UNIVERSIDAD
DE LA REPUBLICA
URUGUAY

Aprendizaje Automático para Competencias Deportivas

Proyecto de grado

Christian Maidana

Vicente Ferreyra

Tutores:

Dr. Andrés González Ramírez (ISEF),

Dr. Guillermo Moncecchi,

Dra. Lorena Etcheverry

Ingeniería en Computacion
Facultad de Ingeniería
Universidad de la República

Montevideo – Uruguay

Octubre de 2021

Agradecimientos

A nuestras familias y amigos por el apoyo brindado durante nuestra carrera y elaboración de este proyecto. Además agradecemos a nuestros tutores, Andrés, Guillermo y Lorena por acompañarnos y guiarnos durante el transcurso de este proyecto. Christian quiere agradecer también a su esposa por el apoyo durante todo este proceso.

RESUMEN

La Ciencia de Datos (CD) es un campo interdisciplinario que utiliza métodos científicos, procesos, algoritmos y sistemas para extraer conocimiento u obtener un mejor entendimiento de datos estructurados o no estructurados. La CD aplicada al deporte es relevante tanto para la academia como para la industria, dado que permite realizar análisis que antes no eran posibles, teniendo como consecuencia impacto en la toma de decisiones, como por ejemplo el fichaje de jugadores, predicción de resultados, etc. En este proyecto se elabora una investigación de los desarrollos existentes referentes a la CD aplicada al deporte, tanto estudios académicos como innovaciones del ámbito privado, realizando un análisis más exhaustivo sobre el fútbol. A su vez, se detallan algunas fuentes de datos disponibles que podrían usarse para la aplicación de CD.

Con el fin de generar una herramienta que aporte información de utilidad para los entrenadores, se elabora un prototipo para la predicción de distintos atributos (tiros de esquina y tiros al arco). A partir de datos obtenidos de Whoscored de las temporadas 2016 a 2019/2020 y el uso de tres clasificadores K Nearest Neighbors, Random forest y Regresión lineal, se logran resultados auspiciosos llegando a una *accuracy* de 87 % para tiros de esquina y 62 % para tiros al arco.

Palabras claves:
Inteligencia artificial, Aprendizaje automático, Deporte, Fútbol.

Tabla de contenidos

Capítulo 1

Introducción

La Ciencia de Datos (CD) es un campo interdisciplinario que utiliza métodos científicos, procesos, algoritmos y sistemas para extraer conocimiento u obtener un mejor entendimiento de datos estructurados o no estructurados. Algunos de los campos que trata la CD son el Aprendizaje Automático (AA), el modelado estadístico o el análisis de datos de alta dimensionalidad [1].

Todo proceso dentro de la CD implica la limpieza, el procesamiento y la posterior manipulación de los datos para realizar análisis más avanzados. Habitualmente, los resultados obtenidos son revisados en busca de patrones que permiten obtener información fundamentada que no era visible a priori, permitiendo tener un mejor conocimiento del dominio.

La CD aplicada al deporte es relevante tanto para la academia como para la industria, dado que permite realizar análisis que antes no eran posibles, teniendo como consecuencia impacto en la toma de decisiones, como por ejemplo el fichaje de jugadores, predicción de resultados, etc. Actualmente, es posible encontrar una variedad de datos muy extensa sobre los distintos deportes provenientes de sensores colocados en los jugadores, cámaras que registran los encuentros o incluso obtención manual por parte de humanos. El problema radica en qué hacer con esos datos y cómo obtener información valiosa que permita una mejor toma de decisiones, por ejemplo, para prevenir lesiones o mejorar aspectos tácticos y técnicos, y como consecuencia, obtener mejores resultados deportivos [2].

La realidad actual indica que no es suficiente con lo que se observa, sino que es necesario comprender, interpretar e incluso intentar predecir lo que sucederá, para poder generar consecuencias favorables. Los métodos de CD son útiles para aplicar a este tipo de problemas [3].

La CD es aplicada en diferentes deportes, entre los que encontramos fútbol, baloncesto, balonmano, béisbol, etc. Se pueden encontrar trabajos que van desde la predicción de resultados en el balonmano masculino [4] al análisis del rendimiento de equipos y jugadores de la "Major League baseball" (MLB), con el fin de encontrar sus fortalezas y debilidades [5].

Si bien se indaga en diferentes deportes, el presente trabajo se centra en la CD aplicada al fútbol. Al igual que en los otros deportes, existen diferentes trabajos y empresas privadas que aplican CD para lograr un beneficio deportivo, ya sea mediante la predicción de resultados, análisis de jugadores para la optimización del reclutamiento y la prevención de lesiones o el análisis general del equipo mediante distintos indicadores [6]-[8].

El objetivo del fútbol es bastante simple de entender: introducir la pelota en el arco del equipo rival y evitar que el rival lo haga en el arco propio. Sin embargo, en la práctica resulta ser un deporte muy complejo dada la cantidad de variables que interactúan: capacidades técnicas y tácticas, estado físico, económico, social y psicológico, tanto propio como del equipo rival, sumado a la planificación previa, la estrategia que se utiliza y los ajustes que realiza el entrenador durante el encuentro. El análisis y la optimización de estas variables procura minimizar los riesgos y maximizar los resultados positivos obtenidos, aunque no resulta fácil dada la dificultad a la hora de modelarlas [9].

En la actualidad, para la mayoría de los clubes del ámbito local, este tipo de análisis se realizan mediante la observación de partidos previos en los cuales los clubes invierten mucho tiempo y recursos. Sin embargo, estos análisis están condicionados por la subjetividad del observador y a veces no son suficientes. [10]

A partir de aquí surge la motivación de este proyecto: ayudar a los entrenadores del ámbito local a la toma de decisiones. Es decir, poder dar información

a partir del procesamiento de un elevado número de datos que permitan aportar soluciones técnicas y tácticas para mejorar los resultados deportivos.

El objetivo general de este proyecto fue generar una herramienta que permita dar información contextual a los entrenadores sobre los distintos atributos que se presentan en los encuentros.

Para alcanzar este objetivo se establecieron los siguientes objetivos específicos:

- Relevar los antecedentes de la aplicación de CD en el deporte, especialmente en el fútbol
- Identificar las fuentes de datos públicas existentes
- Recopilar un conjunto de datos con suficiente nivel de detalle para poder ser utilizado
- Crear un prototipo que permita la predicción de diferentes atributos de un partido y su relación con otros atributos

Para lograr el primer objetivo, se estudió primero el estado actual de la CD en el deporte, tanto a nivel académico como a nivel industrial, con el fin de conocer qué técnicas se utilizan actualmente y definir qué se podría implementar en el medio local.

En cuanto a los objetivos 2 y 3, se buscaron fuentes de datos públicas y se formó un conjunto de datos con granularidad a nivel de partido, lo cual permitió poder utilizarlo más adelante.

Para la creación del prototipo se implementaron tres modelos de AA para predecir la cantidad de tiros al arco y de tiros de esquina generados por un equipo durante un partido. Estos modelos fueron entrenados y evaluados con el conjunto de datos antes formado.

El presente trabajo se organizó en los siguientes cinco capítulos: introducción, antecedentes, presentación del problema, implementación de la solución y conclusiones.

En el próximo capítulo se profundiza en los antecedentes relevados durante el transcurso del trabajo, tanto estudios académicos como innovaciones del ámbito privado que utilizan CD en los deportes antes mencionados y un análisis más exhaustivo sobre el fútbol. A su vez, se detallarán las fuentes de datos existentes.

Capítulo 2

Antecedentes

En esta sección primero, se presentan trabajos con foco en varios deportes como balonmano, fútbol, básquetbol y béisbol y luego se detallan algunos trabajos que utilizan la CD aplicada al fútbol. También se reseñan emprendimientos que ofrecen servicios de CD aplicada al deporte.

2.1. Ciencia de datos aplicada al deporte

Para poder tener una primera visión global de los avances y la disponibilidad de datos se evaluaron trabajos relacionados a diferentes deportes. Dentro de la variedad de objetivos con los que se utiliza la CD en el mundo del deporte se pueden remarcar predicción de resultados, mejora en el fichaje, prevención de lesiones y comprensión de las situaciones del juego para ayudar a la toma de decisiones.

Uno de los objetivos más importantes y transversal a todos los deportes es la predicción de resultados. Si bien no existe una fórmula que determine qué factores conducen a la victoria de un equipo, se puede ver que a través del análisis de varios años de estadísticas es posible encontrar tendencias. Uno de los trabajos estudiados busca encontrar y analizar las debilidades de diferentes equipos y jugadores de la "Major League baseball" (MLB) y ordenarlos en un *ranking* de acuerdo a esto, para luego poder predecir futuros resultados. En este trabajo se utilizó una fuente de datos construida a partir de 10 años de información pública y del pasado y en promedio se obtuvo un 60 % de acierto en la predicción del equipo ganador para cada equipo [5].

También se analizó un trabajo que busca predecir el resultado del equipo ganador para balonmano masculino. Allí se integran redes neuronales con algoritmos de selección de características y obtuvo un 88,24 % de acierto. La fuente de datos utilizada consistió en los datos de los mundiales de balonmano de 2016 y 2018. A su vez, en este trabajo no solo se predijeron resultados sino que se evaluó el peso de cada característica con respecto a la clase objetivo. Esto es interesante ya que a pesar de no tener certeza del resultado del partido, el cual tiene un alto coeficiente de azar, esta información puede ser utilizada por un entrenador para ponderar algunas características al momento de formar y entrenar un plantel. Entre los atributos más relevantes se mencionan la eficacia del arquero, los lanzamientos, calidad de juego (asistencias y pérdidas de posesión) y la eficacia de ataque [4].

De la misma manera que hemos visto para béisbol y balonmano, para el baloncesto, uno de los proyectos analizados intenta predecir el equipo ganador en partidos de "National Basketball Association" (NBA). En este caso se predijo el equipo ganador de un partido en la NBA, obteniendo porcentajes de acierto de hasta un 74 % para el algoritmo Random Forest. En este trabajo se utilizó la base de datos del portal Basketball Reference, particularmente los años 2015/2016 [11].

De la mano con la mejora en el rendimiento pero centrando el objetivo en el equipo, existe una investigación que utiliza la técnica de Deep Imitation Learning para modelar situaciones defensivas. Como fuente de datos utiliza imágenes de 100 partidos considerando cada jugada de ataque del equipo rival. El algoritmo parte con los jugadores en la posición que comienza la jugada y a partir de ahí intenta predecir las posiciones que minimizan la probabilidad de gol del rival. Al ser en tiempo real no es posible su aplicación directamente en el partido. Sin embargo, el entrenador puede recibir un informe con estas situaciones y modificar los entrenamientos [12].

Existen también empresas privadas que desarrollaron aplicaciones comerciales de CD aplicada al deporte. A modo de ejemplo Olocip [13] es una empresa española que cuenta con herramientas para análisis y predicción en distintos deportes como fútbol, basquetbol o tenis. Esta empresa se publicita mencio-

nando que sus herramientas de análisis de datos permiten mejorar los equipos. Sin embargo, no se pudo encontrar información de que técnicas utilizan ni con que datos trabajan. Una de las aplicaciones, relacionada al basquetbol, es una plataforma de scouting y análisis de rendimiento predictivo que permite analizar a cada jugador de forma contextualizada y predecir su rendimiento basado en el equipo de destino, utilizando análisis y técnicas descriptivas basadas en el rendimiento pasado del jugador. Es decir, considerando las estadísticas históricas del jugador el sistema predice que rendimiento tendría en un equipo destino. Cabe mencionar que esta no es la única empresa que provee este tipo de soluciones. Más adelante en el trabajo se comentan otras.

2.2. Ciencia de datos y fútbol

Para tener un objetivo más acotado de análisis, se optó por seleccionar como deporte objetivo el fútbol, ya que es el deporte comúnmente denominado más popular en Uruguay y el deporte colectivo más practicado a nivel nacional [14].

La predicción del resultado de un partido no es sencilla debido a que hay una gran cantidad de factores que se deben tener en cuenta y que además muchas veces no se pueden representar cuantitativamente, predecir dichos resultados es un proceso complejo [15]. Existen numerosos artículos y trabajos que intentan resolver el problema en base a datos estadísticos de partidos previos y algoritmos de aprendizaje automático [16].

Uno de los trabajos analizados, realizado en 2020, intenta predecir el ganador o perdedor de un partido y los atributos más relevantes para estas predicciones, tomando como fuente de datos los partidos de la copa del mundo 2018 con sus 75 atributos reportados por FIFA, de los cuales los más relevantes fueron media de pases intentados, promedio de distancia cubierta, promedio de entrega de pelota en ataque. Este trabajo utiliza como algoritmo de clasificación una red neuronal supervisada, obteniendo 72.7 % y 83.3 % de acierto para predicción de derrotas y victorias respectivamente. En base a los atributos considerados más relevantes, es posible utilizar el resultado del peso de los atributos para adaptar el entrenamiento, las tácticas y el análisis del rival para

mejorar el rendimiento [2].

Con el mismo enfoque encontramos otro trabajo donde se predicen los resultados de la Liga Española, utilizando los datos de las temporadas 2012-13, 2014-15 y 2015-16 aplicando distintos algoritmos de aprendizaje automático como KNN, regresión logística, random forest y Support vector machine (SVM). En este trabajo, se generaron atributos sintéticos como los días desde el partido anterior y días hasta el siguiente partido o la importancia del partido. Se concluyó que la estrategia de agrupar los partido ("clusters") con respecto a una serie de atributos obtiene mejores resultados, en este caso la técnica utilizada para generar los "clusters" fue KNN [17].

Con la misma meta en común, en el año 2017 tuvo lugar el Soccer Prediction Challenge, un concurso que consistió en predecir el resultado de un conjunto de partidos a partir de modelos entrenados con partidos pasados. Una de las particularidades de este concurso es que los partidos que se planteaba predecir eran en el futuro, es decir, no se tenía ningún tipo de información acerca de ellos al momento de entrenar los modelos. El conjunto de datos utilizado por el modelo ganador, al igual que el utilizado por todos los demás candidatos, consiste de un conjunto con los resultados de más de 200.000 partidos de diferentes ligas del mundo a lo largo de 17 años. De cada partido se sabe solamente el nombre de los equipos participantes y el resultado final. El modelo ganador fue un modelo basado en atributos calculados. La clave del éxito de este modelo fue la obtención de nuevas atributos a partir de la información que se tenía, entre las cuales se pueden ver posición en la tabla, importancia del partido (dependiendo las posiciones de los equipos participantes), estado actual del equipo (dependiendo de los últimos cinco partidos), entre otras [6].

Dejando de lado el objetivo del equipo ganador, es posible enfocar el análisis en generar información que sea útil para entrenadores o analistas deportivos. A modo de ejemplo mencionamos el índice Soccer Power Index (SPI) calculado y mantenido por el portal FiveThirtyEight [7], este es un estimado de la fortaleza general de un equipo, que indica cómo le iría al equipo en un partido contra un rival promedio en un estadio neutral. Antes de empezar cada temporada se calcula el valor de SPI de pretemporada (Preseason SPI rating), el cual se basa en el resultado de la temporada pasada y el valor del mercado del equipo,

el cual es tomado de TransferMarket [18]. Luego de cada partido este valor se ajusta basado en la performance del equipo y la fortaleza del oponente . En la figura 2.1 se presenta un ejemplo de este indicador para la Liga Española.

LA LIGA
2022-23

Spain
Updated Sept. 18, 2022, at 4:58 p.m.

	TEAM RATING			AVG. SIMULATED SEASON		END-OF-SEASON PROBABILITIES			
TEAM	SPI	OFF.	DEF.	GOAL DIFF.	PTS.	EVERY POSITION (↓20th 1st↓)	RELEGATED	QUALIFY FOR UCL	WIN LA LIGA
Real Madrid 18 pts	87.1	2.7	0.5	+50	85		<1%	95%	48%
Barcelona 16 pts	88.7	2.7	0.4	+56	84		<1%	93%	41%
Villarreal 11 pts	79.3	2.2	0.6	+26	67		<1%	49%	4%
Atlético Madrid 10 pts	79.3	2.1	0.6	+23	66		<1%	47%	3%
Athletic Bilbao 13 pts	74.1	1.9	0.7	+13	61		<1%	27%	1%
Real Betis 15 pts	69.9	1.8	0.8	+7	60		<1%	24%	1%
Real Sociedad 10 pts	73.5	1.9	0.7	+5	58		2%	20%	<1%
Valencia 9 pts	71.0	1.8	0.7	+7	54		3%	12%	<1%
Osasuna 12 pts	68.1	1.7	0.8	-2	53		4%	9%	<1%
Celta Vigo 7 pts	68.9	1.8	0.8	-5	51		7%	7%	<1%

Figura 2.1: Tabla de la Liga Española con las posiciones actuales y el valor SPI de cada equipo.

Otra métrica que ha tomado gran relevancia en los últimos años es la de goles esperados, del inglés *expected goals* (xG). Esta métrica, popularizada por OptaSports [19], corresponde a los goles que se espera que un jugador o equipo haga en un partido. Mide la calidad de una oportunidad de gol en base a la probabilidad de que el equipo convierta un gol desde una posición en particular del terreno durante un momento particular del encuentro. Cada oportunidad generada adopta un valor entre cero y uno, siendo uno un caso con alta probabilidad de terminar en gol y cero una ocasión donde es prácticamente imposible anotar. Estas probabilidades se suman dando como resultado el valor de xG. Esta métrica utiliza datos de tiros recabados previamente como los presentados en la figura 2.2, de los cuales se considera su ubicación en el campo como uno de los atributos fundamentales así como con qué parte del cuerpo se realizó el disparo, ya que no es lo mismo intentar un remate de cabeza desde el punto penal que un remate con la pierna hábil o inclusive un remate de penal.

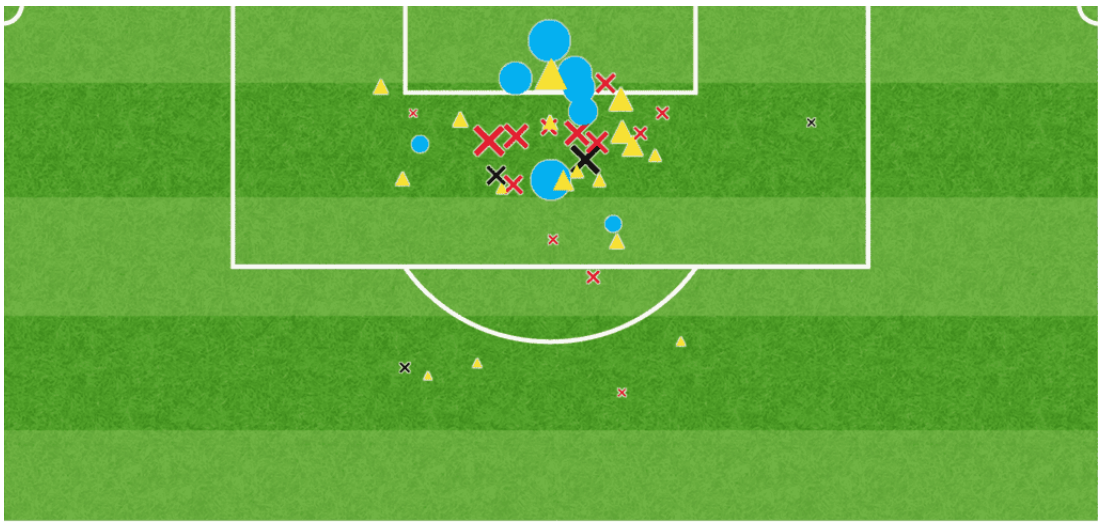

Figura 2.2: Tabla de tiros al arco donde los círculos representan goles, triángulos atajadas del arquero, cruces rojas tiros fuera y cruces negras tiros bloqueados. El tamaño de la figura indica el valor xG que tenía esa jugada. Imagen tomada de Opta [20]

Este modelo permite analizar la performance que tuvieron los jugadores en base a sus goles y los xG. Si un jugador tuvo más goles que xG quiere decir que aprovechó mejor las ocasiones de gol de lo que se esperaba y viceversa. A modo de ejemplo si consideramos los goles esperados y goles efectuados de los delanteros Papiss Demba Cissé y Luis Suarez, con la información presentada en la tabla 2.1, vemos que Cissé logró concretar más goles de los esperados mientras que Suarez de los 15 xG logró concretar 8, es decir 7 menos de los esperados. Para saber qué influyó en que la performance fuera mejor o peor a la esperada es necesario analizar el contexto de cada disparo, ya que un nivel bajo de goles puede deberse a un muy buen nivel del guardametas rival así como una buena presión defensiva que provoque disparos incómodos. Este modelo de análisis es posible extrapolarlo a un equipo completo logrando así una métrica de xG que permita brindar un contexto previo a un partido. Esto puede ser de vital importancia para un entrenador permitiendo así plantear una buena defensa estratégica. Muchos trabajos han creado distintos modelos de xG utilizando información de ubicación, distancia, velocidad de disparo, ángulo de disparo, tipo de disparo, entre otros atributos [20]-[23].

En adición a lo mencionado anteriormente respecto al rendimiento individual, podemos encontrar productos como los que ofrece Olocip [8]. Entre

Jugador	Tiros	xG	Goles
Papiss Demba Cissé	25	4.5	10
Luis Suarez	110	15	8

Tabla 2.1: Comparación de goles esperados con los goles concretados para los jugadores Luis Suarez y Papiss Demba Cissé en la temporada 2011/2012 de la Premier League. Datos obtenidos de [20]

ellas se encuentra TCT-Scout, que utiliza aprendizaje automático para aprender de los patrones de rendimiento de jugadores en temporadas anteriores y poder hacer predicciones del rendimiento futuro, poniéndolos en contexto de una determinada competición, equipo, entrenador, edad o sistema de juego, entre otras, mejorando al scouting de jugadores. Otro producto ofrecido por está empresa es una herramienta de análisis pre partido, la cual utilizando datos estadísticos de los partidos previos y considerando las acciones de cada jugador genera un puntaje que se asigna al equipo. Asimismo, la herramienta indica qué jugadores han sido los más influyentes en cada uno de los equipos y predice de cuáles esperar un buen rendimiento .

De acuerdo con The Correspondent, un caso de éxito de este modelo de scouting y ponderación de jugadores, fue el del equipo danes FC Midtjylland, que utilizó un modelo de scouting generado en el club para reclutar jugadores y logró ganar su primera liga en 2015. Lamentablemente, el modelo desarrollado no es público. Sin embargo se pudo probar su éxito para mejorar los futuros fichajes de jugadores de un club. Los autores del artículo mencionan que la incorporación del jugador Sparv, llamado “the no stats all-star”, cuyo juego no era tan vistoso pero sí muy efectivo, es decir, el jugador no tenía números altos como balones divididos ganados o distancia recorrida pero esto debido a que su posicionamiento en la cancha era tan bueno que no necesitaba correr mucho y tampoco necesitaba ganar los balones divididos ya que los anticipaba antes que ocurrieran. El modelo utilizado indicaba que este jugador de la segunda división de Alemania estaba al nivel de jugar en la “Premier League” por lo que entró en el radar del equipo y terminó siendo una pieza clave para la obtención del título. [24], [25].

2.2.1. Fuentes de datos

Para poder aplicar técnicas de CD es necesario saber con qué datos se cuenta y qué datos se pueden utilizar. Se optó por analizar las fuentes de datos disponibles para formar un conjunto de datos con suficiente nivel de detalle para poder ser utilizado.

Para poder determinar el alcance real del proyecto y de qué manera se obtendrán los datos, es necesario realizar un relevamiento de los datos disponibles, qué empresas se encargan de la recolección de datos y la accesibilidad de los mismos. Existen varias formas para obtener datos relacionados al deporte, ya sea la obtención directa de los datos, es decir, mediante la observación de los encuentros y anotación de los eventos deseados, mediante el consumo de APIS existentes o la utilización de técnicas de *scrapping* para obtener datos de una página que los presente públicamente.

Con respecto a datos históricos y estadísticas de partidos, existen APIs como ElenaSport [26] o APIFootball [27], que cuentan con una cobertura muy amplia para ligas de todo el mundo, con la información por partido de goles, tarjetas, cambios y alineaciones. Es posible también encontrar APIs con más detalles de información como BeSoccer [28] u otra variante de APIFootball [29], las cuales agregan información minuto a minuto de todos los eventos que suceden en un partido. Lamentablemente, el uso de estas aplicaciones es costoso y no serán tomadas en cuenta en este trabajo.

Una de las grandes empresas en lo que respecta a la recolección de datos es Opta [19]. Esta empresa cuenta con cobertura en tiempo real de un gran número de competiciones de todo el mundo. A nivel de partido, brinda las estadísticas del partido (corners, posesión de balón, tiros, etc) e información sobre los eventos que ocurren en el partido con su correspondiente posición y momento en el que ocurren (tiros libres, tiros al arco, etc).

Existen otras empresas dedicadas a la recolección de datos de varios deportes, y publicación de API que puedan ser consumidas por diferentes sitios, por ejemplo Enetpulse [30] y DataFactory [31].

En lo que respecta a datos de posicionamiento en encuentro y su respectivo análisis, la empresa PlayerMaker [32] desarrolló un dispositivo que se coloca en el botín de los jugadores y permite realizar análisis técnicos (cantidad de toques realizados, porcentaje de pierna utilizada, posesión de balón, etc), tácticos (posición en la cancha, mapa de calor, etc) y físico (distancia recorrida, velocidad máxima, etc). En base a estos datos el equipo puede realizar un entrenamiento más personalizado explotando sus fortalezas o mejorando sus debilidades. Estos dispositivos son utilizados también para prevenir lesiones .

A nivel mundial muchas empresas consumen los datos y widgets que proporcionan estas empresas. Algunos ejemplos son LiveFootball [33], WhoScored [34], Footmob [35] o TyCSports [36].

En particular la web Whoscored presenta información de partidos para varias ligas de todo el mundo, donde para algunas de ellas es posible ver información de cada partido con sus respectivas estadísticas. Se pueden ver los datos de las alineaciones de cada equipo, pases realizados, porcentaje de pases acertados, remates, goles y mucha más información de cada incidencia del partido. En lo que respecta al resumen del partido se puede obtener las zonas de ataque, zonas de tiro y zonas de juego. La página no cuenta con una API para acceder a los datos pero si con una interfaz interactiva donde es posible visualizar los tiros al arco como en la figura 2.3 y otras estadísticas como los pases, faltas, etc.

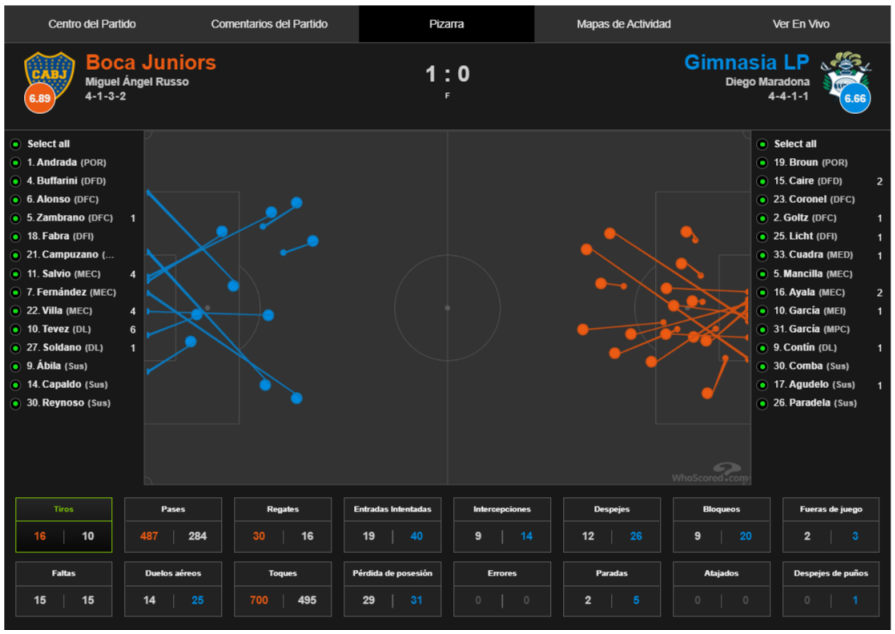

Figura 2.3: Representación de los tiros de un partido en Whoscored. Captura tomada de la página Whoscored [34]

A modo de resumen de los antecedentes y conjuntos de datos, se pudo ver que la mayoría de los antecedentes encontrados y analizados, ya sean trabajos académicos o emprendimientos privados, se centran en la predicción de resultados, aunque todos abordando el problema desde distintas perspectivas, ya sea utilizando resultados previos, información de posición en la tabla, relevancia de los partidos o centrándose en el plantel actual, tomando como datos de entrada el valor de mercado del equipo y de los jugadores participantes. Conocer el estado del arte, los métodos que se utilizan y el rumbo que está tomando la ciencia de datos en el deporte es de vital importancia.

Existe una gran cantidad de problemas que se pueden abordar en lo que respecta a la CD aplicada al deporte. En el siguiente capítulo se describe el problema específico que se propuso abordar, se analizan las fuentes de datos existentes, donde se selecciona la más adecuada para alcanzar los objetivos planteados para luego detallar la implementación y los resultados obtenidos.

Capítulo 3

Predicción de atributos de un partido de fútbol

En el capítulo anterior se mostraron algunos antecedentes de la aplicación de CD en el deporte, donde se ahondó más en el fútbol. Como se mencionó, existe una amplio espectro de aplicaciones y áreas para estas técnicas que van desde las apuestas hasta la prevención de lesiones. En este capítulo se presentará una aplicación de CD en el fútbol, la cual fue elegida como el problema a atacar en el transcurso del trabajo.

De un partido de fútbol es posible obtener mucha información como ocasiones de gol, tiros al arco, tarjetas, tiros de esquina, pases, etc. El problema seleccionado para trabajar es la predicción de algunos de estos atributos, en particular de tiros al arco y tiros de esquina generados por un equipo dado un conjunto de datos conocidos a priori, como pueden ser datos estadísticos pasados. Además, interesa saber qué atributos influyen más en el resultado de la predicción, ya que tener solamente un valor predicho puede no ser tan útil como saber qué atributos influyen sobre este valor. Se utilizan únicamente datos del pasado ya que se espera poder utilizar este modelo para toma de decisiones futuras, por lo que no se pueden utilizar datos como el resultado final o las estadísticas del encuentro a predecir.

Los atributos a predecir fueron seleccionados en conjunto con un equipo de entrenadores y especialistas en el deporte y son considerados relevantes más allá del resultado final, el cual contiene un alto grado de azar. Sin embargo,

la estrategia diseñada en este trabajo es posible aplicarla a cualquier atributo del partido.

A su vez, como se mencionó anteriormente, se buscó relación entre estos atributos y el resto de los atributos que conforman un partido de fútbol, con la finalidad de comprenderlos y poder darle a un entrenador herramientas para la toma de decisiones. Es decir, poder dar la posibilidad de contextualizar los datos indicando cuales influyen positivamente y cuáles negativamente.

A continuación se profundiza en los atributos seleccionados para el prototipo, se describe cómo fue la obtención de datos, por qué se eligió este conjunto y una descripción del resultado final. Por último se mencionan y definen los métodos y técnicas que fueron utilizados en la etapa de implementación.

3.1. Atributos a predecir

Los conjuntos de datos son la materia prima para la aplicación de CD. Cada conjunto está compuesto por instancias que corresponden a los datos que se disponen para hacer un análisis, en este caso una instancia corresponde a un partido. Cada instancia está compuesta por atributos que la describen. En el caso de está aplicación de CD los atributos corresponden a las estadísticas del partido como tiros de esquina, tiros al arco, pases, etc [37].

Uno de los problemas a abordar en el proyecto consiste en investigar a partir de los datos disponibles, cuales son los los atributos que se podrían predecir. Atributos como los tiros al arco, los tiros de esquina o tarjetas mostradas fueron considerados a priori como mejores atributos para analizar dado que el rango de posibles valores se encuentra bastante acotado por la realidad del juego. Finalmente al consultar con expertos en el área se seleccionaron los atributos **tiros de esquina** y **tiros al arco**.

No se intentará predecir el resultado final ya que tiene un alto grado de incertidumbre: una pelota puede pegar en el palo y entrar o pegar en el palo y salir, cambiando el resultado final. Pensando a largo plazo, si un equipo logra tener más situaciones de gol partido a partido, la probabilidad de convertir

goles y ganar los encuentros aumenta. Los atributos elegidos para predecir apuntan a maximizar la cantidad de situaciones de gol más allá de los goles convertidos.

Si bien en el presente trabajo se enfoca en estos atributos, la idea de analizarlos va más allá de eso: se desea comprender su relación sobre otros aspectos del partido, si estos se ven directamente afectados por otros atributos y cómo lo hacen y por último si es posible que un entrenador tome medidas para afectar sobre ellos. Para poder llevar esto a cabo se utilizaron técnicas y métodos de aprendizaje automático presentadas en la sección 3.3

3.2. Conjunto de datos

El conjunto de datos utilizado se obtuvo de la página Whoscored. Se consideraron todos los datos disponibles para la liga Argentina, que consisten de 5 temporadas comprendidas entre 2016 y 2020, con un total de 1667 partidos y 34 equipos participantes. Cada partido fue representado por 45 atributos de los cuales 22 son estadísticas extraídas directamente del partido, 18 son atributos calculados y los restantes 5 son atributos descartados (fecha, resultado, temporada, árbitro y estadio). A continuación fundamentaremos la elección de esta fuente de datos.

Luego de relevadas las fuentes de datos existentes y las distintas alternativas a seguir, se optó por formar un conjunto de datos a partir de fuentes de datos existentes. El estudio realizado y la solución implementada, puede posteriormente ser adaptada a otro conjunto de datos si así se requiere, o incluso es posible extender la información recabada en el actual. En adición, se busca un tipo de datos que sea aplicable al medio local. Dado que se desea que el resultado de este trabajo sea aplicable por un entrenador del medio local, no se consideran datos de GPS o similares dado que su uso aún no esta extendido en el fútbol de Uruguay.

Como se intenta predecir y analizar atributos, es necesario que el conjunto de datos tenga información estadística con las situaciones del partido, y no por ejemplo, solo el resultado final. A la hora de elegir el conjunto de datos se

descartaron todos aquellos que no tuvieran información acerca de los tiros al arco y los tiros de esquina, ya que son los atributos sobre los cuales se trabajó. Finalmente, para elegir el conjunto de datos final se priorizaron la cantidad de atributos sobre cada partido y la cantidad total de partidos.

Como se mencionó en la sección 2.2.1, existen varias alternativas para la obtención de datos. Luego de un análisis exhaustivo de varias de las opciones disponibles, se optó en una primera instancia por la utilización de una API. Sin embargo, no fue posible encontrar una gratuita que sea útil y luego de infructíferas consultas a distintos proveedores de datos pagos, solicitando licencias para poder consumir sus servicios en pro de la investigación académica, se tuvo que descartar esta opción, decantando en la utilización de datos públicos obtenidos mediante la programación de un scrapper.

De las fuentes de datos relevadas Whoscored es la que contiene más detalle de cada partido así como una gran cantidad de partidos. Esta página web fue fundada en 2008, se encarga de ofrecer resultados en directo, resultados finales y ratings de jugadores de las principales ligas y campeonatos de fútbol, entre los que se encuentran las principales ligas de Europa, la MLS (liga principal de Estados Unidos), el Brasileirão y la liga Argentina.

Cabe destacar, que WhoScored no recaba los datos por sí mismo, sino que utiliza servicios de estadísticas detalladas provistos por Opta. Por lo cual, en una posible continuación de este trabajo, sería de gran utilidad poder llegar a un acuerdo con la empresa directamente para que suministre los datos sin tener que realizar todo el procesamiento [19].

La cantidad de datos sobre un partido varía y depende directamente de la liga que se esté evaluando. Por ejemplo, la Liga Española y demás competiciones Europeas presentan datos a un muy buen nivel de detalle desde 2010 mientras que las ligas sudamericanas lo hacen a partir de 2016, reportando para los años anteriores solo estadísticas a nivel general del partido, como lo son alineaciones, goles, sustituciones y tarjetas. Con respecto a la Liga Uruguaya, este sitio reporta solamente el histórico de resultados, información que no cumple los requerimientos planteados.

El fútbol se practica de formas distintas en diferentes países. Sin embargo, las diferencias son menos visibles si se comparan países de un mismo continente, es decir, la diferencia en estilos de juego entre España y Uruguay es mucho más grande que si se compara con Argentina. Teniendo en cuenta que se desea obtener resultados aplicables al medio local, se decidió utilizar los datos de aquella liga en donde los equipos tienen el estilo más parecido al fútbol uruguayo, en este caso la liga Argentina. Para obtener la mayor cantidad de partidos posibles, se utilizaron todos los datos disponibles entre las temporadas 2016/2017 y 2019/2020, obteniendo en total un aproximado de 1700 partidos [38], [39].

Con respecto a los datos de Whoscored, cada partido contiene la información correspondiente a los equipos involucrados con sus jugadores, la formación inicial, entrenador, cambios realizados, resultado parcial y final, estadio, fecha, árbitro y un listado de eventos donde se indica cada evento que ocurrió en dicho partido, con su ubicación física y temporal, los jugadores involucrados y los tipo de eventos. Donde tipo de evento indica si fue un pase, tiro al arco, gol, falta, etc. En total existen 219 tipos de eventos distintos . Para dar una idea sobre la cantidad de datos que se manejan, cada partido contiene aproximadamente 1200 eventos.

3.2.1. Estrategia de obtención y procesamiento de datos

Fue necesario definir una estrategia para la obtención, consolidación, almacenamiento y exportación de datos definida en la figura 3.1. Para la obtención se tuvo que crear un scrapper que recorra la página web y realice las descargas de cada partido iterativamente, este proceso esta detallado en la sección 3.4.1. Luego para la consolidación de datos fue necesario definir que atributos se iban a persistir en la base de datos, estos atributos se detallan en la sección 3.2.2. Como tercer paso luego de tener la base de datos generada fue necesario exportar estos datos en formato de texto para la aplicación de CD, en este paso se computaron todos los atributos calculados los cuales se detallan en la sección 3.2.2.1.

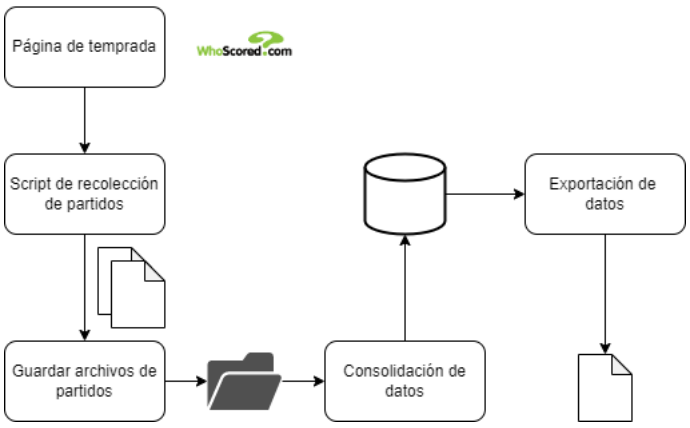

Figura 3.1: Diagrama de pipeline general del proceso desde la obtención de datos de Whoscored hasta su exportación en un archivo CSV.

3.2.2. Descripción del conjunto de datos recopilado

Como se mencionó en la sección anterior, los datos extraídos de la fuente WhoScored contenían información a un nivel de detalle muy bueno. Previamente a almacenar nuestro conjunto de datos en una base de datos, fue necesario procesar cada partido para obtener un conjunto de datos acorde al problema planteado. A partir de los eventos de cada partido se extrajeron estadísticas aprovechables para el problema trabajado. Durante este paso se trabajó la representación de un partido, definiendo qué atributos se mantenían y cuáles eran descartados, ya sea por ser irrelevantes o por quedar fuera del alcance del proyecto.

Se definió un partido como un conjunto de atributos que involucraron a los equipos participantes, sus alineaciones y datos estadísticos del encuentro, como tiros al arco, tiros de esquina generados, entre otros. A su vez, se calcularon algunos atributos que no se encontraban en la fuente original, como el historial de puntos o la racha actual del equipo. El conjunto final de atributos se presenta en la tabla 3.1

Otra de las entidades consideradas importantes fueron los tiros, para los cuales se definió una estructura para almacenar toda la información de cada disparo, como el equipo que realizó el tiro y la posición en la cancha. Esto para tener la posibilidad de trabajar directamente sobre estos datos. La estructura generada para almacenar estos datos se presenta en la tabla 3.2

Atributo	Descripción	Tipo
Temporada	Nombre de la temporada	Categórico
Arbitro	Nombre del árbitro	Categórico
Fecha	Fecha del encuentro	Fecha
Equipo local	Nombre del equipo local	Categórico
Equipo visitante	Nombre del equipo visitante	Categórico
Estadio	Nombre del estadio	Categórico
Cantidad de pases intentados*		Numérico
Cantidad de pases acertados*		Numérico
Cantidad de tiros intentados*		Numérico
Cantidad de tiros acertados*		Numérico
Cantidad de tarjetas amarillas*		Numérico
Cantidad de tarjetas rojas*		Numérico
Cantidad de faltas cometidas*		Numérico
Cantidad de tiros de esquina*		Numérico
Posesión de la pelota*	Valor entre 0 y 100 %	Numérico
Formación*	Formación utilizada por el equipo	Categórico
Resultado	Texto con resultado	Categórico
Histórico de puntos*	Puntos del equipo en el actual torneo	Numérico
Posición*	Puesto en el que se encuentra el equipo en el actual torneo	Numérico
Puntos en las últimas 5 fechas*	Acumulado de puntos en los últimos 5 partidos	Numérico
Diferencia de puntos	Diferencia de puntos entre el equipo local y el visitante al momento del partido	Numérico
Hace cinco partidos *	Resultado obtenido cinco fechas atrás	Categórico
Hace cuatro partidos *	Resultado obtenido cuatro fechas atrás	Categórico
Hace tres partidos *	Resultado obtenido tres fechas atrás	Categórico
Hace dos partidos *	Resultado obtenido dos fechas atrás	Categórico
Partido anterior *	Resultado obtenido una fecha atrás	Categórico

Tabla 3.1: Atributos relevados para cada partido. Aquellos atributos con un * son los que se repiten para el equipo local y el equipo visitante

3.2.2.1. Atributos calculados

Los datos obtenidos contienen información estadística de cada partido, pero no aporta información sobre el contexto, es decir, cómo viene cada equipo, en qué posición de la tabla se encuentra cada uno, cuál es su historial más

Atributo	Descripción	Tipo
Local	Indica si fue el equipo local o visitante	Boolean
A puerta	Indica si fue al arco o no	Boolean
X	Coordenada en el eje X	Númerico
Y	Coordenada en el eje Y	Númerico
Minuto	Minuto del partido en que ocurrió	Númerico
Segundo	Segundo del partido en que ocurrió	Númerico

Tabla 3.2: Atributos relevados para cada tiro

reciente de resultados, etc. Se decidió generar nuevos atributos. Se calcularon los puntos y posición de cada equipo en el torneo en base a los resultados previos, dentro de ese torneo. Además por cada partido se calculó la diferencia de puntos con la que llegaba cada equipo (diferencia de puntos), ya que a priori podría ser que un equipo que está mejor posicionado en la tabla tenga más ocasiones de gol. Otro de los atributos calculados fue la sumatoria de puntos de los últimos cinco partidos por cada equipo.

Adicionalmente se computó la racha previa. Para esto se generaron cinco atributos nuevos con el resultado de los últimos cinco partidos, tomados como valores categóricos posibles: Ganó, Empató o Perdió. Se decidió tomar estos atributos ya que a priori se presume que un equipo con una racha positiva de victoria tenderá a seguir de esta manera, lo que decanta posiblemente en más ocasiones de gol, mientras que para rachas negativas ocurriría lo contrario. En particular, para los primeros cinco partidos de cada equipo ocurre que no existe todo el registro histórico, por lo que fue necesario calcularlo. Por ejemplo, para el primer partido de un equipo existe ningún dato previo de racha, para el segundo solamente el dato para el partido anterior de la racha y así sucesivamente. Para esto, se tomaron dos criterios distintos: para los primero cuatro datos de la racha histórica se tomó el valor con mas ocurrencias para ese equipo dentro del conjunto de entrenamiento, mientras que, siguiendo la presunción de que los equipos tienen a continuar con una racha, para el restante resultado previo se toma el valor más probable entre los cuatro resultados ya calculados de la racha histórica.

El conjunto de datos obtenido luego de procesar todos los archivos descargados cuenta con 1667 partidos. Cada partido hace referencia a uno de esos

archivos, tomados de los torneos desde 2016 hasta 2019-2020, con un total de 34 equipos participantes, todos de la liga Argentina, los cuales se almacenan en una base de datos relacional.

3.2.3. Procesamiento del conjunto de datos

Una vez extraída la información de la base de datos y cargada en un archivo de texto plano, fue necesario realizar un trabajo de procesamiento de datos para mejorar la calidad del conjunto de datos y adaptarlo al problema a estudiar. A continuación se detallan las técnicas utilizadas para manejar atributos categóricos y la estandarización de los datos.

3.2.3.1. Atributos descartados

Previamente a utilizar el conjunto de datos se descartaron algunos atributos que no se consideraron relevantes para esta aplicación. Sin embargo, son parte del conjunto de datos, y podrían ser utilizados en un futuro desarrollo. El atributo temporada y fecha no se utilizaron ya que no se consideró relevante tener en cuenta el factor tiempo. Otro de los atributos que no se consideró fue el estadio y el árbitro. Sin embargo, en caso de querer calcular la cantidad de tarjetas podría ser interesante considerar el árbitro ya que la cantidad de tarjetas en un partido muchas veces está relacionado con el estilo de quien imparte justicia en el encuentro. El otro atributo que no se tuvo en cuenta es el resultado del partido ya que se pretendía estudiar la correlación entre atributos sin la dependencia del resultado.

3.2.3.2. Atributos categóricos

Los atributos categóricos son atributos cuyos valores pertenecen a un conjunto discreto y finito, que puede o no ser no numéricos. Para poder utilizar estos atributos en los métodos de aprendizaje automático seleccionados fue necesario transformar esos textos en un valor numérico. Existen al menos dos posibles técnicas a aplicar. En este caso se utilizó el algoritmo “one-hot-encoding”. Se crearon nuevos atributo por cada etiqueta diferente. En cada instancia, si el valor del atributo original es “i”, el atributo correspondiente al i-ésimo valor valdrá 1, y el resto valdrán 0. En este caso los atributos categóricos existentes son formación, racha de partidos y equipo (local o visitante).

Para los tres casos se utilizó "one-hot-encoding", ya que no es posible considerar un orden entre los distintos tipos de formación ni en los posibles resultados así como tampoco en los equipos involucrados en el partido.

Por ejemplo, si tenemos el atributo formación con tres posibles valores (4-3-3, 4-4-2 y 5-3-2), aplicando "one-hot-encoding" el resultado obtenido sería el presentado en la Tabla 3.3.

Nro	Formación	Formación4-3-3	Formación4-4-2	Formación5-3-2
1	4-3-3	1	0	0
2	5-3-2	0	0	1
3	4-4-2	0	1	0
4	4-3-3	1	0	0

Tabla 3.3: Ejemplo de aplicar "one-hot-encoding" sobre la columna formación, con tres posibles valores. Luego para la aplicación de CD se elimina la columna Formación.

3.2.3.3. Normalización de datos

Dada la heterogeneidad de las escalas de los datos, estos deben ser normalizados antes de poder trabajar con ellos. A modo de ejemplo, el rango de datos para la cantidad de tiros al arco (entre 0 y 40 aprox.) en un partido es muy distinto al rango de la cantidad de pases(entre 100 y 500 aprox.), por lo cual para algoritmos como KNN, donde se considera un vector de N dimensiones (con N igual a la cantidad de atributos) el análisis se vería sesgado. Para evitar este posible sesgo se utilizó la estandarización de cada atributo. En este caso se utilizó una función que estandariza los atributos restando la media (μ) y dividiendo sobre la desviación estándar (σ) 3.1 [40].

$$z = \frac{(x - \mu)}{\sigma} \tag{3.1}$$

3.2.4. Particionamiento de conjunto de datos

Debemos asegurar, para evitar el sobreajuste, que la evaluación del modelo se realice en un conjunto de datos distinto a aquel sobre el cual se entrenó. Es

decir, si entrenamos nuestro algoritmo con los mismos datos que lo evaluamos entonces podremos obtener un muy buen resultado pero al probarlo con datos que no pertenezcan a este conjunto su resultado no sería tan bueno, muy probablemente el modelo está memorizando los datos de entrenamiento, sin poder generalizar. Por lo tanto, se separan los datos en un conjunto de entrenamiento y uno de evaluación. Se utilizaron 80 % de partidos para entrenamiento y el restante 20 % para evaluación, dando como resultado 1334 partidos de datos para entrenamiento y 333 partidos para evaluación.

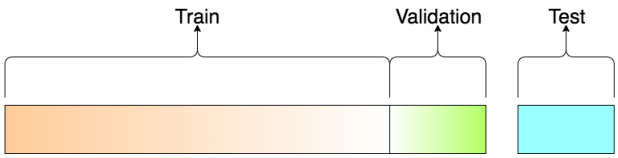

Figura 3.2: Separación conjunto de datos. Imagen tomada de notas del curso Aprendizaje Automático [41].

Las particiones fueron generadas de manera aleatoria para evitar que las agrupaciones u ordenamientos presentes en el conjunto original puedan dar lugar a distribuciones distintas. Por ejemplo, no sería una buena opción tomar para los datos de entrenamiento los primeros partidos ordenados por fecha ya que los estilos de juego varían con el tiempo, así como los equipos varían sus resultados y estilos de juego.

En caso de necesitar ajustar los parámetros del modelo utilizado, es posible utilizar varias técnicas de validación. Entre las que se encuentran: separar una parte adicional a partir de los datos de entrenamiento llamada conjunto de validación 3.2, y validación cruzada la cual preserva el conjunto de evaluación y trabaja directamente sobre el conjunto de entrenamiento, sin la necesidad de generar otra partición. Este último método fue el utilizado en este proyecto debido a que permite disminuir la varianza de los resultados (ya que se obtiene como el promedio de varias evaluaciones), aunque es más costoso en términos de tiempo de entrenamiento.

3.3. Aprendizaje automático

El campo del Aprendizaje Automático (AA) plantea la pregunta de cómo construir programas que aprendan automáticamente mediante la experiencia, es decir que a partir de un conjunto de datos, el cual es considerado "experiencia" para el algoritmo, y una métrica para evaluar el rendimiento, un algoritmo de AA es capaz de obtener mejores resultados a medida que adquiere más experiencia. En la actualidad existen muchas aplicaciones utilizando este campo de la ciencia de la computación y la Inteligencia Artificial [42].

Entre los métodos de AA se encuentran dos grandes grupos: Aprendizaje Supervisado y Aprendizaje No Supervisado. En la primera categoría se conoce el valor final o etiqueta de las instancias mientras que en el segundo es desconocido. Los métodos supervisados se subdividen en métodos de clasificación y de regresión dependiendo de si el conjunto de atributos finales es discreto o no. Por otro lado, los métodos de aprendizaje no supervisado se utilizan normalmente para tareas de agrupamiento, ya que no se conocen los posibles valores finales. Durante este proyecto se utilizaron tres técnicas populares de AA supervisado, dos clasificadores y un método de regresión. Para el caso del método de regresión se aproximó el resultado al valor entero más cercano.

Si bien la cantidad de tiros de esquina o tiros al arco es siempre un número entero y se podrían tomar cada número como una categoría, existe el concepto de distancia entre los distintos valores y es relevante. Dicho esto, se considera el problema como un problema híbrido y se utilizaron tanto métodos clasificación como métodos de regresión. Para el caso de los métodos de regresión que retornan un número real fue redondeado al entero más cercano.

3.3.1. Técnicas y métodos utilizados

Se utilizaron tres métodos de AA supervisado durante el proyecto: Regresión Lineal, KNN y Random Forest.

3.3.1.1. Regresión lineal

La regresión lineal es una forma de aprendizaje supervisado donde, a partir de un vector $x^T = (x_1, x_2, \ldots, x_n)$ con n atributos (o variables) se busca construir una función (hipótesis) $h_\theta(x) : R^n \to R$ que prediga la salida $y \in R$ (llamada variable o atributo de salida), continua. En este caso los n atributos serían los datos estadísticos del partido mientras que el y sería el atributo a predecir. Además de para predicción, se utilizó la regresión lineal para evaluar los pesos que tiene cada atributo de partido en las variables a calcular, lo cual brinda una información de mucha utilidad para el entrenador, pudiendo sacar conclusiones sobre cuáles son las variables más preponderantes en el atributo a analizar. Además de esto, se utilizó la regresión lineal como algoritmo de predicción.

3.3.1.2. KNN

El algoritmo más básico basado en instancias es KNN (K-nearest neighbors), este método representa las instancias como un punto en un espacio n-dimensional. Es un algoritmo de distancia ponderada altamente efectivo para muchos problemas de clasificación. Es robusto ante datos de entrenamiento con ruido y efectivo con un conjunto de datos suficientemente grande.

En este caso la información proporcionada son partidos representados por los atributos seleccionados, los cuales representan un punto en un espacio de M dimensiones, y cuya clase es el atributo a predecir. Y para realizar la clasificación se toman los N vecinos más cercanos, los cuales se obtienen a partir del cálculo de distancia euclidiana.

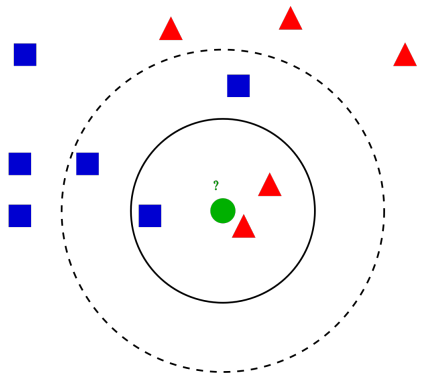

Figura 3.3: Ejemplo del algoritmo KNN. El ejemplo que se desea clasificar es el círculo verde. Para k = 3 este es clasificado con la clase triángulo, ya que hay solo un cuadrado y 2 triángulos, dentro del círculo que los contiene. Si k = 5 este es clasificado con la clase cuadrado, ya que hay 2 triángulos y 3 cuadrados, dentro del círculo externo.

En particular para este algoritmo fue necesario optimizar el parámetro K, considerando los posibles valores 1, 3, 5 y 7. Para poder hacerlo se utilizó validación cruzada aplicando el método de k-fold presentado en 3.3.1.4.

3.3.1.3. Random Forest

Random Forest es un popular algoritmo de AA de aprendizaje supervisado utilizado para resolver problemas de clasificación y regresión. Se basa en la combinación de árboles de decisión. En lugar de depender de un árbol de decisión, toma la predicción de cada árbol y, en función de los votos mayoritarios de las predicciones, predice el resultado final. Entre las ventajas que presenta este tipo de clasificador es su robustez y el buen control del sobreajuste. En la imagen 3.4 se presenta un ejemplo de un algoritmo Random forest realizando una clasificación para una instancia dada.

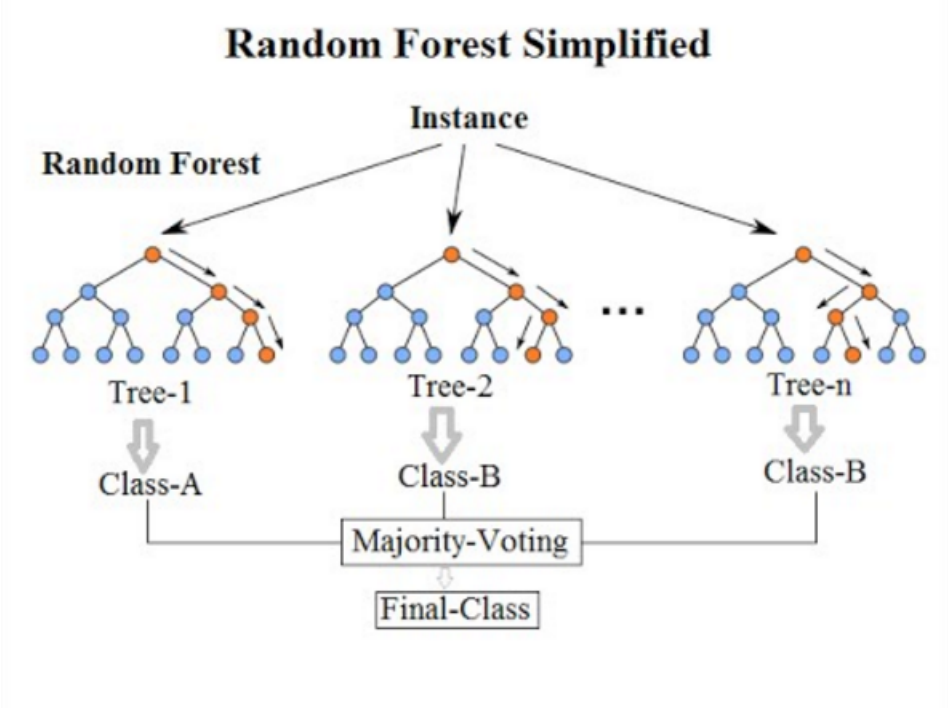

Figura 3.4: Ejemplo de algoritmo Random Forest con N clasificadores, clasificando una instancia. Donde el primer árbol los clasifica con la clase A, el segundo y el árbol n-ésimo con la clase B. Imagen extraida de IBM [43]

En particular para este algoritmo fue necesario optimizar varios parámetros. Así como para KNN, se utilizó validación cruzada aplicando el método de k-fold.

- Núm árboles: Cantidad de clasificadores generador (100, 250, 500)
- Función: Gini o Entropy. Corresponde a la función para medir la calidad de la partición, es decir el criterio para la selección de atributos y generación de los árboles de decisión [44].
- Bootstrap: Indica si utilizar o no todo el conjunto de datos para la generación de clasificadores. False: Utiliza todos los datos de entrenamiento para generar cada clasificador. True: Utiliza la misma cantidad de datos de entrenamiento pero con reemplazo para generar cada clasificador.
- Máx. atributos: Cantidad máxima de atributos a considerar, los valores posibles son 'sqrt' o 'log2'. Dependiendo de qué opción se tome la cantidad máxima de atributos que considerará cada clasificador será la raíz cuadrada del total o el logaritmo en base 2 del total respectivamente [45].

3.3.1.4. Validación cruzada

Se utilizó validación cruzada para ajustar los parámetros de los algoritmos. En particular para este proyecto se utilizó el método K-fold, que consiste en dividir el conjunto de entrenamiento en K partes (llamados folds), entrenar el modelo con $k-1$ folds y evaluar en el restante. Este proceso se repite cambiando la parte elegida 3.5

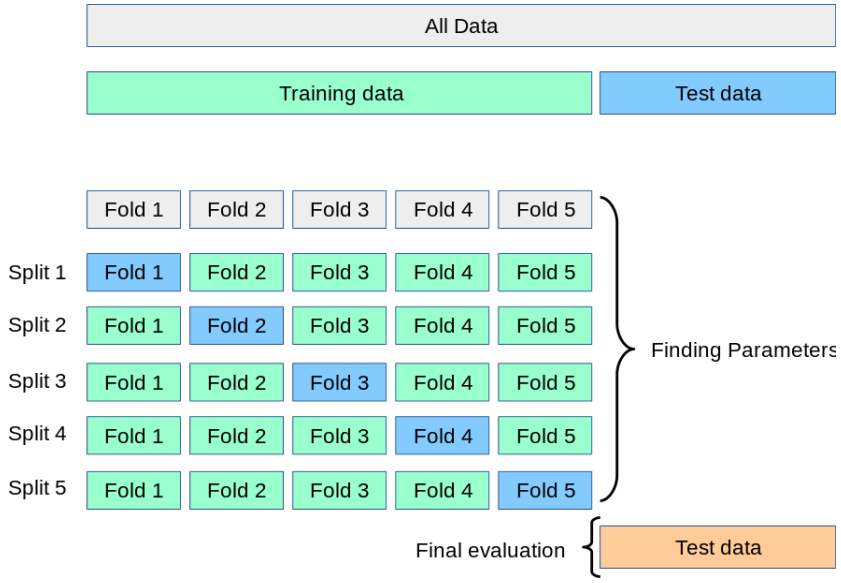

Figura 3.5: Validación cruzada utilizando K-Fold. Imagen tomada de notas del curso Aprendizaje Automático [41]

Se optó por generar 5 Folds, debido a que es el número de folds que se generan por defecto y además que como resultado obtenemos un 80 % de datos para entrenamiento y 20 % para validación, es decir de las 1334 filas de datos de entrenamiento se utilizarán 1067 para entrenamiento y 267 para evaluación.

3.3.1.5. Métricas de evaluación

Para evaluar la performance se decidió utilizar una métrica conocida como Accuracy (Acierto). la misma consiste en calcular el cociente entre la cantidad de predicciones acertadas sobre la cantidad de predicciones realizadas.

Si bien esto permite tener una idea de cuantas predicciones correctas realiza el modelo, no aporta información acerca de qué tan malas son las predicciones en las que no se acierta. Llevando esto a nuestro objetivo de ayudar a un entrenador a tomar decisiones. La planificación de un entrenador es muy similar

si sabe que el equipo rival va a generar 10 tiros de esquina o si sabe que va a generar 11. Sin embargo es muy distinta a si sabe que el contrario tendrá 1 o 2. Dicho esto, saber si los valores de tiros al arco y tiros de esquina van a ser bajos o altos es más importante que saber el número exacto.

A partir de esto se definieron dos nuevas métricas basadas en la ya mencionada Accuracy: Accuracy 1 y Accuracy 2, las cuales toman como correctas aquellas predicciones que se encuentren a distancia 1 o 2 del resultado verdadero respectivamente, es decir, tienen cierta tolerancia al fallo.

La definición de las métricas es la siguiente:

- **Accuracy**: Cociente entre predicciones acertadas sobre predicciones realizadas.
- **Accuracy 1**: Cociente entre la cantidad de predicciones acertadas sobre predicciones realizadas, considerando acertadas aquellas predicciones cuya diferencia con el valor esperado sea menor o igual a 1.
- **Accuracy 2**: Cociente entre la cantidad de predicciones acertadas sobre predicciones realizadas, considerando acertadas aquellas predicciones cuya diferencia con el valor esperado sea menor o igual a 2.

El uso de estas tres métricas permitió comparar los resultados obtenidos por los tres modelos y a su vez, gracias a las últimas métricas definidas, se puede tener una idea de si las predicciones fallidas se encuentran cerca o no del resultado esperado. Por ejemplo, si un clasificador predecía 4 y el valor real era 5, entonces el resultado para accuracy sería falso mientras que para accuracy 1 y accuracy 2 el resultado sería verdadero. Para el mismo ejemplo, si el clasificador hubiese predicho 3 entonces el resultado sería falso para accuracy, accuracy 1 y verdadero para accuracy 2.

3.4. Implementación de la solución

En esta sección se detallan las tecnologías utilizadas en el trabajo, desde la recolección de datos hasta el análisis y la representación de los resultados. La implementación consistió en tres grandes etapas: obtención, procesamiento y

análisis de los datos. A continuación se profundiza en cada una de las etapas mencionadas.

3.4.1. Obtención de datos

En la sección 3.2.2 se describe el proceso de obtención de datos a grandes rasgos, en esta sección nos enfocaremos en los aspectos propios de la implementación. Los datos se obtienen de la web Whoscored, que si bien sus datos son públicos no cuenta con una API para obtener la información de manera directa, por lo que fue necesario implementar un scrapper. Este programa fue implementado en Javascript y se ejecutó mediante la extensión de Firefox GreaseMonkey [46]. El proceso de ejecución consiste en acceder dentro de Whoscored a la página principal del torneo, aquí se comienza a ejecutar el proceso que obtiene los partidos y se almacenan en una lista, se redirige el navegador a la dirección web del primer partido donde el scrapper procesa los datos y los descarga a un servidor local, se itera sobre toda la lista secuencialmente hasta descargar todos los partidos. La figura 3.6 presenta un esquema de este proceso.

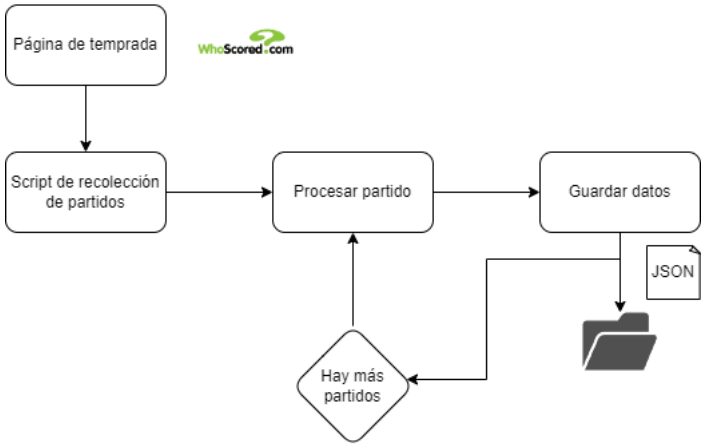

Figura 3.6: Proceso de obtención de datos desde Whoscored a un servidor local.

3.4.2. Procesamiento de datos

Como se describe en la sección 3.2.2, los datos obtenidos como archivos json debieron ser procesados, reduciendo los aproximadamente 1200 eventos

por partido, a estadísticas aprovechables para esta aplicación de ciencia de datos. Para realizar este procesamiento se llevaron a cabo dos tareas: migración de los datos JSON a una base de datos relacional y exportación de datos en formato CSV.

Se optó por la utilización de una base de datos relacional como paso intermedio para tener una capa de persistencia que permita realizar otro tipo de análisis en el futuro, en caso de ser necesario. En este punto se utilizó un ORM de Python llamado SQLAlchemy. Un ORM(Object Relational Mapping), es una forma de representar la información que se encuentra en la base de datos relacional como objetos en el lenguaje de programación utilizado, en este caso Python. La gran ventaja de utilizar un ORM es la facilidad para manejar los datos, ya que se pueden utilizar las funciones que el lenguaje provee.

Para poder utilizar una base de datos relacional fue necesario diseñarla teniendo en cuenta necesidades encontradas. El diagrama de clases se presenta en la imagen 3.7. La solución desarrollada, permitió ejecutar un proceso donde indicando una carpeta con los archivos y el nombre del campeonato, se extrae la información estadística del partido filtrando el listado de evento por los tipos de eventos de intereses, de donde se calcula el total de pases intentados y logrados, el total de tarjetas, las faltas, etc. por cada equipo participante. Además de las estadísticas se almacena el identificador numérico del partido el cual es fundamental para evitar tener datos duplicados. La fecha, el estadio, los equipos participantes con sus respectivas formaciones iniciales, el árbitro, el resultado y una lista de los tiros al arco realizados indicando el equipo si fue a puerta y su posición física (coordenada x e y) y temporal (minuto, segundo), en la tabla 3.4 se detallan los campos y tipos de datos almacenados por cada tiro al arco.

Id	Nombre	
Local	Indica si fue el equipo local o visitante	Boolean
A puerta	Indica si fue al arco o no	Boolean
X	Coordenada en el eje X	Numérico
Y	Coordenada en el eje Y	Numérico
Minuto	Minuto del partido en que ocurrió	Numérico
Segundo	Segundo del partido en que ocurrió	Numérico

Tabla 3.4: Atributos relevados para cada tiro

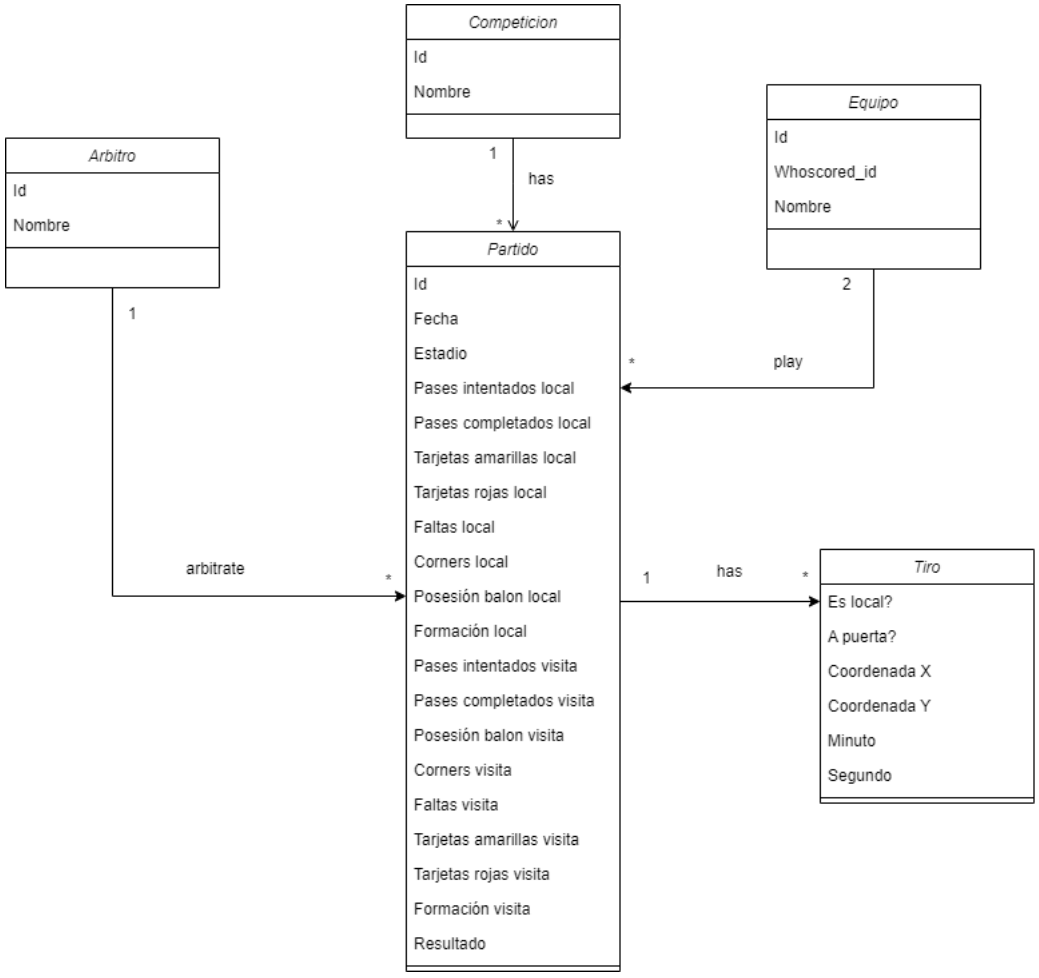

Figura 3.7: Diagrama de clases. Con las entidades de Partido, competición, equipo, árbitro y tiro.

El último paso del pipeline de datos es la exportación de los datos a un archivo CSV. Para esto se obtiene toda la información de los partidos de la base de datos consolidando la información del partido con sus equipos asociados, su competición, el árbitro y los tiros efectuados. En este proceso también se realizó el cálculo de los atributos como las posiciones en la Liga, los puntos actuales y la racha actual de cada equipo, para poder ser utilizados por los modelos de AA. Los datos fueron procesados ordenados por fecha desde la más pasada a la más reciente. Esto permitió poder realizar el cálculo de los puntos en el campeonato actual, así como los datos de racha de equipo en los últimos cinco partidos. Se decidió no almacenar los atributos calculados en la base de datos porque en caso de agregar un nuevo partido se deberían volver a calcular. Finalmente, el proceso almacena la información de cada partido

como una línea del archivo CSV de salida. En total el archivo cuenta con 1667 líneas, representando los 1667 partidos obtenidos de WhoScored. Cada línea tiene 40 columnas de las cuales 22 son estadísticas extraídas directamente del partido mientras que los restantes 18 son atributos calculados, en la tabla 3.1 se encuentra el listado de los atributos, incluidos los que no se consideran en la descarga que fueron especificados en 3.2.3.1.

3.4.3. Análisis de datos

Para el análisis de datos y la implementación del prototipo final, se realizó un notebook utilizando Python. Esta solución fue implementada en Google Colab [47], un producto de Google Research que permite escribir y ejecutar código arbitrario de Python en el navegador. Esta herramienta es muy utilizada para tareas de aprendizaje automático y análisis de datos. El notebook se conecta a un servidor de Google Compute Engine que utiliza Python 3, y cuenta con 12GB de memoria RAM y un disco duro de 100GB.

Se optó por utilizar Colab ya que permite implementar una solución en un entorno colaborativo, donde el equipo pudo programar en simultáneo, los resultados se pueden visualizar de una manera sencilla con funcionalidades de Python y permite guiar al lector por cada paso realizado y cada algoritmo ejecutado.

La solución implementada toma como entrada el archivo CSV con información de los partidos antes mencionado y realiza algunas tareas extras de procesamiento de datos, como cálculo de valores faltantes, manejo de atributos categóricos mediante la aplicación de "one-hot-encoding" 3.2.3.2 y la estandarización de los datos utilizando la funcionalidad de la librería "scikit-learn" llamada "StandarScaler" [40].

El hecho de que cada partido está representado mediante 40 atributos hace que sea difícil relevar nuevos partidos por lo que se decidió generar un nuevo conjunto de datos utilizando los atributos más importantes, con el fin de que las conclusiones obtenidas fueran más fáciles de aplicar por un entrenador del medio local ya que sería más económico generar un conjunto de datos con me-

nos cantidad de atributos pero que igualmente sus resultados sean similares a la versión con todos los atributos. Para obtener los atributos más relevantes se utilizó Regresión Lineal (RL). Si se toma un atributo como objetivo y se aplica RL, se obtiene el peso de cada atributo con respecto al objetivo, es decir, que tanto influye. Se realizó este procedimiento para los tiros al arco y los tiros de esquina generados y se formaron nuevos conjuntos de datos con los 8 atributos más relevantes.

Una vez terminado el procesamiento de datos se predijo el valor de los tiros al arco y de los tiros de esquina generados utilizando los tres modelos presentados: Regresión Lineal, KNN y Random Forest, tanto para el conjunto de datos completo como para los reducidos.

En este capítulo se presentó el problema, se dio una definición de AA, de los métodos y métricas utilizados y se detalló la implementación de la solución. En el capítulo 4 se mostrará cómo se utilizaron estos métodos para la predicción y comprensión de los tiros al arco y los tiros de esquina generados por un equipo, se comparan los diferentes modelos utilizando las métricas presentadas y se mostrarán los resultados obtenidos.

Capítulo 4

Resultados

En este capítulo se mostrarán los resultados obtenidos de la aplicación de los métodos presentados en el capítulo anterior, utilizando el conjunto de datos trabajado. Este capítulo está dividido en dos secciones, una para la predicción de tiros de esquina, y otra para la predicción de tiros al arco, donde se presenta la distribución de datos, el análisis del peso de cada atributo para todos los atributos y para el conjunto de atributos reducido, el ajuste de parámetros y finalmente el análisis y comparación de los resultados obtenidos por los diferentes modelos utilizando las métricas definidas anteriormente.

Como fue mencionado en capítulos anteriores, la elección de los atributos a predecir fue en conjunto con expertos locales del área, los cuales fueron considerados los principales interesados durante el transcurso del proyecto. Estos atributos fueron cantidad de tiros de esquina y cantidad de tiros al arco.

4.1. Línea base

Para tener un punto de referencia para los resultados obtenidos por los modelos entrenados se utilizó un clasificador que retorna siempre el valor con mas ocurrencias en el conjunto de entrenamiento. Los resultados fueron comparados utilizando las métricas presentadas y se generaron gráficas para entender mejor los resultados obtenidos.

4.2. Predicción de cantidad de tiros de esquina

El primer atributo evaluado fueron los tiros de esquina. Los tiros de esquina son generados cuando la pelota se va por la línea final impulsada por el equipo rival, lo que significa que se está jugando en su territorio y cerca de su arco. Generar más tiros de esquina, más allá del hecho de que son jugadas que pueden decantar en un gol por sí mismas, significa que se está atacando más, por lo tanto es algo deseable. A su vez, si se sabe que el equipo rival genera más tiros de esquina, se puede planificar la defensa de otra manera, enfocada en el juego aéreo.

4.2.1. Distribución de datos

A partir del conjunto de datos se estudió la cantidad de tiros de esquina en cada partido, obteniendo como resultado que los tiros de esquina se concentraban en valores próximos al máximo, el cual es 4. Dando como indicio que nuestro clasificador línea base obtendría un porcentaje elevado para las métricas seleccionadas. En lo que respecta al rango de datos, los valores de cantidad de tiros de esquina van desde 0 a 19.

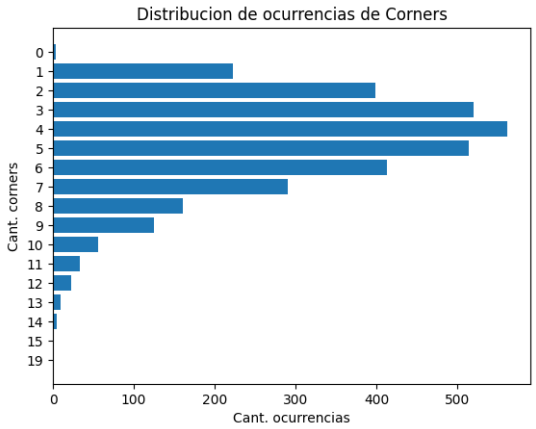

Figura 4.1: Distribución de cantidad de ocurrencias de tiros de esquina por partido

4.2.2. Ajuste de parámetros

Para poder optimizar los parámetros de KNN y Random Forest fue necesario aplicar validación cruzada. Para el primero el único parámetro a opti-

mizar es la cantidad de vecinos K. Los posibles valores seleccionados son los comúnmente utilizados 1, 3, 5 y 7. Como resultado se obtuvo la tabla 4.1 presentando los mejores valores en las métricas utilizadas para $K = 5$.

K	Accuracy	Accuracy 1	Accuracy 2
1	**0.13**	0.37	0.59
3	**0.13**	0.38	0.59
5	0.12	**0.41**	**0.62**
7	**0.13**	0.39	**0.62**

Tabla 4.1: Resultados obtenidos para el cálculo de los mejores hiperparámetros de KNN.

Para el caso de Random Forest se debió probar considerando todas las posibles combinaciones. En total fueron 24 ejecuciones obteniendo como resultado de la tabla 4.2. Analizando la tabla se puede observar que no necesariamente utilizar más de arboles es mejor, ya que en este caso el valor intermedio elegido fue el que tuvo mejores resultados. Esto puede deberse a que el hecho de utilizar más cantidad de árboles puede sobreajustar el clasificador a los datos de entrenamiento, obteniendo peores resultados durante la evaluación. El mejor criterio a utilizar fue "gini", obtiene levemente mejores resultados en cada caso y además cuenta con la ventaja de ser más rápido [48]. Se utilizaron todos los datos de entrenamiento sin reemplazo al momento de generar cada clasificador, debido a que el valor de bootstrap que obtuvo mejores resultados fue "false". Finalmente, el mejor valor para el parámetro de máxima cantidad de atributos fue "sqrt", esto quiere decir que se toma la raíz cuadrada de la cantidad de atributos de entrada. Por ejemplo, si el conjunto de datos cuenta con 40 atributos entonces el método tomará $sqrt(40) = 6$ atributos para generar cada árbol.

Est.	Criterio	Feat.	Bootstrap	Accuracy	Accuracy 1	Accuracy 2
100	gini	log2	True	0.16	0.48	0.72
100	gini	log2	False	0.17	0.49	0.72
100	gini	sqrt	True	0.16	0.48	0.72
100	gini	sqrt	False	0.18	**0.51**	0.74
100	entropy	log2	True	0.15	0.49	0.70
100	entropy	log2	False	0.17	0.48	0.72
100	entropy	sqrt	True	0.16	0.48	0.74
100	entropy	sqrt	False	0.16	0.49	0.73
250	gini	log2	True	0.15	0.48	0.75
250	gini	log2	False	0.17	0.49	0.75
250	gini	sqrt	True	0.18	0.48	0.75
250	**gini**	**sqrt**	**False**	**0.19**	**0.51**	**0.77**
250	entropy	log2	True	0.16	0.47	0.73
250	entropy	log2	False	0.16	0.49	0.74
250	entropy	sqrt	True	0.17	0.5	0.76
250	entropy	sqrt	False	0.16	**0.51**	0.74
500	gini	log2	True	0.16	0.49	0.74
500	gini	log2	False	0.16	0.49	0.76
500	gini	sqrt	True	0.17	0.49	0.75
500	gini	sqrt	False	0.17	0.49	**0.77**
500	entropy	log2	True	0.15	0.48	0.73
500	entropy	log2	False	0.16	0.5	0.75
500	entropy	sqrt	True	0.15	0.49	0.76
500	entropy	sqrt	False	0.16	0.5	0.76

Tabla 4.2: Resultados obtenidos para el cálculo de los mejores hiperparámetros de Random forest.

En las figuras 4.2, 4.3, 4.4 y 4.5 se pueden ver gráficas *boxplot* de la información detallada en la tabla antes mostrada. Se realizó una gráfica para cada una de todas las dimensiones utilizadas en la validación cruzada.

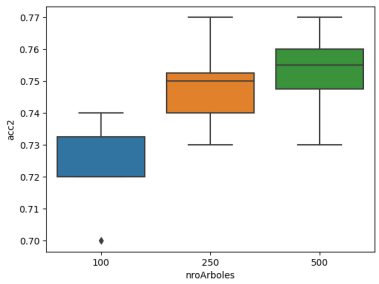

Figura 4.2: Boxplot de cantidad de arboles y accuracy 2

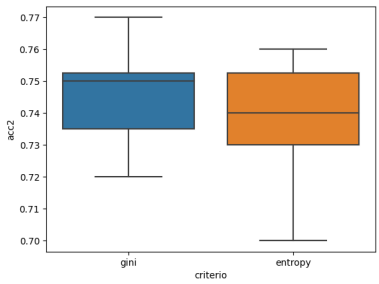

Figura 4.3: Boxplot de criterio y accuracy 2

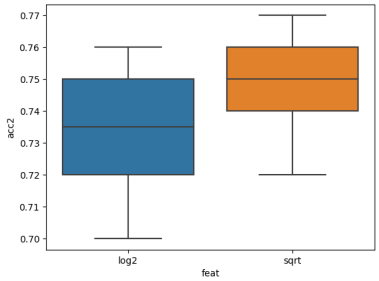

Figura 4.4: Boxplot de features y accuracy 2

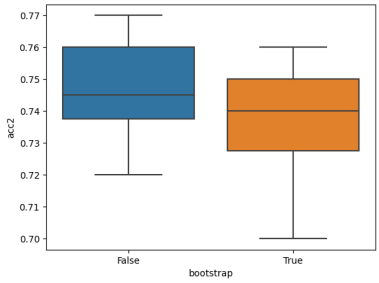

Figura 4.5: Boxplot de bootstrap y accuracy 2

4.2.3. Análisis de resultados

Luego de analizado el conjunto de datos se ejecutaron los clasificadores implementados. En la tabla 4.3 se muestran los resultados de las métricas sobre la ejecución de cada uno de los clasificadores, evaluando para accuracy coincidencias exactas, $accuracy_1$ coincidencias con tolerancia 1, y $accuracy_2$ coincidencias con tolerancia 2.

Método	Accuracy	Accuracy 1	Accuracy 2
Línea base	0.16	0.47	0.73
KNN	0.13	0.38	0.59
Random Forest	0.18	0.49	0.76
Regresión Lineal	**0.19**	**0.70**	**0.86**

Tabla 4.3: Resultados obtenidos por los métodos para predicción de tiros de esquina en las tres métricas utilizadas

Para la medida de accuracy los resultados fueron muy bajos para todos

los métodos. Sin embargo KNN fue el único que estuvo por debajo de la línea base. Lo mismo ocurre con KNN para las demás métricas por lo cual no es posible considerarlo como un buen estimador para esta aplicación de ciencia de datos. Al evaluar en las métricas de $accuracy_1$ y $accuracy_2$ los resultados mejoran considerablemente, los métodos de Random Forest y Regresión Lineal obtuvieron mejores resultados que nuestra línea base. KNN obtuvo los resultados más bajos: esto puede ser debido a que como este método considera con el mismo peso cada atributo, los puntos generados que fueran similares no necesariamente corresponden a iguales valores de clase objetivo. Como se puede ver en la tabla 4.8, muchos de los atributos no tienen influencia sobre la clase objetivo.

Para Random Forest se obtuvieron mejores resultados que la línea base. A diferencia de KNN, Random Forest no considera todos los atributos a la hora de generar sus clasificadores: esto puede ser uno de los factores que lo llevó a obtener mejores resultados. Considerando regresión lineal, este método sobresalió sobre los demás obteniendo valores más acertados en las distintas métricas utilizadas la diferencia con los demás métodos radica en que a pesar de utilizar todos los atributos le asocia a cada uno un ponderador, que resulta en: prácticamente no considerar los atributos que no tienen correlación (asignado un valor de ponderador cercano a 0) y dando una relevancia a los atributos que sí influyen ya sea de manera positiva o negativa.

En las figuras 4.6 y 4.7 se muestran las matrices de confusión para los tiros de esquina utilizando KNN y Random Forest. Se puede observar que en ambos casos los resultados se acumulan en la esquina superior izquierda, consecuencia de que los modelos predijeron valores menores a 7 en la mayoria de los casos. Los valores en la matriz de confusión de Random Forest se ubican mas cerca de la diagonal, lo que significa que los resultados fueron mejores.

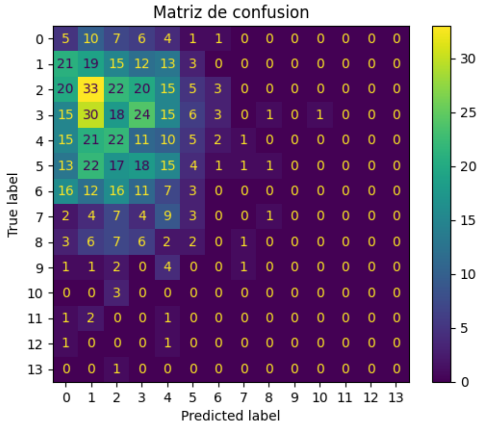

Figura 4.6: Matriz de confusion para los tiros de esquina utilizando el método KNN con $K = 5$

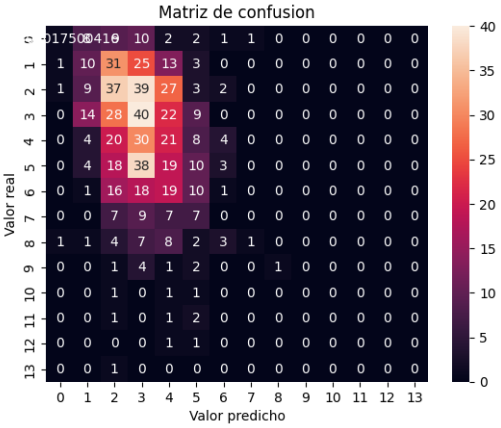

Figura 4.7: Matriz de confusion para los tiros de esquina utilizando el método Random Forest con 250 estimadores

4.2.4. Peso de atributos

Mediante la aplicación de Regresión Lineal se obtuvo el peso de cada atributo sobre la clase objetivo, en este caso los tiros de esquina. Esto se consideró como un punto importante ya que permitió visualizar cómo influía cada atributo sobre la cantidad de tiros de esquina. En la tabla 4.8 se presentan los pesos obtenidos para cada atributo. Se pudo observar que, aunque muchos de los atributos tenían una correlación casi nula sobre la cantidad de tiros de esquina, existían unos pocos cuya correlación es considerable, como por ejemplo la diferencia de ranking entre equipos. Esto se interpretó como que si un equipo que está mejor posicionado en la tabla de posiciones enfrenta a uno que está

abajo, el primero tendría una gran cantidad de tiros de esquina, y que esto aumenta a medida que la diferencia entre sus posiciones aumenta.

En base al vector generado por regresión lineal se formó un conjunto de datos reducido para el que se consideraron los 8 atributos que más incidían sobre la cantidad de tiros de esquina. La imagen 4.9 muestra el peso de los 8 atributos más relevantes, ya sea positiva o negativamente. Atributos como los pases intentados están correlacionados positivamente con la cantidad de tiros de esquina. Esto puede significar que si un equipo realiza más pases tiene más probabilidad de generar tiros de esquina o que los equipos que generan más tiros de esquina también son los que realizan más pases, mientras que atributos como la posesión de balón del rival afecta de manera negativa.

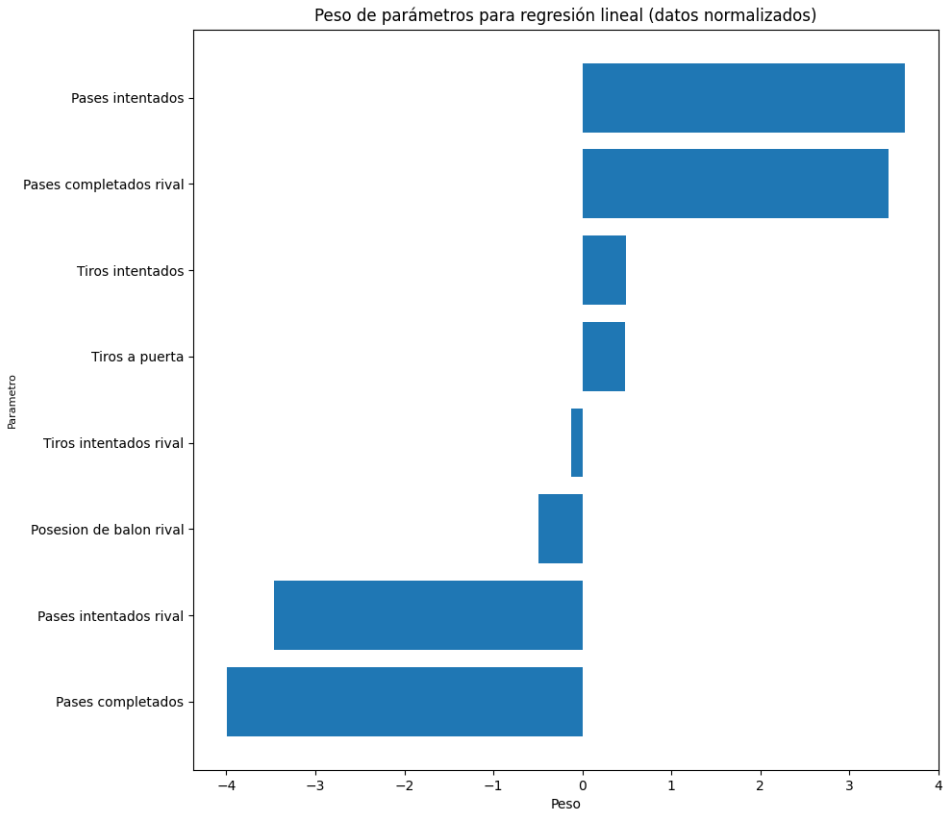

Figura 4.9: Peso de los 8 atributos más relevantes en la clase Tiros de esquina ordenados de mayor a menor

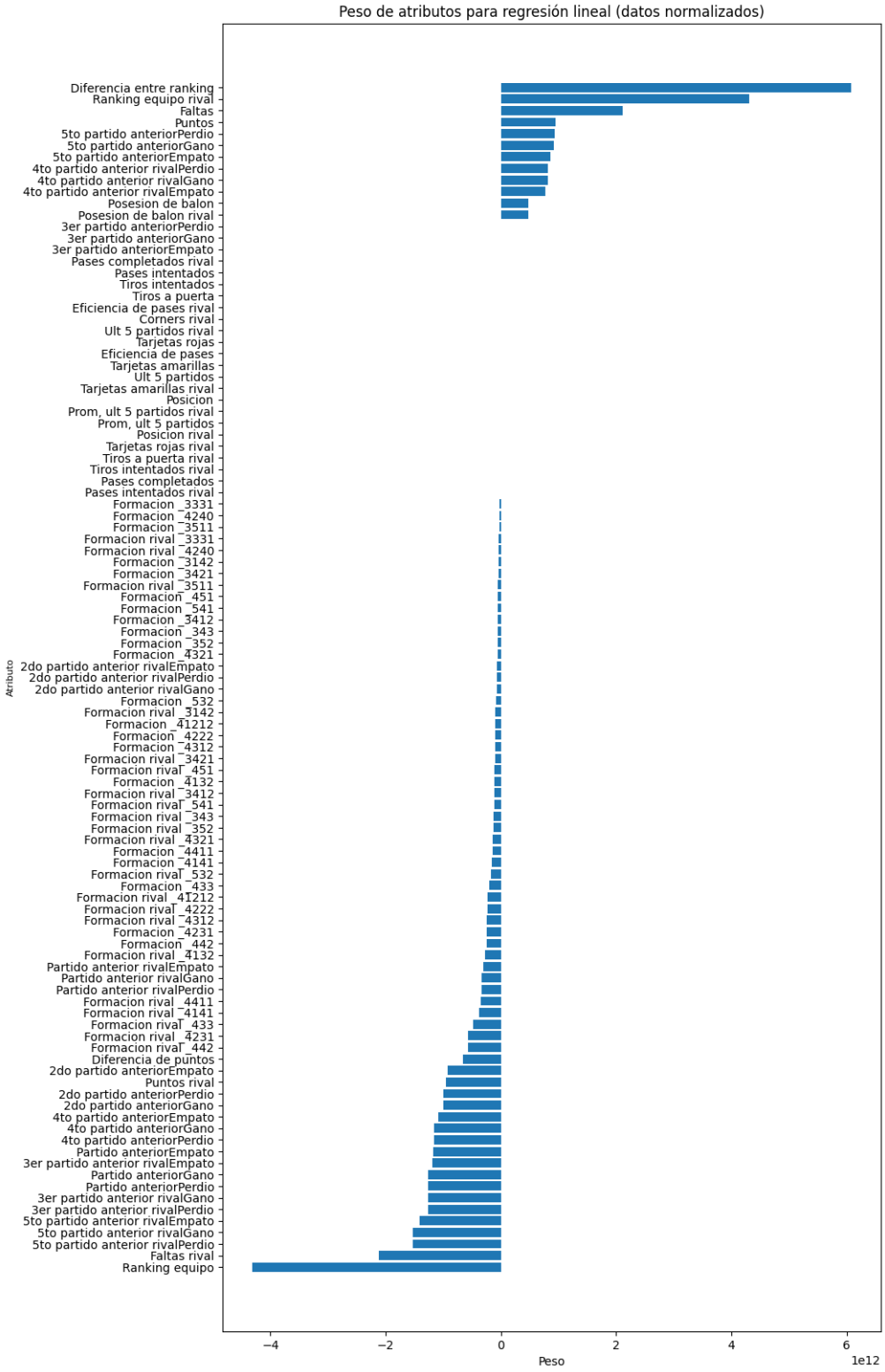

Figura 4.8: Peso de todos los atributos en la clase Tiros de esquina ordenados de mayor a menor

4.2.5. Regresión lineal - Evaluación atributos más relevantes

Para evaluar si era necesaria la utilización de todos los atributos o si era posible generalizar mejor el problema considerando un conjunto reducido de atributos, se formó un subconjunto con los 8 atributos más relevantes y el atributo a predecir. Se ejecutaron pruebas para el método que obtuvo mejores resultados, en este caso regresión lineal, considerando solamente los 8 atributos más relevantes. En este caso los que más influyen en la cantidad de tiros de esquina son:

- Pases intentados
- Pases completados del rival
- Tiros intentados
- Tiros a puerta
- Tiros intentados del rival
- Posesión de balón del rival
- Pases intentados del rival
- Pases completados

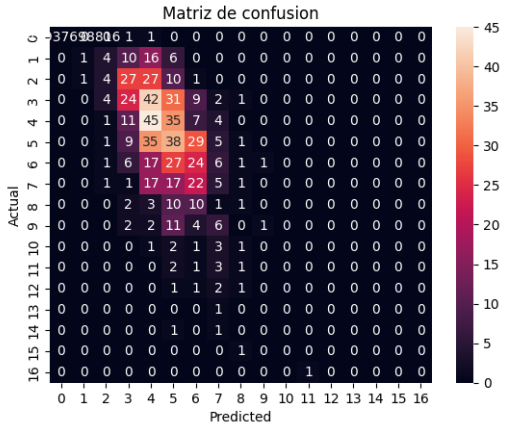

Figura 4.10: Resultados para el método Regresión Lineal con 8 atributos

Como resultado se obtuvo la matriz de confusión 4.10 donde vemos que las predicciones se acumulan entre los valores seis y veintidós sobre la esquina superior izquierda, como en el caso anterior. Sin embargo en este caso los resultados se vuelvan más sobre la diagonal, es decir, los resultados tienden a

ser más precisos. Asimismo, los resultados son más precisos 4.4 con respecto a regresión lineal utilizando todos los atributos.

Método	Accuracy	Accuracy 1	Accuracy 2
Línea base	0.16	0.47	0.73
KNN	0.13	0.38	0.59
Random Forest	0.18	0.49	0.76
Regresión Lineal	0.19	**0.70**	0.86
Regresión Lineal 8 atr.	**0.21**	**0.70**	**0.87**

Tabla 4.4: Resultados obtenidos para los métodos y regresión lineal con 8 atributos más relevantes, para predicción de tiros de esquina

Considerando los tres métodos utilizados para la clasificación y predicción se pudo concluir que el método que logró mejores predicciones fue el de regresión lineal, logrando resultados de casi 90 % de acierto para la métrica $accuracy_2$. Un resultado interesante que se obtiene es que al considerar menos atributos, el método obtiene mejores resultados, por lo cual es posible implementar en la práctica este predictor sin la necesidad de tener la totalidad de los datos de un torneo.

Asimismo, este método nos permite generar un vector de pesos, el cual nos indica de qué manera influye cada atributo en la cantidad de tiros de esquina generados, esto último podría ser una herramienta de utilidad para los entrenadores ya que permite identificar variables que anteriormente no se tenían en cuenta para optimizar la cantidad de tiros de esquina que genera el equipo, así como también estimar la cantidad de tiros de esquina que tendrá su equipo para poder aprovechar al máximo esas ocasiones. De manera análoga, es posible también estimar cuántos tiros de esquina tendrá el rival en el próximo encuentro y enfocar los entrenamientos de la semana considerando este dato.

Para este caso, que nuestro equipo tenga más pases intentados aumenta la probabilidad de generar tiros de esquina. Un resultado contraintuitivo es que los pases completados influyen negativamente en este atributo. Sin embargo consultando esto con profesionales del fútbol comentaron que tiene lógica este resultado, debido a que tener menos pases completados puede asociarse con intersecciones realizadas por el rival cuyo rebote puede finalizar en un balón

fuera del terreno de juego, propiciando un saque de banda o un tiro de esquina.

4.3. Predicción de cantidad de tiros al arco

Otro atributo interesante a evaluar fue la cantidad de tiros generados por un equipo. Así como ocurre con la cantidad de tiros de esquina, tener más tiros decanta en tener más situaciones de gol, por lo tanto conocer este atributo y poder dar la posibilidad a un entrenador de accionar sobre un equipo para maximizar esta variable podría ser de gran importancia.

4.3.1. Distribución de datos

A partir del conjunto de datos se estudió la cantidad de tiros al arco en cada partido, obteniendo como resultado que los tiros al arco se concentraban mayoritariamente en valores cercanos al máximo. Dando como indicio que nuestro clasificador línea base obtendría un porcentaje elevado para las métricas seleccionadas. En lo que respecta al rango de datos, los valores de cantidad de tiros al arco van desde 0 a 49.

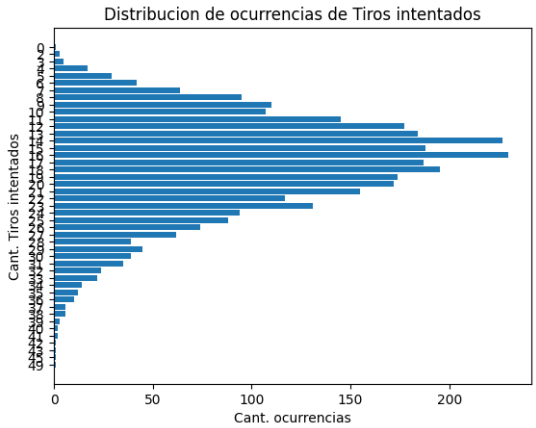

Figura 4.11: Distribución de cantidad de ocurrencias de tiros por partido

4.3.2. Ajuste de parámetros

Para poder optimizar los parámetros de KNN y Random Forest fue necesario aplicar validación cruzada. Para el primero el único parámetro a opti-

mizar es la cantidad de vecinos K, los posibles valores seleccionados son los comúnmente utilizados 1, 3, 5 y 7. Donde los resultados fueron los presentados en la tabla 4.5 obteniendo los mejores valores en las métricas utilizadas para $K = 1$.

K	Accuracy	Accuracy 1	Accuracy 2
1	**0.07**	**0.14**	**0.24**
3	0.04	**0.14**	0.22
5	0.04	**0.14**	0.23
7	0.04	**0.14**	0.23

Tabla 4.5: Resultados obtenidos para el cálculo de los mejores hiperparámetros de KNN.

Para el caso de Random Forest se debió probar considerando todas las posibles combinaciones, en total fueron 24 ejecuciones obteniendo como resultado de la tabla 4.6. Analizando la tabla se puede observar que en este caso utilizar más de árboles fue mejor, ya que en este caso el clasificador con más estimadores fue el que tuvo mejores resultados. Esto puede deberse al hecho de que para tiros al arco tenemos aproximadamente el doble de clases que para tiros de esquina. El mejor criterio a utilizar fue "gini", obtiene levemente mejores resultados en cada caso y además cuenta con la ventaja de ser más rápido [48]. Se utilizaron todos los datos de entrenamiento sin reemplazo al momento de generar cada clasificador, debido a que el valor de bootstrap que obtuvo mejores resultados fue "false". Finalmente, el mejor valor para el parámetro de máxima cantidad de atributos fue "sqrt", esto quiere decir que se toma la raíz cuadrada de la cantidad de atributos de entrada. Por ejemplo, si el conjunto de datos cuenta con 40 atributos entonces el método tomará $sqrt(40) = 6$ atributos para generar cada árbol.

4.3.3. Análisis de resultados

Luego de analizado el conjunto de datos se ejecutaron los clasificadores implementados. En la tabla 4.7 se muestran los resultados de la ejecución de cada uno de los clasificadores, evaluando para accuracy coincidencias exac-

Est.	Criterio	Feat.	Bootstrap	Accuracy	Accuracy 1	Accuracy 2
100	gini	log2	True	0.07	0.2	0.33
100	gini	log2	False	0.08	0.19	0.32
100	gini	sqrt	True	0.07	0.17	0.32
100	gini	sqrt	False	0.07	0.2	0.33
100	entropy	log2	True	0.07	0.19	0.3
100	entropy	log2	False	0.07	0.2	0.32
100	entropy	sqrt	True	0.07	0.19	0.31
100	entropy	sqrt	False	0.08	**0.23**	0.36
250	gini	log2	True	0.07	0.2	0.34
250	gini	log2	False	0.07	0.2	0.33
250	gini	sqrt	True	**0.09**	0.2	0.35
250	gini	sqrt	False	0.08	0.21	0.36
250	entropy	log2	True	0.08	0.2	0.34
250	entropy	log2	False	0.07	0.2	0.32
250	entropy	sqrt	True	0.08	0.21	0.33
250	entropy	sqrt	False	0.08	0.21	0.35
500	gini	log2	True	0.08	0.2	0.34
500	gini	log2	False	0.08	0.21	0.34
500	gini	sqrt	True	0.08	0.2	0.35
500	**gini**	**sqrt**	**False**	**0.09**	0.21	**0.38**
500	entropy	log2	True	0.07	0.2	0.33
500	entropy	log2	False	0.07	0.21	0.32
500	entropy	sqrt	True	0.07	0.21	0.35
500	entropy	sqrt	False	0.08	0.21	0.35

Tabla 4.6: Resultados obtenidos para el cálculo de los mejores hiperparámetros de Random forest.

tas, $accuracy_1$ coincidencias con tolerancia 1, y $accuracy_2$ coincidencias con tolerancia 2.

Método	Accuracy	Accuracy 1	Accuracy 2
Línea base	0.07	0.18	0.31
KNN	0.07	0.15	0.24
Random Forest	0.08	0.21	0.36
Regresión Lineal	**0.32**	**0.50**	**0.62**

Tabla 4.7: Resultados obtenidos por los métodos para predicción de tiros al arco en las tres métricas utilizadas

De igual manera que para la cantidad de tiros de esquina, para la medida de accuracy los resultados fueron muy bajos para todos los métodos. Sin

embargo KNN fue el único que estuvo por debajo de la línea base. Lo mismo ocurre con KNN para las demás métricas por lo cual no es posible considerarlo como un buen estimador para esta aplicación de ciencia de datos. Al evaluar en las métricas de $accuracy_1$ y $accuracy_2$ los resultados mejoran considerablemente, los métodos de Random Forest y Regresión Lineal obtuvieron mejores resultados que nuestra línea base.

En las figuras 4.12 y 4.13 se pueden ver las matrices de confusión de las ejecuciones de los modelos KNN y Random Forest, en ambos casos utilizando los hiperparámetros óptimos. Para el caso de KNN, los resultados se encuentran dispersos lo que da la pauta de que los resultados no fueron buenos, sin embargo, para Random Forest, los valores se encuentran proximos a la diagonal, nuevamente indicando que los resultados fueron buenos.

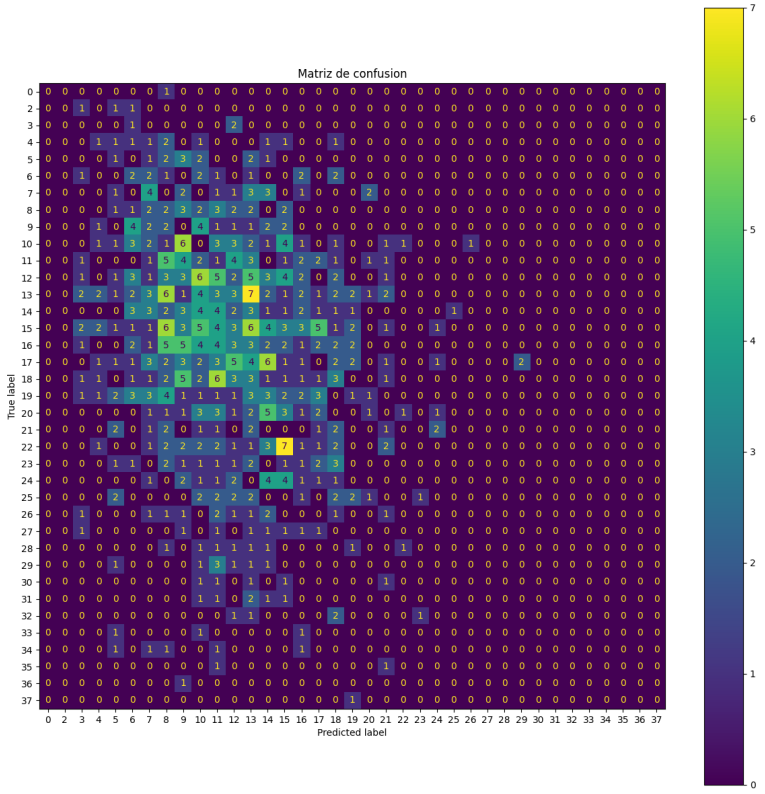

Figura 4.12: Matriz de confusión para tiros al arco del método KNN con $K = 1$

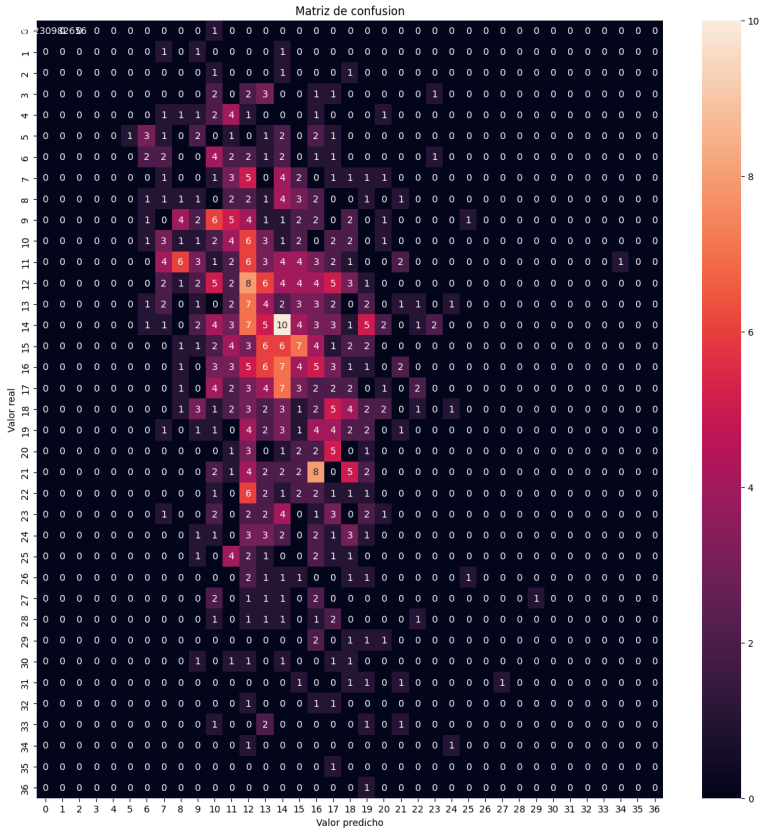

Figura 4.13: Matriz de confusión para tiros al arco del método Random Forest 500 estimadores

4.3.4. Peso de atributos

Mediante la aplicación de Regresión Lineal se obtuvo el peso de cada atributo en la clase objetivo. En la imagen 4.14 es posible apreciar que muchos atributos no tiene una correlación relevante con la cantidad de tiros, mientras que otros tienen una correlación interesante sobre nuestro atributo objetivo. Para el caso de los 8 atributos más relevantes sin considerar los atributos categóricos, no se tuvieron en cuenta para esta sección ya que de ser necesario considerar uno era necesario tomar todas las columnas referentes al mismo atributo. Como resultado se obtuvo el vector representado en la imagen 4.15 qué atributos como pases completados tienen una correlación positiva con la cantidad de tiros mientras que atributos como la posesión de balón tienen una correlación negativa. Esto puede significar que si un equipo completó más pases tiene más probabilidad de finalizar jugadas con un tiro al arco, mientras que los equipos que mantienen mucho la posesión atributos como la posesión

de balón del rival afecta de manera negativa.

4.3.5. Regresión lineal - Evaluación atributos más relevantes

Para evaluar si era necesaria la utilización de todos los atributos o si era posible generalizar mejor el problema considerando un conjunto reducido de atributos, se formó un subconjunto con los 8 atributos más relevantes y el atributo a predecir. Se ejecutaron pruebas para el método que obtuvo mejores resultados, en este caso regresión lineal, considerando solamente los 8 atributos más relevantes. En este caso los que más influyen en la cantidad de tiros son:

- tiros de esquina
- Pases completados
- Pases intentados rival
- Es locatario
- Posesión de balón
- % de pases acertados
- Pases intentados
- Pases completados del rival

Método	Accuracy	Accuracy 1	Accuracy 2
Línea base	0.07	0.18	0.31
KNN	0.07	0.15	0.24
Random Forest	0.08	0.21	0.36
Regresión Lineal	**0.32**	**0.50**	**0.62**
Regresión Lineal 8 atr.	0.14	0.29	0.42

Tabla 4.8: Resultados obtenidos para los métodos y regresión lineal con 8 atributos más relevantes, para predicción de tiros al arco

Considerando los tres métodos utilizados para la clasificación y predicción podemos concluir que el método que logra mejores predicciones es el de regresión lineal, de la misma forma que ocurre con la predicción de cantidad de tiros de esquina. Sin embargo, en este caso la utilización de todos los atributos obtuvo mejores resultados que utilizar los más relevantes y la métrica de aciertos

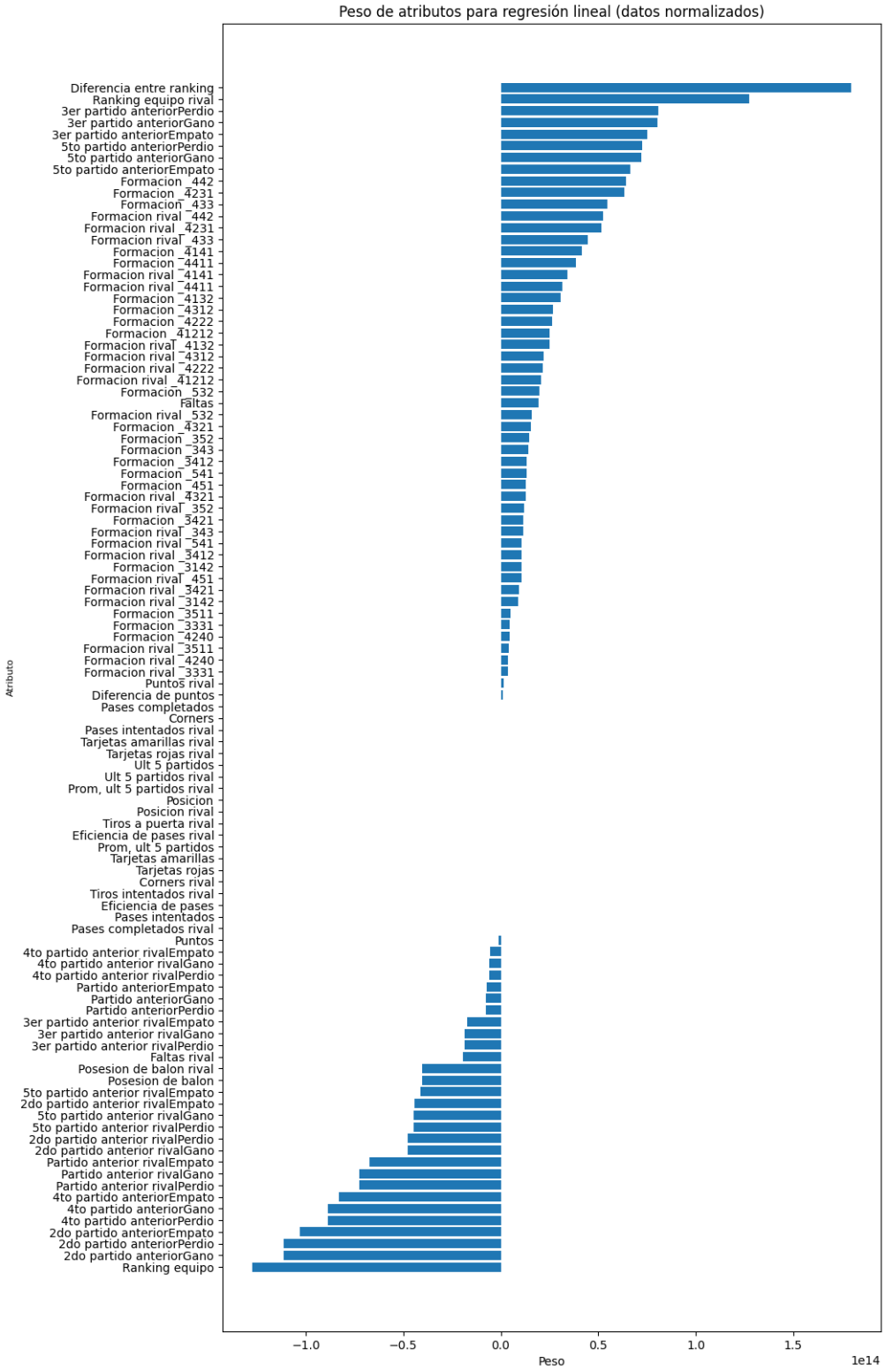

Figura 4.14: Peso de todos los atributos en la clase Tiros intentados ordenados de mayor a menor

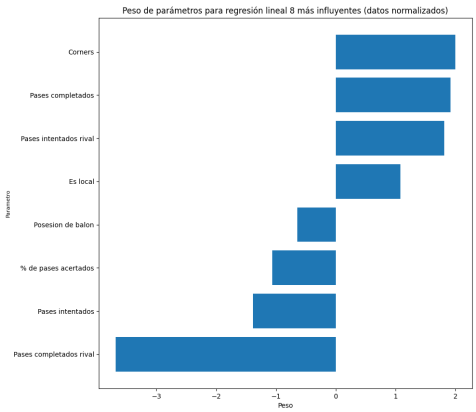

Figura 4.15: Peso de los 8 atributos más relevantes en la clase Tiros de esquina ordenados de mayor a menor

de $accuracy_2$ no supera el 70 % para ningún método. Por otro lado utilizando este método es posible generar un vector de pesos, el cual nos indica de qué manera influye cada atributo en la cantidad de tiros generados, esto último podría considerarse una herramienta de utilidad para los entrenadores ya que permite identificar variables que anteriormente no se tenían en cuenta para optimizar la cantidad de tiros de esquina que genera el equipo. Para este caso, que nuestro equipo tenga más tiros de esquina, pases intentados o inclusive el hecho de ser local aumenta la probabilidad de generar tiros.

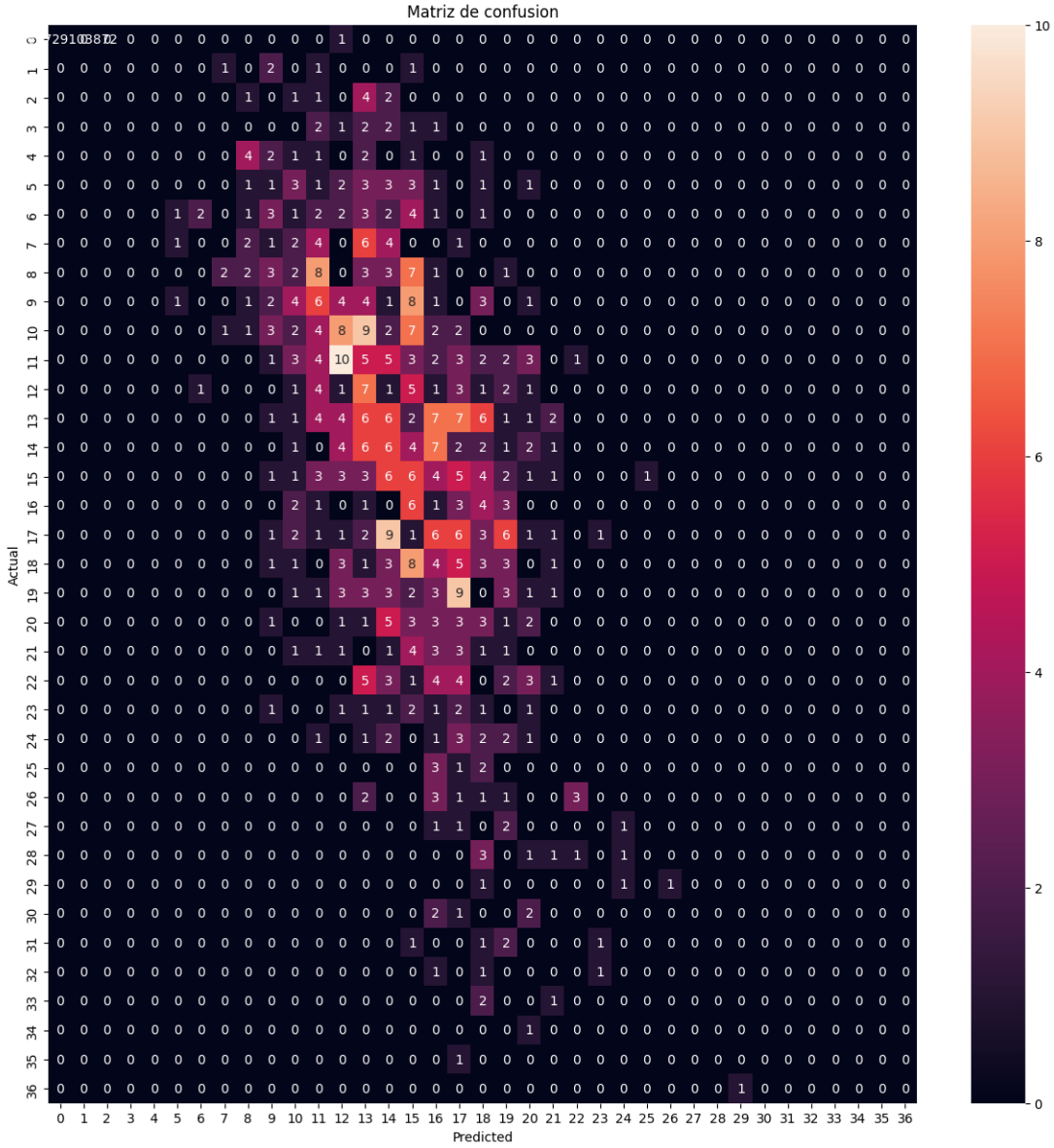

Figura 4.16: Resultados para el método Regresión Lineal con 8 atributos

Capítulo 5

Conclusiones

En este proyecto se propuso abordar una temática muy amplia como es la aplicación de la ciencia de datos en el deporte. El objetivo general fue la creación de una herramienta que brinde a un entrenador información para la planificación de sus encuentros. A partir de lo visto en el capítulo 4 un entrenador puede saber qué atributos debe potenciar para aumentar la generación de tiros al arco y tiros de esquina y utilizando el prototipo implementado podría tener una idea de cómo será el partido qué desea planificar. Además, en base a los resultados obtenidos un entrenador puede estimar a priori la cantidad de tiros de esquina que tendrá su equipo o el equipo rival con un 87 % de precisión ± 2. Mientras que para los tiros al arco se puede estimar con 62 % de precisión ± 2.

En la fase exploratoria fue posible encontrar información sobre investigaciones académicas realizadas sobre distintos deportes enfocadas principalmente en la predicción de resultados, mientras que en el ámbito privado las aplicaciones encontradas cubren un espectro mayor, entre las cuales se encuentra el *scouting*, prevención de lesiones, comprensión de las situaciones del juego para ayudar a la toma de decisiones, etc. Sin embargo, hasta donde llega nuestro conocimiento, no se encuentran trabajos que aborden la problemática de este proyecto, es decir, realizar un estudio para mejorar la planificación de un partido, y en consecuencia mejorar los resultados de un equipo.

No abundan las fuentes de datos abiertas con gran información a nivel de partido. Si bien fue posible generar un conjunto de datos no pudo ser a partir

de encuentros del medio local. Durante el transcurso de la búsqueda de datos se pudo apreciar que la mayoría de los conjuntos de datos abiertos disponibles se enfocan más en el resultado que en la información acerca de lo acontecido durante el partido, lo cual es coherente con lo antes mencionado acerca de los trabajos estudiados. Para poder obtener datos al nivel deseado fue necesario programar un scrapper que descargara los datos de una página web.

Entre los objetivos específicos del proyecto se planteó elaborar un conjunto de datos que sea útil para la aplicación de CD y centrar las bases para que futuros trabajos puedan basar sus investigación en los datos obtenidos y el desarrollo realizado. Se considera que el conjunto de datos generado puede ser útil para otros trabajos, e incluso es fácilmente extensible si se desea agregar más partidos. Tanto los archivos de cada partido, la base de datos generada y el conjunto de datos final están disponibles para que puedan ser utilizados en futuros desarrollos. En adición, la herramienta para procesar los archivos de partidos, que permite cargarlos en una base de datos también está disponible con su respectivo manual de utilización.

Se pudo implementar con éxito un prototipo de predicción de tiros al arco y tiros de esquina, el cual tuvo buenos resultados utilizando las métricas con tolerancia. A pesar de solo evaluar los tiros de esquina y tiros al arco, el prototipo creado permite seleccionar cualquier atributo de los disponibles como objetivo para generar un análisis predictivo y calcular el peso de cada atributo sobre dicha clase objetivo. Llevando esto a la realidad de nuestro proyecto, podría ser utilizado por un entrenador del medio local, donde el único requisito necesario será tener información acerca del equipo rival. La utilización del conjunto de atributos reducido y la similaridad de resultados con el conjunto de datos original dan la pauta de que con poca información se puede obtener contexto acerca del partido qué está por suceder.

5.1. Trabajo a futuro

Este proyecto se basó principalmente en la exploración de los trabajos existentes y las fuentes de datos disponibles en los cuales el equipo invirtió una cantidad considerable de tiempo. Por lo cual, como trabajo a futuro se plantea la posibilidad de utilizar los conocimientos recabados, la herramienta desa-

rrollada y el conjunto de datos generados para continuar con el desarrollo de nuevas funcionalidades y nuevos análisis. Por ejemplo como se menciona en el capítulo de antecedentes existen indicadores como xG y SPI, estos podrían ser generados utilizando el conjunto de datos obtenido, e inclusive se podrían utilizar estos indicadores como nuevos atributos a la hora de predecir.

Complementando la teoría, se plantea que sería interesante poner en práctica en el medio local esta herramienta para asistir a un entrenador y evaluar los resultados obtenidos.

Referencias bibliográficas

[1] A. Liu, "Data science and data scientist," *IBM. Descargado de http://www. researchmethods. org/DataScienceDataScientists. pdf*, 2015.

[2] A. Hassan, A.-R. Akl, I. Hassan y C. Sunderland, "Predicting wins, losses and attributes' sensitivities in the soccer world cup 2018 using neural network analysis," *Sensors*, vol. 20, n.º 11, pág. 3213, 2020.

[3] C. Soto-Valero y col., "Application of automated learning methods for analyzing and predicting sports outcomes.," *RETOS-Neuvas Tendencias en Educacion Fisica, Deporte y Recreacion*, n.º 34, págs. 377-382, 2018.

[4] I. Cabrera Quercini y A. González, "Análisis de los indicadores de éxito en handball masculino a través de redes neuronales artificiales," en *XIII Congreso Argentino y VIII Latinoamericano de Educación Física y Ciencias (Ensenada, 30 de septiembre al 4 de octubre de 2019)*, 2019.

[5] C. S. Valero, "Predicting Win-Loss outcomes in MLB regular season games–A comparative study using data mining methods," *International Journal of Computer Science in Sport*, vol. 15, n.º 2, págs. 91-112, 2016.

[6] Daniel Berrar Philippe LOPES Jesse Davis Werner Dubitzky, *The 2017 Soccer Prediction Challenge*, https://osf.io/ftuva/, Online; accedido 27/10/2021, 2017.

[7] FiveThirtyEight, *La Liga Predictions, FiveThirtyEight*, https://projects.fivethirtyeight.com/soccer-predictions/la-liga/, Online; accedido 08/11/2021, 2021.

[8] Olocip, *Futbol Olocip, Tu partner tecnológico de Inteligencia Artificial*, https://olocip.com/futbol/, Online; accedido 08/11/2021, 2021.

[9] M. A. Gómez-Ruano, "Current approaches to performance analysis in sport," *RICYDE. Revista Internacional de Ciencias del Deporte*, vol. 14, n.º 51, págs. 3-4, 2018.

[10] *El análisis de datos aplicado a los deportes avanza en Uruguay — la diaria*, Online: accedido 25/09/2022. dirección: https://ladiaria.com.uy/deporte/articulo/2022/6/el-analisis-de-datos-aplicado-a-los-deportes-avanza-en-uruguay/.

[11] J. Morate Vázquez, "Predicción de equipo ganador en el baloncesto," B.S. thesis, 2016.

[12] H. M. Le, P. Carr, Y. Yue y P. Lucey, "Data-driven ghosting using deep imitation learning," 2017.

[13] Olocip, *Baloncesto, Olocip, Tu partner tecnológico de Inteligencia Artificial*, https://olocip.com/baloncesto/, Online; accedido 08/11/2021, 2021.

[14] E. de tendencia de consumos, *HÁBITOS DEPORTIVOS DE LOS URUGUAYOS*, https://www.biosistemas.com.uy/content/habitos-deportivos-de-los-uruguayos, Online; accedido 3/11/2021, 2021.

[15] J. Hucaljuk y A. Rakipović, "Predicting football scores using machine learning techniques," en *2011 Proceedings of the 34th International Convention MIPRO*, 2011, págs. 1623-1627.

[16] R. Rein y D. Memmert, "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science," *SpringerPlus*, vol. 5, n.º 1, págs. 1-13, 2016.

[17] E. Radío Gallego y col., "Análisis de resultados de encuentros de fútbol mediante técnicas de machine learning," B.S. thesis, 2017.

[18] Transfermarkt, *Transfermarkt*, https://www.transfermarkt.es/, Online; accedido 08/11/2021, 2021.

[19] O. Sports, *Opta Sports*, https://www.statsperform.com/opta-feeds/, Online; accedido 08/11/2021, 2021.

[20] S. Green, *Assessing The Performance of Premier League Goalscorers - Stats Perform*, https://www.statsperform.com/resource/assessing-the-performance-of-premier-league-goalscorers/, Online; accedido 08/11/2021, 2012.

[21] J. Whitmore, *What are expected goals(xG)?* https://theanalyst.com/na/2021/07/what-are-expected-goals-xg/, Online; accedido 08/11/2021, 2021.

[22] M. Caley, *Shot Matrix 1: Shot Matrix and expected goals*, https://cartilagefreecaptain.sbnation.com/2013/11/13/5098186/shot-matrix-i-shot-location-and-expected-goals, Online; accedido 08/11/2021, 2013.

[23] A. S. Analysis, *Shot Matrix 1: Shot Matrix and expected goals*, https://www.americansocceranalysis.com/expected-goals-2-explanation/, Online; accedido 08/11/2021, 2014.

[24] A. Rathke, "An examination of expected goals and shot efficiency in soccer," *Journal of Human Sport and Exercise*, vol. 12, n.º 2, págs. 514-529, 2017.

[25] M. D. Hoog, *How data, not people, call the shots in Denmark*, https://thecorrespondent.com/2607/How-data-not-humans-run-this-Danish-football-club/517995289284-77644562, Online; accedido 08/11/2021, 2015.

[26] ElenaSport, *The final API for football*, https://elenasport.io/, Online; accedido 10/11/2021, 2021.

[27] api-football, *RESTful API for Football data*, https://www.api-football.com/, Online; accedido 08/11/2021, 2021.

[28] BeSoccer, *La API de Futbol más completa*, https://api.besoccer.com/es, Online; accedido 08/11/2021, 2021.

[29] A. Football, *Livescore API*, https://apifootball.com/, Online; accedido 08/11/2021, 2021.

[30] EnetPulse, *Sports Data Widgets, Custom Sports API Service*, https://www.enetpulse.com/, Online; accedido 11/11/2021, 2021.

[31] DataFactory, *Convertimos datos en historias*, https://www.datafactory.la/, Online; accedido 11/11/2021, 2021.

[32] Playermaker, *Playermaker, The Number 1 Wearable Tracker for Soccer*, https://playermaker.com/, Online; accedido 08/11/2021, 2021.

[33] LiveFootball, *Live Football Scores, Fixtures and Results*, https://www.livescore.com/en/, Online; accedido 11/11/2021, 2021.

[34] Whoscored, *Estadísticas de Fútbol, Resultados de Fútbol en Directo*, https://es.whoscored.com/, Online; accedido 11/11/2021, 2021.

[35] FootMob, *The Pulse of Football*, https://www.fotmob.com/, Online; accedido 11/11/2021, 2021.

[36] T. Sports, *Noticias del deporte del canal líder en Argentina*, https://www.tycsports.com/, Online; accedido 11/11/2021, 2021.

[37] *Conceptos básicos de Machine Learning*, Online: accedido 01/09/2022. dirección: https://cleverdata.io/conceptos-basicos-machine-learning/.

[38] D. Benitez, *Fútbol europeo y sudamericano: un mismo deporte, muchas diferencias*, https://www.abc.com.py/periodismo-joven/2020/02/24/futbol-europeo-y-sudamericano-un-mismo-deporte-muchas-diferencias/, Online; accedido 02/08/2022, 2022.

[39] A. Marin, *La diferencia entre el fútbol en Europa y Sudamérica es abismal*, https://www.marca.com/claro-mx/opinion/opinion/2021/06/22/60d1f57f22601d1e468b459b.html, Online; accedido 02/08/2022, 2022.

[40] sklearn, *StandardScaler*, https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html, Online; accedido 18/06/2022, 2022.

[41] G. Moncecchi, *Curso Aprendizaje Automático*, https://eva.fing.edu.uy/enrol/index.php?id=43, Online: accedido 02/08/2022, 2022.

[42] T. M. Mitchell y T. M. Mitchell, *Machine learning*, 9. McGraw-hill New York, 1997, vol. 1.

[43] scikit learn documentation, *What is Random Forest? IBM*, https://www.ibm.com/cloud/learn/random-forest#\protect\leavevmode@ifvmode\kern+.2222em\relax~\protect\leavevmode@ifvmode\kern+.2222em\relaxtext=Randomforest%20is%20a%20commonly,both%20classification%20and%20regression%20problems., Online: accedido 02/08/2022, 2022.

[44] *1.10. decision trees*, Online: accedido 01/09/2022. dirección: https://scikit-learn.org/stable/modules/tree.html#tree-mathematical-formulation.

[45] ——, *Sklearn Linear Regresion*, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html, Online: accedido 02/08/2022, 2022.

[46] A. Lieuallen, *Greasemonkey*, https://addons.mozilla.org/es/firefox/addon/greasemonkey/, Online; accedido 07/12/2021, 2021.

[47] Google, *Colaboratory*, https://colab.research.google.com/, Online; accedido 31/05/2022, 2021.

[48] *Decision Trees: Gini vs Entropy*, Online: accedido 01/09/2022. dirección: https://quantdare.com/decision-trees-gini-vs-entropy/.

APÉNDICES

Apéndice 1

Arquitectura del sistema

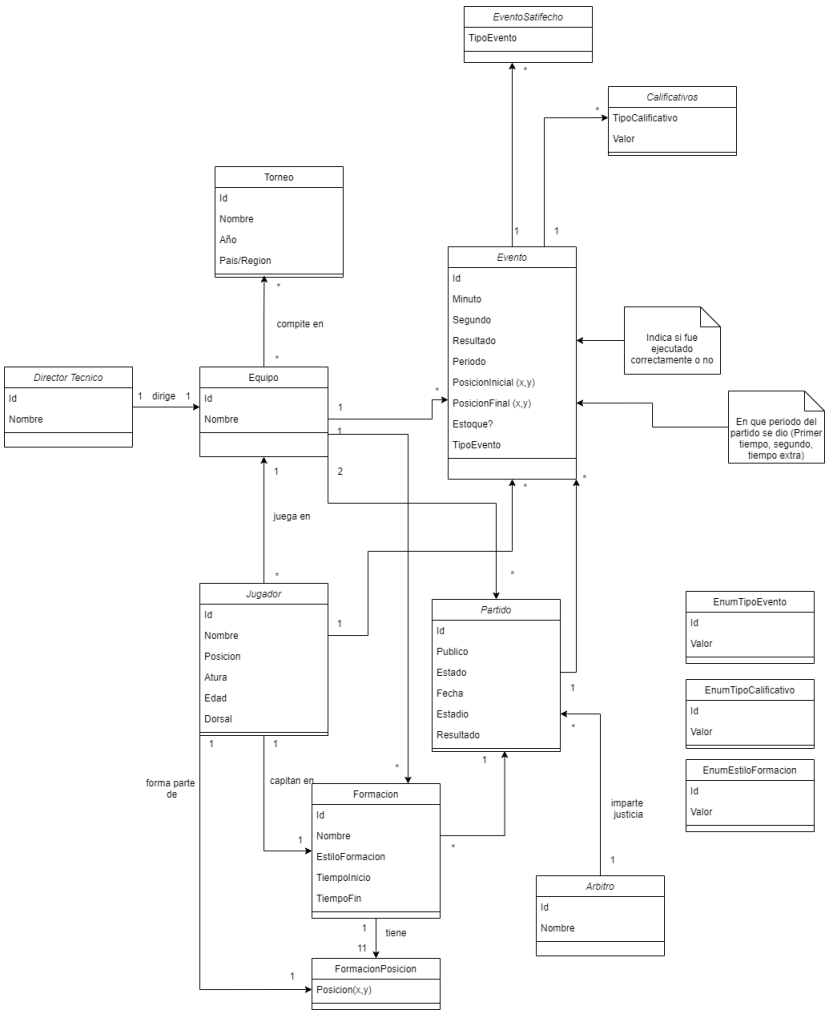

Figura 1.1: Diagrama de clases de partido Whoscored

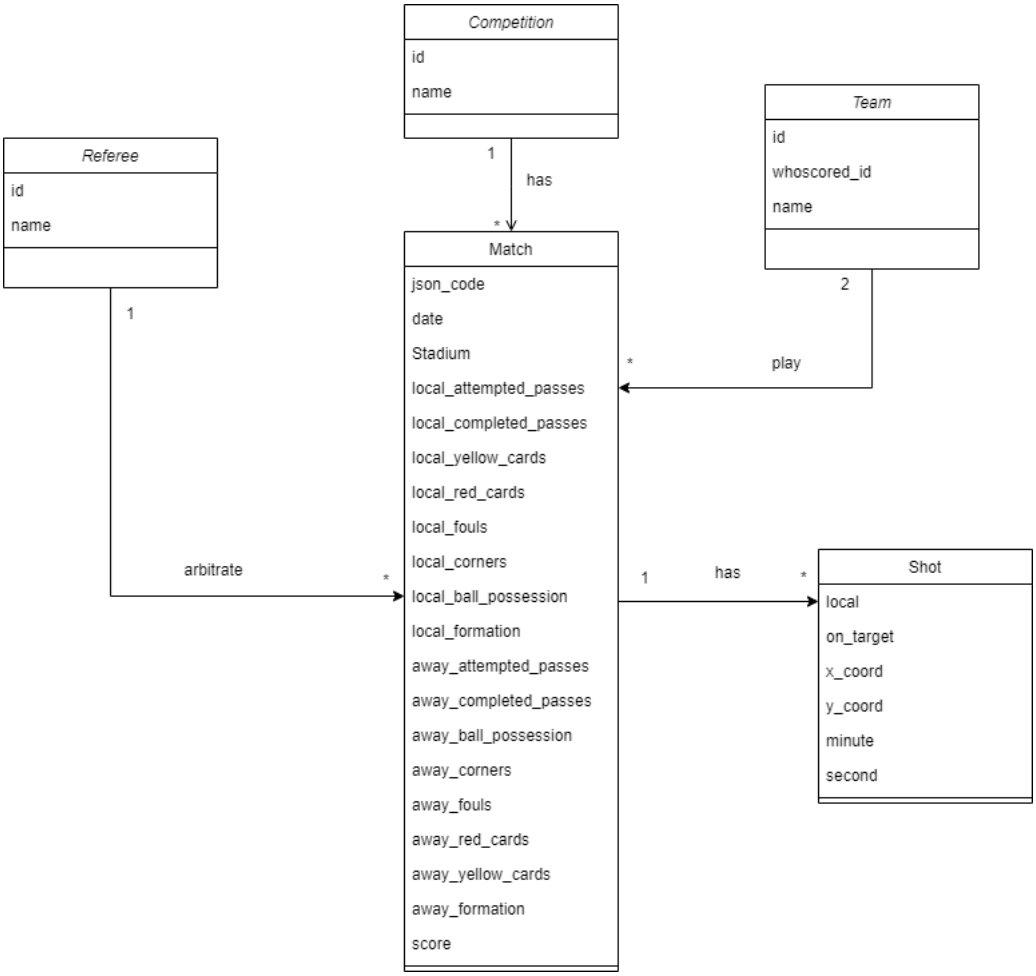

Figura 1.2: Diagrama de clases de solucion