecte hem seguit la metodologia *CRISP-DM* [1] utilitzant principalment *python* com a llenguatge de programació. S'ha fet ús de diverses llibreries com poden ser *pandas, matplotlyb,fbm,sklearn*, entre altres. El projecte es dividix en dos anàlisis, el principal i més innovador, en el qual fem ús de la difusió anòmala per a obtindre informació de les distintes trajectòries del baló durant un partit. L'anàlisi secundària es basa a observar els distints estils de joc de les seleccions per mitjà de grafos i teoria de xarxes, en este cas, xarxes de passes. En quant a la part de difusió anòmala s'han utilitzat models de xarxes neuronals convolucionals que han sigut triats després de realitzar diverses proves i descartar els pitjors models.

Paraules clau: anàlisi esportiu, Big Data, machine learning, trajectòria, difusió anòmala, xarxes neuronals convolucionals, grafos, teoria de xarxes

Resumen

El análisis deportivo es un campo el cual ha ido creciendo mucho estos últimos años. De la mano del *Big Data* se han empezado a hacer muchos más estudio y con más posibilidades. En el mundo del fútbol la Ciencia de Datos junto con el machine learning ha cambiado por completo muchos de los análisis que existían y se han creado otros nuevos. En este proyecto hemos seguido la metodología *CRISP-DM* [1] utilizando principalmente *python* como lenguaje de programación. Se ha hecho uso de diversas librerías como pueden ser *pandas, matplotlyb,fbm,sklearn*, entre otras. El proyecto se divide en dos análisis, el principal y más innovador, en el cual hacemos uso de la difusión anómala para obtener información de las distintas trayectorias del balón durante un partido. El análisis secundario se basa en observar los distintos estilos de juego de las selecciones mediante grafos y teoría de redes, en este caso, redes de pases. En cuanto a la parte de difusión anómala se han utilizado modelos de redes neuronales convolucionales que han sido elegidos tras realizar varias pruebas y descartar los peores modelos.

Paraules clau: análisis deportivo, Big Data, machine learning, trayectoria, difusión anómala, redes neuronales convolucionales, grafos, teoría de redes

Abstract

Sports analysis is a field which has been growing a lot in recent years. With the help of Big Data, many more studies have begun to be carried out and with more possibilities. In the world of football, Data Science together with machine learning has completely changed many of the existing analyses and new ones have been created. In this project we have followed the methodology *CRISP-DM* [1] using mainly *python* as programming language. We have made use of several libraries such as *pandas, matplotlyb,fbm,sklearn*, among others. The project is divided into two analyses, the main and most innovative, in which we make use of anomalous diffusion to obtain information on the different trajectories of the ball during a match. The secondary analysis is based on observing the different styles of play of the teams using graphs and network theory, in this case, passing networks. As for the anomalous diffusion part, convolutional neural network models have been used, which have been chosen after carrying out several tests and discarding the worst ones.

Key words: sports analysis, Big Data, machine learning, trajectory, anomalous diffusion, convolutional neural networks, graphs, network theory

Índice general

Índice de figuras

CAPÍTULO 1
Introducción

Vivimos una era de ciencia e innovación con nuevas formas de diferenciarse de la competencia. El fútbol no es la excepción. Ante este nuevo panorama, muchos equipos de fútbol decidieron apostar por la Big Data como un elemento diferenciador para competir con éxito en el nuevo escenario global. Actualmente los equipos disponen de sofisticados *softwares* y dispositivos para medir el desempeño de sus futbolistas.

El análisis de los partidos de fútbol ha ido evolucionando mucho con el paso del tiempo. Aunque no podemos saber exactamente cuando empezaron a analizarse los partidos de fútbol, sí se tiene constancia de que el 18 de marzo de 1950 , Charles Reep utilizó una libreta para anotar información como los goles, jugadas o posiciones de los jugadores en el partido disputado entre el Swidon Town y el Bristol Rovers, dos equipos ingleses.

Hoy en día el análisis va más allá de anotar los datos en una libreta. La llegada del Big Data ha facilitado mucho el estudio de los partidos, por lo que se ha convertido en un aspecto fundamental para todos los equipos, ya que un buen análisis puede dar mucha ventaja en los encuentros.

En los últimos años numerosos estudios van dirigidos al desarrollo de herramientas predictivas como pueden ser modelos de regresión o redes neuronales. Estas herramientas se utilizan para ofrecer un análisis sobre el rendimiento futuro de determinados futbolistas y las posibilidades de ciertos fichajes, de análisis y el uso de la estadística para mejorar el rendimiento de los deportistas. En el artículo [7], publicado en Istmo, se hablaba de usar aplicaciones para analizar en tiempo real el vídeo de los partidos y algunos indicadores como el porcentaje de pases acertados o la potencia de cada jugador. Como resultado, generaban informes para facilitar la labor del equipo técnico.

Una técnica que aún no se está aplicando es el estudio de las trayectorias caracterizando los procesos de difusión anómala subyacentes, bien sea de la pelota, o de los jugadores individual o colectivamente. Este tipo de difusión se ha utilizado principalmente para describir distintos escenarios físicos

1.1 Motivación

El fútbol es el segundo deporte más practicado del mundo, con aproximadamente 1000 millones de personas, solo por detrás de la natación [5]. Aunque no es el deporte más practicado si es el más seguido, con unos 4000 millones de aficionados lo que representa un impacto social y económico enorme, generando pasiones y sobre todo generando mucho dinero. Este potencial económico conlleva mucha presión por los resultados y por ello hay mucho interés en controlarlos.

Desde hace mucho tiempo se analizan los partidos para optimizar el rendimiento de los equipos, mejorar los resultados deportivos y con ellos los económicos. Estos análisis del comportamiento de los jugadores y equipos son cada día más frecuentes y basados en numerosos y diversos aspectos del juego. Estos análisis se pueden enfocar en como plantea y juega el partido un equipo [4]

Este trabajo basado en el análisis de las trayectorias del balón supone una forma más de observación que pretende obtener indicadores que supongan una mejora en el rendimiento de los equipos y de sus resultados deportivos.

Este sistema de análisis va más allá de las estadísticas básicas y profundiza en temas relacionados con las dinámicas internas del juego, como son las trayectorias del balón.

En otros deportes como el béisbol o el baloncesto se ha utilizado el análisis de datos para optimizar el rendimiento de los equipos, como por ejemplo podemos ver en la película *Moneyball*. Esta película se centra en un equipo de béisbol que está en una mala racha y el gerente del equipo decide contratar a un joven que mediante estadísticas y análisis de datos empieza a hacer fichajes con un presupuesto reducido. Estos nuevos jugadores no son los más famosos, pero mediante el análisis de sus características crean un equipo lo más completo y competitivo posible. Queremos seguir esta linea de análisis para intentar optimizar el rendimiento de equipos de fútbol. Esta visión del análisis deportivo es muy común hoy en día y en el fútbol hay muchos estudios que intentan hacer fichajes que puedan ser útiles en el equipo con poco presupuesto, como podemos ver en [25]

Mediante técnicas de teorías de redes, la cual ya se ha empezado a utilizar en el análisis deportivo [6] y las propiedades de la difusión asociadas a las trayectorias, queremos ver si somos capaces de caracterizar los diferentes juegos de los equipos con el fin de introducir nuevas medidas que puedan ser útiles a la hora de analizar los partidos de fútbol.

1.2 Objetivos

El objetivo principal del proyecto es estudiar las trayectorias del balón en un campo de fútbol aplicando modelos de clasificación y regresión para poder caracterizar el estilo de juego de diferentes equipos y cómo esto evoluciona en función del desarrollo de los partidos. Nuestros datos los obtenemos de *Statsbomb* el cual nos proporciona datos de distintas competiciones y partidos. Estos datos tienen la particularidad de que en ellos se encuentran coordenadas de las acciones, es

decir en que parte del campo ocurre cada evento. Para estudiar estas trayectorias vamos a utilizar modelos los cuales vamos a alimentar con datos ficticios.

Para generar los datos partimos con la hipótesis de que el movimiento de la pelota no es aleatorio en el transcurso del partido, puesto que hay intenciones tanto de avanzar hacia la portería contraria, como eventualmente de hacer cambios de lado del juego. Para ello, generaremos datos sintéticos con los que entrenar los modelos, en particular, utilizando modelos de movimiento fraccionario Browniano.

Estos comportamientos se pueden estimar a partir del cálculo del estudio de la varianza media de los desplazamientos del balón a lo largo del tiempo (Mean Square Displacement) denotado por $MSD(t)$. En el caso de que el movimiento fuera aleatorio, el $MSD(t)$ se comporta como Dt^{α}, donde t representa el tiempo, D el coeficiente de difusividad, y α el exponente de difusión.

Cuando $\alpha = 1$ nos encontramos ante un movimiento Browniano (aleatorio), cuya distribución viene dada por una Gaussiana, no obstante, dada la dinñamica del balón entendemos que dicho movimiento no es aleatorio y queremos caracterizar como es el proceso que lo rige en cada caso.

Otro de nuestros objetivos será clasificar a los equipos según las trayectorias del balón que dibujen durante el partido. Las clasificaremos en 2 clases, si son superdifusivas en las cuales $\alpha > 1$ o si son subdifusivas donde $\alpha < 1$. Durante el proyecto amos a utilizar movimiento fraccionario Browniano *(fractional Brownian motion)* en la generación de trayectorias. Este movimiento viene determinado en principio por el exponente de *Hurst*.

El α o exponente de difusión anómala que hemos mencionado anteriormente es igual a 2 veces el exponente *Hurst* el cual determina el tipo de movimiento generado por el proceso. El exponente de *Hurst* se denomina "índice de dependencia.º "índice de dependencia a largo plazo". Este indica la tendencia de nuestra trayectoria a seguir con la misma dinámica, es decir, si viene de un pase corto seguir con un pase corto o cambiar a uno largo.

Además, intentaremos predecir qué α tienen las trayectorias y estudiar las diferencias entre equipos y momentos del partido (si van ganando, empatando, o perdiendo). Esperamos que esta aproximación pueda ser incorporada en un futuro para la clasificación de trayectorias y pueda ser de utilidad para mejorar el rendimiento de los equipos de fútbol.

1.3 Estructura de la memoria

A continuación vamos a hacer una breve explicación de los distintos capítulos que forman la memoria del proyecto. Estos capítulos son:

- **Introducción:** en esta sección introducimos el proyecto, dando información sobre la motivación del trabajo y explicando brevemente el mundo del análisis de fútbol y la nueva técnica(difusión anómala) que vamos aplicar para intentar obtener información relevante y que pueda servir de utilidad.

- **Estado del arte:** en este apartado damos contexto de como es hoy en día el análisis de partidos de fútbol y como el *machine learning* y la ciencia de datos forman parte de este mundo. También contextualizamos para qué sirve y en qué experimentos se ha hecho uso de la difusión anómala.
- **Metodología:** una vez finalizado el estado del arte podemos plantearnos como empezar el proyecto. En esta sección explicamos cuáles van a ser las fases de nuestro proyecto y que metodología hemos seguido. También se incide en que tecnología nos hemos apoyado para poder hacer el trabajo.
- **Métodos utilizados:** en esta sección explicamos que información vamos a utilizar en el proyecto, es decir, los datos y que métodos vamos a emplear sobre ellos para obtener distintos resultados que más adelante estudiaremos para obtener información relevante.
- **Modelos físicos de difusión:** el análisis deportivo es muy común hoy en día por lo que existen muchos tipos diferentes de análisis que se pueden hacer. En este proyecto la parte novedosa es la difusión, por lo que en este apartado hacemos una explicación de qué son y cómo funcionan los modelos de difusión, en concreto, la difusión anómala. También explicamos los distintos tipos de desplazamiento que pueden tener las trayectorias de las partículas (en nuestro caso el balón de fútbol) cuando se propagan.
- **CNN:** las redes neuronales convolucionales son el principal modelo de *machine learning* que vamos a utilizar, por lo que en este apartado hacemos una explicación de qué son y cómo surgieron. También explicamos su arquitectura y las distintas capas que pueden formarla. Para finalizar en este apartado damos información de cómo se han creado los distintos modelos de redes neuronales convolucionales que se han utilizado en el proyecto.
- **Resultados:** tras aplicar los modelos y las técnicas de análisis que utilizamos en el proyecto mostramos los resultados. Este apartado incluye gráficas que ayudan a explicar que resultados hemos obtenido ya sea de la parte de difusión o de la parte del análisis más usual.
- **Conclusiones:** en esta sección escribimos las conclusiones que hemos obtenido tras estudiar los resultados del apartado anterior. Damos una visión general de como ha ido el trabajo y si ha alcanzado nuestras expectativas.
- **Anexos:** en el apartado de anexos complementamos el contenido del trabajo aportando información que puede ser útil pero no es imprescindible, por lo que no se incluye en la parte principal de la memoria. También relacionamos el trabajo con las ODS u objetivos de desarrollo sostenible.

CAPÍTULO 2
Estado del arte

2.1 Difusión

Hoy en día la difusión se ha utilizado en una gran variedad de campos, desde la biología hasta el mercado de valores, pero principalmente esta se utiliza en la física, química y biología [34].

En los últimos años, los avances en las técnicas de fluorescencia han aumentado enormemente la disponibilidad de trayectorias de alta precisión de moléculas individuales en sistemas vivos [27], lo que ha producido un impulso creciente para desarrollar métodos de cuantificación de la difusión anómala [18]. Las técnicas de fluorescencia son metodologías biofísicas que explotan el fenómeno de la fluorescencia para examinar y analizar las interacciones proteína-proteína, proteína-ácido nucleico, ligando-receptor y ligando-lípido. Estas técnicas también son útiles en el estudio de la conformación y orientación de las proteínas, así como de las constantes de difusión y unión.

Con respecto a la difusión anómala, esta se ha utilizado para describir muchos escenarios físicos, como la difusión de proteínas dentro de las células o a través de medios porosos [15]. En los experimentos típicos destinados a comprender la difusión, los datos disponibles consisten en trayectorias, como una molécula en una célula, una cotización en el mercado de valores o un animal que busca comida en su entorno. El objetivo es extraer de estas trayectorias información sobre las propiedades del trazador y del medio en el que se produce su movimiento, es decir, inferir el exponente de difusión anómala . En nuestro caso tenemos un objetivo bastante parecido pero aplicado a datos deportivos, más concretamente a trayectorias del balón en un partido de fútbol.

2.2 Análisis fútbol

Hablando del análisis de los partido de fútbol, los analistas de datos utilizan muchas información, como la media de cualquier dato (goles, fueras de juego, pases...) , para crear nuevas métricas como los goles o pases esperados que puede haber en un partido. También se estudia la posición de los jugadores en el terreno de juego para obtener información que pueda ser útil.

Hoy en día hay muchas empresas como *Driblab* o *Wyscout* que utilizan el poder de los datos y la estadística en el fútbol. Estas empresas disponen de bases de datos con información de más de 200 equipos y 200.000 jugadores. Esto hace que puedan realizar análisis mucho más completos y tienen la posibilidad de hacer cosas que no se puede con unos datos más escasos.

Mediante la ciencia de datos se investiga y se publican cada vez más desarrollos de algoritmos de *machine learning* tendentes a analizar el juego de cada futbolista y reubicarlo en la posición óptima de acuerdo con sus características [10], o poder predecir una lesión, aunque uno de los principales usos del *machine learning* es intentar predecir los resultados de los partidos [16].

Otros programas analizan de forma dinámica las tácticas y el posicionamiento dentro del campo. Mediante una serie de modelos estadísticos, estudian el comportamiento real del equipo durante los partidos

La lista de estudios de ciencia de datos aplicados al fútbol es cada vez más larga; motivada, en gran medida, por el interés deportivo y económico y a la financiación de los clubes.

Este trabajo va más allá de estos datos y métricas, intentando sacar información útil de las diferentes trayectorias que se dan acabo durante el partido. Los datos los obtenemos del *github* de *Statsbomb*. *Statsbomb* es una empresa creada con el objetivo de recopilar y analizar datos deportivos de la manera más completa posible. Para este proyecto utilizamos datos recopilados por ellos, en los que en encontramos información organizada por eventos, cada evento es algo que ha pasado durante un partido. Principalmente vamos ha hacer uso de las coordenadas de los jugadores (su posición en el campo) para dibujar las trayectorias.

Aparte de utilizar los datos proporcionados por *Statsbomb* vamos a tener que hacer uso de datos sintéticos, es decir, generados artificialmente ya que para entrenar los modelos hacen falta muchas más trayectorias de las que disponemos en la base de datos. En el proceso de generar los datos es donde vamos a hacer uso de la difusión anómala.

CAPÍTULO 3
Metodología

3.1 Fases del proyecto

El ciclo de vida de un proyecto es una serie de etapas por las que pasa un proyecto de principio a fin. El número y la secuencia de los ciclos está determinado por la gestión y varios otros factores, como las necesidades de las organizaciones involucradas en el proyecto, la naturaleza del proyecto y su aplicación. Las fases tienen puntos de inicio, fin y control definidos y están limitadas por el tiempo. El ciclo de vida del proyecto se puede definir y modificar de acuerdo con las necesidades y aspectos de la organización. Aunque cada proyecto tiene un comienzo y un final claros, las metas, los productos y las actividades específicas varían ampliamente. Los ciclos de vida forman la base de las actividades a realizar en el proyecto, independientemente del trabajo específico a realizar.

En los proyectos en los que los datos tienen un papel esencial, es muy común utilizar la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) la cual surgió en dos empresas (DaimlerChrysler y SPSS) pioneras en la aplicación de minería de datos a los procesos de negocio.

La metodología CRISP-DM consta de seis fases distintas, las cuales dependen entre sí de manera secuencial y cíclica. Como podemos ver en la Figura 3.1 las seis fases son: *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation* y *deployment*.

Las fases mencionadas anteriormente son fundamentales en el proyecto y cada una tiene un papel diferente que ayuda a completar el proyecto. En cada una de las fases realizamos lo siguiente:

- **Business understanding (comprensión del negocio):** Esta es la primera fase con la que debe comenzar cualquier proyecto de minería de datos, obteniendo una comprensión profunda del problema en cuestión, identificando las necesidades y objetivos del proyecto desde una perspectiva comercial y luego traduciéndolos en objetivos técnicos y planes de proyecto.

  Primero, establece los criterios para medir el éxito del proyecto. Luego se realiza una evaluación de la situación actual para determinar el contexto y los requisitos del problema, incluidos los negocios y la minería de datos. Fi-

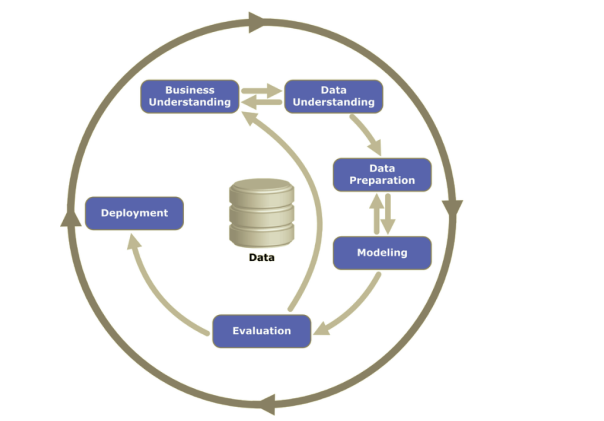

Figura 3.1: Metodoloía CRISP-DM

nalmente, ejecute el plan del proyecto, considerando qué pasos se tomarán y qué procedimientos se utilizarán en cada paso.

- **Data understanding (comprensión de los datos):** En esta fase, los datos se recopilan y exploran inicialmente para establecer la primera conexión con el problema. Esta fase suele ser crítica en los proyectos, ya que la mala interpretación de los datos puede aumentar el tiempo total del proyecto y también reducir la garantía de éxito. Hay varias tareas para completar esta etapa. En nuestro caso los datos ya estaban recopilados en el Github de Statsbomb [37], por lo que únicamente tuvimos que observar y analizar los datos para ver que información podíamos obtener de ellos.

- **Data preparation (preparación de los datos):** La fase de preparación de datos incluye todas las operaciones necesarias para construir el conjunto de datos final (los datos utilizados por las herramientas de modelado) a partir de los datos sin procesar originales. Las tareas incluyen seleccionar tablas, registros, atributos y transformar y limpiar datos para herramientas de modelado. En este proyecto, como explicamos en la sección 4.1, además de los datos obtenidos de *Statsbomb*, también generamos datos. Estos datos los generamos mediante difusión anómala para poder tener una cantidad de trayectorias que sea considerable a la hora de entrenar los modelos ya que con los datos reales no era suficiente y no podríamos obtener unos resultados fiables.

- **Modeling (modelado):** En esta parte del proyecto, seleccionamos y aplicamos las técnicas de modelado relevantes para el problema ,ajustando sus parámetros y valores óptimos. Existen distintos tipos de modelo, en nuestro caso aplicaremos redes neuronales como explicamos en el apartado 6.3

- **Evaluation (evaluación):** Una vez terminamos con la fase del modelado, pasamos a evaluar dichos modelos para ver si cumplen con nuestras expectativas. Si las cumplen pasaríamos a la siguiente fase (despliegue). En caso contrario deberíamos volver a las fases anteriores con el objetivo de encontrar mejores resultados. Estos resultados los podemos ver en el capítulo 8 de la memoria.

- **Deployment (despliegue):** En general, la generación de modelos no es el final de un proyecto. A menudo el objetivo del modelo es mejorar el conocimiento de los datos, pero el conocimiento resultante debe organizarse y presentarse para que lo use el cliente. Como hemos comentado, el modelado no siempre es el final del proyecto, dependiendo de los requerimientos que tengamos, en esta fase podemos desde hacer un informe básico hasta realizar un análisis en profundidad de los datos y resultados que hemos obtenido.

3.2 Tecnología empleada

Antes de empezar el proyecto surgió la duda de con que lenguaje de programación íbamos a trabajar ya que teníamos varias opciones, *python* y *R*. Ambos lenguajes tenían ventajas e inconvenientes, pero finalmente nos decidimos por *python*, por el mejor domino de este lenguaje que teníamos, y por su facilidad y accesibilidad a muchas librerías que nos podían ser útiles.

A lo largo del proyecto hemos utilizado diversas librerías, algunas generales o más básicas como podrían ser *numpy*, *math*, *pandas* o *random*, y otras más específicas como:

- **matplotlyb:** esta librería la hemos utilizado para generar algunas gráficas que se muestran en la memoria.
- **fbm:** utilizamos esta librería en el apartado 4.1.2 ya que dispone de los métodos exactos para simular el movimiento browniano fraccionario (fBm) o el ruido gaussiano fraccionario (fGn) en *python*.
- **sklearn:** de esta librería utilizamos dos modulos para realizar el pre-modelado.
  - **model_selection:** utilizado para dividir los datos en train y test mediante la función *train_test_split*.
  - **preprocessing:** de este modulo de la librería utilizamos dos funciones, *LabelEncoder* y *OneHotEncoder* las cuales sirven para codificar los datos para poder utilizarlos en el modelo de clasificación.
- **mcfly:** igual que en la librería *sklearn*, en esta,para aplicar el modelo de clasificación que se explica en el apartado 6.3.2 utilizamos dos módulos
  - **find_architecture:** este módulo proporciona la funcionalidad principal de esta librería, la búsqueda de una arquitectura de modelo óptima.
  - **modelgen:** como se puede deducir del nombre, este módulo se utiliza para generar los modelos.

    El flujo de trabajo el cual seguimos es que mediante la función *generate_models* de *modelgen* generamos y compilamos los modelos y mediante la función *train_models_on_samples* de *find_architecture* entrenamos esos modelos.

- **keras:** *keras* es una biblioteca de redes neuronales de alto nivel la cual utilizamos para crear el modelo de redes neuronales convolucionales de regresión. Para ello hacemos uso de los siguientes modulos:
  - **models:** de este módulo utilizamos la función *Sequential* ya que es la más apropiada para un modelo y poder añadir las capas que creamos necesarias.
  - **layers:** las diferentes capas que forman nuestro modelo las importamos de este módulo. Como explicamos en el apartado 6.3.1 hacemos uso de capas convolucionales ,capas LSTM bidireccionales y la función de activación mediantes las funciones de keras *layers.Conv1D*, *layers.Bidirectional* y *layers.Dense* respectivamente.

Para finalizar queremos destacar el uso del entorno *jupyter notebook* a la hora de crear el modelo implementado con *keras* ya que nos facilitaba el trabajo por su interfaz y su compatibilidad con *Google Colab* ya que este nos permitía acelerar mucho el tiempo de cómputo debido a que ofrecen el uso de una GPU (unidad de procesamiento gráfico) gratuita.

CAPÍTULO 4
Métodos utilizados

4.1 Preparación de los datos

Los datos que vamos a utilizar en el proyecto son las distintas trayectorias que sigue la pelota durante el partido en el campo de fútbol. Estas trayectorias pueden ser de diferentes longitudes y están formadas por puntos con coordenadas X e Y. La longitud de la trayectoria viene marcada por la cantidad de puntos que tenga. La trayectoria esta marcada por los pases que hacen los jugadores, un punto es desde donde se realiza el pase y otro es donde lo recibe. Si entre dos pases el jugador se ha desplazado en el campo una distancia considerable, hemos decidido que sea de 5 m, contamos ese desplazamiento como un pase por lo que crearíamos dos nodos diferentes en la trayectoria.

Actualmente estamos estudiando las trayectorias en un plano, es decir, no utilizamos la altura a la que pueda estar el balón en cada momento de la trayectoria. En un futuro nos gustaría poder estudiar si la altura de los pases influye en el juego de los equipos y si podemos llegar a utilizar esta información para mejorar el rendimiento de los equipos.

4.1.1. Datos reales

Durante el proyecto utilizamos distintos tipos de datos, ya sean datos sintéticos o datos reales. Estos últimos los obtenemos del github de una empresa llamada *Statsbomb*.

StatsBomb fue fundado por Ted Knutson en 2013, creado como un *blog* de análisis de fútbol, ha ido creciendo hasta convertirse en una empresa de análisis deportivo, la cual hoy en día es en un lugar de referencia para muchos analistas.

Para poder realizar análisis deportivo se necesitan datos y StatsBomb también se dedica a eventos de partidos de fútbol. Los datos de que encontramos en el *github* estan en formato *JSON* y se organizan de la siguiente manera:

- **Competitions:** como vemos en la figura 4.1 este es un archivo en el cual encontramos información sobre las distintas competiciones las cuales *StatsBomb* ha recopilado datos. Lo más importante es el identificador que nos permite conectar la información con los siguientes archivos.

```
{
  "competition_id" : 49,
  "season_id" : 3,
  "country_name" : "United States of America",
  "competition_name" : "NWSL",
  "competition_gender" : "female",
  "competition_youth" : false,
  "competition_international" : false,
  "season_name" : "2018",
  "match_updated" : "2021-11-06T05:53:29.435016",
  "match_updated_360" : "2021-06-13T16:17:31.694",
  "match_available_360" : null,
  "match_available" : "2021-11-06T05:53:29.435016"
}, {
  "competition_id" : 2,
  "season_id" : 44,
  "country_name" : "England",
  "competition_name" : "Premier League",
  "competition_gender" : "male",
  "competition_youth" : false,
  "competition_international" : false,
  "season_name" : "2003/2004",
  "match_updated" : "2021-11-14T22:29:00.646120",
  "match_updated_360" : "2021-06-13T16:17:31.694",
  "match_available_360" : null,
  "match_available" : "2021-11-14T22:29:00.646120"
}, {
  "competition_id" : 55,
  "season_id" : 43,
  "country_name" : "Europe",
  "competition_name" : "UEFA Euro",
  "competition_gender" : "male",
  "competition_youth" : false,
  "competition_international" : true,
  "season_name" : "2020",
  "match_updated" : "2022-02-01T17:20:34.319496",
  "match_updated_360" : "2021-11-11T13:54:37.507376",
  "match_available_360" : "2021-11-11T13:54:37.507376",
  "match_available" : "2022-02-01T17:20:34.319496"
}
```


Figura 4.1: Datos de las competiciones

- **Matches:** en los distintos archivos de tipo partido 4.2, encontramos el identificador de la competición a la que pertenecen, un identificador propio para enlazarlo con los distintos eventos que ocurren en él e información de dicho partido, ya sea el resultado, equipos que juegan, localización...

```
{
  "match_id" : 3795187,
  "match_date" : "2021-07-03",
  "kick_off" : "21:00:00.000",
  "competition" : {
    "competition_id" : 55,
    "country_name" : "Europe",
    "competition_name" : "UEFA Euro"},
  "season" : {
    "season_id" : 43,
    "season_name" : "2020"},
  "home_team" : {
    "home_team_id" : 911,
    "home_team_name" : "Ukraine",
    "home_team_gender" : "male",
    "home_team_group" : null,
    "country" : {
      "id" : 238,
      "name" : "Ukraine"
    },
    "managers" : [ {
      "id" : 2303,
      "name" : "Andrii Shevchenko",
      "nickname" : null,
      "dob" : "1976-09-29",
      "country" : {
        "id" : 238,
        "name" : "Ukraine"}}]},
  "away_team" : {
    "away_team_id" : 768,
    "away_team_name" : "England",
    "away_team_gender" : "male",
    "away_team_group" : null,
    "country" : {
      "id" : 68,
      "name" : "England"
    },
    "managers" : [ {
      "id" : 277,
      "name" : "Gareth Southgate",
      "nickname" : null,
      "dob" : "1970-09-03",
      "country" : {
        "id" : 68,
        "name" : "England"}}],
  "home_score" : 0,
  "away_score" : 4,
  "match_status" : "available",
  "match_status_360" : "available",
  "last_updated" : "2021-07-04T16:39:20.746",
  "last_updated_360" : "2022-08-04T12:00",
  "metadata" : {
    "data_version" : "1.1.0",
    "shot_fidelity_version" : "2",
    "xy_fidelity_version" : "2"
    }
```


Figura 4.2: Datos de los partidos

- **Events:** en estos archivos se incluye los diferentes eventos o acciones que ocurren durante un partido, el cual podemos saber ya que el nombre del archivo es el identificador de este. Estos eventos como vemos en la figura 4.3 van desde el saque inicial, pasando por pases, tiros cambios hasta llegar el final del partido.

```
{
  "id" : "4bd02a7e-9ab5-4c10-993b-48875aa54b73",
  "index" : 3,
  "period" : 1,
  "timestamp" : "00:00:00.000",
  "minute" : 0,
  "second" : 0,
  "type" : {
    "id" : 18,
    "name" : "Half Start"
  },
  "possession" : 1,
  "possession_team" : {
    "id" : 768,
    "name" : "England"
  },
  "play_pattern" : {
    "id" : 1,
    "name" : "Regular Play"
  },
  "team" : {
    "id" : 768,
    "name" : "England"
  },
  "duration" : 0.0,
  "related_events" : [ "22da729b-6b2f-4fdc-9e91-587af4be8807" ]
}, {
  "id" : "22da729b-6b2f-4fdc-9e91-587af4be8807",
  "index" : 4,
  "period" : 1,
  "timestamp" : "00:00:00.000",
  "minute" : 0,
  "second" : 0,
  "type" : {
    "id" : 18,
    "name" : "Half Start"
  },
  "possession" : 1,
  "possession_team" : {
    "id" : 768,
    "name" : "England"
  },
  "play_pattern" : {
    "id" : 1,
    "name" : "Regular Play"
  },
  "team" : {
    "id" : 785,
    "name" : "Croatia"
  },
  "duration" : 0.0,
  "related_events" : [ "4bd02a7e-9ab5-4c10-993b-48875aa54b73" ]
},
```


Figura 4.3: Datos de los eventos de un partido

- **360:** los datos 360, se relacionan con un evento por su identificador. Estos datos son las coordenadas X e Y del jugador, es decir, que posición ocupan en el campo en el momento que ocurre la acción.

Entre estos datos que recopila *StatsBomb* , como hemos comentado anteriormente, están los datos 360, los cuales como vemos en la imagen 4.4son las coordenadas de muchos eventos que ocurren durante el partido, como pueden ser pases, tiros, faltas... Estos datos se encuentrarn en formato *JSON*, por lo que mediante la librería *pandas* de *python* nos ayudamos para poder manejarlos de manera más fácil y cómoda.

```
{
  "event_uuid" : "d690cf5f-0a7f-41fa-ac3a-2eb39013b879",
  "visible_area" : [ 12.4538540810486, 80.0, 42.4041452136704, 0.973607039024174, 66.0345288484566,
                     4.37940871333336, 75.6004736776445, 80.0, 12.4538540810486, 80.0 ],
  "freeze_frame" : [ {
    "teammate" : true,
    "actor" : true,
    "keeper" : false,
    "location" : [ 38.69952, 48.246754 ]
  }, {
    "teammate" : true,
    "actor" : false,
    "keeper" : false,
    "location" : [ 40.42738, 33.483665 ]
  }, {
    "teammate" : true,
    "actor" : false,
    "keeper" : false,
    "location" : [ 50.8696, 66.65857 ]
  }, {
    "teammate" : true,
    "actor" : false,
    "keeper" : false,
    "location" : [ 51.59526, 49.97813 ]
  }, {
    "teammate" : false,
    "actor" : false,
    "keeper" : false,
    "location" : [ 52.347244, 46.105434 ]
  }, {
    "teammate" : true,
    "actor" : false,
    "keeper" : false,
    "location" : [ 54.72594, 31.745445 ]
  }, {
    "teammate" : true,
    "actor" : false,
    "keeper" : false,
    "location" : [ 59.490192, 20.745625 ]
  }, {
    "teammate" : false,
    "actor" : false,
    "keeper" : false,
    "location" : [ 60.68872, 58.97332 ]
  }, {
    "teammate" : true,
    "actor" : false,
    "keeper" : false,
    "location" : [ 69.42701, 33.667328 ]
  } ]
}
```


Figura 4.4: Datos de los partidos

Para obtener las trayectorias que necesitamos, hemos utilizado las coordenadas de los pases que ocurren a lo largo del partido, teniendo en cuenta que deben ser seguidos como podemos ver en las figura 4.5 y **??**. En estos gráficos vemos los distintos pases que tiene la trayectoria, de las cuales podemos seguir el orden viendo el número que se encuentra en los vértices.

En el GitHub de *StatsBomb* encontramos datos de varias competiciones, como la liga española, o inglesa, la *Champions League* o la Eurocopa. Podíamos elegir también entre distintas temporadas y finalmente nos decidimos por la Eurocopa de 2020 ya que era una de las competiciones que más datos 360 tenía y tener más datos nos facilitaba los análisis y poder entrenar los modelos. Por todo estos es que estaremos trabajando con las trayectorias de las distintas selecciones que participaron en esta competición.

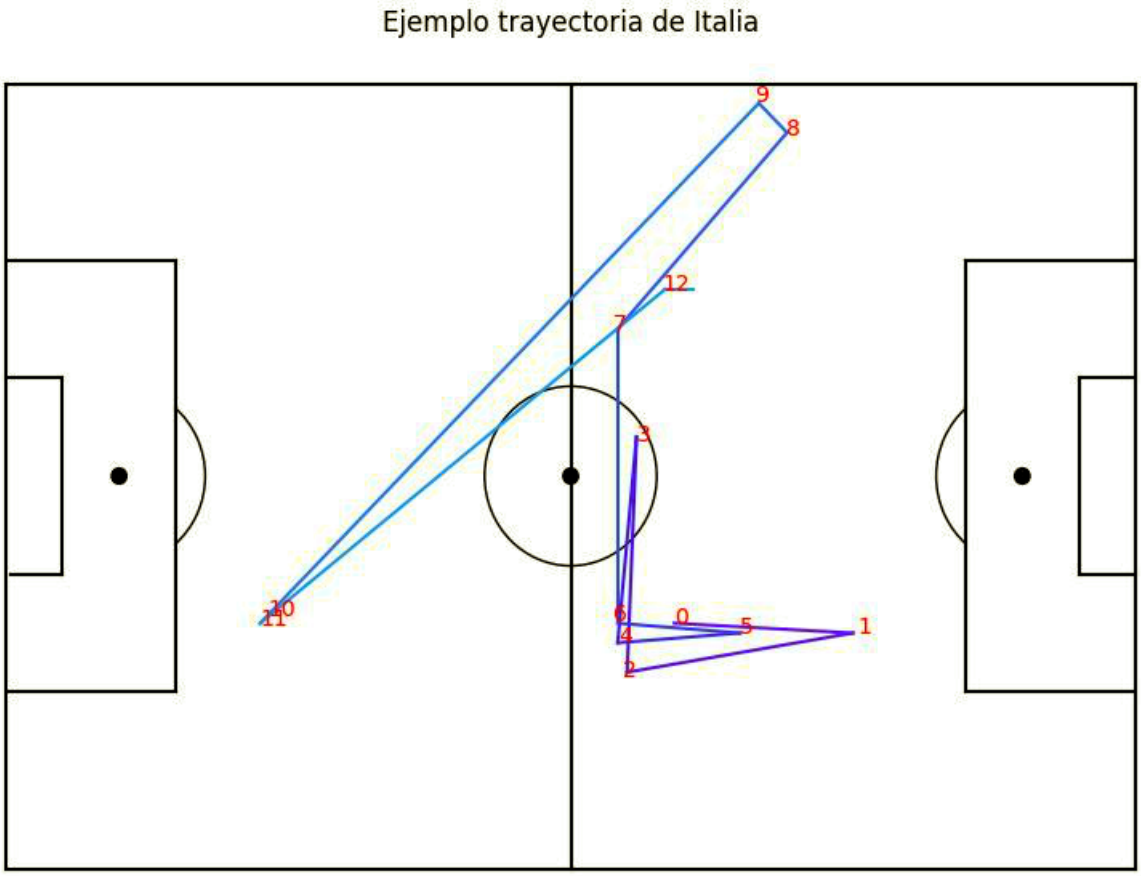

Figura 4.5: Trayectoria real de Italia en un partido de fútbol

4.1.2. Generación de los datos

A parte de los datos que obtenemos de *Statsbomb*, en nuestros modelos hemos tenido que utilizar datos generados para poder entrenarlos, y así más adelante utilizar los modelos entrenados en nuestros datos reales. Hemos tenido que generar datos artificialmente ya que con los datos reales no teníamos suficientes trayectorias para poder alimentar los modelos y que funcionaran de una manera adecuada.

Para generar las trayectorias que vamos a utilizar para entrenar los modelos, hemos seguido lo indicado por Kowalek [20], donde se generan datos sintéticos para el análisis de distintos modelos de difusión. Al generar las trayectorias se podían cambiar ciertas partes del código como son:

- **Alpha(α):** este es el parámtro que más hemos variado para tener una gran variedad ya que queríamos etiquetar o clasificar las trayectorias según el valor de este parámetro.
- **N:** este parámetro es la longitud de las trayectorias generadas, las cuales han oscilado entre una longitud de 10 y 35.

- **SNR:** el SNR o *Signal-to-noise ratio* es una medida utilizada en ciencia e ingeniería que compara el nivel de una señal deseada con el nivel de ruido de fondo. El valor de este parámetro lo hemos definido entre 1 y 2 para así tener trayectorias con poco ruido o mucho ruido.

En nuestro caso hemos tenido que adaptar el código ya que las trayectorias que necesitábamos debían estar confinadas en las dimensiones de un campo de fútbol. Como los campos de fútbol pueden tener tamaños diferentes, hemos utilizado nuestros datos reales para poner límites a las trayectorias, siendo estos 120 para la coordenada X (largo del campo) y 80 para la coordenada Y (ancho del campo) como podemos ver en la figura 4.6.

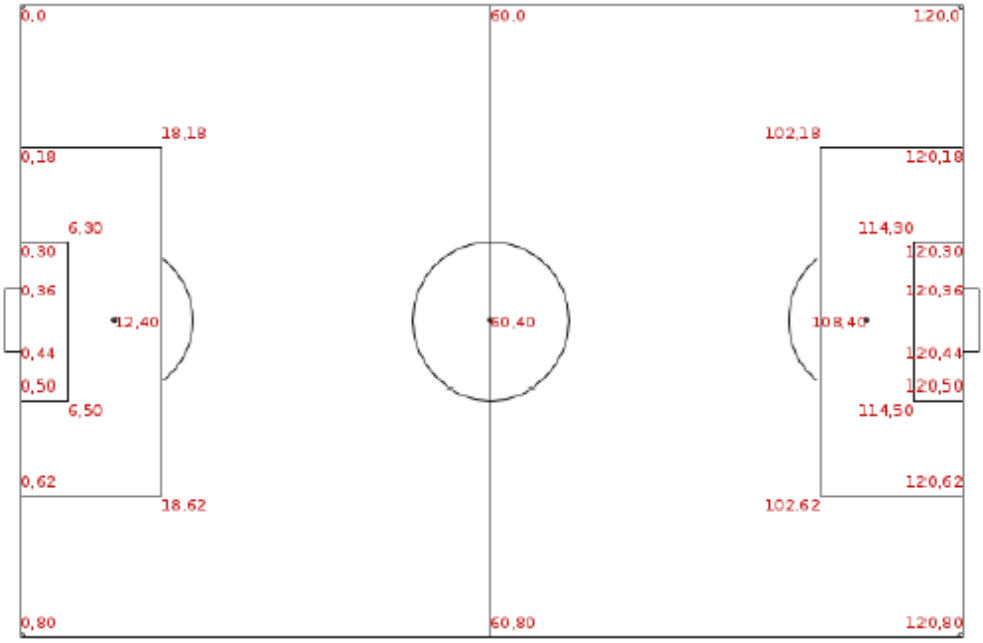

Figura 4.6: Medidas campo fútbol

Estas trayectorias confinadas las hemos generado basándonos en el movimiento fraccionario browniano (FBM). Este modelo permite generar trayectorias cuyo MSD (desplazamiento cuadrático medio) se comporta proporcionalmente a t^{α}, siendo t el tiempo, con $0 < \alpha < 2$. En el caso $\alpha = 0$ tenemos no hay movimiento, para $\alpha = 1$ tenemos un movimiento aleatorio equiparable al ruido Gaussiano y en el caso $\alpha = 2$ tenemos un modelo completamente difusivo, que se comportaría como la ecuación del calor en dos dimensiones. Los casos en los que el MSD se comporta como t^{α}, con $\alpha \neq 1$, se consideran difusión anómala **??**. En otras palabras, existe una relación no lienal entre la varianza de los desplazamientos entre pases y el tiempo trasncurrido.

El FBM es una generalización del movimiento browniano. A diferencia del movimiento browniano clásico, los incrementos de fBm no tienen por qué ser independientes. fBm es un proceso gaussiano de tiempo continuo que comienza en cero, tiene esperanza igual a cero a lo largo del tiempo y tiene la siguiente función de covarianza:

$$E[B_H(t)B_H(s)] = \frac{1}{2}(|t|^{2H} + |s|^{2H} - |t-s|^{2H}),$$

donde H, el índice Hurst o parametro Hurst, es un número real entre 0 y 1. El exponente Hurst describe la irregularidad del movimiento resultante, cuanto más alto sea el valor más suave es el movimiento. Existe una relación simple entre el índice Hurst y el exponente de difusión anómala α, tal que

$$2H = \alpha$$

Para simular el movimiento browniano fraccional hemos utilizado una librería implementada en Python llamada **fbm**.

Al generar las trayectorias existen varios parámetros que podemos cambiar para poder crear distintos tipos. El exponente α es uno de ellos, el cual puede tener un valor dentro del rango (0.1, 1.9) con paso 0.1 lo que nos deja con 19 posibles valores distintos de α. Además de las distintas α, generamos trayectorias con distintas longitudes N, que van desde 10 hasta 35 y distintos coeficientes de difusión D el cual puede ser entre 2 y 5. Hemos elegido este rango de coeficientes para que las trayectorias se adapten bien a las dimensiones del campo de fútbol ya que con coeficientes más bajos las trayectorias con esas longitudes se centraban en un espacio muy reducido por lo que no eran representativas de lo que es la realidad. Por otro lado, las trayectorias las generamos con esas longitudes para que sean del mismo tamaño que las trayectorias reales que vamos a utilizar. Los principales motivos de haber elegido estos tamaños son, el número de trayectorias reales que tenemos en nuestros datos con esas longitudes, ya que de un tamaño mayor teníamos muy pocas y no era de mucha utilidad estudiarlas. El otro motivo es que las trayectorias con menor longitud muy pocas veces son buenas jugadas en el fútbol, por lo que si queremos mejorar el rendimiento del equipo hay que estudiar las trayectorias que, a nuestro entender, son buenas jugadas durante el partido.

Las trayectorias que tienen longitud menor a 35 las completamos con ceros para así poder utilizar todas las trayectorias en el mismo modelo. Este procesado de los datos es llamado *padding* el cual es necesario para poder alimentar los modelos de una manera correcta y que estos puedan funcionar.

Una vez generadas las trayectorias añadimos ruido Gaussiano a las mismas. Para generar dicho ruido utilizamos la siguiente fórmula:

$$Q = \frac{\sqrt{D\Delta t}}{\sigma},$$

donde el valor de sigma lo elegimos aleatoriamente entre un intervalo de valores $SNR = [1, 2]$ utilizando la función **random.uniform** de la librería **random** de Python. El ruido lo añadimos para hacer más realistas las trayectorias del balón ya que en un partido real estas siempre se pueden ver afectadas por el viento o pueden ser desviadas por un rival, lo que es un factor que es bueno tenerlo en cuenta en nuestro análisis. Antes de añadir el ruido nos tenemos que asegurar de que no sobrepasamos los límites del campo en ningún punto de la trayectoria.

Una vez generadas las trayectorias que queremos las vamos a separar en distintos datasets, los datos de entrenamiento, los datos de test y los datos de validación ya que va a ser necesario para aplicar los modelos de *machine learning* 4.2 que queremos utilizar para realizar los análisis.

Para dividir los datasets utilizamos la función **model_selection.train_test_split** de la librería **sklearn**. Hacemos una primera división en la que el 30 % de los datos los utilizaremos como el dataset de test. Del 70 % restante, el 30 % lo utilizaremos como dataset de validación y el resto como el dataset de entrenamiento.

Cuando dividimos los datos en distintos datasets aprovechamos para etiquetar las trayectorias de dos maneras, una para regresión y otra para clasificación. Para clasificación etiquetamos las trayectorias en dos clases distintas, las cuales se dividen en si son superdifusivas ($\alpha > 1$ o subdifusivas($\alpha < 1$)[2]. Cuando las etiquetamos para regresión utilizamos como etiqueta el α con el que se ha generado la trayectoria, por lo que podemos tener trayectorias etiquetadas con α distintos, los valores de los cuales están en el rango de valores de α con los que se han generado las trayectorias.

4.2 Machine learning

El machine learning o aprendizaje automático es una disciplina del campo de la inteligencia artificial mediante la que utilizando distintos algoritmos se dota a los ordenadores o programas informáticos a ser capaces de identificar patrones y ser capaces de predecir o clasificar datos.

En el aprendizaje automático existen tres tipos diferentes de aprendizaje, el supervisado, el no supervisado y el aprendizaje por refuerzo. En nuestro caso vamos a utilizar aprendizaje supervisado ya que las etiquetas de nuestros datos son los distintos parámetros que pueden cambiar a la hora de generar las trayectorias, como pueden ser la longitud (N), el α o el SNR.

4.2.1. Aprendizaje supervisado

Los algoritmos de aprendizaje supervisado cuentan con un aprendizaje previo basado en un sistema de etiquetas asociadas a unos datos que les permiten tomar decisiones o hacer predicciones. El algoritmo se entrena a partir de muestras con sus etiquetas asociadas y va aprendiendo a identificar patrones, para así poder predecir o clasificara un dato.

Un ejemplo es un detector de spam que etiqueta un e-mail como spam o no dependiendo de los patrones que ha aprendido del histórico de correos (remitente, relación texto/imágenes, palabras clave en el asunto, etc.).

En nuestro caso vamos a utilizar las trayectorias generadas con difusión anómala, cuya etiqueta es el α con el que se ha generado dicha trayectoria. Estas etiquetas las utilizaremos para regresión. En el método de clasificación utilizaremos 2 etiquetas según sean superdifusivas o subdifusivas.

CAPÍTULO 5
Modelos físicos de difusión

A la hora de estudiar los modelos físicos de difusión, el enfoque más elemental para analizar los movimientos es comenzar a analizar el *Desplazamiento Cuadrático Medio* de las trayectorias, ya que su crecimiento cambia a lo largo del tiempo. Además, este es un enfoque básico para clasificar las trayectorias [30, 32] y para distinguir entre trayectorias normales y anómalas. Debido a la naturaleza estocástica de las trayectorias, conviene analizarlas conjuntamente.

5.1 Medidas de difusión

Definiremos el MSD (desplazamiento cuadrático medio) y el TAMSD (desplazamiento cuadrático medio promediador en el tiempo) como medidas elementales para estudiar cómo se propagan las trayectorias de las partículas. Formulamos estos modelos en el espacio tridimensional $\mathbb{R}^3$. Descripciones similares se pueden proporcionar fácilmente a $\mathbb{R}$ y $\mathbb{R}^2$.

5.1.1. Difusión según el desplazamiento cuadrático medio (MSD)

El *desplazamiento cuadrático medio* de una trayectoria indica el tamaño medio de una región *explorada* por el movimiento de un conjunto de partículas. Dadas N partículas, $x^1, x^2, ...x^N$, cuyas posiciones en el instante t vienen dadas por $x^i(t)$, el MSD se calcula como

$$MSD(t) = \frac{1}{N}\sum_{i=1}^{N}\left|\left|x^i(t) - x^i(0)\right|\right|^2. \tag{5.1}$$

Con el MSD podemos adivinar si las partículas se propagan sólo por difusión o si hay alguna fuerza de transporte adicional que actúa sobre ellas. Según la estimación del MSD, un movimiento puede clasificarse en una de estas 4 clases [31].

- *Difusión libre (FD)*: El movimiento puede dividirse en paseos aleatorios microscópicos idénticamente distribuidos, siguiendo una función de densidad normal proporcionada por la segunda ley de Fick que establece que la tasa de concentración de una sustancia evoluciona a lo largo del tiempo de

forma proporcional a la segunda derivada respecto al espacio de la concentración. En esta situación, $\langle x_i(t)^2 \rangle = 2Dt$, para $i = 1, 2, 3$, y $MSD(t) = 6Dt$, donde D es la constante de difusión del proceso y t representa el tiempo, aumentando linealmente respecto al tiempo [2].

- *Difusión Dirigida (DD)*: Cuando $\text{MSD}(t) = v^2t^2$, donde v es la magnitud de la velocidad de la partícula.
- *Difusión confinada (CD)*: Cuando $\text{MSD}(t) = R_C^2 \left(1 - e^{-6Dt/R_C^2}\right)$, donde R_C es el radio dentro del cual la partícula está confinada.
- *Difusión anómala (AD)*: Cuando el MSD no se difunde normalmente, la variación puede tener una varianza más lenta o más rápida respecto a un proceso de difusión normal. En el primer caso el movimiento se denomina *subdifusivo* y en el segundo *superdifusivo* [2]. El MSD sigue $\text{MSD}(t) = 6Dt^\alpha$ con $0 < \alpha < 1$ para la subdifusividad y $1 < \alpha < 2$ para la superdifusividad.

5.1.2. Desplazamiento cuadrático medio promediado en el tiempo (TAMSD)

Cabe mencionar que cuando se realizan experimentos y se obtienen datos, se puede disponer de muchas trayectorias pero estas son cortas [27]. El supuesto de ergodicidad establece que podemos recuperar las propiedades de un proceso con sólo observar la evolución de una muestra aleatoria suficientemente larga. Por tanto, se puede entender que la observación de una partícula a lo largo de un tiempo suficientmente largo, puede dar una descripción del movimiento equiparable a la que obtendríamos mediante el MSD.

Dadas N posiciones consecutivas de una partícula x^0, denotadas por $x^0(j\Delta t)$ con $j = 1, \ldots, N$, obtenidas con un intervalo de tiempo Δt, entonces el *desplazamiento medio cuadrático promediado a lo largo tiempo o time-average mean square displacement (TAMSD)* para un intervalo de tiempo $n\Delta t$ se obtiene como

$$\rho(n\Delta t) = \frac{1}{N-n} \sum_{i=1}^{N-n} \left|\left| x^0((i+n)\Delta t) - x^0(i\Delta t) \right|\right|^2. \tag{5.2}$$

Bajo el supuesto de ergodicidad, el TAMSD converge al MSD cuando N tiende a ∞. Saxton y Jacobson demostraron que la TAMSD para los 4 tipos de difusión mencionados son [35]:

$$\textit{Difusión Libre (ND)}: \quad \rho_{ND}(n\Delta t) = 6Dn\Delta t \tag{5.3}$$

$$\textit{Difusión Dirigida (DD)}: \quad \rho_{DD}(n\Delta t) = R_c^2 \left[1 - A_1 \exp\left(\frac{-6A_2 Dn\Delta t}{R^2 c}\right)\right], \tag{5.4}$$

$$\textit{Difusión Confinada (CD)}: \quad \rho_{CD}(n\Delta t) = 6Dn\Delta t + (vn\Delta t)^2 \tag{5.5}$$

$$\textit{Difusión Anómala (AD)}: \quad \rho_{AD}(n\Delta t) = 6D(n\Delta t)^\alpha \tag{5.6}$$

donde $\alpha < 1$, v es la velocidad, R_c es el radio dentro del cual la partícula está confinada, y A_1, A_2 caracterizan el lugar de confinamiento.

La presencia de ruido dificulta el ajuste de los datos con los modelos y no se puede obviar al procesar cualquier tipo de datos obtenidos de manera experimental. Sólo, con valores muy pequeños Δt, podríamos intentar obtener estimaciones razonables. Sin embargo, cuando se trata de trayectorias cortas, Δt no se puede fijar. Por ello, es interesante trabajar con trayectorias largas de una sóla partícula en vez de con muchas trayectoria cortas y, tomando en cuanta las primeras, plicar machine learning directamente a partir de los datos brutos sin limitarnos al ajuste de modelos estadísticos.

5.1.3. Difusión Anómala

Se han propuesto varios modelos estocásticos para describir los procesos de difusión anómalos [29]. Los más destacados fueron recientemente considerados en el AnDi Challenge [33, 32]:

1. *Movimiento Browniano (BM) y Movimiento Browniano Fraccionario (FBM)*: Si algunas partículas de un fluido se mueven aleatoriamente chocando entre sí, entonces se dice que las partículas siguen un movimiento browniano [36]. El movimiento de una partícula que sigue un movimiento browniano puede ser descrito por una función de densidad gaussiana $B_H(t)$ para $t \geq 0$. También podemos referirnos a este movimiento como un proceso de Wiener. Su covarianza estará dada por

$$E(B_H(t), B_H(s)) = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right) \tag{5.7}$$

   donde H, que se conoce como exponente de Hurst, es un número real entre 0 y 1. Si la función de densidad de movimiento de las partículas tiene una covarianza con $H \neq 1/2$, entonces hablamos de un *Fractional Brownian Motion (FBM)* [26]. Para $H > 1/2$, resulta en superdifusión y en subdifusión para $H < 1/2$.

2. *Annealed Transient Time Motion (ATTM)*: En este modelo, el movimiento de la partícula consiste en la composición de trozos de longitud aleatoria con difusividades aleatorias. Aunque es un movimiento browniano localmente, el movimiento es no ergódico [28, 38]. Esto se ha observado en la difusión de los receptores de la membrana celular.

3. *Continuous-Time Random Walk (CTRW)*: Una partícula se mueve siguiendo un paseo (o proceso) aleatorio si su trayectoria consiste en una secuencia de movimientos aleatorios. Si combinamos este movimiento con tiempos de espera aleatorios entre cada paso, lo llamamos un *Continuous-Time Random Walk (CTRW)*.

4. *Lévy Walk (LW)*: Una partícula sigue un Lévy Walk (LW) si su movimiento combina un movimiento aleatorio con pasos intermedios cuya longitud sigue asintóticamente una ley de potencia [19]. Esto se identifica con una variación en la distribución de los tiempos de espera, que puede tener muchas colas en lugar de seguir una función exponencial decreciente.

4. *Movimiento Browniano Escalado (SBM)*: Se trata de un proceso estocástico intermitente con un coeficiente de difusión dependiente del tiempo $D(t) \approx t^{\alpha-1}$ with $\alpha > 0$ [3, 21].

Para estos modelos, los exponentes anómalos se situarán en el intervalo [0.05,1.95], ya que los valores inferiores a 0.05 apenas muestran un movimiento. Cabe mencionar que cada movimiento puede tener exponentes en un rango diferente, dependiendo de las propiedades de difusión. Por un lado, los modelos subdifusivos, como ATTM y CTRW, presentan valores de $\alpha \leq 1$. Por otro lado, los modelos superdifusivos, como LW, presentan valores de $\alpha > 1$. Los FBM no pueden tener un comportamiento balístico $\alpha < 2$, y finalmente, los SBM pueden presentar valores en todo el rango. Se han considerado varios métodos para estimar este exponente que se basan principalmente en el TAMSD [17]

CAPÍTULO 6
Arquitectura del modelo

Para tratar los problemas de clasificación y de regresión, vamos a utilizar un modelo basado en la combinación de varias capas de redes convolucionales y de redes neuronales recurrentes (LSTM bidireccionales), que han dado buenos resultados en el análisis de trayectorias y en el análisis de series temporales [12, 24]

6.1 CNN

Las redes neuronales convolucionales son un tipo de red neuronal artificial utilizadas más comúnmente para analizar imágenes visuales. Son versiones regularizadas de perceptrones multicapa. Un perceptrón es una neurona artificial, y, por tanto, una unidad de red neuronal. El perceptrón efectúa cálculos para detectar características o tendencias en los datos de entrada. Se trata de un algoritmo para el aprendizaje supervisado de clasificadores binarios. Los perceptrones multicapa son generalmente redes completamente conectadas, es decir, cada neurona en una capa está conectada a todas las neuronas en la siguiente capa. La *conectividad total* de estas redes las hace propensas al sobreajuste de datos, por lo que hay que tener cuidado para no obtener un rendimiento deficiente en el conjunto de datos de validación.

En 1980, K. Fukushima [11] introdujo los fundamentos de las redes neuronales convolucionales o *convolutional neuronal networks* (CNN) los cuales se basan en el Neocognitron. Años más tarde, en 1998, el modelo fue mejorado por Y. LeCun [22] al introducir el aprendizaje basado en la propagación hacia atrás para así poder entrenar el modelo correctamente. En el año 2012, fueron refinadas por D. Ciresan [9] y otros, y fueron implementadas para una unidad de procesamiento gráfico (GPU) consiguiendo así resultados impresionantes.

6.2 Arquitectura de las CNN

Las redes neuronales convolucionales están formadas por distintas capas, donde cada capa tiene una función distinta dentro del modelo. Una capa en un modelo de aprendizaje profundo es una estructura o topología de red en la arquitectura

del modelo, que toma información de las capas anteriores y luego pasa información a la siguiente capa. Existen tres tipos de generales de capas, como podemos ver en la figura 6.1, la capa de entrada (input layer), las capas ocultas (hidden layers) y la capa de salida (output layer).

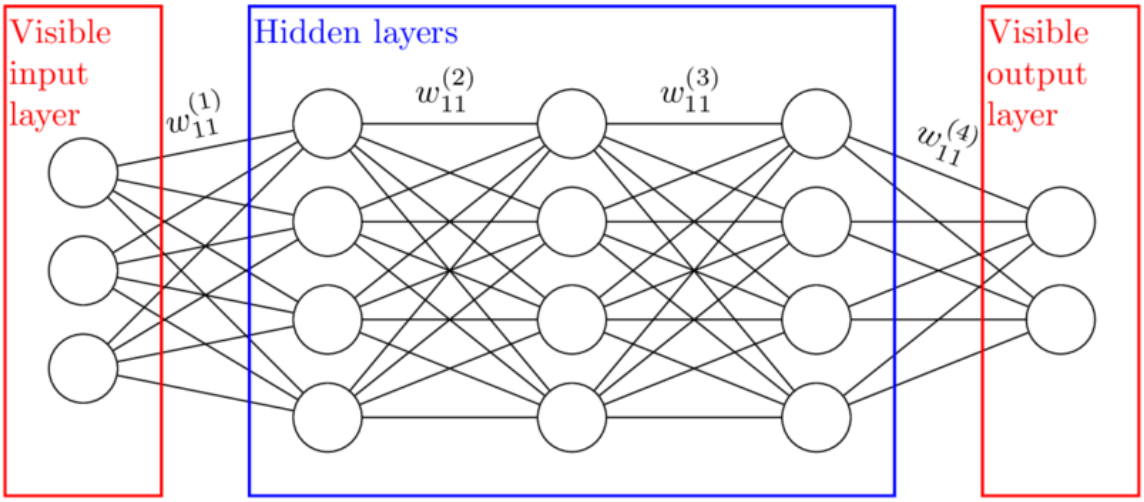

Figura 6.1: Capas de una red neuronal convolucional

Dentro de las capas ocultas, existen varios tipos de ellas como se puede obser var en la figura 6.2, como son las capas convolucionales, las capas de agrupación (pooling layers) , las capas totalmente conectadas (fully conected layers), las capas ReLU, etc. DEtallamos en qué consisten las mismas a continuación

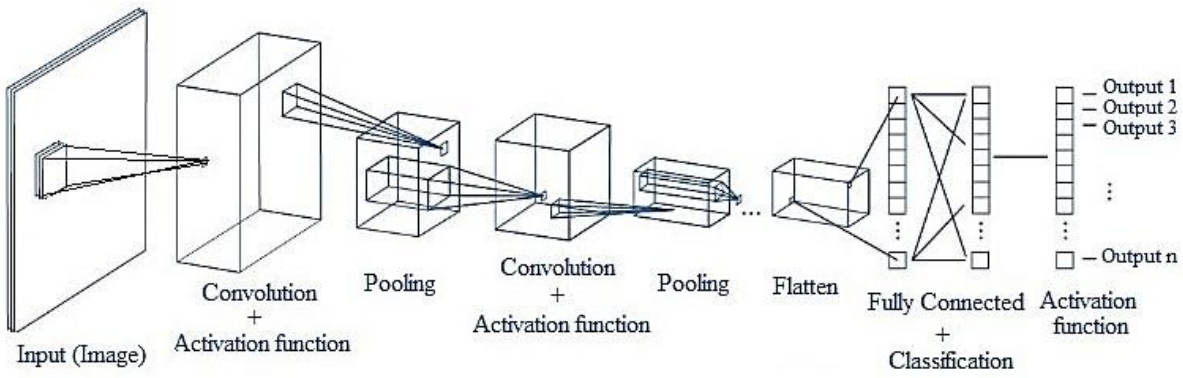

Figura 6.2: Arquitectura de una red neuronal convolucional

- Las *capas convolucionales* son los principales bloques de construcción utilizados en las redes neuronales convolucionales. Una convolución es la simple aplicación de un filtro a una entrada (input) que resulta en una activación. La aplicación repetida del mismo filtro a una entrada da como resultado un mapa de activaciones llamado mapa de características, que indica las ubicaciones y la fuerza de una característica detectada en una entrada, como podrían ser una imagen, o en nuestro caso trayectorias de un balón de fútbol.

- El papel de las *capas de agrupación* es reducir la dimensionalidad de los mapas de características para disminuir la número de parámetros y cálculos en la red. Esto tiene el efecto de hacer que los mapas de características submuestreados resultantes sean más robustos a los cambios. Dos métodos comunes de agrupación son la agrupación promedio (*average pooling*) y la agrupación máxima (*max pooling*) que resumen la presencia promedio de una característica y la presencia más activada de una característica, respectivamente.
- Las *capas totalmente conectadas* conectan cada neurona de una capa con cada neurona de otra capa. Es lo mismo que una red neuronal multicapa tradicional (MLP). La matriz aplanada pasa por una capa totalmente conectada para clasificar los datos.

LSTM

Las Long short-term memory (LSTM) [14] son redes neuronales recurrentes (RNN) que contienen múltiples puertas (gates) y son capaces de aprender dependencias a largo plazo debido a que su estructura tiene la capacidad de recordar información por periodos largos de tiempo, olvidar información innecesaria y exponer cuidadosamente la información en cada paso del tiempo. Han resultado de gran utilidad para el procesamiento del lenguaje natural. Han resultado muy adecuadas para clasificar, procesar y hacer predicciones basadas en datos de series temporales, ya que puede haber desfases de duración desconocida entre los eventos importantes de una serie temporal [23].

Las LSTM están formadas por puertas. Una unidad LSTM común se compone de una célula, una puerta de entrada, una puerta de salida y una puerta de olvido. La célula recuerda valores a lo largo de intervalos de tiempo arbitrarios y las tres puertas regulan el flujo de información que entra y sale de la célula como podemos ver en la figura 6.3

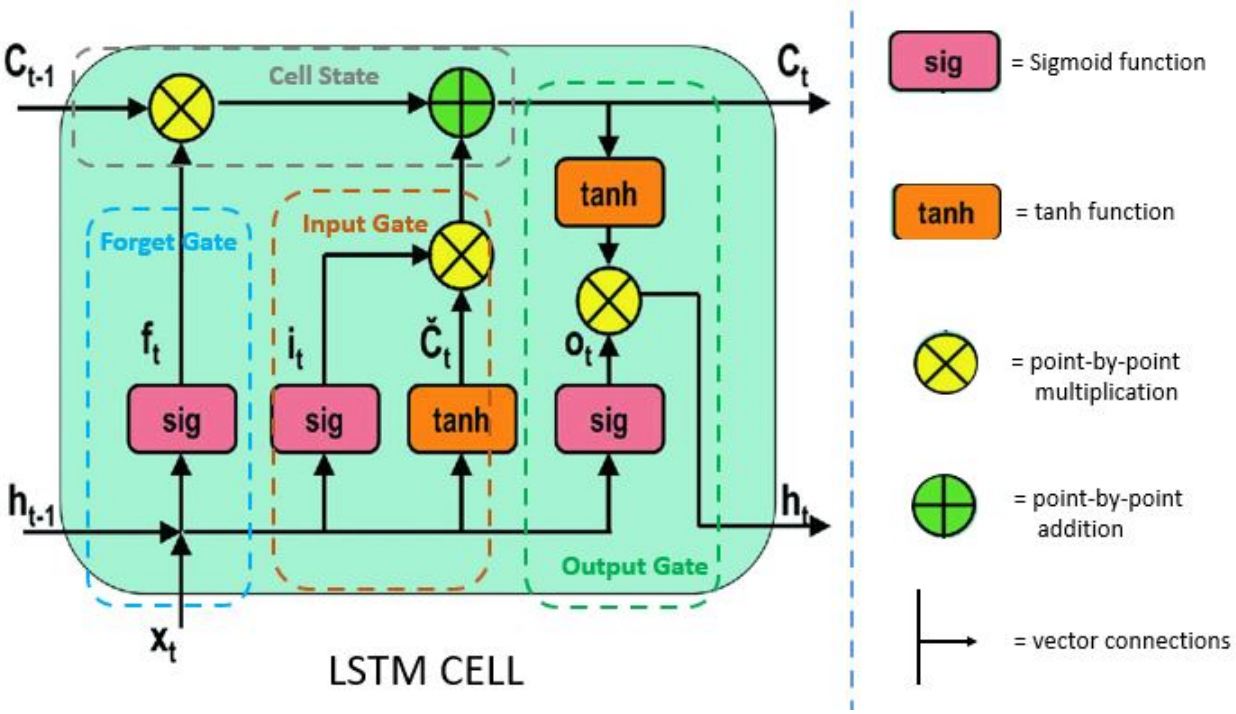

Figura 6.3: Arquitectura LSTM

Una mejora de las mismas son las LSTM bidireccionales [14, 23]. Una red neuroanl recurrente bidireccional (BDRNN) se constituye de dos capas (una capa que aprende representaciones previas y otra que retrocede en el el tiempo, lo cual nos ayuda a aprender de representaciones futuras), mejorando así el rendimiento de la RNN.

Las LSTM bidireccionales entrenan dos LSTM sobre la la secuencia de entrada en lugar de una sola. La primera funciona de manera normal, como hemos comentado anteriormente, y la segunda es una copia invertida de la secuencia de entrada como podemos ver en la figura 6.4, lo que puede proporcionar un contexto adicional a la red y un aprendizaje más rápido y completo sobre el problema.

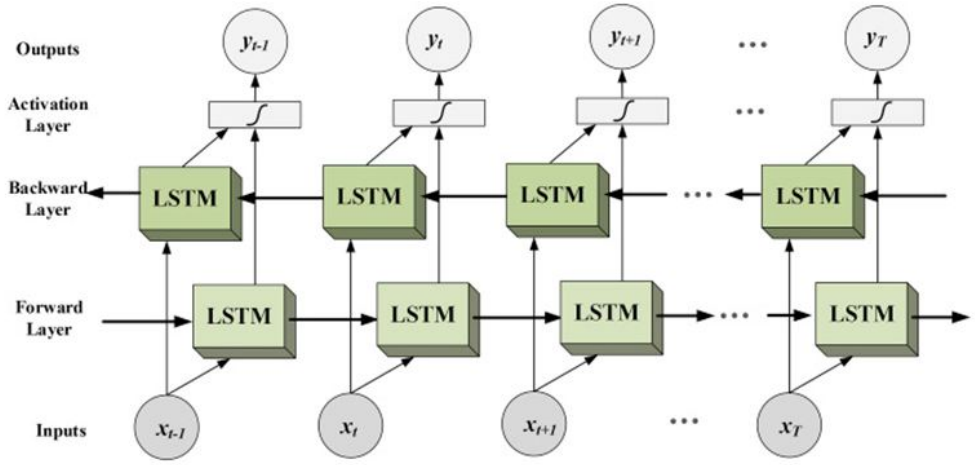

Figura 6.4: Estructura LSTM bidireccional

6.3 Implementación de los modelos

Para clasificar las trayectorias reales o para poder predecir que α deberían tener hemos utilizado redes neuronales convolucionales. Las hemos implementado utilizando dos paquetes de Python diferentes.

Para poder clasificar las trayctorias en si son superdifusivas cuando $\alpha > 1$ o subdifusivas cuando $\alpha < 1$ hemos utilizado la librería **Mcfly**, en cambio, para el modelo de regresión hemos utilizado la librería **keras**.

6.3.1. Modelo implementado con Keras

Los modelos implementados con **keras**,una librería de Python, tienen la ventaja de poder crearlos desde cero añadiendo las capas que queramos. Existen varios tipos de capas y funciones que podemos utilizar para construir un modelo que funcione bien con nuestros datos.

Después de crear y probar varios modelos llegamos a la conclusión de que la mejor opción era la siguiente:

- Optimizador **Adam**
- 2 capas convolucionales
- 5 capas LSTM bidireccionales
- Función de activación lineal

Como optimizardor utilizamos el algoritmo **Adam**, el cual es un método de descenso de gradiente estocástico que se basa en la estimación adaptativa de momentos de primer y segundo orden. Tras varias pruebas, el ratio de aprendizaje(learning rate) del optimizador lo dejamos en 0.001 ya que era el ratio de aprendizaje que mejor resultados nos daba.

Las primeras capas del modelo son dos capas convolucionales, las cuales vienen seguidas de 5 capas LSTM bidireccionales. Entre cada bloque utilizamos una capa de **Dropout** la cual ayuda a evitar el sobreajuste. Finalmente la salida de la ultima capa LSTM, pasa los datos a una capa densa totalmente conecta, en la cual utilizamos una función de activación ReLu. La salida de la misma alimenta una ultima capa densa totalmente conectada, pero esta vez utilizando una función de activación lineal.

6.3.2. Modelo implementado con Mcfly

Para crear los modelos primero vamos a generar una lista de modelos *Keras*(librería utilizada anteriormente), no entrenados, adaptados a la forma de los datos de entrada utilizando el parámetro *x_shape* . Para ello vamos a hacer uso de la función *generate_models*, a la cual mediante el parámetro *number_of_models* le decimos cuantos modelos queremos generar, en nuestro hemos decidido utilizar 5 por la falta de tiempo. Para especificar el número de clases distintas que

tenemos en las etiquetas de los datos utilizamos *number_of_classes*. A la hora de generar los modelos existen 4 tipos de arquitecturas de red disponibles en la librería *Mcfly* : las CNN, DeepConvLSTM, InceptionTime y ResNet.

Las arquitecturas de los modelos generados en el paso anterior con la función *generate_models*, se comparan con la función *train_models_on_samples* entrenándolos con los datos de entrenamiento y evaluando los modelos en el subconjunto de validación los cuales hemos creado previamente. Al buscar el modelo de mejor rendimiento, *Mcfly* realiza una búsqueda aleatoria en el espacio de los hiperparámetros. Elegimos implementar la búsqueda aleatoria, porque es simple y bastante eficaz . Esto nos ayudará a elegir el mejor modelo candidato.

El modelo que hemos utilizado para clasificar las trayectorias en superdifusivas o subdifusivas ha sido un modelo CNN, ya que al compararlo con los otros 3 tipos de arquitecturas era el que mejor resultados nos daba.

CAPÍTULO 7
Redes de pases

Este proyecto se ha centrado principalmente en la prueba de los modelos de difusión para estudiar las trayectorias y pases de un partido de fútbol e intentar obtener información que pueda resultar útil.

Con las trayectorias y los datos que tenemos del *github de Statsbomb* se pueden realizar muchos otros análisis, pero para seguir con la dinámica de las trayectorias de los pases hemos decidido estudiar los pases entre los jugadores de un equipo, en este caso, de una selección europea.

Para poder estudiar la relación de pases entre los jugadores primero vamos a introducir varios conceptos:

- **Grafo:** Un grafo es una estructura que consiste en un conjunto de objetos en el que algunos pares de objetos están relacionados. Existen distintos tipos de grafos, estos pueden ser grafos no dirigidos o grafos dirigidos. En un grafo no dirigido las relaciones o aristas son simétricas, es decir, que no tienen dirección como podemos ver en la imagen 7.1. En cambio los grafos dirigidos si que tienen dirección, es decir que las aristas, en este caso arcos o flechas, tienen un sentido definido como podemos ver en la figura 7.2. Los grafos están compuestos por varios elementos:

  - **Nodo:** es la unidad fundamental de un grafo.
  - **Arista:** es la relación que existe entre dos nodos del grafo.
  - **Grado:** el grado de un nodo es la cantidad de aristas que están conectadas a él. En el caso de los grafos dirigidos se dividen los grados en grado de entrada y grado de salida.
  - **Pesos:** existen grafos los cuales están ponderados. En este tipo de grafos cada arista tiene asignado un valor, el cual es llamado peso. Principalmente este tipo de grafos sirve para medir o registrar la importancia de los distintos enlaces lo que permite capturar más información que los grafos sin pesos.

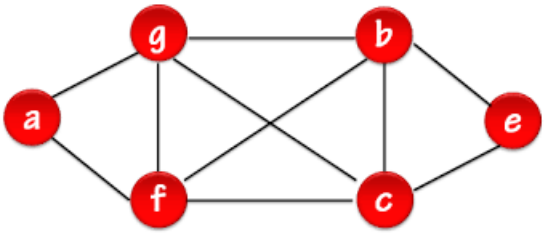

Figura 7.1: Grafo no dirigido

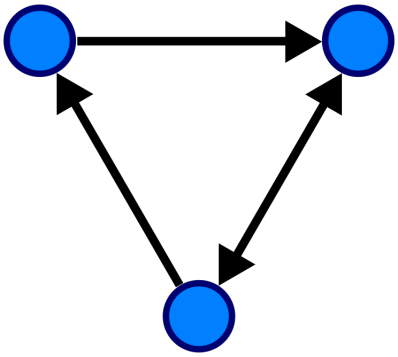

Figura 7.2: Grafo dirigido

En nuestro caso, como podemos ver en la figura 7.3, los nodos son los distintos jugadores de un equipo que están en el terreno de juego, y las aristas son los pases que han realizado entre ellos. Las aristas tienen distintos tamaños dependiendo del número de pases que se hayan realizado, es decir, cuanto más pases haya habido entre dos jugadores más gruesas serán las aristas por lo que podemos decir que los pases son los pesos de estas.

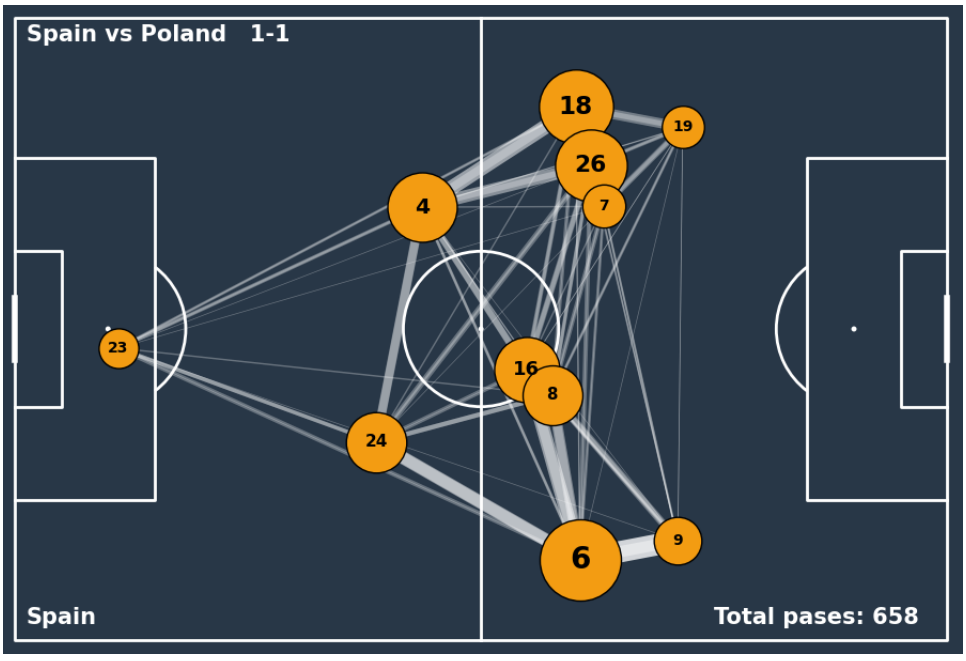

Figura 7.3: Red de pases de España en el partido España-Polonia

Para ubicar los nodos en el campo utilizamos la media de las coordenadas desde donde cada jugador ha realizado o recibido un pase. El ejemplo más claro es el portero, este ha realizado pases desde distintos sitios, como pueden ser desde un pase en el área pequeña para realizar un saque de portería hasta un pase en la frontal del área para ganara un balón divido. Hemos decidido ubicar los nodos de esta manera ya que creemos que es la forma más representativa de mostrar el juego de un equipo. En el ejemplo podemos ver como España ha planteado un partido en el que el juego principalmente se ha desarrollado en el centro del campo.

También hay que destacar el tamaño de los nodos, lo que nos ayuda a identificar los jugadores más relevantes o más influyentes (jugadores que más han participado en el juego) durante el partido. Esta información puede ser útil a la hora de analizar al rival para así poder neutralizar de una manera más eficiente al rival.

CAPÍTULO 8
Resultados

8.1 Difusión anómala

Al aplicar el modelo implementado con *keras* obtenemos para cada trayectoria que le pasamos un α. Hemos decidido estudiar los resultados de varias maneras. En primer lugar vamos a ver las diferencias que puede haber entre las distintas selecciones. Como podemos ver en los histogramas 8.1 y 8.2, la mayoría de alphas se concentran en valores cercanos a 1.8, pero la diferencia más notoria es la cantidad de trayectorias que genera cada equipo.

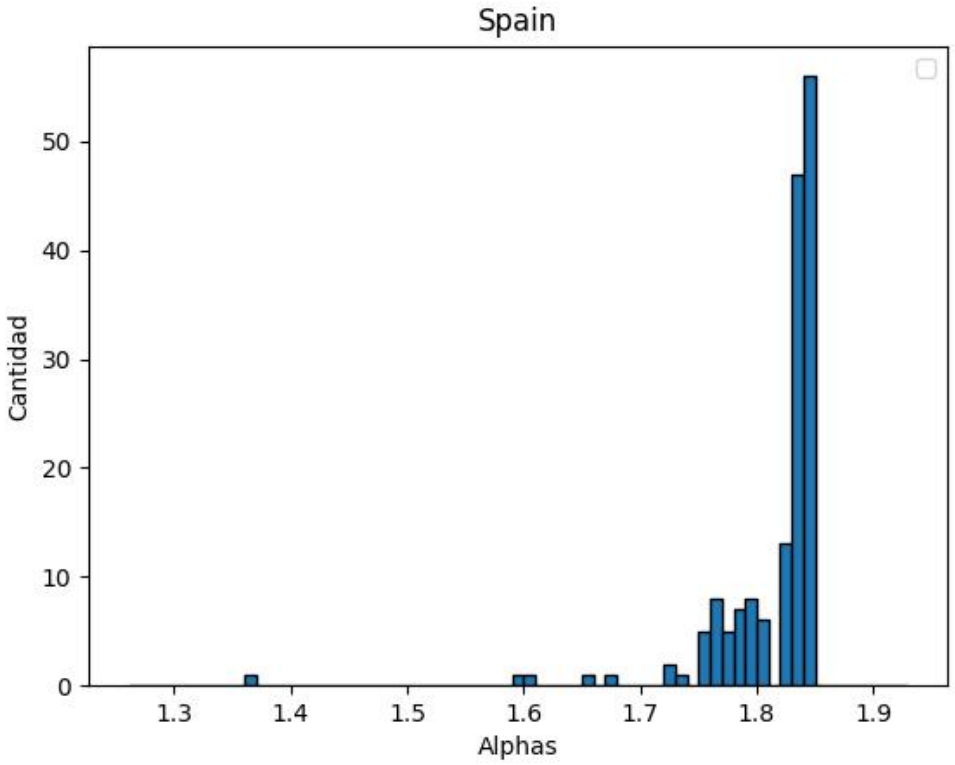

Figura 8.1: Histograma trayectorias España

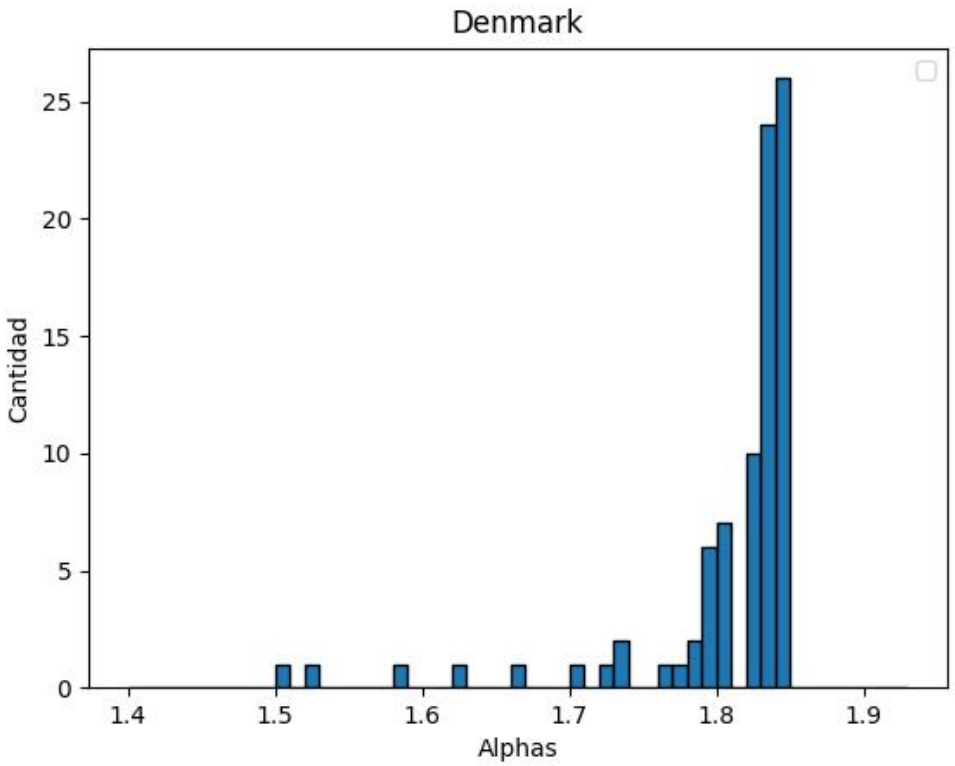

Figura 8.2: Histograma trayectorias Dinamarca

Ambas selecciones llegaron a semifinales, por lo que el número de partidos jugados no es un inconveniente a la hora de compararlas. Vemos que España con 163 trayectorias, tiene prácticamente el doble que Dinamarca la cual tiene 86 trayectorias, teniendo en cuenta que deben de ser de mínimo 10 pases, por lo que podemos deducir que España tiende a tener trayectorias más largas que muchas otras selecciones.

8.2 Teoría de Redes

Al aplicar la teoría de redes a datos de trayectorias y pases de fútbol se puede obtener información muy útil para estudiar el estilo de juego y el planteamiento del partido que tiene un equipo. Como podemos ver en la figura 8.3 este equipo tiene un estilo de juego muy marcado el cual consiste en buscar al delantero con pases largos. En el gráfico también mostramos las alineaciones, cambios y goles que han sucedido durante el partido ya que puede ser útil a la hora de analizar el partido.

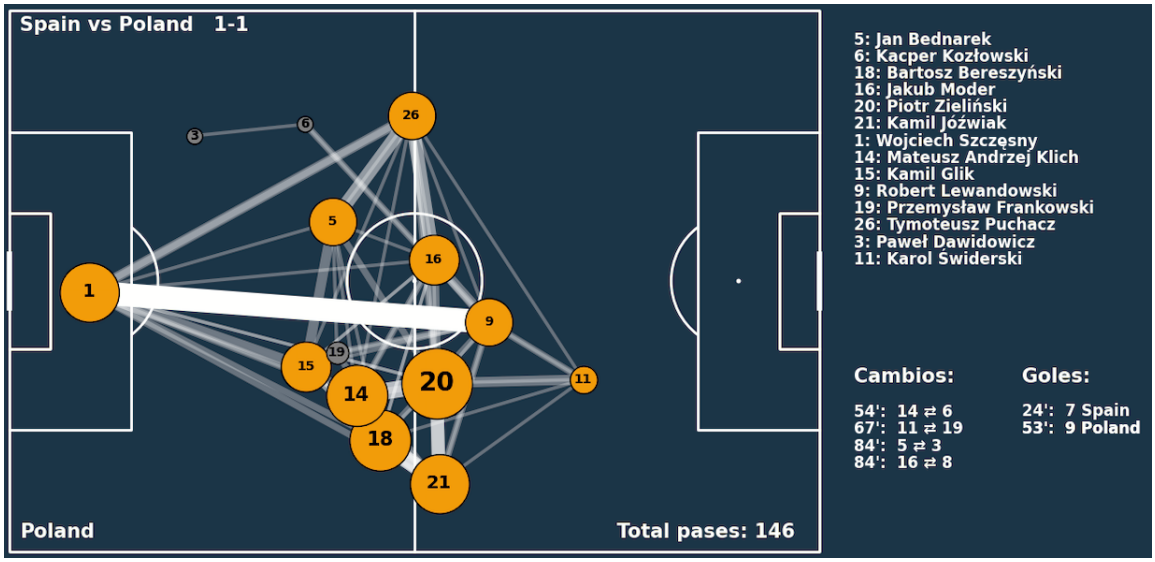

Figura 8.3: Red de pases Polonia

Estas redes de pases nos permiten analizar el estilo de juego de un equipo, lo que puede ser muy útil para plantear el partido dependiendo del oponente. Estos análisis previos a un partido pueden ser una pieza clave en las victorias de un equipo ya que permiten anticiparse al rival y plantear un juego que funcione muy bien contra el estilo del otro equipo.

CAPÍTULO 9
Página web

Tras realizar todos los análisis pensamos en cual podría ser una buena idea para mostrar toda la información que hemos obtenido. Tenemos la idea de crear un página web en la cual se muestren los análisis.

Esta página web tendrá 4 opciones principales en las que se mostrará la información de distintas formas. La página web estara estructurada de la siguiente manera:

- **Página de inicio:** como podemos ver en la figura 9.1 en esta pantalla encontramos el menú principal de la web.

Figura 9.1: WEB: Página de inicio

En esta pantalla podemos elegir entre 4 opciones, en cada una de las cuales se muestra información de distintas formas. Estas opciones son el análisis de un equipo, el análisis de un partido, la comparación de dos equipos y

finalmente el análisis realizado con difusión anómala. En cada opción se muestra lo siguiente:

- **Análisis Equipo:** como se entiende del nombre, en esta opción se muestra información del análisis de un solo equipo durante toda la Eurocopa. Como vemos en la figura 9.2 en la opción de análisis de un equipo se muestra distinta información de una selección, la cual podríamos elegir en un desplegable ubicado en la esquina superior derecha. En primer lugar, en el lado izquierdo, tenemos los jugadores convocados para el torneo, junto con el número de pases completados a lo largo de los partidos jugados en la competición. Junto a la anterior lista encontramos datos de la selección como el entrenador, presidente o año de fundación. En la parte central mostramos datos como los partidos jugados, goles a favor, goles de cabeza, tarjeta recibidas pases totales... La parte principal del análisis esta formada por 2 gráficos. El gráfico de la izquierda el cual muestra la posición en el campo desde donde se realizaron todos los tiros de la selección durante la competencia. Este gráfico puede ser útil para estudiar a un equipo y ver desde que zona del campo es más probable que disparen a puerta y así tener más cuidado a la hora de defender. El segundo gráfico esta basado en la teoría de redes, en el cual se muestra un grafo con distintos jugadores del equipo seleccionado. Como durante todo el torneo han jugado más de 11 jugadores mostramos en el gráfico los jugadores que más titularidades han tenido en cada posición a lo largo de todos los partidos. En el grafo los nodos son los jugadores del equipo y las aristas son los pases que se han realizado entre ellos. El grosor de las aristas es directamente proporcional a la cantidad des pases.

Figura 9.2: WEB: Análisis Equipo

- **Análisis Partido:** uno de los análisis más comunes el el deporte es el análisis de un partido. En esta pantalla podemos encontrar los datos más generales del partido, como son el resultado, alineaciones y cambios y goles que han ocurrido durante los 90 minutos. En cuanto a los gráficos seguimos con la dinámica de estudiar las trayectorias y pases. En este caso vamos a estudiar mediante redes de pases a los dos equipos, más concretamente a los jugadores que salieron desde el inicio, es decir el 11 titular. Como podemos ver en el ejemplo 9.3, hay una clara diferencia en los estilos de juego de los dos equipos. La primera gran diferencia es la cantidad de pases que han realizado los dos equipos, 658 España y 146 Polonia. Para ver como han planteado el partido las dos selecciones nos podemos fijar en el grosor de las aristas del grafo. Mientras que la selección de Polonia ha tenido un estilo muy marcado, el cual consistía en buscar a Robert Lewandoski (se puede ver claramente en la arista entre el nodo 1 y el nodo 9,es decir, el portero y el delantero), España ha decidido jugar más tocando y manteniendo el control del esférico. Se puede ver que los laterales de la selección, Jordi Alba(18) y Marcos Llorente(6), han tenido una gran influencia en el juego, triangulando con los centrales y mediocentros de la selección. Con este análisis pretendemos ver las diferencias en los estilos de juego de los equipos e intentar ver que estilos son más efectivos en general y cuales son mejor para enfrentarse a cada estilo de juego.

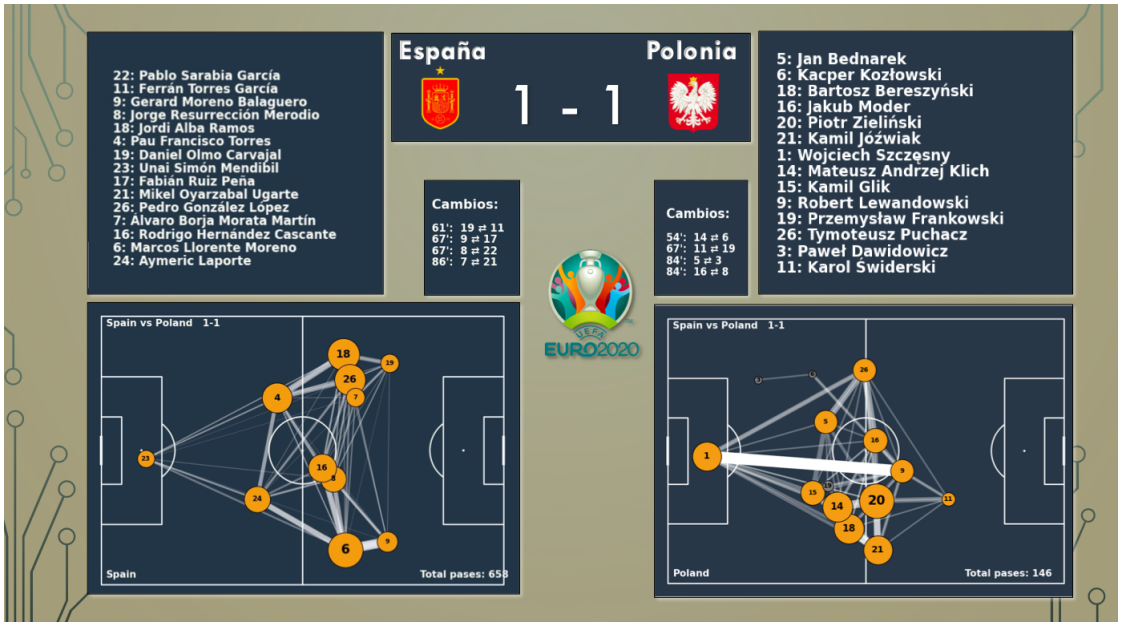

Figura 9.3: WEB: Análisis Partido

- **Comparación Equipos:** en esta opción de la página web se muestra la información de dos selecciones a lo largo de la competición 9.4. En el

centro mostramos los datos generales del equipo, como son los goles, partidos jugados, partidos ganados, perdidos o empatados. Los gráficos muestran el mapa de calor de los equipos durante el torneo, esto nos permite entender que estilo de juego han tenido y por que partes del campo han participado más con el balón. Para obtener los gráfico se han utilizado las coordenadas de los jugadores en el momento de dar o de recibir un pase.

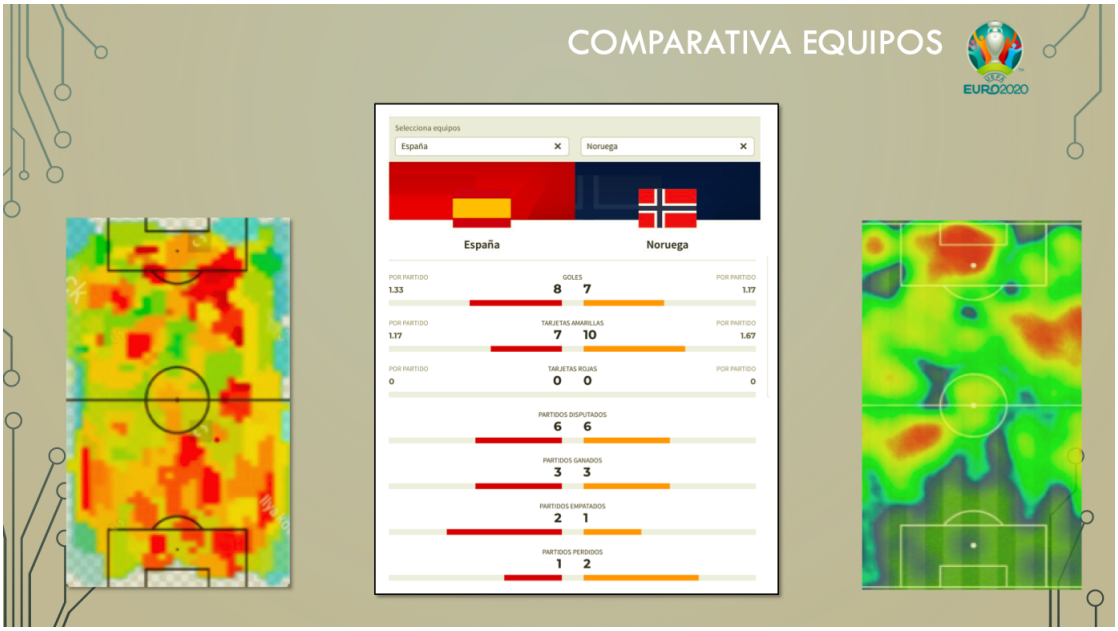

Figura 9.4: WEB: Comparativa Equipos

- **Análisis Difusión Anómala:** durante el proyecto el objetivo principal era estudiar las trayectorias del balón mediante la difusión anómala. Por eso en la pagina web creamos una pantalla en la que se muestra el análisis utilizando la difusión de una selección. Mostramos un histograma, con la cantidad de trayectorias que se han etiquetado con un α en concreto. Estos α se han determinado mediante un modelo de redes neuronales convolucionales, el cual ha sido alimentado con trayectorias artificiales creadas mediante difusión anómala. Como vemos en la figura 9.5 en esta opción de la página web encontramos un histograma con las trayectorias agrupadas según el exponente de difusión anómala (α) que el modelo de redes neuronales ha asignado a cada trayectoria. En el ejemplo podemos ver como todas las trayecotorias de la selección española durante la Eurocopa son superdifusivas, es decir, con un $\alpha > 1$

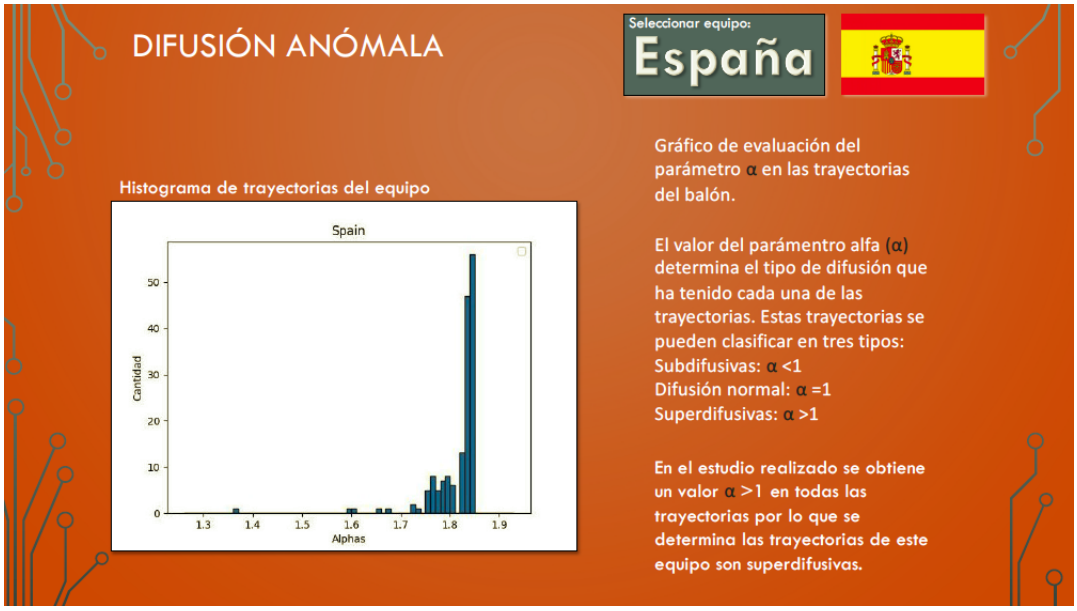

Figura 9.5: WEB: Difusión Anómala

El objetivo de la creación de esta página web sería facilitar a los usuarios información de la Eurocopa 2020 de una manera fácil, cómoda y bastante visual para que la gente pueda entender cuales han sido algunos de los puntos que han hecho que una selección haya tenido un buen o mal rendimiento.

Para finalizar aclarar que esto es una idea de como podría ser la página web, por lo que el diseño y organización podrían cambiar igual que se podría añadir información que creamos que pudiera ser útil o quitar información que no nos resulte convincente o necesaria para la web.

CAPÍTULO 10
Conclusiones

Tras todas las pruebas realizadas con los modelos de redes neuronales, ya sean los utilizados para regresión o clasificación hemos podido obtener alguna información, que en un futuro, con algo más de tiempo y profundizando y mejorando los modelos podría llegar a ser de utilidad en el mundo del fútbol.

En primer lugar hay que destacar que todas las trayectorias de todos los equipos eran superdifusivas. Como podemos observar en la figura 10.1 esto tendría sentido ya que el desplazamiento cuadrático medio aumenta con el tiempo. Esto lo podemos relacionar con la verticalidad del juego, es decir, los equipos en algún momento van a querer atacar, por lo que deberán ir hacia delante en el terreno de juego y no quedarse estancados en una misma zona.

Al aplicar los modelos de difusión no obtuvimos los resultados que nos hubieran gustado, ya que como vemos en las figuras 8.1 y 8.2 la mayoría de trayectorias se concentran en valores de α entre 1.7 y 1.9. Con estos resultados la idea sería en un futuro ampliar la cantidad de α en ese rango y estudiar las diferencias que pueda haber entre los equipos.

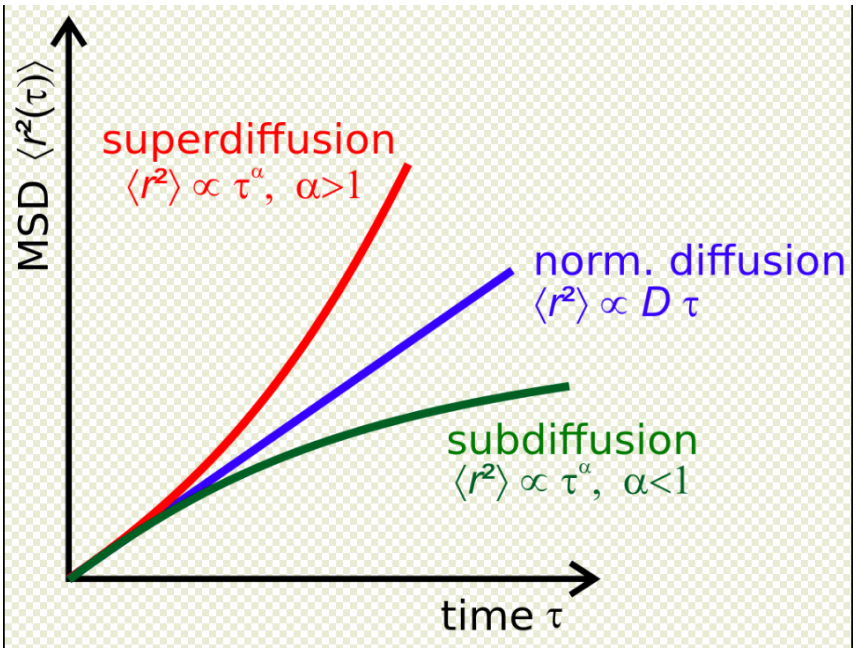

Figura 10.1: Desplazamiento cuadrático medio para diferentes tipos de difusión anómala

CAPÍTULO 11
Trabajos futuros

Tras finalizar este proyecto ha habido muchas cosas que nos gustaría haber hecho pero la falta de tiempo ha sido un inconveniente. En primer lugar, como he dicho anteriormente, realizar un análisis más concreto mediante la difusión anómala ya que en este primer análisis no hemos podido sacar información realmente determinante.

Por otro lado, nos gustaría implementar la idea que tenemos de la página web ya que creemos que podría ser de utilidad y bastante interesante. Para ello deberíamos perfeccionar el diseño, como se ve estéticamente y cambiar o añadir información que creamos que pudiera ser útil visualizarla en la web.

Si este análisis en un futuro se perfecciona y se obtienen resultados interesantes o que puedan servir a la hora de estudiar el juego de un equipo de fútbol se podría aplicar estos modelos de difusión anómala en otros deportes e intentar también obtener buenos resultados en esos deportes.

Bibliografía

[1] Metodología CRISP-DM https://www.dataprix.com/files/Metodologia_CRISP_DM.pdf

[2] S.B. Alves, G.F. de Oliveira Jr., L.C. de Oliveira, T. Passerat-de Silans, M. Chevrollier, M. Oriá, and H.L.D.S. Cavalcante. Characterization of diffusion processes: Normal and anomalous regimes. *Physica A*, 447:392 – 401, 2016.

[3] A.S. Bodrova, A.V. Chechkin, and I.M. Sokolov. Scaled brownian motion with renewal resetting. *Phys. Rev. E*, 100(1), Jul 2019.

[4] A.M. Barrero, I.M. Gutiérrez, M.F. Prieto Análisis del modelo de juego en un equipo de fútbol profesional de la Bundesliga de Alemania. Estudio caso (Analysis of the game model in a professional football team in the German First Division. Case study). *Retos*, 39, 628–634.

[5] Beach Volley Tour. 10 Deporte más practicados del mundo https://beachvolleytour.es/10-deportes-mas-practicados-en-el-mundo/ Consultado 23 de noviembre de 2022.

[6] J.M. Buldú, J.M. Busquets, J.d Martínez, J.H. Herrera-Diestra, et al. Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. *Front. Psychol.* 9, 1900 (2018).

[7] S. Cacho-Elizondo y J.D. Lázaro. Big Data en el Futbol. El Nuevo Juego. 365. 23-26 (2019).(2019).

[8] Cadena Ser. *Récord mundial: 91.553 personas en el Clásico femenino en el Camp Nou* https://cadenaser.com/2022/03/30/un-clasico-de-record-la-asistencia-al-camp-nou-puede-ser-historica/

[9] D.C. Cireşan, U. Meier, L.M. Gambardella, J. Schmidhuber . Deep Big Multilayer Perceptrons for Digit Recognition. In: Montavon, G., Orr, G.B., Müller, KR. (eds) Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, vol 7700. Springer, Berlin, Heidelberg (2012).

[10] Drilab. Análisis de ubicación: el mejor destino posible para Moisés Caicedo https://www.driblab.com/es/ubicacion-de-jugadores/analisis-de-ubicacion-el-mejor-destino-posible-para-moises-caicedo/ Consultado el 23 de noviembre de 2022.

[11] K. Fukushima. https://www.fi.edu/laureates/kunihiko-fukushima. Consultado el 23 de noviembre de 2022.

[12] Ò. Garibo-i-Orts, A. Baez-Boscá, M.A. García March. Efficient recurrent neural network methods for anomalously diffusing single particle short and noisy trajectories. *J. Phys. A: Math. Theor.* 54.50: 504002 (2021).

[13] M. Hellmann, J. Klafter, D.W. Heermann, and M. Weiss. Challenges in determining anomalous diffusion in crowded fluids J. Phys: Condens. Matter, 23(23) 234113 (2021).

[14] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735– 1780, 1997.

[15] M. Javanainen, H. Hammaren, L. Monticelli, J.H. Jeon et al. Anomalous and normal diffusion of proteins and lipids in crowded lipid membranes. *Faraday discussions*, 161, 397-417 (2013).

[16] A. Joseph, N.E. Fenton, and M. Neil. Predicting football results using Bayesian nets and other machine learning techniques, *Knowledge-Based Systems*, 19(7), 544-553 (2006).

[17] E. Kepten, I. Bronshtein, and Y. Garini. Improved estimation of anomalous diffusion exponents in single-particle tracking experiments. *Phys. Rev. E*, 87(5):052713, 2013.

[18] E. Kepten, A. Weron, G. Sikora, K. Burnecki, Y. and Garini. Guidelines for the fitting of anomalous diffusion mean square displacement graphs from single particle tracking experiments. *PLoS ONE* 10, e0117722 (2015).

[19] J. Klafter and G. Zumofen. Lévy statistics in a hamiltonian system. *Phys. Rev. E*, 49:4873–4877, Jun 1994.

[20] P. Kowalek, H. Loch-Olszewska, and J. Szwabiński. Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach. *Phys. Rev. E*, 100(3):032410, 2019.

[21] S.C. Lim and S.V. Muniandy. Self-similar Gaussian processes for modeling anomalous diffusion. *Phys. Rev. E*, 66:021114, Aug 2002.

[22] Y.A. LeCun, L. Bottou, G.B. Orr, K.R. Müller. Efficient BackProp. In: Montavon, G., Orr, G.B., Müller, KR. (eds) Neural Networks: Tricks of the Trade. *Lecture Notes in Computer Science*, 7700. Springer, Berlin, Heidelberg (2012).

[23] Z.C. Lipton, J.n Berkowitz, and C. Elkan. A critical review of recurrent neural networks for sequence learning. arXiv:1506.00019, 2015.

[24] M.A. Lozano, O. Garibo i Orts, E. Piñol, M. Rebollo et al. Open data science to fight COVID-19: Winning the 500k XPRIZE Pandemic Response Challenge. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases ECML-PKKD 2021* (pp. 384-399). Springer, Cham.

[25] Moneyball Method: Using Data to Build a Football Dream Team (On a Budget) https://www.graphext.com/post/the-moneyball-method-using-data-to-build-a-football-dream-team-on-a-budget

[26] B.B. Mandelbrot and J.W. Van Ness. Fractional brownian motions, fractional noises and applications. *SIAM Rev.*, 10(4):422–437, 1968.

[27] C. Manzo and M.F. Garcia-Parajo. A review of progress in single particle tracking: from methods to biophysical insights. *Rep. Prog. Phys.* 78, 124601 (2015).

[28] P. Massignan, C. Manzo, J. A. Torreno-Pina, M. F. García-Parajo, M. Lewenstein, and G. J. Lapeyre. Nonergodic subdiffusion from Brownian motion in an inhomogeneous medium. *Phys. Rev. Lett.*, 112:150603, Apr 2014.

[29] R. Metzler, J.-H. Jeon, A.G. Cherstvy, and E. Barkai. Anomalous diffusion models and their properties: non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. *Phys. Chem. Chem. Phys.*, 16(44):24128–24164, 2014.

[30] X. Michalet. Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. *Phys. Rev. E*, 82:041914, Oct 2010.

[31] N. Monnier, S.M. Guo, M. Mori, J. He, P. Lénárt, and M. Bathe. Bayesian approach to msd-based analysis of particle motion in live cells. *Biophys. J.*, 103(3):616–626, 2012.

[32] G. Muñoz-Gil, G. Volpe, M.A. Garcia-March, E. Aghion et al. Objective comparison of methods to decode anomalous diffusion. *Nat. Commun.* 12.1 1-16 (2021).

[33] G. Muñoz-Gil, G. Volpe, M.A. García-March, R. Metzler, M. Lewenstein, C. Manzo The Anomalous Diffusion challenge.: single trajectory characterisation as a competition. *Emerging Topics in Artificial Intelligence 2020*. Vol. 11469. SPIE, 2020.

[34] F.A. Oliveira, R.M.S. Ferreira, L.C. Lapas, and M.H.. Vainstein, Anomalous diffusion: A basic mechanism for the evolution of inhomogeneous systems. *Front. Phys.* 7 (2019): 18.

[35] M.J Saxton and K. Jacobson. Single-particle tracking: applications to membrane dynamics. *Annual review of biophysics and biomolecular structure*, 26(1):373–399, 1997.

[36] H. Scher and E.W. Montroll. Anomalous transit-time dispersion in amorphous solids. *Phys. Rev. B*, 12:2455–2477, Sep 1975.

[37] Statsbomb. Github. https://github.com/statsbomb/open-data

[38] A. Weron, K. Burnecki, E.J. Akin, L. Solé et al. Ergodicity breaking on the neuronal surface emerges from random switching between diffusive states. *Sci. Rep.*, 7(1):1–10, 2017.

APÉNDICE A
Anexo I: Objetivos de desarrollo sostenible

Objetivos de Desarrollo Sostenible	Alto	Medio	Bajo	No procede
ODS 1. **Fin de la pobreza.**		X		
ODS 2. **Hambre cero.**			X	
ODS 3. **Salud y bienestar.**	X			
ODS 4. **Educación de calidad.**				X
ODS 5. **Igualdad de género.**		X		
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**	X			
ODS 9. **Industria, innovación e infraestructuras.**			X	
ODS 10. **Reducción de las desigualdades.**		X		
ODS 11. **Ciudades y comunidades sostenibles.**				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17. **Alianzas para lograr objetivos.**				X

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

Los Objetivos de Desarrollo Sostenible (ODS), sucesores de los Objetivos de Desarrollo del Milenio (ODM), fueron establecidos por la Asamblea General de las Naciones Unidas el 25 de septiembre de 2015 para abordar los principales desafíos globales. Un total de 193 países han elegido compromisos de desarrollo comprometidos a responder a las necesidades globales actuales establecidas en un programa de 17 objetivos (que incluyen 169 metas) que deben alcanzarse para 2030

Alguno de los ODS se pueden relacionar con el mundo del fútbol, y más concretamente con el análisis de los partidos, estos son:

- **ODS1. Fin de la pobreza:** El fútbol es el deporte más practicado del mundo, por ende, es el que más repercusión tiene. Gracias a esto es un deporte que tiene muchas facilidades a la hora de practicarlo ya que prácticamente en cualquier pueblo tienen equipo de fútbol. Por otro lado, debido a su alta visibilidad, muchas empresas están invirtiendo en él, lo que ha hecho que en los últimos años el fútbol se haya ido privatizando y encareciendo poco a poco, lo que es un problema para los aficionados con menos solvencia económica. Se debería intentar desprivatizar las retransmisiones y hacerlo apto para todo típo de público independientemente de su nivel económico.
- **ODS3. Salud y bienestar:** Para las personas, la salud es un aspecto fundamental en su vida, ya se la salud física o mental. Practicar algún deporte ayuda a mantener una buena salud física, y sentirse bien físicamente, ayuda a la salud mental. Además de practicarlo, el fútbol ayuda a mucha gente a evadir sus problemas mientras esta viendo un partido. Por todo esto, es necesario fomentar la práctica de algún deporte ya que tienen muchos aspectos positivos.
- **ODS5. Igualdad de género:** La igualdad de género implica que todas las personas tengan los mismos derechos, recursos y oportunidades independientemente de su identidad de género. El mundo del fútbol ha sido, históricamente, un deporte principalmente practicado por hombres, pero en los últimos años, el fútbol femenino ha ido ganando repercusión, como pudimos ver en uno de los últimos clásicos femeninos en el cual llenaron casi todo el *Camp Nou* [8]. Poco a poco, aunque aún falta bastante, las mujeres van teniendo las mismas oportunidades que los hombres en este deporte.
- **ODS8. Trabajo decente y crecimiento económico:** Un crecimiento económico inclusivo y sostenido puede impulsar el progreso, crear empleos decentes para todos y mejorar los estándares de vida. El fútbol en los últimos años ha ido creando muchos puestos de trabajo distintos, lo que ha conllevado un crecimiento económico muy grande en muchos países. No solo trabajan los futbolistas, si no que también se crean puestos de trabajo en el cuerpo técnico, medios de comunicación, árbitros, mantenimiento de las instalaciones deportivas... Todo esto influye en el ODS1, ya que la creación de trabajo ayuda a mejorar la calidad económica de muchas personas.

- **ODS10. Reducción de las desigualdades:** Reducir la desigualdad y no dejar a nadie atrás es fundamental para lograr los Objetivos de Desarrollo Sostenible.

  La desigualdad dentro y entre los países es una preocupación constante. A pesar de algunos signos positivos de una reducción de la desigualdad en algunas áreas, como una reducción de la desigualdad de ingresos en algunos países y un estatus comercial preferencial que favorece a los países de bajos ingresos, la desigualdad persiste.

  En el fútbol aún existe desigualdad entre algunos países, ya sea en términos de instalaciones y material o por la cantidad de recursos obtenidos por la retransmisión de los partidos. Aunque aún exista esta desigualdad el fútbol ha permitido igualar las posibilidades que tienen todos los países de poder jugar, como estamos viviendo ahora, un mundial de fútbol independientemente de la diferencia económica entre países.

APÉNDICE B
Anexo II: Relación de los estudios con el proyecto

Para finalizar el proyecto vamos a relacionar las distintas competencias transversales y conocimientos que hemos trabajado a lo largo del grado de Ciencia de datos. Durante los distintos cursos y asignaturas que hemos cursado se han trabajado varias competencias que nos han resultado útiles para realizar este proyecto.

Una de las competencias que me ha resultado más necesaria para realizar este trabajo es el *Pensamiento crítico* ya que al ir avanzando en el proyecto he tenido que interpretar información como la de la difusión anómala porqué no había impartido ninguna asignatura anteriormente relacionada con este proceso. También analizar la información y obtener alguna conclusión de esta ha sido posible gracias a la evolución que he tenido en el pensamiento crítico durante distintas asignaturas del grado como pueden ser proyecto I, II y III o las prácticas que realicé al finalizar el cuarto curso académico.

La *Comunicación efectiva* ha sido uno de mis puntos débiles siempre y aunque aún me queda mucho por mejorar a lo largo de los distintos trabajos que he realizado he ido mejorando aplicando todos lo aprendido en este proyecto de final de grado.

Aunque se han aplicado muchas de las competencias transversales a lo largo de este proyecto la más importante sin ninguna duda es la de *Comprensión e integración* de los conocimientos y habilidades que hemos adquirido durante estos cuatro cursos. El conocimiento principal que se ha aplicado en el trabajo es el de la programación en *python*, ya que todos los análisis se han realizado con este lenguaje de programación. Además de la programación toda la parte de grafos y redes que hemos utilizado para analizar los partidos de fútbol se ha podido implementar gracias a los conocimientos obtenidos en asignaturas como *Modela Discreto y Teoría de la Información*. Todos estos conocimientos obtenidos se han podido aplicar de una manera correcta gracias a las distintas asignaturas en las que trabajamos la competencia transversal de *Aplicación y pensamiento práctico*.

Siguiendo con lo aprendido en la carrera, destacar algunos conocimientos como el análisis exploratorio o la visualización de datos que han sido de mucha utilidad a la hora de completar el trabajo ya que han mejorado aspectos muy primordiales como son el preprocesado de los datos y la visualización de los resultados.

Finalmente destacar la competencia de *Innovación, creatividad y emprendimiento* ya que se ha aplicado difusión anómala a datos futbolísticos y se puede decir que estos modelos no se utilizan en el mundo del análisis deportivo. Por la innovación que esto conlleva era más complicado encontrar información de otros proyectos que pudiera estar relacionada y esto ha hecho que la creatividad y emprendimiento tuvieran un papel importante en el tabajo.