Escuela Politécnica Superior de Jaén
Máster Universitario en Ingeniería Informática

UNIVERSIDAD DE JAÉN
Escuela Politécnica Superior (Jaén)

Trabajo Fin de Máster

MODELOS PREDICTIVOS PARA LA PREVENCIÓN DE LESIONES EN EL FÚTBOL

Alumno: Stumpf, Marcos

Tutores: Prof. D. María José Gacto Colorado
Prof. D. Pedro González García
Dpto.: Informática

Abril, 2018

Índice

Índice de Ilustraciones

Índice de Tablas

1. INTRODUCCIÓN

El análisis de datos es un campo muy estudiado en el mundo de los negocios desde la introducción del ordenador y de las bases de datos, sin embargo ganó relevancia y fuerza con la aparición del término *big data* en el año 1997 en una publicación sobre visualización de datos (Cox & Ellsworth, 1997) donde decía:

"La visualización ofrece un desafío interesante para los sistemas informáticos, los conjuntos de datos generalmente son bastante grandes, lo que reduce las capacidades de la memoria principal, el disco local e incluso el disco remoto. Llamamos a esto el problema del big data. Cuando los conjuntos de datos no caben en la memoria principal (en el núcleo), o cuando no encajan ni siquiera en el disco local, la solución más común es adquirir más recursos."

No obstante, Nikola Tesla ya había previsto en 1926 que los humanos serian capaces de acceder y analizar volúmenes masivos de datos en el futuro usando dispositivos de pequeño tamaño.

En los deportes, la analítica fue introducida algunas décadas después teniendo como precursor el béisbol, como se describe en el libro *Moneyball* (Lewis, 2004), que años después fue transformado en película de gran éxito (Miller, 2011). Desde entonces, la analítica venía siendo empleada de forma exclusiva en el análisis estadístico del desempeño de equipos y deportistas. Sólo en la ultima década es cuando comenzó a ser empleado en el área de la medicina deportiva por un simple motivo, que el nivel de exigencia y superación está aumentando continuamente en los deportes, haciendo que los atletas necesiten mejorar su nivel físico. Esto genera una amenaza constante sobre los atletas, entrenadores y equipos, como es el riesgo de lesiones. Este riesgo no sólo amenaza a nivel físico y psicológico al propio atleta, sino también al rendimiento del equipo, puesto que limita las opciones y ventajas estratégicas sobre el oponente y puede impactar en las finanzas del club (gastos de rehabilitación, productividad, etc.). Por estas razones la ocurrencia de lesiones y su gestión se convierte en un desafío común en los deportes.

Para reforzar la importancia de la prevención (Alvarez, 2017), en febrero de 2015, FIFA y la *International Football Association Board* (IFAB) si reunieron para anunciar la introducción del *FIFA Quality Programme for Electronic Performance and Tracking Systems,* llamado EPTS, que establece criterios de calidad, procedimientos de prueba y beneficios preventivos médicos con el propósito de utilizar tales dispositivos durante los partidos (FIFA, 2015). Un video explicativo se presenta en la ilustración 1.1.

Ilustración 1.1 Video explicando el programa EPTS (link)

Otra iniciativa de gran impacto creada por la FIFA fue el programa de prevención de lesiones llamado “The 11”. Este programa fue creado teniendo como base diversos estudios científicos y su objetivo es crear métodos específicos de entrenamiento enfocados a mejorar los grupos musculares más importantes, es decir, aquellos correlacionados con lesiones, según los estudios. Según datos de la FIFA, el porcentaje de lesiones está entre 10 y 50 lesiones por cada 1.000 horas de juego, donde los problemas más comunes son:

- Lesiones musculares y esguinces de tobillo
- Lesiones por uso excesivo
- Lesiones de rodilla
- Dolor de hueso púbico por distintas causas en varones
- Conmociones cerebrales

Durante la Copa Mundial de la FIFA Alemania 2006 ™, se produjeron una media de 2.3 lesiones por partido. Más de la mitad de ellas provocaron una pérdida de tiempo de diferente duración, durante la que el jugador no pudo entrenar, competir o ambas (FIFA, 2007).

De esta forma, este trabajo experimental se centra en la prevención de lesiones, un tema cada vez más importante en el mundo del deporte. Sin embargo, como no tenemos aún la posibilidad de recolectar datos históricos médicos de partidos jugados, nuestro foco se centrará en los datos históricos médicos provenientes de sesiones de entrenamiento. Para ello se podrán utilizar distintas fuentes de datos, principalmente procedentes de GPS. Estos datos se analizarán mediante modelos predictivos para la detección de las variables que permitan predecir el riesgo de lesión de un atleta. Esta herramienta permitirá a entrenadores, preparadores y fisioterapeutas disponer de un sistema de ayuda a la decisión para la gestión del riesgo de lesión de los atletas.

2. OBJETIVOS

El propósito de este trabajo es el diseño y desarrollo de nuevos modelos predictivos de lesiones en el fútbol profesional a partir de los datos obtenidos de los entrenamientos. En particular, se investiga la presencia de un patrón semanal a través de entrenamientos durante la temporada (es decir, a corto plazo), la importancia de las características para describir entrenamientos de fútbol y la probabilidad de predicción de lesiones.

De esta forma, el objetivo general se desglosa en los siguientes objetivos específicos:

- Analizar de forma cuantitativa la relación entre las lesiones y las distintas variables recogidas de cada atleta.
- Construir modelos para la predicción de lesiones, que puedan ser aplicados e incluso extendidos por el personal profesional de un equipo de fútbol.
- Proporcionar modelos, puntos de referencia e ideas para ayudar a los científicos a trabajar en un futuro en el área de predicción de lesiones.

3. MATERIAL Y METÓDOS

Para alcanzar esto objetivo con los mejores resultados en términos de desempeño se utilizan dos enfoques:

- Enfoque A: no realizar ningún filtro previo de los atributos o características (es decir las columnas de la tabla). Esta será nuestra línea de base.
- Enfoque B: realizar un filtro previo de los atributos o características a través del empleo de la prueba de hipótesis de Wilcoxon Mann-Whitney U.

Sujetos: Treinta y siete jugadores profesionales de fútbol que compiten en la Serie A brasileña (edad = 31 ± 11 años, altura = 181 ± 14 cm, masa corporal = 78,63 ± 20,33 kg) participaron en el estudio durante el período de competición de la temporada 2017 (20 semanas). Siete defensas centrales, cuatro laterales, doce centrocampistas y diez delanteros fueron analizados. Los porteros no se han incluido en el estudio.

Procedimiento: La actividad física de los jugadores durante cada sesión de entrenamiento fue monitorizada utilizando un sistema de posición global (GPS) diferencial de 1 Hz integrado (QStarz, Taiwan). Cada jugador portaba el llavero (GPS) en su pantalón corto. Todos los dispositivos estaban conectados 1 hora antes del entrenamiento, y todos los jugadores usaban el mismo dispositivo GPS para cada sesión de entrenamiento. Después de la grabación, los datos fueron descargados a un ordenador utilizando el software de gestión deportiva QSports™ (QSports, 2013), recortándose los datos al horario de inicio y fin del entrenamiento para saber el volumen exacto de la sesión de entrenamiento. Este software cuenta con una estructura de base de datos donde usted puede almacenar y analizar su entrenamiento y actividades basado en estadísticas significativas para una mejor revisión con varios dispositivos deportivos.

El Ceará Sporting Club, representado por el fisioterapeuta Giovanni Bruno Sala Ramirez, dio permiso para usar estos datos para la investigación, de acuerdo con la política de privacidad.

Descripción de los datos: Se registraron un total de 131 sesiones de entrenamiento en equipo. Las variables de la carga de entrenamiento registrada fueron:

#	Variable	Descripción
1	**EDA**	Edad del jugador (en años)
2	**LPA**	Variable objetivo (clase binaria) que indica la lesión de un jugador después del entrenamiento
3	**PPRE**	Peso del jugador antes del entrenamiento (en kilogramos)
4	**PPOS**	Peso del jugador después del entrenamiento (en kilogramos)
5	**DIF**	Diferencia entre el peso antes y después del entrenamiento (en kilogramos)
6	**ALT**	Altura del jugador (en centímetros)
7	**IMC**	Índice en la escala de masa corporal
8	**PGC**	Porcentual de grasa corporal
9	**MCM**	Índice de masa corporal magra (en kilogramos)
10	**PSR**	Índice de percepción subjetiva de recuperación
11	**PSE**	Índice de percepción subjetiva de esfuerzo
12	**DOL**	Índice en la escala de dolor
13	**VTR**	Volumen de entrenamiento (en minutos)
14	**SPR**	Número de carreras en alta velocidad (sprint > $5{,}5\ m \cdot s^{-2}$)
15	**VEM**	Velocidad media del jugador (en kilómetros por hora)
16	**DIT**	Distancia máxima recorrida (en metros)
17	**FCMAX**	Frecuencia cardiaca máxima (en latidos por minuto)
18	**FCMED**	Frecuencia cardiaca media (en latidos por minuto)
19	**PMFC**	Porcentual medio de la frecuencia cardiaca máxima
20	**Z5**	Cantidad de minutos en la zona de frecuencímetro a 90% de la frecuencia cardiaca máxima
21	**Z4**	Cantidad de minutos en la zona de frecuencímetro a 80% de la frecuencia cardiaca máxima
22	**Z3**	Cantidad de minutos en la zona de frecuencímetro a 70% de la frecuencia cardiaca máxima
23	**Z2**	Cantidad de minutos en la zona de frecuencímetro a 60% de la frecuencia cardiaca máxima
24	**Z1**	Cantidad de minutos en la zona de frecuencímetro a 50% de la frecuencia cardiaca máxima
25	**TRIMP**	Impulso de entrenamiento calculado a través de la multiplicación de la duración del ejercicio en cada zona por la intensidad (ej.: Z1=1x duración en minutos, Z2=2x duración en minutos…)
26	**UA**	Unidades arbitrarias, es decir la carga de entrenamiento calculada a través de la multiplicación del volumen de entrenamiento por la percepción subjetiva de esfuerzo
27	**POS**	Posición del jugador (ej.: defensa central, centrocampista, lateral, delantero)
28	**MJUG**	Histórico de minutos jugados por el jugador
29	**PART**	Variable binaria que indica si el jugador participó en el partido

Tabla 3.1 Descripción de los datos de exposición recogidos de GPS y de registros de la comisión técnica.

Además, nuevas variables fueron extraídas de cálculos con las variables existentes.

#	Variable	Descripción
30	PMAX	Potencia máxima
31	VO2MAX	Consumo máximo de oxigeno
32	PMVO2	Porcentual medio del oxígeno máximo

Tabla 3.2 Descripción de los datos calculados de las variables existentes

Análisis estadístico: Primero realizamos un análisis básico para obtener la frecuencia entre clases (equilibrada o no) y algunas estadísticas básicas como la media, la desviación típica, el mínimo, el máximo, la varianza y la curtosis. Después seguimos con la evaluación del conjunto de datos (*dataset*) a través del calculo de la tasa media de error, mediante validación cruzada utilizando un árbol de decisión como predictor, cogiendo el porcentaje de errores y dividiendo su desviación típica por la media. Un resultado < 1, indica que el conjunto de datos tiene calidad para continuar el análisis.

La siguiente fase del proceso es el preprocesamiento, donde buscamos valores perdidos y anomalías, aplicando tratamientos distintos (quitar lineas o rellenar valores) y también realizamos la ingeniería de las características (*feature engineering* en inglés) creando nuevas variables en el conjunto de datos.

A partir de este punto es cuando dividimos el proceso en dos enfoques. En el primer enfoque (llamado de línea de base) no hacemos nada y pasamos directamente a la siguiente tarea (selección de características o reducción de la dimensión). En el segundo enfoque realizamos el análisis de importancia de las características y se calcula el valor-p utilizando la prueba de hipótesis de Wilcoxon Mann-Whitney U antes de pasar a la siguiente tarea (selección de características o reducción de la dimensión).

Seguimos con la tarea de selección de características utilizando tres técnicas (correlación de puestos de Spearman, PCA y eliminación de características hacia atrás) para detectar las características relevantes en la predicción de lesiones. Este proceso también reduce la dimensionalidad del espacio de características para

superar el riesgo de sobreajuste y hace que el modelo de aprendizaje automático sea más fácil de interpretar por investigadores y expertos de área (Gareth, Witten, Hastie, & Tibshirani, 2013). Para la técnica no supervisada de reducción de la dimensión, PCA, como tenemos variables en escalas diferentes se utilizó como escalador estándar el metodo *z-score*, con la diferencia respecto al metodo min-max de que esta tiene un aumento de la varianza, obteniendo mejores resultados para PCA al analizar la matriz de covarianza.

Antes de seguir para la fase siguiente (modelado) el conjunto de datos se normalizó utilizando el escalador estándar min-max (excepción del PCA). La razón por la cual se utilizó esta escala es para mejorar la solidez de las características (desviación estándar muy pequeña) y obtener la distribución de valores cerca de una forma gaussiana. Los datos fueron estandarizados para cada sujeto con el fin de reducir la variabilidad intra-sujeto. Además, ciertos algoritmos requirieron la normalización de datos para estandarizar la distancia entre características, mejorando así la precisión de la clasificación.

En la fase de modelado se realizó la validación cruzada en diversos conjuntos de prueba para evaluar el proceso de aprendizaje automático. El coeficiente de correlación de matthews (*Matthews Correlation Coefficient (MCC)* en inglés) que llamaremos MCC, el *recall* sobre la clase positiva (lesión) y la tasa de falsos positivos o falsa alarma (*False Positive Rate (FPR)* en inglés), se calcularon también para estimar la precisión del algoritmo.

El MCC (1) es un coeficiente de correlación entre las clasificaciones binarias observadas y predichas que devuelve un valor entre -1 y +1. Un coeficiente de +1 representa una predicción perfecta; un coeficiente 0 indica que no es mejor que la predicción aleatoria; y un coeficiente -1 indica desacuerdo total entre predicción y observación. El MCC se puede calcular directamente a partir de la matriz de confusión usando la fórmula:

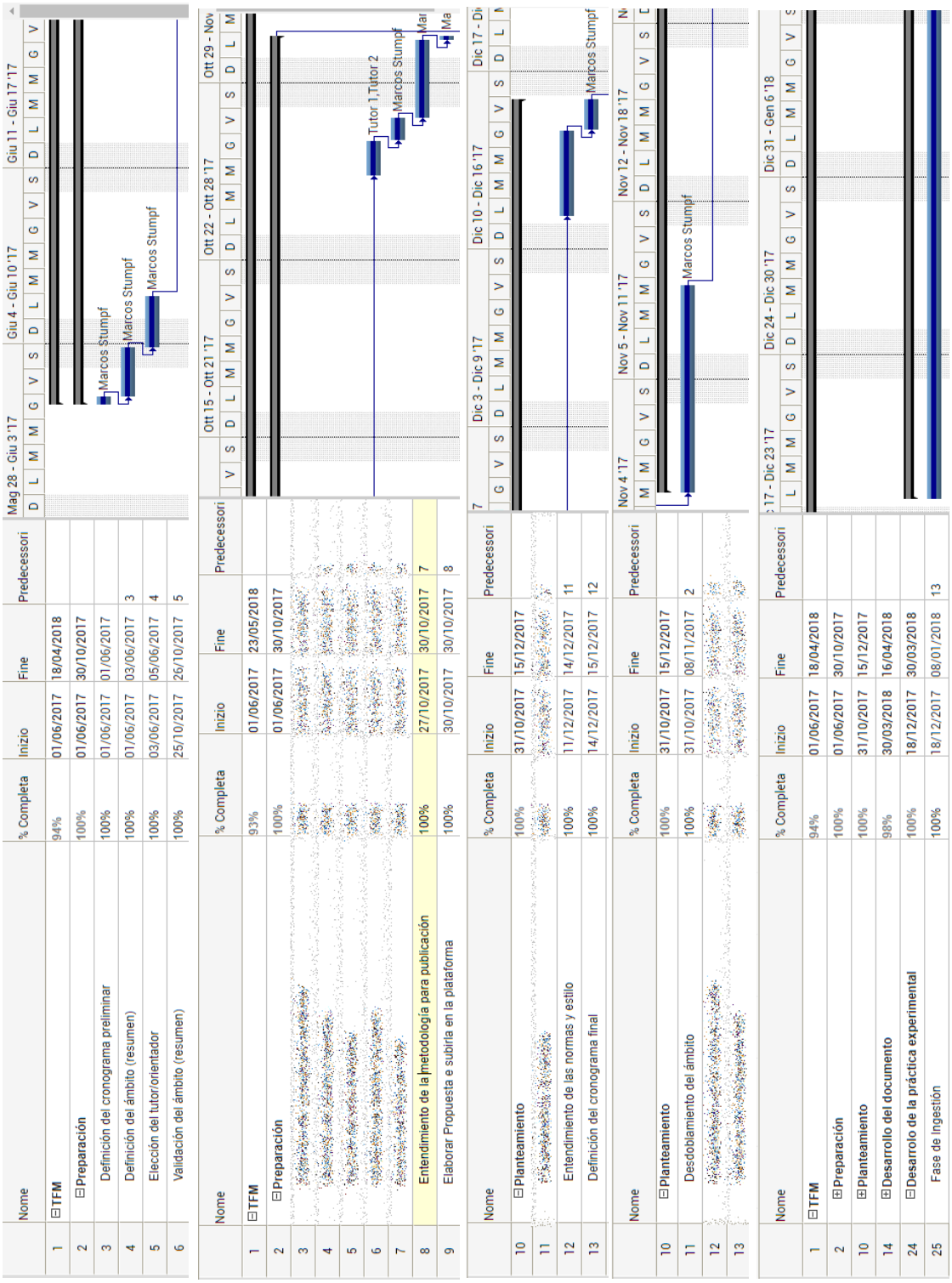
$$MCC = \frac{TP \times TN \ - FP \ \times FN}{\sqrt{(TP \ + FP)(TP \ + FN)(TN + FP)(TN + FN)}} \quad \textbf{(1)}$$

Donde: $MCC = \ matthews\ correlation\ coefficient$

$$TP = true\ positive$$
$$TN = true\ negative$$
$$FP = false\ positive$$
$$FN = false\ negative$$

Algunos científicos afirman que el coeficiente de correlación de Matthews es la puntuación individual que más información ofrece para establecer la calidad de predicción del clasificador binario utilizando la matriz de confusión (Boughorbel, Jarray, & M., 2017). Estas métricas de calidad del modelo fueron elegidas debido a que tratamos con una muestra con gran desbalanceo entre clases, por lo que las métricas más populares, como por ejemplo la exactitud, fueron descartadas por privilegiar la clase mayoritaria (no lesión), que en este caso no es la clase en la que estamos más interesados. Como meta de selección del modelo se estableció un Recall superior al 80% y un FPR inferior al 20%.

Cronograma: Realizamos la planificación de las actividades utilizando la herramienta Gantter, que es una API de terceros utilizada dentro del entorno Google Drive. Esta herramienta permite exportar el cronograma en formato Excel y Project. El cronograma se muestra a continuación.



	Nome	% Completa	Inizio	Fine	Predecessori
1	TFM	94%	01/06/2017	18/04/2018	
2	Preparación	100%	01/06/2017	30/10/2017	
3	Definición del cronograma preliminar	100%	01/06/2017	01/06/2017	
4	Definición del ámbito (resumen)	100%	01/06/2017	03/06/2017	3
5	Elección del tutor/orientador	100%	03/06/2017	05/06/2017	4
6	Validación del ámbito (resumen)	100%	25/10/2017	26/10/2017	5

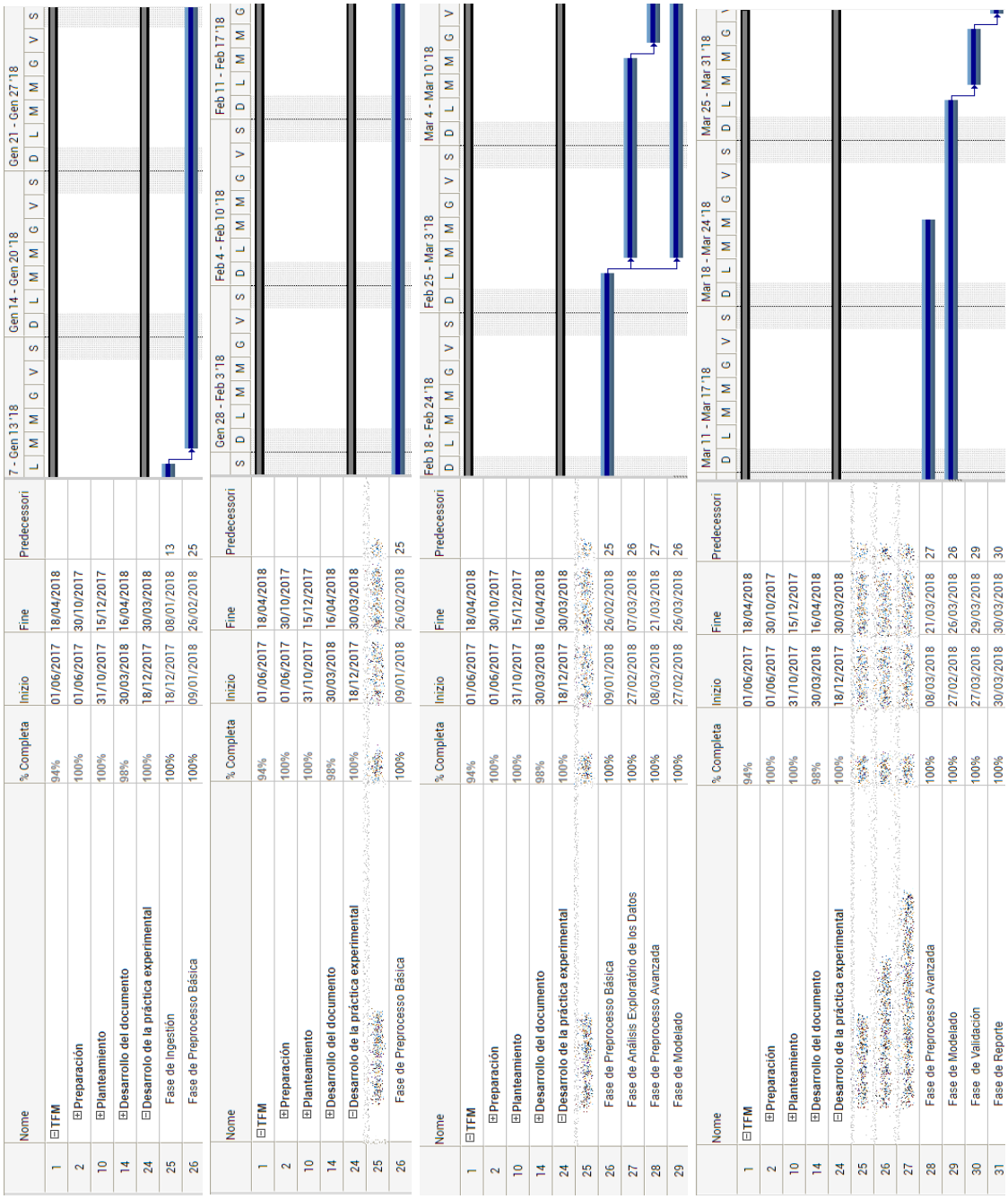
	Nome	% Completa	Inizio	Fine	Predecessori
1	TFM	93%	01/06/2017	23/05/2018	
2	Preparación	100%	01/06/2017	30/10/2017	
3					
4					
5					
6					
7					
8	Entendimiento de la metodología para publicación	100%	27/10/2017	30/10/2017	7
9	Elaborar Propuesta e subirla en la plataforma	100%	30/10/2017	30/10/2017	8

	Nome	% Completa	Inizio	Fine	Predecessori
10	Planteamiento	100%	31/10/2017	15/12/2017	
11					
12	Entendimiento de las normas y estilo	100%	11/12/2017	14/12/2017	11
13	Definición del cronograma final	100%	14/12/2017	15/12/2017	12

	Nome	% Completa	Inizio	Fine	Predecessori
10	Planteamiento	100%	31/10/2017	15/12/2017	
11	Desdoblamiento del ámbito	100%	31/10/2017	08/11/2017	2
12					
13					

	Nome	% Completa	Inizio	Fine	Predecessori
1	TFM	94%	01/06/2017	18/04/2018	
2	Preparación	100%	01/06/2017	30/10/2017	
10	Planteamiento	100%	31/10/2017	15/12/2017	
14	Desarrollo del documento	98%	30/03/2018	16/04/2018	
24	Desarrolo de la práctica experimental	100%	18/12/2017	30/03/2018	
25	Fase de Ingestión	100%	18/12/2017	08/01/2018	13

Ilustración 3.1 Cronograma del proyecto (Etapas de Preparación, Planteamiento y Desarrollo)



	Nome	% Completa	Inizio	Fine	Predecessori
1	⊟**TFM**	94%	01/06/2017	18/04/2018	
2	⊞Preparación	100%	01/06/2017	30/10/2017	
10	⊞Planteamiento	100%	31/10/2017	15/12/2017	
14	⊞Desarrollo del documento	98%	30/03/2018	16/04/2018	
24	⊟Desarrolo de la práctica experimental	100%	18/12/2017	30/03/2018	
25	Fase de Ingestión	100%	18/12/2017	08/01/2018	13
26	Fase de Preprocesso Básica	100%	09/01/2018	26/02/2018	25

	Nome	% Completa	Inizio	Fine	Predecessori
1	⊟**TFM**	94%	01/06/2017	18/04/2018	
2	⊞Preparación	100%	01/06/2017	30/10/2017	
10	⊞Planteamiento	100%	31/10/2017	15/12/2017	
14	⊞Desarrollo del documento	98%	30/03/2018	16/04/2018	
24	⊟Desarrolo de la práctica experimental	100%	18/12/2017	30/03/2018	
25					
26	Fase de Preprocesso Básica	100%	09/01/2018	26/02/2018	25

	Nome	% Completa	Inizio	Fine	Predecessori
1	⊟**TFM**	94%	01/06/2017	18/04/2018	
2	⊞Preparación	100%	01/06/2017	30/10/2017	
10	⊞Planteamiento	100%	31/10/2017	15/12/2017	
14	⊞Desarrollo del documento	98%	30/03/2018	16/04/2018	
24	⊟Desarrolo de la práctica experimental	100%	18/12/2017	30/03/2018	
25					
26	Fase de Preprocesso Básica	100%	09/01/2018	26/02/2018	25
27	Fase de Análisis Exploratório de los Datos	100%	27/02/2018	07/03/2018	26
28	Fase de Preprocesso Avanzada	100%	08/03/2018	21/03/2018	27
29	Fase de Modelado	100%	27/02/2018	26/03/2018	26

	Nome	% Completa	Inizio	Fine	Predecessori
1	⊟**TFM**	94%	01/06/2017	18/04/2018	
2	⊞Preparación	100%	01/06/2017	30/10/2017	
10	⊞Planteamiento	100%	31/10/2017	15/12/2017	
14	⊞Desarrollo del documento	98%	30/03/2018	16/04/2018	
24	⊟Desarrolo de la práctica experimental	100%	18/12/2017	30/03/2018	
25					
26					
27					
28	Fase de Preprocesso Avanzada	100%	08/03/2018	21/03/2018	27
29	Fase de Modelado	100%	27/02/2018	26/03/2018	26
30	Fase de Validación	100%	27/03/2018	29/03/2018	29
31	Fase de Reporte	100%	30/03/2018	30/03/2018	30

Ilustración 3.2 Cronograma del proyecto (Etapa de Desarrollo)

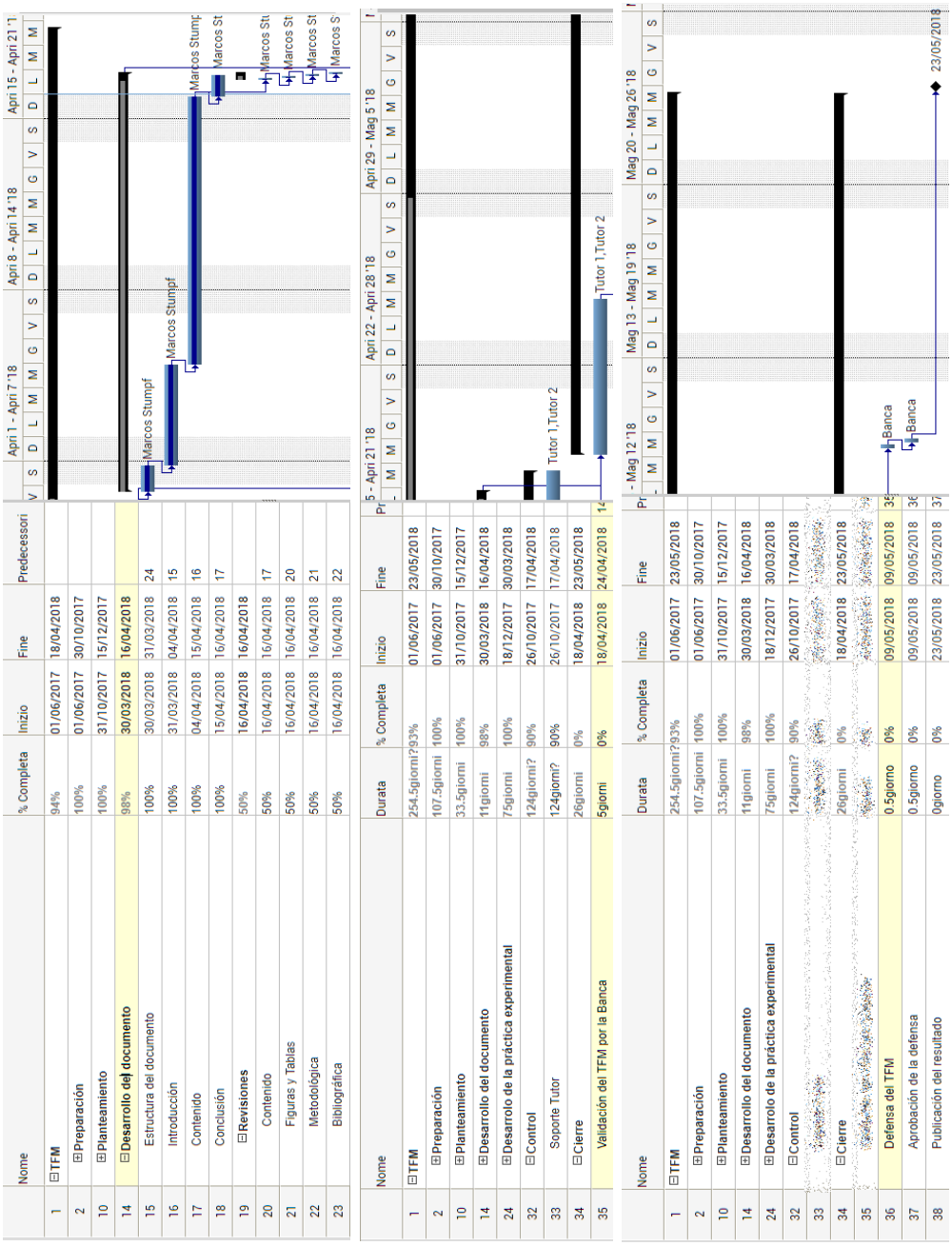


Ilustración 3.3 Cronograma del proyecto (Etapas de Desarrollo, Control y Cierre)

Análisis de costes: Para el desarrollo de este TFM utilizamos *software* con licencias académicas como en el caso de Tableau, de fuentes abiertas como en el caso de Knime y Python y versiones de prueba como en el caso de Minitab. La herramienta Excel ya disponía de licencia propia del investigador. Las horas con personal, en gran parte involucraron al propio investigador, teniendo a continuación las horas de los tutores y también las horas del profesional que cedió el conjunto de

datos. Sin embargo, las tablas a continuación muestran el presupuesto estimado para el proyecto.

Ítem	Tipo	Licencia	Cantidad	Costo unitario (€ euros)	Duración (meses)	Costo total (€ euros)
Knime	Software	Fuente abierta	1	0,00	07	0,00
Python	Software	Fuente abierta	1	0,00	07	0,00
Tableau	Software	Académica	1	59,47[1]	07	416,29
MiniTab	Software	Prueba gratis	1	1.595,00[4]	07	930,42
Office	Software	Personal	1	5,95[1]	07	41,65
Notebook	Hardware	-	1	440,00	07	51,35[5]
Investigador	Mano de obra	-	1	15,00[2]	07	16.800,00[3]
					TOTAL	**18.239,71**

Tabla 3.3 Presupuesto del proyecto.

[1] costo al mes de la versión paga. [2] costo hora de la persona. [3] 160 horas / mes. [4] costo anual. [5] costo de depreciación en el periodo.

Tarea	Horas	Costo (€ euros)
Preparación	108	1.620,00
Planteamiento	96	1.440,00
Desarrollo	812	12.180,00
Control	16	240,00
Cierre	88	1.320,00
TOTAL	**1.120**	**16.800,00**

Tabla 3.4 Horas y costo de trabajo del investigador por etapa.

4. ANTECEDENTES

Este capítulo comienza presentando información general sobre las técnicas y métodos que se utilizaron en la investigación descrita en este trabajo, así como la análisis de deportes y presenta otros campos científicos relacionados con el análisis de deportes. A continuación, se analiza la evolución de los estudios sobre lesiones y del análisis deportivo hasta los días actuales, antes de pasar al problema de la lesión en el fútbol con la ayuda del aprendizaje automático y discutir cómo se relaciona esto con este trabajo.

4.1. Ciencia de Datos

Según (Dhar, 2013) y (Leek, 2013), ciencia de datos *(data science* en inglés) es un campo interdisciplinario de métodos, procesos, algoritmos y sistemas para

extraer conocimiento o información de los datos en diversas formas, ya sea estructuradas o no.

El término *data science* fue utilizado por primera vez en 1997, durante conferencia inaugural para la cátedra Harry C. Carter por el professor Jeff Wu, que pidió que la estadística pasase a llamarse ciencia de dato y que las personas que trabajan en estadística se pasasen a denominar científicos de datos *("data scientists"* en inglés).

Sin embargo, de acuerdo con (Hayashi, 1998), la ciencia de datos emplea técnicas y teorías extraídas de muchos campos dentro del ámpio área de matemáticas, estadística, ciencias de la información y ciencias de la computación, en particular de los subdominios de aprendizaje automático *(machine learning* en inglés), datos en gran escala (*big data* en inglés), clasificación, análisis de clusters, cuantificación de incertidumbre, ciencia computacional, minería de datos (*data mining* en inglés), bases de datos, computación paralela, visualización y análisis de negocios. Como descripción profesional del trabajo, un científico de datos es alguien con habilidades relacionadas con la mayoría de estas áreas, a diferencia de, por ejemplo, un estadístico o experto en bases de datos que se especializa en uno solo de ellos.

4.2. Aprendizaje Automático

Según (Hayashi, 1998), el aprendizaje automático (*machine learning* en inglés) es un subcampo de las ciencias de la computación que evolucionó del reconocimiento de patrones y de la teoría del aprendizaje computacional en inteligencia artificial. El aprendizaje automático explora el estudio y la construcción de algoritmos que puedan aprender de sus errores y hacer predicciones sobre los datos. Tales algoritmos operan construyendo un modelo a partir de conjuntos de datos a fin de hacer predicciones o decisiones guiadas por los datos en lugar de simplemente siguiendo instrucciones programadas inflexibles y estáticas (Bishop, 2008).

Algunas partes del aprendizaje automático están íntimamente relacionadas (y muchas veces superpuestas) a la estadística computacional; una disciplina que se

centra en cómo hacer predicciones a través del uso de ordenadores, con investigaciones enfocando en las propiedades de los métodos estadísticos y su complejidad computacional. Tiene fuertes lazos con la optimización matemática, que produce métodos, teoría y dominios de aplicación para este campo. El aprendizaje automático se utiliza en una variedad de tareas de computación donde es impracticable crear y programar algoritmos explícitos. Ejemplos de estas aplicaciones incluyen filtrado de spam, reconocimiento óptico de caracteres (OCR), procesamiento de lenguaje natural, motores de búsqueda, diagnósticos médicos (que es el foco de este trabajo), bioinformática, reconocimiento de voz, reconocimiento de escritura, visión computacional y locomoción de robots (Wernick, Brankov, & Strother, 2010).

En el campo del análisis de datos (*data analysis* en inglés), el aprendizaje automático es un método usado para planificar modelos complejos y algoritmos que se prestan a hacer predicciones de uso comercial, lo que se conoce como análisis predictivo (*predictive analysis* en inglés). Estos modelos analíticos permiten a investigadores, científicos de datos, ingenieros, y analistas "producir decisiones y resultados confiables y repetibles" y descubrir "ideas ocultas" a través del aprendizaje de las relaciones y tendencias históricas en los datos (SAS, 2016). Se pueden distinguir los sigjientes tipos de modelos (Bishop, 2008):

- Modelos geométricos, construidos en el espacio de instancias y que pueden tener una, dos o múltiples dimensiones. Si hay un borde de decisión lineal entre las clases, se dice que los datos son linealmente separables, como en la Ilustración 4.1.

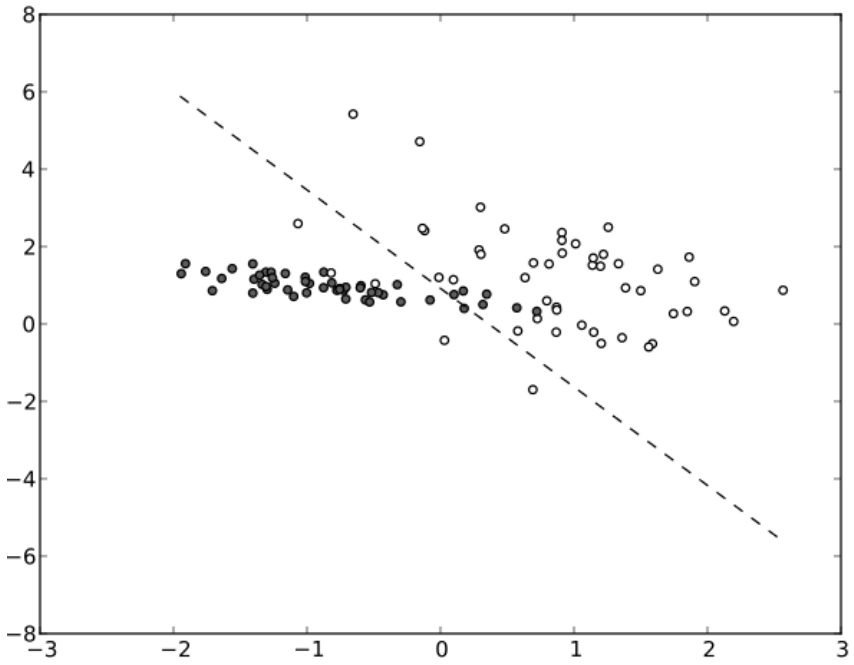


Ilustración 4.1 Una representación de un modelo geométrico.
(De Qwertyus - Trabajo propio, CC0, https://goo.gl/L577tr)

- Modelos probabilisticos, que intentan determinar la distribución de probabilidades que describen la función que enlaza a los valores de las características con valores determinados. Uno de los conceptos claves para desarrollar modelos probabilisticos es la estadística bayesiana, como en la Ilustración 4.2.

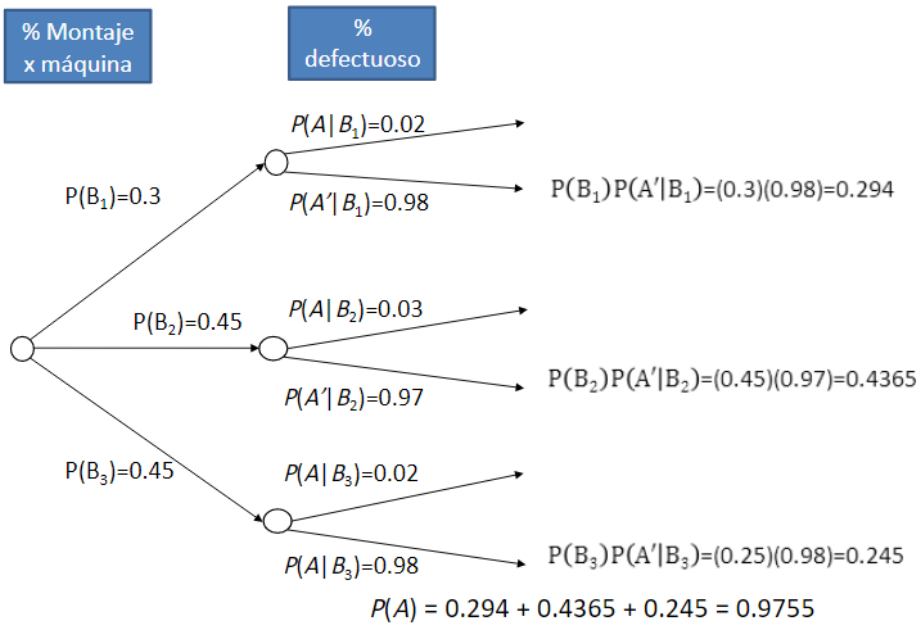


Ilustración 4.2 Una representación de un modelo probabilístico.

- Modelos lógicos, que transforman y expresan las probabilidades en reglas organizadas en forma de árboles de decisión, como en la Ilustración 4.3.

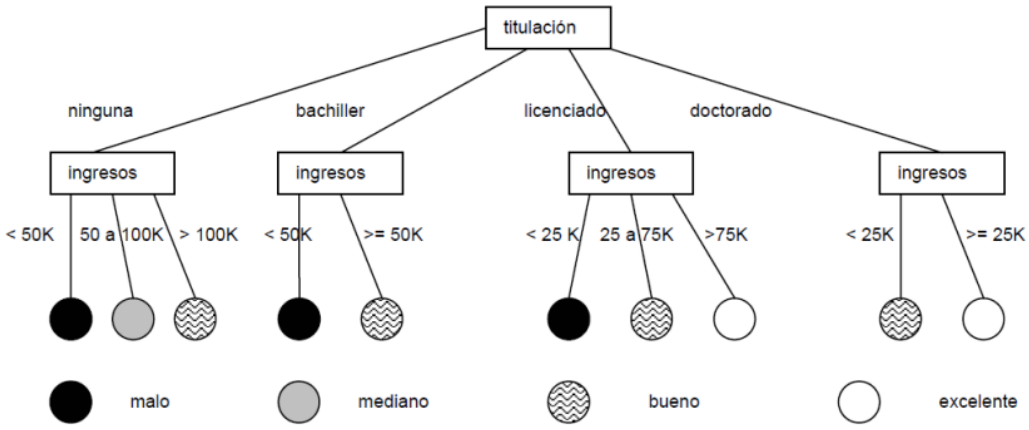


Ilustración 4.3 Una representación de un modelo lógico.

Alternativamente, pueden caracterizarse por su modus operandi (Flach, 2012):

- Modelos de Agrupamiento, que dividen el espacio de instancia en segmentos; en cada segmento se aprende un modelo muy simple (por ejemplo, constante).
- Modelos de Gradiente, que aprenden un único modelo global sobre el espacio de la instancia. Los clasificadores geométricos, como las máquinas de soporte vectorial (SVM) son modelos de gradientes.

En la Ilustración 4.4 tratamos de relacionar los modelos ya explicados con algunos tipos de algoritmos utilizados para un mejor entendimiento de los siguientes capítulos. Aquellos que están más cerca unos de otros comparten las mismas características. Los algoritmos más a la derecha (en rojo) son modelos lógicos, los algoritmos en la parte superior izquierda (en verde) son modelos geométricos y algoritmos en la parte inferior izquierda (en azul), modelos probabilisticos.

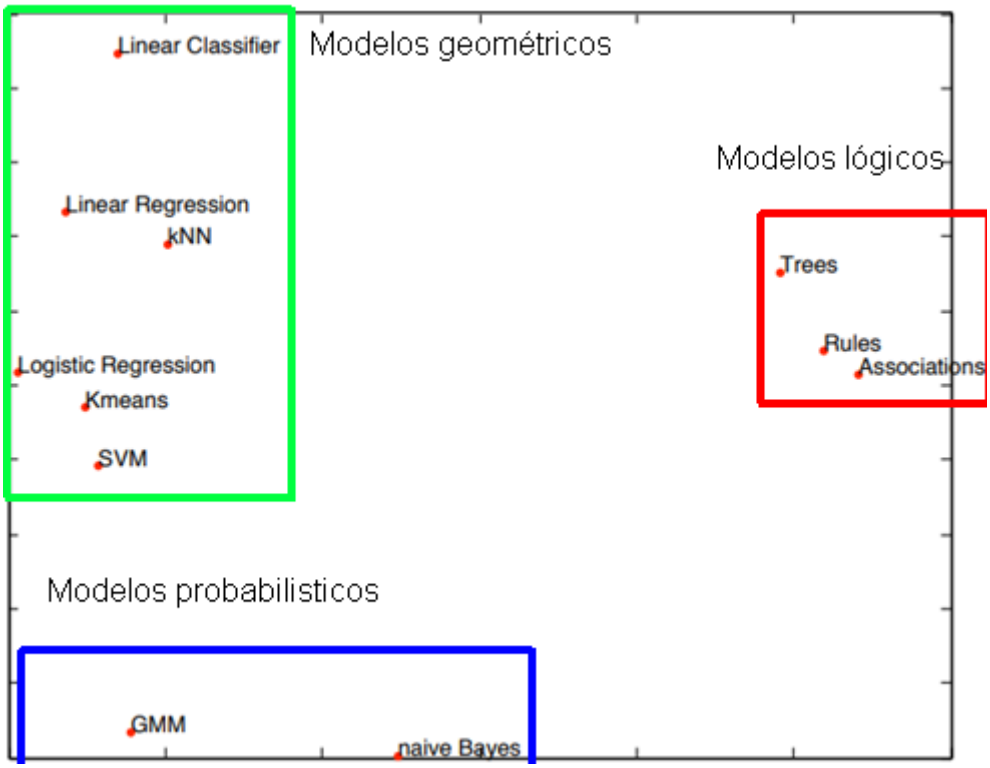


Ilustración 4.4 Mapa que relaciona algunos algoritmos con el tipo de modelo asociado.

Todos los modelos trabajan sobre datos en bruto recolectados, que son etiquetados o no (cuando trabajamos con aprendizaje no supervisado) por sus atributos o características, que es un tipo de medida realizada sobre cualquier instancia a medir. Este tipo de medida puede estar compuesta por atributos predictores y atributos clase (cuando trabajamos con aprendizaje supervisado). Los atributos asignan el espacio de instancias a un conjunto de valores o dominio de atributos. Los valores del dominio pueden ser números, valores binarios o un conjunto textual cualquiera. Además son clasificados en variables categóricas (que contienen un número finito de categorías o grupos distintos no pudiendo tener un orden lógico), discretas (que tienen un número finito de valores entre dos valores numéricos cualesquiera) y continuas (que tienen un número infinito de valores entre dos valores cualesquiera de tipo numérica o de fecha / hora).

4.2.1. Minería de Datos

El aprendizaje automático a veces se confunde con la minería de datos, que es un sub-campo que se centra más en el análisis exploratorio de datos (*exploratory data analysis (EDA)* en inglés) y se conoce como aprendizaje no supervisado (Mannila, 1996). Este es el proceso de explorar grandes cantidades de datos en

busca de estándares consistentes, como reglas de asociación o series temporales, para detectar relaciones sistemáticas entre variables, detectando así nuevos subconjuntos de datos (Maimon & Rokach, 2010).

4.2.2. Tipos de Algoritmos

Los diferentes algoritmos de aprendizaje automático se agrupan en una taxonomía en función de su salida. Los principales tipos de algoritmos son:

- Aprendizaje supervisado, el algoritmo produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. Un ejemplo de este tipo de algoritmo es el problema de clasificación, donde el sistema de aprendizaje trata de etiquetar (clasificar) una serie de vectores utilizando una entre varias categorías (clases). La base de conocimiento del sistema está formada por ejemplos etiquetados (Kotsiantis, 2007).
- Aprendizaje no supervisado, todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema. No se tiene información sobre las categorías de esos ejemplos. Por lo tanto, en este caso, el sistema tiene que ser capaz de reconocer patrones para poder etiquetar las nuevas entradas (Duda, Hart, & Stork, 2000).
- Aprendizaje semi supervisado, estos tipos de algoritmos combinan los dos algoritmos anteriores para poder clasificar de manera adecuada. Se tiene en cuenta los datos marcados y los no marcados. En muchas situaciones prácticas, el costo de etiquetar es bastante alto, ya que requiere expertos humanos capacitados para hacerlo. Entonces, en ausencia de etiquetas en la mayoría de las observaciones, aunque presentes en pocos, los algoritmos semi supervisados son los mejores candidatos para la construcción del modelo (Abney, 2007).
- Aprendizaje por refuerzo, el algoritmo aprende observando el mundo que le rodea. Su información de entrada es el *feedback* o retro alimentación que obtiene del mundo exterior como respuesta a sus acciones, la Ilustración 4.5 representa de forma genérica este concepto. Por lo tanto, el sistema aprende mediante prueba y error. Algunas

aplicaciones de los algoritmos de aprendizaje de refuerzo son los juegos de tablero con computadora, manos robóticas y coches que conducen de forma autónoma (Sutton & Barto, 1998).

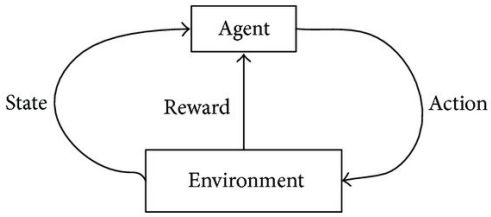


Ilustración 4.5 Representación en alto nivel del flujo de aprendizaje por refuerzo.

Después de repasar los conceptos sobre modelos, atributos, tipos y categorías de estos atributos, además de los principales tipos de algoritmos, se presenta la Ilustración 4.6 donde relaciona cada uno de estos términos de forma estructurada, e incluyendo un ejemplo de su aplicación práctica en el mundo real.

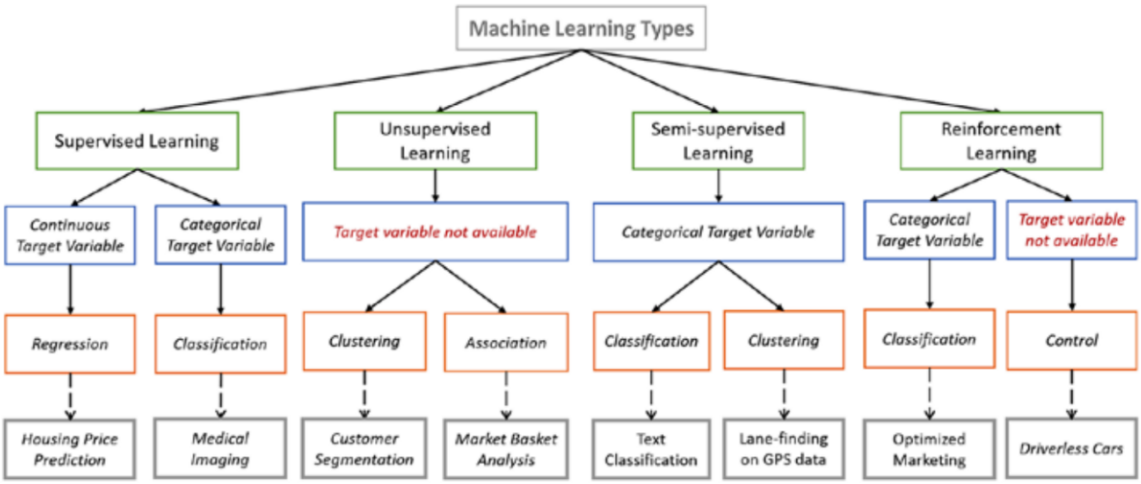


Ilustración 4.6 Representación de la relación entre los tipos de algoritmos y atributos con su aplicación práctica.

4.2.3. Enfoques

En este apartado se describirán los 5 mejores algoritmos que se han utilizado trabajo experimental. Aunque para el trabajo experimental hayamos realizado pruebas con 10 algoritmos, como se indica más adelante en la sección de informes (5.1.6), el motivo de la elección de 5 algoritmos para este TFM, fue el de seguir uno de sus objetivos específicos, que es la construcción de modelos que puedan ser aplicados y extendidos; es decir, aquellos modelos que presentan en sus métricas

valores que satisfagan en parte las metas, de Recall superior al 80% y FPR inferior al 20%, para su posible aplicación en el área.

- Árboles de decisión. Es un modelo de predicción utilizado en diversos ámbitos que van desde la inteligencia artificial hasta la Economía. Dado un conjunto de datos se fabrican diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas como en la Ilustración 4.7, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema (Flach, 2012).

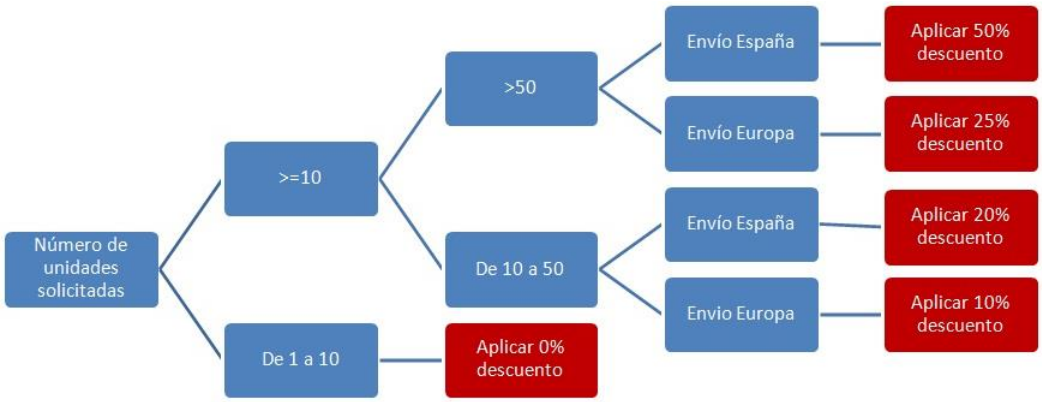


Ilustración 4.7 Mapa que relaciona algunos algoritmos con el tipo de modelo asociado.

  - Bosques aleatorios (*random forest* en inglés (Ho, 1995)), es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de ellos. Es una modificación sustancial de *bagging* que construye una larga colección de árboles no relacionados y luego los promedia. La idea esencial del *bagging* es promediar muchos modelos ruidosos pero aproximadamente imparciales, y por tanto reducir la varianza. Los bosques de decisión aleatoria corrigen el hábito de los árboles de decisión de sobre ajustarse a su conjunto de entrenamiento. Cada árbol es construido usando el siguiente algoritmo:
    - Sea *N* el número de casos de prueba, *M* es el número de variables en el clasificador;

- Sea *m* el número de variables de entrada a ser usado para determinar la decisión en un nodo dado; m debe ser mucho menor que *M*;
- Elegir un conjunto de entrenamiento para este árbol y usar el resto de los casos de prueba para estimar el error;
- Para cada nodo del árbol, elegir al azar m variables en las cuales basar la decisión. Calcular la mejor partición del conjunto de entrenamiento a partir de las *m* variables.

Para la predicción, el árbol empuja hacia abajo un nuevo caso. Luego se le asigna la etiqueta del nodo terminal donde termina. Este proceso se repite por todos los árboles en el ensamblado, y la etiqueta que obtenga la mayor cantidad de ocurrencias es devuelta como predicción (Ilustración 4.8). En muchos problemas el rendimiento del algoritmo *random forest* es muy similar a la del *boosting*, y es más simple de entrenar y ajustar. Como consecuencia, el *random forest* es popular y ampliamente utilizado, pues para un conjunto de datos lo suficientemente grande produce un clasificador muy preciso (Fernandez-Delgado, Cernadas, Barro, & Amorim, 2014).

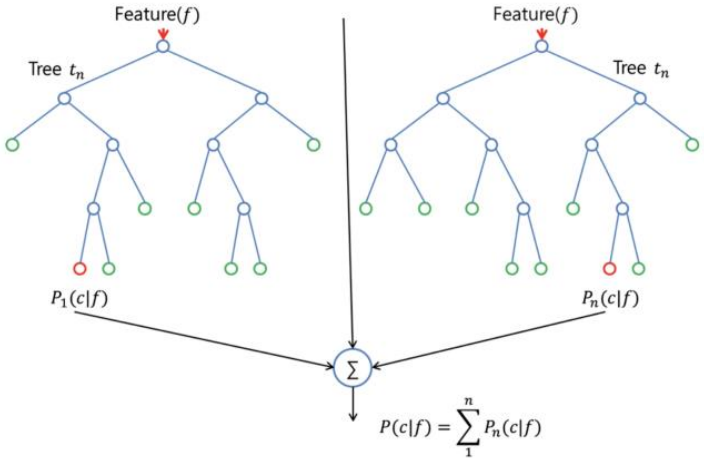


Ilustración 4.8 Representación simple de la estructura de los bosques aleatorios.

- *Boosting*. De forma similar al *Bagging*, cada clasificador es entrenado usando un conjunto de entrenamiento diferente. La principal diferencia con el *Bagging* es que los conjuntos de datos re-muestreados se

construyen específicamente para generar aprendizajes complementarios y la importancia del voto se basa en el rendimiento de cada modelo, en lugar de la asignación del mismo peso para todos los votos. Esencialmente, este procedimiento permite aumentar el rendimiento por encima de un umbral arbitrario, simplemente añadiendo modelos más débiles. Dada la utilidad de este hallazgo, *Boosting* es considerado uno de los descubrimientos más significativos en el aprendizaje automático (Lantz, 2013).

    - AdaBoost (*Adaptive Boosting* en inglés (Tsai, Hsu, & Yen, 2014)), es una *c*ombinación de las ideas de Bagging y Boosting y no requiere un gran conjunto de entrenamiento como el *Boosting*. Inicialmente, cada ejemplo de formación de un determinado conjunto de entrenamiento tiene el mismo peso, entonces AdaBoost llama a un clasificador débil repetidamente en iteraciones. Para cada llamada, la distribución de pesos se actualiza para indicar la importancia del ejemplo en el conjunto de datos utilizado para la clasificación. En cada iteración, los pesos de cada ejemplo clasificado incorrectamente se incrementan (o alternativamente, los pesos clasificados correctamente se decrementan), para que entonces el nuevo clasificador trabaje en más ejemplos. Por lo tanto, la decisión final se basa en la votación ponderada de los clasificadores individuales.

- Lógica difusa (*Fuzzy logic* en inglés (Ahlawat, Gautam, & Sharma, 2014)). Es la forma de lógica multivalorada en la que los valores lógicos de las variables pueden ser cualquier número real entre 0 (FALSO) y 1 (VERDADERO). En cambio, en la lógica booleana, los valores lógicos de las variables pueden ser sólo 0 y 1. La lógica difusa se ha extendido para tratar con el concepto de verdad parcial, donde el valor verdadero puede comprender entre completamente verdadero y completamente falso. Las implementaciones de la lógica difusa permiten que los estados indeterminados puedan ser tratados por dispositivos de control. De este modo, es posible evaluar conceptos no cuantificables (Ilustración 4.9). Casos prácticos: evaluar la temperatura (cálido, frío, medio, etc.), la

sensación de felicidad (radiante, feliz, apático, triste, etc.), la veracidad de un argumento (correcto, incoherente, falso, etc.).

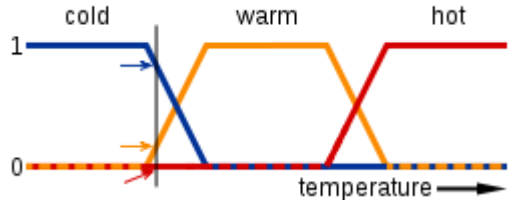


Ilustración 4.9 Representación simple de lógica difusa.

- FURIA (*Fuzzy Unordered Rule Induction Algorithm*) (Huehn & Huellermeier, 2009)), es una modificación y una extensión del aprendiz de reglas del estado de la técnica RIPPER (Cohen, 1995). En particular, FURIA aprende reglas difusas en lugar de reglas convencionales y conjuntos de reglas no ordenadas en lugar de listas de reglas. Además, para tratar con ejemplos descubiertos, hace uso de un método eficiente de estiramiento de reglas. Una de sus principales ventajas está relacionada con la interpretabilidad de las reglas. Sin embargo, cada regla generada presenta su propia distribución de los conjuntos difusos que definen los atributos. De esta manera, la interpretabilidad de la regla establecida como un todo se ve afectada. FURIA es uno de los métodos de clasificación difusa más avanzados actualmente.

- Lista de decisiones (Rivest, 1987), son una representación de funciones booleanas que pueden aprenderse fácilmente a partir de ejemplos. Las listas de decisiones de un solo término son más expresivas que las disyunciones y las conjunciones; sin embargo, las listas de decisiones de un solo término son menos expresivas que la forma normal disyuntiva general y la forma normal conjuntiva. El lenguaje especificado por una lista de decisiones de longitud k incluye como un subconjunto el lenguaje especificado por un árbol de decisión de profundidad k. Las listas de decisiones de aprendizaje se pueden usar para el aprendizaje eficiente de atributos.
  - PART (*Partial* en inglés (Frank & Wittenx, 1998)), es llamado así porque está basado en el árbol de decisión parcial. El método combina dos paradigmas de aprendizaje de reglas, C4.5 (Quinlan

J. R., 1993) y RIPPER (Cohen, 1995), que operan en dos etapas. Primero, inducen un conjunto de reglas inicial y luego lo refinan usando una etapa de optimización bastante compleja que descarta (C4.5) o ajusta (RIPPER) las reglas individuales para que funcionen mejores juntas. El algoritmo genera varias veces árboles de decisión parciales, combinando así estos dos métodos citados anteriormente, creando reglas a partir de árboles de decisión y la técnica de aprendizaje de reglas divide y vencerás. La principal ventaja de PART sobre los otros esquemas discutidos no es el rendimiento sino la simplicidad: al combinar dos paradigmas de aprendizaje de reglas produce buenos conjuntos de reglas sin necesidad de optimización global. A pesar de esta simplicidad, el método produce conjuntos de reglas que se comparan favorablemente con los generados por C4.5 y C5.0 (Quinlan J. R., 1996), y son más precisos (aunque más grandes) que los producidos por RIPPER.

- Red neuronal artificial (*Artificial neural network (ANN)* en inglés (Van Gerven & Bohte, 2017)). Son sistemas de computación vagamente inspirados por las redes neuronales biológicas que constituyen los cerebros de los animales. Dichos sistemas "aprenden" (es decir, mejoran progresivamente el rendimiento en) tareas al considerar ejemplos, generalmente sin programación específica de la tarea. Por ejemplo, en el reconocimiento de imágenes, pueden aprender a identificar imágenes que contienen gatos mediante el análisis de imágenes de ejemplo que han sido etiquetadas manualmente como "gato" o "sin gato" y utilizando los resultados para identificar gatos en otras imágenes. Lo hacen sin ningún conocimiento previo sobre los gatos, por ejemplo, que tienen pelaje, cola, bigotes y cara de gato. En cambio, desarrollan su propio conjunto de características relevantes a partir del material de aprendizaje que procesan. Una ANN se basa en una colección de unidades conectadas o nodos llamadas neuronas artificiales (una versión simplificada de neuronas biológicas en un cerebro animal). Cada conexión (una versión simplificada de una sinapsis) entre las neuronas artificiales puede transmitir una señal de

uno a otro. La neurona artificial que recibe la señal puede procesarla y luego señalar las neuronas artificiales conectadas a ella. En las implementaciones de ANN comunes, la señal en una conexión entre neuronas artificiales es un número real, y la salida de cada neurona artificial se calcula mediante una función no lineal de la suma de sus entradas. Las neuronas artificiales y las conexiones generalmente tienen un peso que se ajusta a medida que avanza el aprendizaje. El peso aumenta o disminuye la intensidad de la señal en una conexión. Las neuronas artificiales pueden tener un umbral tal que solo si la señal agregada cruza ese umbral es la señal enviada. Por lo general, las neuronas artificiales se organizan en capas. Las diferentes capas pueden realizar diferentes tipos de transformaciones en sus entradas. Las señales viajan desde la primera capa (entrada) hasta la última (salida), posiblemente después de atravesar las capas varias veces.

- Red neuronal de retroalimentación (*feedforward neural network* en inglés). En esta red la información siempre se mueve en una dirección; nunca va hacia atrás. Una red neuronal de retroalimentación es una red neuronal artificial en donde las conexiones entre las unidades no forman un ciclo, como presentado en la Ilustración 4.10. Como tal, es diferente de las redes neuronales recurrentes (Zell, y otros, 1994) (Schmidhuber, 2015).

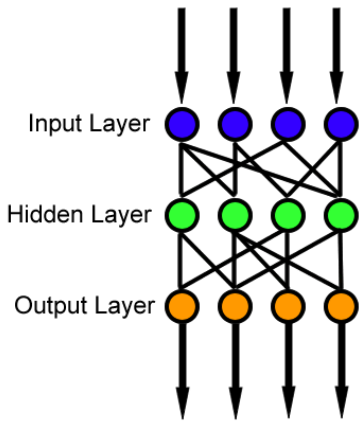


Ilustración 4.10 Representación simple de una red neuronal de retro alimentación.

- Retro propagación (*backpropagation* en inglés (Goodfellow, Bengio, & Courville, 2016)), es un método utilizado en redes neuronales artificiales para calcular un gradiente que se necesita en el cálculo de los pesos que se utilizarán en la red. En el contexto del aprendizaje, la retro propagación es comúnmente utilizada por el algoritmo de optimización del descenso del gradiente para ajustar el peso de las neuronas mediante el cálculo del gradiente de la función de pérdida. Esta técnica también se denomina a veces propagación de errores hacia atrás, porque el error se calcula en la salida y se distribuye de vuelta a través de las capas de la red. La motivación para la retro propagación es entrenar una red neuronal multicapa de manera que pueda aprender las representaciones internas apropiadas para que pueda aprender cualquier asignación arbitraria de entrada a salida. La retro propagación requiere un resultado conocido y deseado para cada valor de entrada; por lo tanto, se considera que es un método de aprendizaje supervisado.
- Rprop, abreviatura de *backpropagation* resistente (Riedmiller & Braun, 1992). Es una heurística de aprendizaje supervisado en redes neuronales artificiales *feedforward*. Este es un algoritmo de optimización de primer orden. De manera similar a la regla de actualización de Manhattan, Rprop toma en cuenta solo el signo de la derivada parcial sobre todos los patrones (no la magnitud), y actúa independientemente sobre cada "peso". Para cada peso, si hubo un cambio de signo de la derivada parcial de la función de error total en comparación con la última iteración, el valor de actualización para ese peso se multiplica por un factor η-, donde η- <1. Si la última iteración se produjo el mismo signo, el valor de actualización se multiplica por un factor de η +, donde η +> 1. Los valores de actualización se calculan para cada peso de la manera anterior, y finalmente cada peso se cambia por su propio valor de actualización, en el sentido opuesto dirección de la derivada parcial de ese peso, para minimizar la función de error total. η + se establece empíricamente en 1.2 y η- en 0.5. Junto al algoritmo

de correlación en cascada y el algoritmo de Levenberg-Marquardt, RPROP es un algoritmo de actualización por lotes y utiliza uno de los mecanismos de actualización de peso más rápidos.

- PCA (*principal component analysis* en inglés (Abdi & Williams, 2010)), es un procedimiento matemático que utiliza una transformación ortogonal (ortogonalización de vectores) para convertir un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de variables linealmente no correlacionadas llamadas de componentes principales. El número de componentes principales es menor o igual al número de variables originales. Esta transformación se define de forma que el primer componente principal tiene la mayor varianza posible (es decir, es responsable del máximo de variabilidad en los datos), y cada componente siguiente, a su vez, tiene la máxima varianza bajo la restricción de ser ortogonal a (es decir, no correlacionado con) los componentes anteriores. Los componentes principales son garantizados independientemente si los datos se distribuyen normalmente (conjuntamente). El PCA es sensible a la escala relativa de las variables originales. El PCA puede hacerse por descomposición en auto valores (valores propios) de una matriz de covarianza, generalmente después de centralizar (y normalizar o utilizar las puntuaciones-Z) la matriz de datos para cada atributo. Con frecuencia, su operación puede ser tomada como reveladora de la estructura interna de los datos, de una forma que mejor explica la varianza en los datos. Ahora, el PCA es más comúnmente utilizado como una herramienta de análisis de exploración de datos y para hacer modelos predictivos.
- Selección del subconjunto de características basadas en la correlación de puestos (*rank correlation* en inglés (Minitab Inc., 2016)), un coeficiente de correlación mide el grado en que dos variables tienden a cambiar juntas. El coeficiente describe tanto la fuerza como la dirección de la relación. La correlación de Spearman o correlación de puestos, evalúa la relación monotónica entre dos variables continuas u ordinales. En una relación monotónica, las variables tienden a cambiar juntas, pero

no necesariamente a un ritmo constante. El coeficiente de correlación de Spearman puede variar en valor de -1 a +1 y se basa en los valores ordenados para cada variable en lugar de los datos brutos. Para que el coeficiente de correlación de Spearman sea +1, cuando una variable aumenta, la otra aumenta en una cantidad constante como en la Ilustración 4.11.

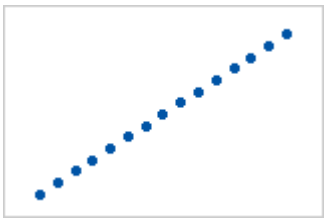


Ilustración 4.11 Spearman = +1

Si la relación es que una variable aumenta cuando la otra aumenta, pero la cantidad no es constante, el coeficiente de Spearman todavía es igual a +1 en este caso, como en la Ilustración 4.12.

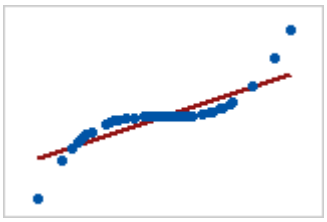


Ilustración 4.12 Spearman = +1

Cuando una relación es aleatoria o inexistente, el coeficiente de correlación es casi cero, como en la Ilustración 4.13.

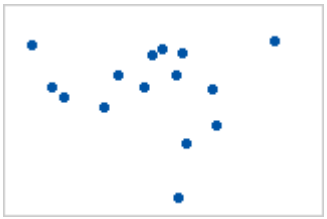


Ilustración 4.13 Spearman = −0.093

Si la relación es una línea perfecta para una relación decreciente, entonces el coeficiente de correlación es -1, como en la Ilustración 4.14.

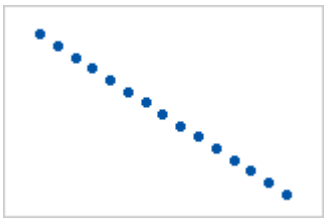


Ilustración 4.14 Spearman = −1

Si la relación es que una variable disminuye cuando la otra aumenta, pero la cantidad no es constante, entonces el coeficiente de correlación de Spearman sigue siendo igual a -1 en este caso, com en la Ilustración 4.15.

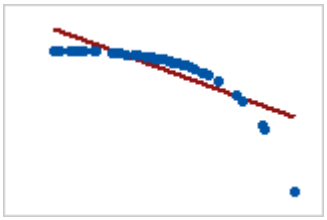


Ilustración 4.15 Spearman = −1

Los valores de correlación de -1 o 1 implican una relación lineal exacta, como la que existe entre el radio y la circunferencia de un círculo. Sin embargo, el valor real de los valores de correlación está en cuantificar relaciones imperfectas. Encontrar que dos variables están correlacionadas a menudo conlleva realizar un análisis de regresión para describir más de este tipo de relación.

- Eliminación de características hacia atrás (*Backward Feature Elimination* en inglés (Knime, 2015)), en esta técnica, en una iteración dada, el algoritmo de clasificación seleccionado se entrena en n características de entrada. Luego eliminamos una función de entrada a la vez y entrenamos el mismo modelo n veces en las características de entrada n. La función de entrada cuya eliminación ha producido el aumento más pequeño en la tasa de error se elimina, dejándonos con características de entrada n-1. La clasificación se repite utilizando las características n-2, y así sucesivamente. Cada iteración k produce un modelo entrenado en las características n-k y una tasa de error e (k). Al seleccionar la tasa

de error tolerable máxima, definimos el número más pequeño de características necesarias para alcanzar ese rendimiento de clasificación con el algoritmo de aprendizaje automático seleccionado.

4.3. Lesiones y analítica de deportes

En 2016, Kampakis decía (Kampakis, 2016):

“Las lesiones son comunes en todos los niveles del fútbol con muchas lesiones que ocurren dentro del juego profesional. Si bien el análisis deportivo es un campo de creciente popularidad, el problema de predecir lesiones en el fútbol ha evadido en gran parte a la comunidad de análisis deportivo.”

En cuanto a la primera declaración no hay duda, pero tomaremos la segunda declaración como base para analizar si hay coherencia, utilizando para ello Internet como herramienta de análisis. Para ello seleccionamos la herramienta de "google trends" y la herramienta de estadísticas de “wikipedia” para medir la audiencia de términos como analítica de los deportes, analítica del fútbol, previsión de partidos y predicción de lesiones, así como vistas de páginas sobre lesiones deportivas y analítica de los deportes.

Al analizar la Ilustración 4.16, podemos ver que el interés por la analítica de los deportes está realmente creciendo en popularidad, pero cuando el término analítico está ligado al fútbol, el interés permanece estabilizado.

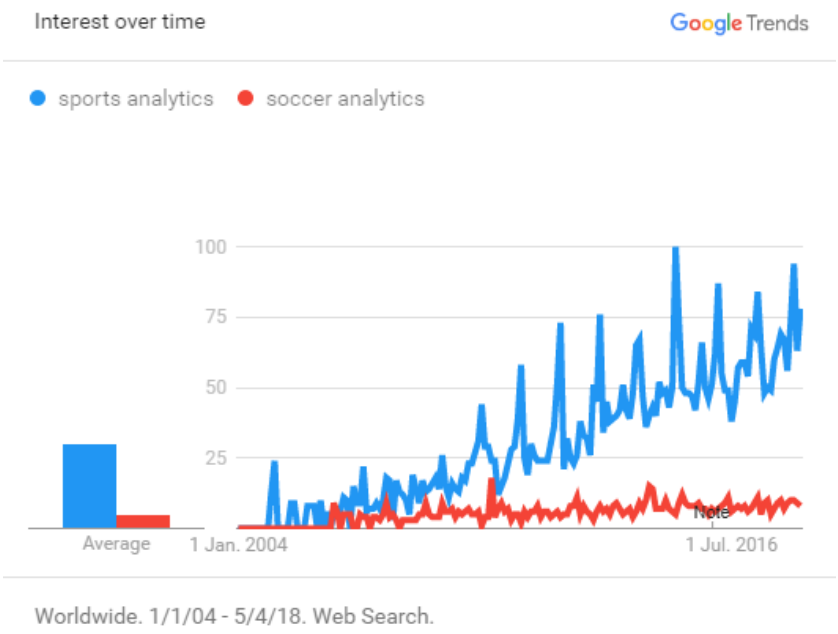


Ilustración 4.16 Frecuencia de términos buscados en Internet.

Por otro lado, el análisis de la Ilustración 4.17, muestra que el interés por la predicción de partidos crece considerablemente, mientras que la predicción de lesiones mantiene un bajo interés, demostrando el potencial de ser explotado.

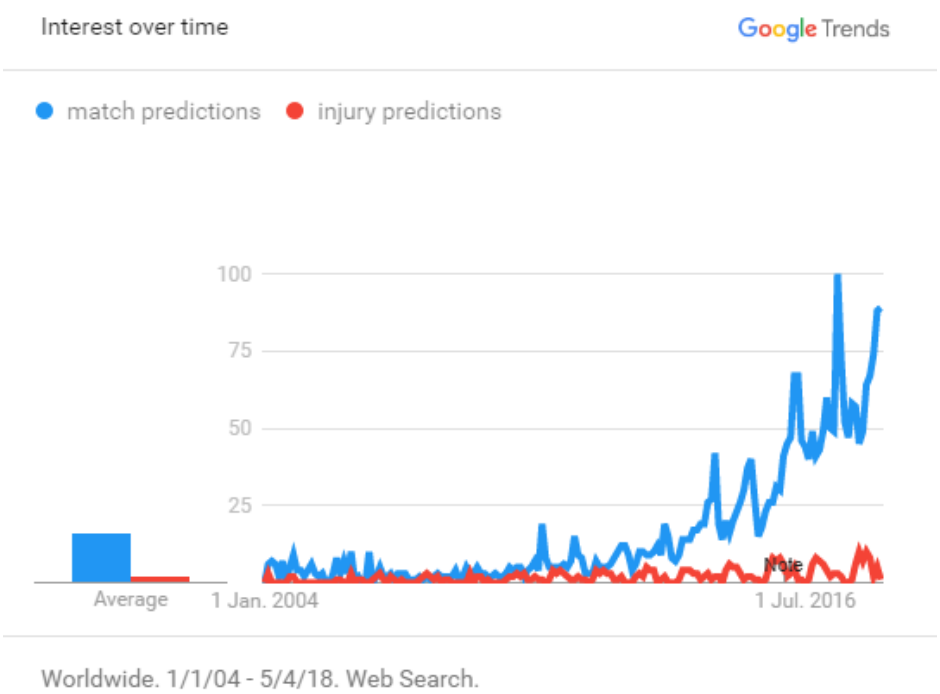


Ilustración 4.17 Frecuencia de términos buscados en Internet.

Por último, en la Ilustración 4.18, podemos ver un crecimiento similar en visitas a páginas relacionadas con la analítica de los deportes, y una alta tasa de visitas a páginas de lesiones deportivas, pero manteniendo un bajo crecimiento.

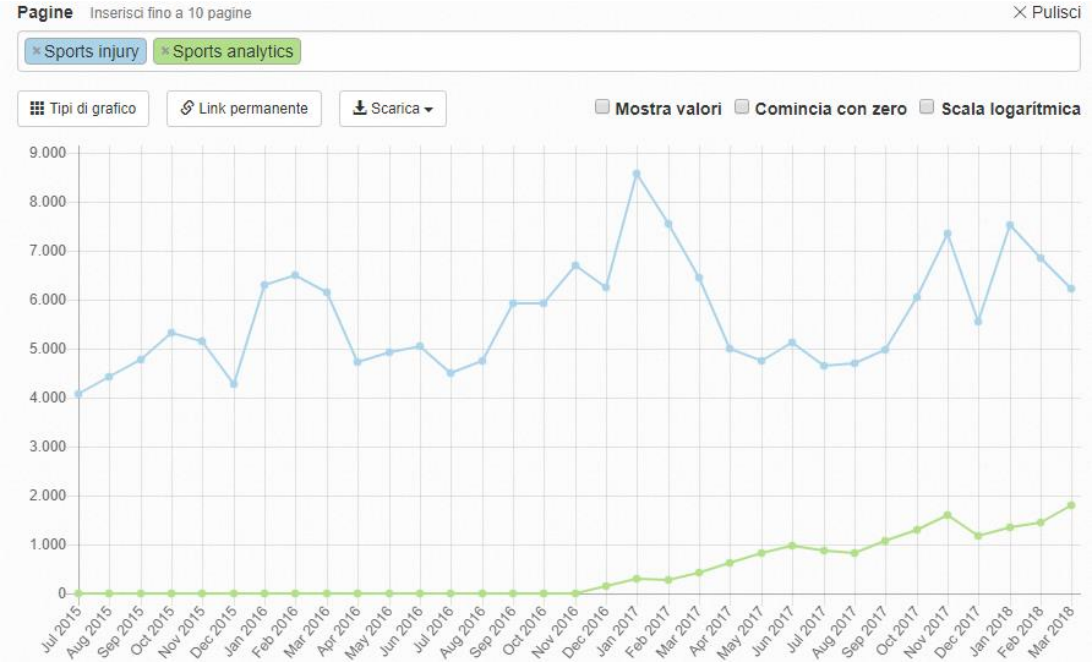


Ilustración 4.18 Frecuencia visualizaciones de páginas en Wikipedia.

4.3.1. Investigación general sobre lesiones de fútbol

Hägglund & Waldén en 2016 publicaron un estudio (Hägglund & Waldén, 2016) a partir de una gran base de datos que comprende más de 100 clubes de elite de 20 países diferentes, con más de 14,000 lesiones registradas, donde identificaron que la tasa de lesión en la elite del fútbol profesional es de 6-9 lesiones por 1.000 horas de juego, siendo mucho más alta durante los partidos (24-30 lesiones / 1.000 horas de juego) que en los entrenamientos (3-5 lesiones / 1.000 horas de entrenamiento).

En términos económicos, los costos de las lesiones deportivas pueden ser divididos en costos directos -el costo del tratamiento médico- y costos indirectos -que pueden ser desde la pérdida de productividad debido a la ausencia de un atleta en un juego, o si se trata de una estrella puede afectar la calidad del juego, haciéndolo menos espectacular y manteniendo a los aficionados lejos del estadio-. La ausencia de un atleta durante partidos importantes puede afectar las posibilidades de victoria del equipo, lo que también puede reducir las ganancias. Además, el historial de lesiones de un atleta también se toma en cuenta durante la ventana de transferencias y podría reducir considerablemente el valor del jugador, o incluso vetarlo. Por lo tanto, el costo de la lesión puede ser mayor que el costo de

simplemente tratar y rehabilitar al jugador. Abotel dice que el costo de las lesiones de los jugadores es asombrosamente alto (Abotel, 2015).

"El costo promedio estimado de las lesiones de los jugadores en las 4 mejores ligas de fútbol profesional en 2015 fue de $ 12.4 millones por equipo. Se estima que cada año los equipos de fútbol pierden un equivalente del 10% -30% de la nómina de jugadores por lesiones."

Esto significa que incluso una pequeña reducción en el número de lesiones podría sumar grandes ahorros. Entonces, ¿se imagina que la mitad de todas las lesiones deportivas profesionales pudieran prevenirse? Los ahorros solo para las ligas masculinas serían más de mil millones de dólares al año.

Pasando del campo económico al médico, las lesiones musculares (representan el 37% de todas las lesiones), especialmente las isquiotibiales y de la ingle, son las más comunes en el fútbol moderno y representan un verdadero desafío para los médicos que trabajan en el campo. Esguinces de articulaciones / ligamentos, predominantemente en el tobillo y rodilla, también son frecuentes y pueden causar una ausencia prolongada de entrenamiento y juegos. Los cuatro grandes grupos musculares de la extremidad inferior (aductores, isquiotibiales, cuádriceps y pantorrilla) comprenden más del 90% de todas las lesiones musculares en el fútbol profesional (Hägglund & Waldén, 2016). Además (Mueller-Wohlfahrt & et al., 2012) indicaron que el 96% de todas las lesiones musculares en fútbol ocurre en situaciones sin contacto, donde la mayoría ocurren durante el juego. Dentro de estos porcentajes, aproximadamente el 25% de las lesiones son graves, según (Hägglund & Waldén, 2016), con una ausencia de más de cuatro semanas. Por otro lado, un porcentaje de las lesiones (9% -34%) puede atribuirse al uso excesivo (Nielsen & Yde, 1989). En el contexto del entrenamiento, en la última década, se demostró que cualquier lesión que potencialmente podría considerarse "relacionada con la carga de entrenamiento" se considera comúnmente como "prevenible" (Gabbett, 2016).

Muchas investigaciones han estudiado qué factores pueden influir en las lesiones. Por ejemplo, (Dallinga, Benjaminse, & Lemmink, 2012) estudiaron qué herramientas de detección se pueden usar para predecir lesiones en las extremidades inferiores en deportes de equipo. Descubrieron que las herramientas

de detección que se concentraban en la funcionalidad muscular y articular podían ser predictivas.

De la misma manera, los factores psicológicos pueden afectar la propensión de un jugador a sufrir lesiones (Junge, 2000), algo que también podría afectar al fútbol (Junge, A; Dvorak, J; Rösch, D, 2000). (Johnson & Ivarsson, 2011) trataron de identificar los factores psicológicos que pueden predecir lesiones en jugadores de fútbol jóvenes y descubrieron una estructura que puede explicar el 23% de la ocurrencia de lesiones. Resultados similares se mantienen para los jugadores de fútbol sénior. Johnson e Ivarsson (2010) informan que los factores psicológicos podrían explicar el 14.6% de la ocurrencia de lesiones. Recientemente (Laux & et al., 2015) examinaron la contribución de las variables de estrés y recuperación del RESTQ-Sport para el riesgo de lesiones en jugadores profesionales de fútbol. Las escalas relacionadas con el estrés: fatiga, perturbación en los intervalos, lesión corporal y calidad del sueño predijeron significativamente lesiones que los atletas sufrieron en el mes después de la evaluación.

4.3.2. Investigación actual sobre predicción de lesiones utilizando aprendizaje automático.

Ante los resultados alentadores alcanzados por diversas investigaciones en el campo de las lesiones, haciendo su prevención posible a través de su previsibilidad mediante el análisis de sus factores de riesgo, científicos del deporte e investigadores observaron la oportunidad de utilizar el poder de la inteligencia artificial, en particular del aprendizaje automático, para ayudarlos en esta compleja tarea. Una de las primeras investigaciones que intentan predecir lesiones utilizando el proceso de aprendizaje automático fueron conducidas por (Talukder, y otros, 2016). Crearon un modelo capaz de predecir con precisión cuándo es más probable que un jugador de la NBA se lesione. Este proceso de aprendizaje automático ha permitido a la comisión técnica identificar el mejor momento para que un equipo de descanso a los jugadores y, en consecuencia, reducir el riesgo de lesiones largas.

Rossi en su trabajo de investigación (Rossi, 2017) analizó datos recogidos de GPS de 23 atletas profesionales de fútbol durante 80 sesiones de entrenamiento. Utilizando árboles de decisión como algoritmo de aprendizaje automático,

consiguiendo un acierto del 60,9% al prever lesiones. Los atributos identificados por él como los más importantes al predecir lesiones fueron el número total de aceleraciones (sprints) por encima de 2 y 3 $m \cdot s^{-2}$ y la distancia en metros cuando la energía metabólica es superior a 25,5 W/Kg por minuto. Este enfoque de aprendizaje automático ha permitido a los equipos de fútbol identificar cuándo deben prestar más atención sus jugadores durante los entrenamientos y los partidos para reducir el riesgo de lesiones, al tiempo que mejora la estrategia del equipo.

En otro trabajo de investigación (Kampakis, 2016), analizó datos de GPS de entrenamientos de jugadores del equipo principal de un club de la primera liga inglesa, e identificó qué variables estaban más correlacionadas con lesiones utilizando diversos algoritmos y técnicas, entre las cuales descatan el análisis supervisado de componentes principales (*Supervised PCA* en inglés) como el mejor método entre todos los probados. Los resultados reforzaron la idea de que los datos recogidos por los GPS durante los entrenamientos pueden ser utilizados para predecir lesiones. Entre las 69 variables utilizadas, fue posible reducir unas pocas fuertemente correlacionadas con las lesiones, como el impacto y la carga de estrés, ambas en la zona 6, que representan carreras a alta velocidad (*sprints*).

Por lo tanto, de acuerdo con los resultados encontrados en estudios previos, es lícito suponer que el análisis de un gran número de variables (es decir, la carga de los jugadores durante el entrenamiento y el partido y los datos derivados del rendimiento) podría mejorar la precisión para predecir lesiones en deportes de equipo. Las futuras investigaciones deberían programarse sobre la base de estas especulaciones.

Actualmente no es posible recopilar datos de GPS durante los partidos con el propósito de informar a la comisión técnica sobre la situación física de sus jugadores en el campo. Como se ha comentado antes, más del 90% de las lesiones ocurren durante los partidos, es decir, que situaciones como ésta limitan el desempeño que un modelo predictivo puede alcanzar para ese problema.

Pero IFAB (*Internation Football Association Board* en inglés) y la FIFA ya se están moviendo para permitir que los jugadores utilicen durante los partidos este tipo de dispositivos, como el GPS, a través de su programa de calidad (*FIFA Quality*

Programme en inglés), creando normativas para que los fabricantes de sistemas electrónicos de rastreo y rendimiento (*electronic performance and tracking systems (EPTS)* en inglés) sean certificados, de la misma forma que se hizo con otras tecnologías como la tecnología de la línea de gol (*goal-line technology (GLT*) en inglés) y árbitro asistente de vídeo (*video assistant referee (VAR)* en inglés).

Este nuevo proyecto en FIFA abarca la gestión de la creación de un centro de datos global para datos de rendimiento de jugadores y equipos, institutos de investigación acompañantes en el área de validación y estandarización de datos, así como la interacción regular con la industria de *wearables* y sistemas de seguimiento (FIFA, 2018). Sin embargo, la decisión se basó en dos condiciones: que, hasta que el EPTS demuestre que trae beneficios médicos evidentes, sus datos no podrán ser utilizados en tiempo real en el área técnica, y que estos sistemas no pongan en peligro a nadie en el campo de juego durante el partido. De hecho, el fútbol se está moviendo hacia la era moderna, le guste o no.

5. PREDICCIÓN DE LESIONES INTRÍNSECAS EN EL FÚTBOL PROFESIONAL UTILIZANDO MEDICIONES GPS Y REGISTROS DE EXPOSICIÓN EN ENTRENAMIENTO

La tecnología GPS portátil puede proporcionar una gran cantidad de información sobre el rendimiento de un atleta. Esta información podría usarse para descubrir signos tempranos de fatiga o sobreentrenamiento que pueden provocar lesiones. Esta investigación compara varios métodos (Random Forest (Ho, 1995), AdaBoost (Tsai, Hsu, & Yen, 2014), FURIA (Huehn & Huellermeier, 2009), RPROP (Riedmiller & Braun, 1992) y PART (Frank & Wittenx, 1998)) para predecir lesiones intrínsecas a partir de datos de GPS. La técnica *Backward Feature Elimination (BFE)* funciona bastante bien en la tarea de reducir la dimensión en los dos enfoques, logrando una estadística MCC media de 0,39 $\pm 0{,}02\ (SEM)$ para el enfoque A (sin filtro previo) y 0,43 $\pm 0{,}02\ (SEM)$ para el enfoque B (con filtro previo – prueba de hipótesis de Wilcoxon) considerando los 5 mejores resultados de la Ilustración 5.33 y Ilustración 5.34, respectivamente. Las técnicas de reducción de la dimensionalidad nos ayudan a extraer las características que más se correlacionan con las lesiones

reduciendo al mismo tiempo el gran número de variables, lo que mejora considerablemente el tiempo de ejecución de los modelos.

5.1. Diseño y métodos

Para planificar este trabajo de investigación, recogemos diferentes enfoques (CRISP-DM, IBM, Microsoft, Blitzstein & Pfister's, OSEMN) para el proceso de análisis de datos y definimos nuestro propio modelo ideal de análisis que se puede observar en la Ilustración 5.1.

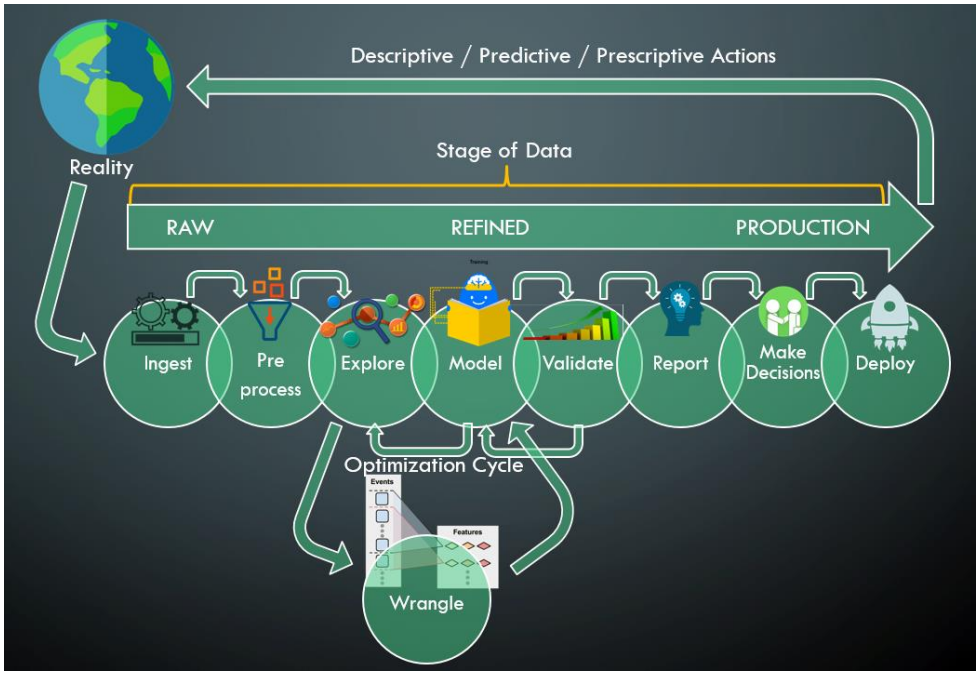


Ilustración 5.1 Representación del modelo ideal para el proceso de ciencia de datos.

Los algoritmos fueron elegidos para esta investigación después de inferir el conjunto de datos, verificando: las características de las variables o atributos (clase o objetivo, categórica, discreta o continua, texto, booleana o numérica), y principalmente su fuerte desequilibrio entre clases. Sabiendo esto se realizaron investigaciones dirigidas a identificar los algoritmos más indicados en este escenario. Los enfoques A y B mencionados en el objetivo de este trabajo, siguen el flujo idéntico con la única diferencia en la conexión entre la tarea de preprocesamiento y reducción de la dimensionalidad, donde para el enfoque B aplicamos un filtro teniendo como base la prueba de hipótesis de Wilcoxon Mann-

Whitney U. Esta diferencia será ilustrada en el momento del paso entre estas dos tareas dentro de la fase de preprocesamiento (5.1.2). En los próximos subcapítulos explicaremos cada etapa del proceso como se muestra en la Ilustración 5.1. De estas utilizamos en este trabajo experimental sólo: recolección, preprocesamiento, análisis exploratorio de datos, modelado, validación e informes. En virtud del club estar participando en competiciones importantes a nivel nacional y regional, las etapas de toma de decisión e implementación fueron inviabilizadas en este trabajo por exigir su participación en el proceso.

5.1.1. Fase de recolección

Como se mencionó en los capítulos anteriores, los datos fueron recolectados de sistemas electrónicos de rastreo y rendimiento (EPTS), en nuestro caso un sistema de posicionamiento global (GPS), y luego compilados por un *software* y transferidos a hojas de cálculo en formato Excel (“.xls”). A partir de estas planillas realizamos un trabajo de análisis previo para verificar la calidad de los datos antes de importarlos en la herramienta; es decir, los datos pueden estar en formatos o estructuras incorrectas para la importación, como podemos percibir en la Ilustración 5.2, por lo que puede ser necesario realizar transformaciones, conversiones, filtros u otros tipos de tratamientos, llegando al formato ideal como en la Ilustración 5.3.

01/09 - Tarde

Titulares Fares Lopes: Trabalho de Posse de Bola (3 x 5 min) + Bola Parada Ofensiva e Defensiva (10 min).
Os Demais: Trabalho de Posse de Bola (3 x 5 min) + Trabalho Físico Técnico Específico (15 min) + Mini Jogo (2 x 10 min) + Bola Parada (2 x 5 min).
Total: 80 min.

		HIDRATAÇÃO			ESCALAS				VOLUME	GPS			FC			
NOME	DM / TRANSIÇÃO	PRÉ	PÓS	DIF	PSR	PSE	DOR		TOTAL	SPRINTS	VEL MÉDIA	DISTÂNCIA TOTAL	BPM - MÁX	MÉD - BPM	% MÉDIA	
		98	97,4	0,61												
		86,3		100,00												
		32,05		100,00												
	TRANSIÇÃO	69		100,00	4		4	Joelho Esquerdo					202			
		89,25		100,00	3		0		30	0	2,91	2382	189	126	67	
		75,7	74,9	1,06	5	7	5		80	9	3,54	5004	192	141	73	
			72,8	#DIV/0!		5			70				201	138	69	
				#DIV/0!		6			80	4	3,19	4300	183	128	70	
		86,4		100,00	5		3	Anteriores coxa	30	0	2,37	1637	192	119	62	
		76,2		100,00	3		5	Posteriores parte distal	30				203		0	
		75,4		100,00	7		4	Posteriores e gluteos	30	2	2,66	2049	191	134	70	
		75,75		100,00					30	0	2,70	1862	193	118	61	
		72,5	72,4	0,14	4	6	0		70	10	4,68	4437	186		0	
	TRANSIÇÃO	64		100,00					35				199		0	
		68,75	67,8	1,38	3	4	3		80	11	3,94	5410	209	156	75	
		87,3	86,5	0,92	4	6	4	Gluteos e Posteriores	80	1	2,63	3527	188	133	71	
	LIBERADO - ASSUNTOS PARTICULARES			#DIV/0!									166		0	

Ilustración 5.2 Hoja de cálculo en formato inadecuado para la importación.

1		ID COLUNA ---------->	1	2	3	4	5	6	7	8	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
2	DIA	NOME	EDA	LES	PPRE	PPOS	DIF	PSR	PSE	DOL	VTR	SPR	VEM	DIT	FCMAX	FCMED	PMFC	Z5	Z4	Z3	Z2	Z1	TRIMP	UA	DES	NASC	POS	MJUG	PREF	ALT	IMC	PART	MCM	PGC	LPA	PREL
3	1	[redacted]	23	0	98	97,4	0,61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28/09/1995	POR	180	93	195	24	0	80,2	18,2%	0	0
4	1	[redacted]	28	0	86,3	0	100,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22/07/1990	POR	3510	86	192	22	0	74,1	14,2%	0	0
5	1	[redacted]	35	0	92,1	0	100,00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25/11/1983	POR	990	82	189	22	0	75,8	17,7%	0	0
6	1	[redacted]	31	0	69	0	100,00	4	0	4	0	0	0	0	202	0	0	0	0	0	0	0	0	0	0	29/04/1987	ATA	987	67	167	20	0	58,9	14,6%	0	0
7	1	[redacted]	20	0	89,3	0	100,00	3	0	0	30	0	2,9	2382	189	126	66,7	0	4	7	12	9	70	0	0	25/04/1998	ATA	459	86	186	23	0	73,6	17,5%	0	0
8	1	[redacted]	22	0	75,7	74,9	1,06	5	7	5	80	9	3,5	5004	192	141	73,4	6	17	13	9	9	164	560	0	22/01/1996	ATA	830	75	168	22	0	62,2	17,8%	0	0
9	1	[redacted]	21	0	0	72,8	0,00	0	5	0	70	0	0	0	201	138	68,7	1	8	4	7	10	73	350	0	19/06/1997	ATA	162	74,7	182	21	0	65,9	11,8%	0	0
10	1	[redacted]	33	0	0	0	0,00	0	6	0	80	4	3,2	4300	183	128	69,9	0	10	20	30	5	165	480	0	01/08/1985	ATA	1700	84	185	23	0	71	15,5%	0	0
11	1	[redacted]	30	0	86,4	0	100,00	5	0	3	30	0	2,4	1637	192	119	62	0	1	6	8	15	53	0	0	20/01/1988	MEI	1229	77	185	21	0	72	16,6%	0	0
12	1	[redacted]	23	0	76,2	0	100,00	3	0	5	30	0	0	0	203	0	0	0	0	0	0	0	0	0	0	16/07/1995	MEI	838	69	183	19	0	66,9	12,2%	0	0
13	1	[redacted]	31	0	75,4	0	100,00	7	0	4	30	2	2,7	2049	191	134	70,2	1	6	7	11	6	78	0	0	21/05/1987	VOL	530	72	177	20	0	64,8	14,1%	0	0
14	1	[redacted]	37	0	75,8	0	100,00	0	0	0	30	0	2,7	1862	193	118	61,1	0	0	6	7	6	38	0	0	24/06/1981	VOL	183	71	177	20	0	64,9	14,3%	0	0
15	1	[redacted]	23	0	72,5	72,4	0,14	4	6	0	70	10	4,7	4437	186	0	0	0	0	0	0	0	0	420	0	10/05/1995	ATA	1781	65	172	19	0	62	14,5%	0	0
16	1	[redacted]	28	0	64	0	100,00	0	0	0	35	0	0	0	199	0	0	0	0	0	0	0	0	0	0	09/02/1990	MEI	1534	62	167	19	0	56,7	11,4%	0	0
17	1	[redacted]	22	0	68,8	67,8	1,38	3	4	3	80	11	3,9	5410	209	156	74,6	2	29	14	8	6	190	320	0	11/06/1996	MEI	505	73	181	20	0	62,9	8,4%	0	0
18	1	[redacted]	30	0	87,3	86,5	0,92	4	6	4	80	1	2,6	3527	188	133	70,7	0	16	21	27	10	191	480	0	14/09/1988	ZAG	2376	84	190	22	0	73,9	15,3%	1	0
19	1	[redacted]	42	0	0	0	0,00	0	0	0	0	0	0	0	166	0	0	0	0	0	0	0	0	0	0	13/01/1976	ATA	2465	70	176	20	0	62	11,4%	1	0
20	1	[redacted]	31	0	0	0	0,00	0	7	0	80	4	4,5	4599	188	0	0	0	0	0	0	0	0	560	0	20/03/1987	MEI	1645	80	181	22	0	68	15,0%	0	0
21	1	[redacted]	118	1	0	0	0,00	0	0	0	0	0	0	0	200	0	0	0	0	0	0	0	0	0	2	0	VOL	0	0	0	0	0	-21,2	0,0%	0	0
22	1	[redacted]	30	1	0	0	0,00	0	0	0	0	0	0	0	208	0	0	0	0	0	0	0	0	0	2	23/01/1988	LAD	1321	65	172	19	0	58,6	9,8%	0	0
23	1	[redacted]	26	0	79,4	0	100,00	3	0	0	30	0	2,2	1643	196	114	58,2	0	2	4	5	10	40	0	0	18/07/1992	ZAG	888	72	175	21	0	66	16,9%	0	0
24	1	[redacted]	31	0	81,4	81,85	-0,55	5	6	4	80	2	3	3972	184	155	84,2	25	25	24	5	0	307	480	0	23/08/1987	ATA	165	79	180	22	0	68,3	16,1%	0	0

Ilustración 5.3 Hoja de cálculo en formato adecuado para la importación.

En la Ilustración 5.4, podemos ver también la importación de los datos y otros tratamientos realizados, como la agrupación de los conjuntos de datos convirtiendo los datos de sesiones diarias en periodicidad semanal. La agregación durante semanas fue elegida por dos razones. Primero, una semana es una división natural para la programación en el fútbol. Cada semana se caracteriza por al menos un partido, y el programa de entrenamiento se diseña semanalmente. En segundo lugar, otras divisiones naturales serían diarias, o sumando semanas (por ejemplo, las últimas dos semanas). El problema con las divisiones diarias (es decir, cada fila en el conjunto de datos representaría un día de capacitación) es doble. Primero, el número de observaciones o líneas sería demasiado grande, pero el número de lesiones aún sería pequeño. En segundo lugar, es poco probable que los datos GPS de un solo día sean indicativos de fatiga o sobreentrenamiento. Es más probable que un promedio de al menos unos pocos días sea más informativo.

Se usó la media como resumen de agregación para agregar las instancias a lo largo de la semana.

El conjunto de datos final consiste en 4559 observaciones y 36 variables, más la variable de respuesta. De las 4559 observaciones, hubo 454 casos (9,96%) de lesiones (correspondientes a las semanas en que el jugador resultó lesionado).

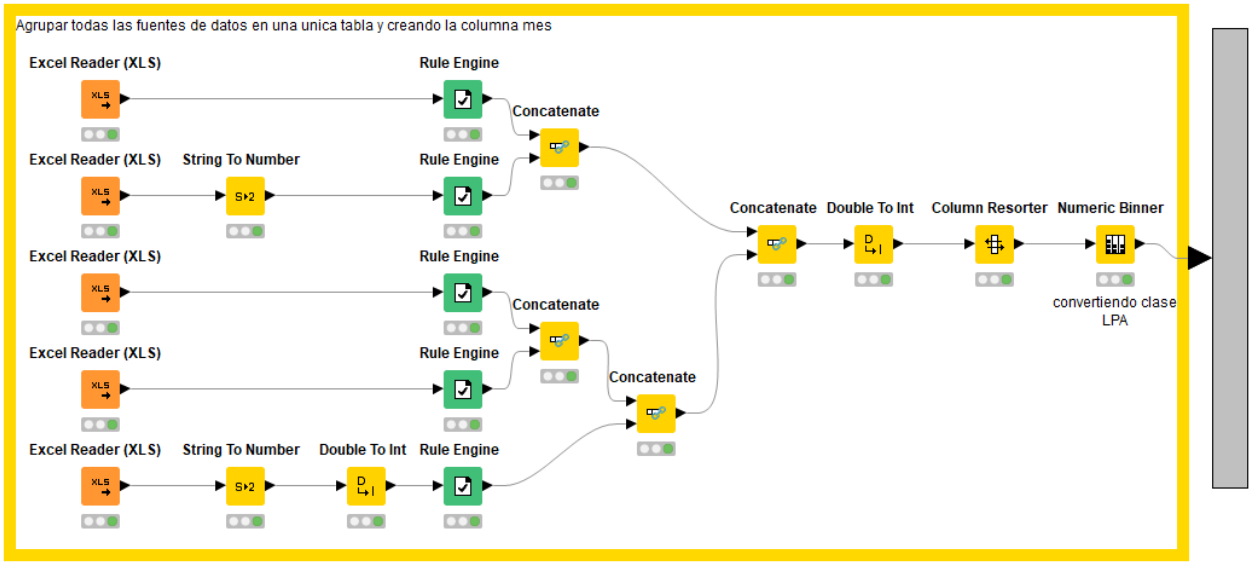


Ilustración 5.4 Flujo del proceso de importación

En la tarea de evaluar la calidad del *dataset*, como explicamos en el capítulo 4 y puede observarse en la Ilustración 5.5, fue posible continuar con el proceso, pues la tasa de error obtenida fue < 1, como puede ver en la Ilustración 5.6.

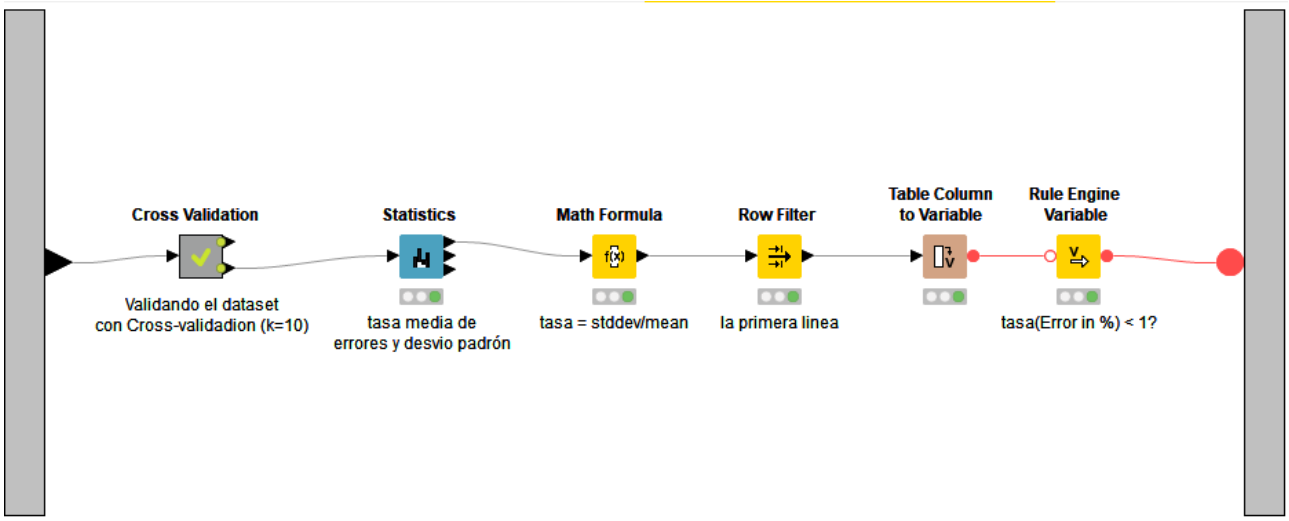


Ilustración 5.5 Flujo de la tarea de evaluación de calidad del conjunto de datos.

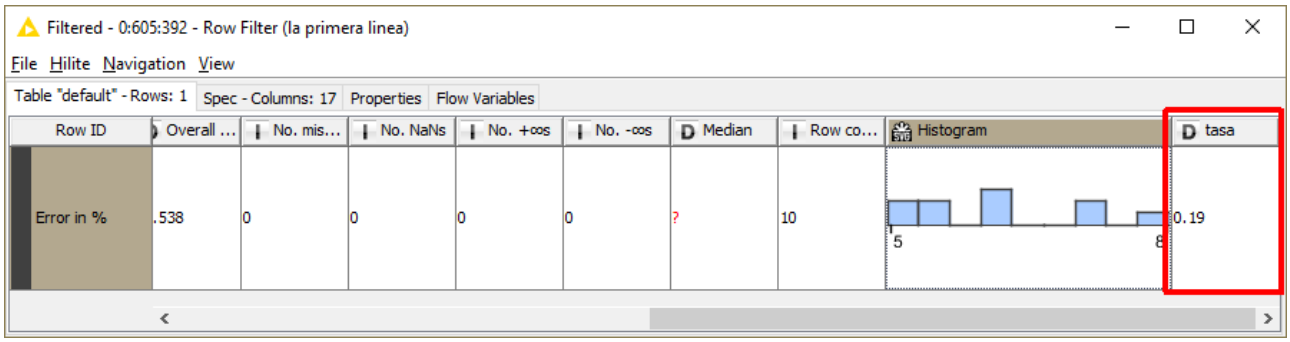


Ilustración 5.6 Resultado de la evaluación de calidad del conjunto de datos.

5.1.2. Fase de preprocesamiento

En la Ilustración 5.34 vemos el proceso de preprocesamiento explicado en el capítulo 4, donde primero quitamos jugadores y posiciones que no contienen valores para las sesiones de entrenamiento, como el caso de los porteros y de otros tres jugadores, y nos quedamos con 3774 observaciones con 358 casos de lesiones (9,49%).

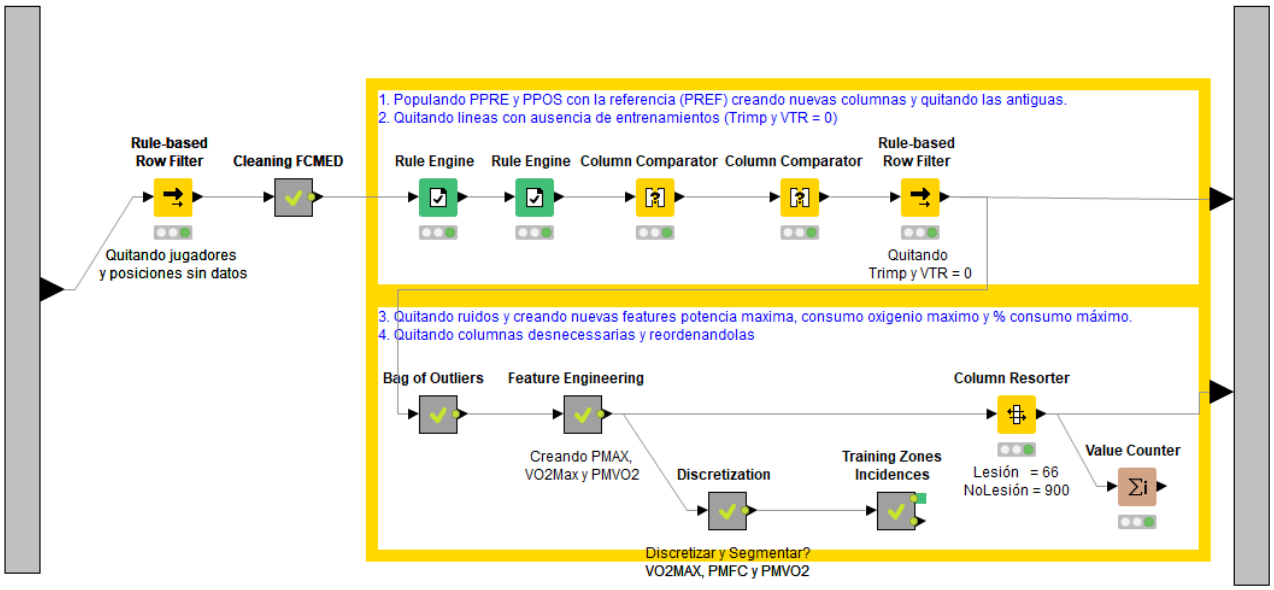


Ilustración 5.7 Resultado de la evaluación de calidad del conjunto de datos.

La próxima tarea del proceso, fue identificar valores perdidos, donde para la variable FCMED buscamos rellenar con la media de cada jugador en las observaciones en que hay valor para FCMED, para la variable peso, como PPRE (peso pre-sesión) y PPOS (peso pos-sesión) también rellenamos tomando en cuenta el peso de referencia del jugador, para las variables TRIMP y VTR quitamos las observaciones donde hay valores 0 para ambas, pues implican en el jugador no ha

realizado ejercicios en la sesión y esto significa también quitar jugadores que continúan lesionados. Al final nos quedamos con 2355 observaciones con 174 casos de lesiones (7,39%).

En la próxima tarea aplicamos la eliminación de ruidos, donde analizamos todas las variables buscando quitar observaciones donde la media de la desviación estándar sea mayor o menor que 2, al final nos quedamos con 966 observaciones con 66 casos de lesiones (6,83%) y 28 variables.

En la última tarea realizamos la ingeniería de las características creando las variables PMAX (2), VO2MAX (3) y PMVO2 (4). A continuación presentamos las principales formulas utilizadas para obtener los valores para estas nuevas variables.

Fórmula de la potencia máxima (PMAX), según (Walker, Halliday, & Resnick, 2012):

$$Pmax = F\ x\ V \quad \textbf{(2)}$$

Donde: $F = fuerza$
$V = velocidad$

Fórmula del consumo de oxígeno máximo (VO2MAX), según estimación de (Uth, Sørensen, Overgaard, & Pedersen, 2005) que dice que la regla de conversión se basó solamente en las mediciones en hombres bien entrenados de entre 21 y 51 años. También informaron que la fórmula es más confiable cuando se basa en la medición real de la frecuencia cardíaca máxima.

$$VO_2max = \left(\frac{HRmax}{HRrest}\right) \times 15.3\ \frac{mL}{Kg.minute} \quad \textbf{(3)}$$

Donde: $HRmax = frecuencia\ cardiaca\ maxima\ (FCmax)$
$HRrest = frecuencia\ cardiaca\ en\ reposo\ (FCrep)$

Fórmula del porcentual medio del oxígeno máximo (PMVO2), según estimación de (Swain, Abernathy, Smith, Lee, & Bunn, 1994) que dice que la ecuación de ACSM

no es capaz de predecir con precisión el VO2max en atletas y que solo los modelos de regresión se correlacionaron moderadamente con los valores realmente medidos de VO2max.

$$PMVO_2 = \left(\frac{\%HRmax - 37}{0.64}\right) \quad \textit{(4)}$$

Donde: $\%HRmax = porcentual\ de\ la\ frecuencia\ cardiaca\ máxima\ (FCmax)$

Aún dentro de la fase de preprocesamiento, aplicamos los enfoques A y B, como explicamos en los capítulos 4 y 5. El enfoque B se diferencia sólo por la prueba de hipótesis, del que presentaremos a continuación su diagrama, parámetros y resultados (Ilustración 5.8, Ilustración 5.9 y Tabla 5.1).

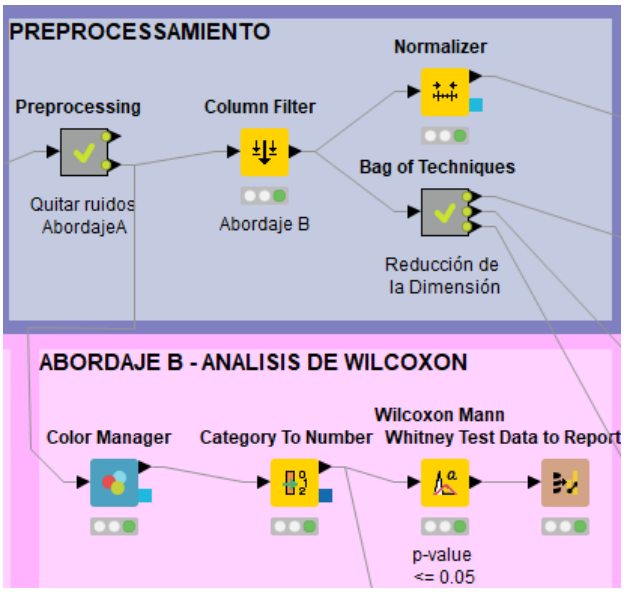


Ilustración 5.8 Diagrama con la prueba de hipótesis y filtro basado en su resultado.

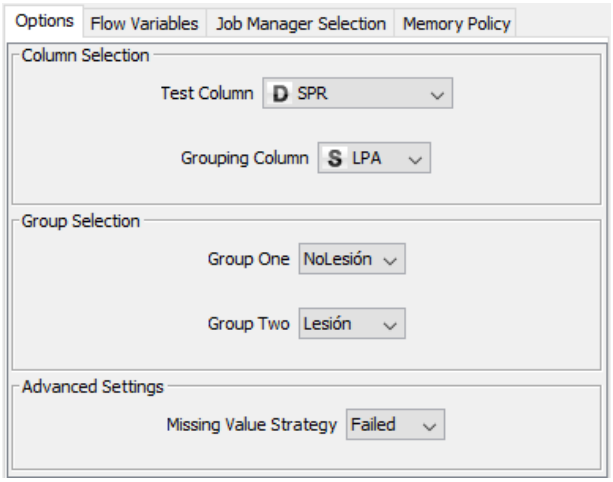


Ilustración 5.9 Parámetros de la prueba de hipótesis para una variable.

Variable	p-Value	H0 rechazado
SPR	0	Sí
VEM	0	Sí
DIT	0,005	Sí
PMAX	0	Sí
FCMED	0,072	No
PMFC	0,624	No
PMVO2	0,624	No
FCMAX	0,119	No
VO2MAX	0,119	No
Z5	0,804	No
Z4	0,506	No
Z3	0,064	No
Z2	0,02	Sí
Z1	0,052	No
TRIMP	0,129	No
UA	0,569	No
MJUG	0	Sí
ALT	0,197	No
IMC	0,012	Sí
MCM	0,638	No
PGC	0,029	No
EDA	0,101	No
DIF	0,12	No
PSR	0,699	No
PSE	0,197	No
DOL	0,571	No
VTR	0,889	No

Tabla 5.1 Resultado de la prueba de hipótesis Wilcoxon Mann-Whitney U

En el enfoque A no realizamos ningún filtro adicional en las variables, mientras que en el enfoque B filtramos las variables permaneciendo sólo aquellas

cuyo resultado presentado en la Tabla 5.1, tiene un p-Value menor o igual a 0,05, lo que significa rechazar la hipótesis nula de que las distribuciones son iguales (medianas) entre las dos muestras (clases) de la misma población.

Seguimos para la próxima tarea, que será aplicar tres técnicas para la reducción de la dimensionalidad del conjunto de datos, las cuales explicamos a continuación el resultado (Tabla 5.2 y Tabla 5.3) de la reducción obtenido en cada técnica y para cada enfoque.

Rank Correlation (RC)	Principal Component Analysis (PCA)	Backward Feature Elimination (BFE)
DIF	PCA dimension 0	EDA
PSR	PCA dimension 1	PSR
PSE	PCA dimension 2	UA
DOL	PCA dimension 3	Z4
VEM	PCA dimension 4	Z1
FCMAX	PCA dimension 5	FCMAX
Z5	PCA dimension 6	ALT
MJUG	PCA dimension 7	MJUG
PGC		PGC

Tabla 5.2 Variables seleccionadas por cada técnica en el enfoque A

Rank Correlation (RC)	Principal Component Analysis (PCA)	Backward Feature Elimination (BFE)
SPR	PCA dimension 0	DIT
Z2	PCA dimension 1	PGC
MJUG	PCA dimension 2	MJUG
IMC	PCA dimension 3	IMC
	PCA dimension 4	

Tabla 5.3 Variables seleccionadas por cada técnica en el enfoque B

Entre las tres técnicas solamente PCA que presenta como resultado componentes en lugar de variables. En la Ilustración 5.10 y Ilustración 5.11 se muestran los gráficos donde demuestra el grado de varianza de los componentes de mayor a menor, y a través de este gráfico realizamos el corte para seleccionar solamente los componentes que presentan varianza relevante; es decir, representan generalmente el 90% o más de los datos.

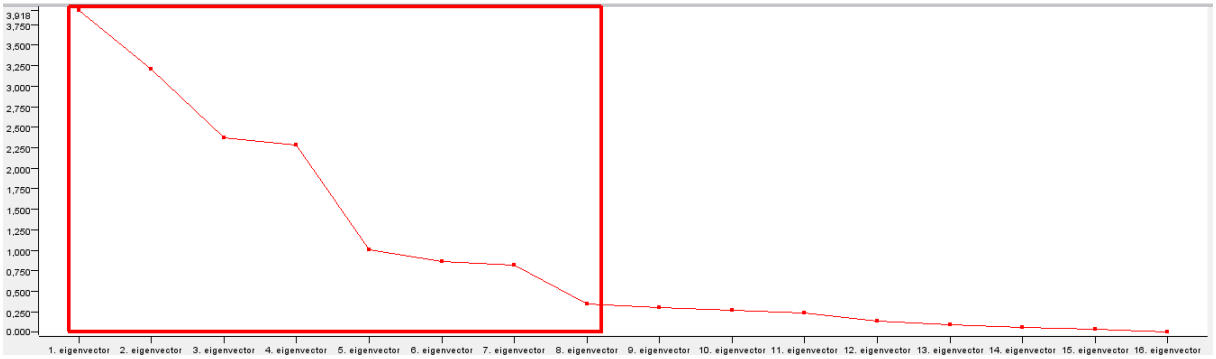


Ilustración 5.10 Grafico con la varianza de los componentes PCA en el enfoque A.

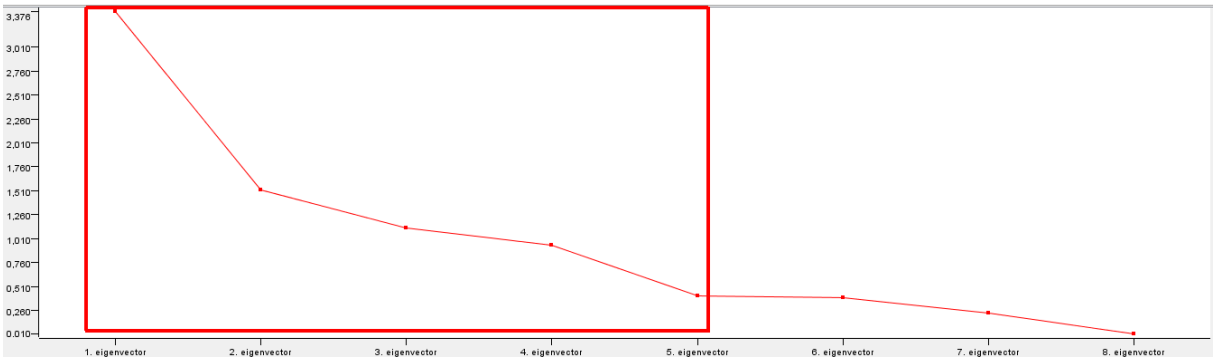


Ilustración 5.11 Grafico con la varianza de los componentes PCA en el enfoque B.

5.1.3. Fase de análisis exploratorio de datos

En estadística, el análisis exploratorio de datos (*Exploratoy Data Analysis (EDA)* en inglés) es un enfoque para analizar datos. Se puede usar o no un modelo estadístico, pero principalmente EDA es para ver lo que los datos pueden decirnos más allá de la tarea de modelado formal o prueba de hipótesis. El análisis de los datos de exploración fue promovido por John Tukey para alentar a los estadísticos a explorar los datos, y posiblemente formular hipótesis que podrían conducir a una nueva recopilación de datos y experimentos. EDA es diferente del análisis de datos inicial (IDA), que se enfoca en verificar las suposiciones requeridas para el ajuste del modelo y las pruebas de hipótesis, maneja los valores perdidos y realiza transformaciones de las variables según sea necesario. Es decir, EDA abarca IDA (Andrienko & Andrienko, 2005).

A partir de esta premisa comenzamos a analizar los datos utilizando algunas formas de visualización como gráficos de dispersión, histogramas y *box plots* con el objetivo de identificar información que pueda ser relevante respecto a la clase lesión.

Comenzamos con el gráfico de dispersión. Un diagrama de dispersión puede sugerir varios tipos de correlaciones entre variables con un cierto intervalo de confianza, siendo también muy útil cuando deseamos ver cómo dos conjuntos de datos comparables aceptan mostrar relaciones no lineales entre variables. Además, si los datos están representados por un modelo mixto de relaciones simples, estas relaciones serán visualmente evidentes como patrones superpuestos (Jarrell, 1994) (Utts, 2005). Entre todas las variables analizadas, llamó la atención la relación entre el número de minutos jugados y una lesión. Al ver la Ilustración 5.12, podemos percibir que jugadores que disputaron menos de 1.000 minutos durante la temporada tuvieron mayor incidencia de lesión. Es importante recordar que durante una temporada un jugador puede participar al menos en 38 partidos teniendo en cuenta sólo el campeonato nacional, pero tenemos en este intervalo de tiempo campeonatos regionales (16 partidos disputados) y copas (10 partidos disputados), lo que aumenta considerablemente este número (64 partidos). Como consecuencia, disminuye de la misma forma el número de minutos en que cada jugador participó por partido en la temporada. Teniendo en cuenta este número total de partidos, cada jugador lesionado habría participado una media de 15,6 minutos por partido.

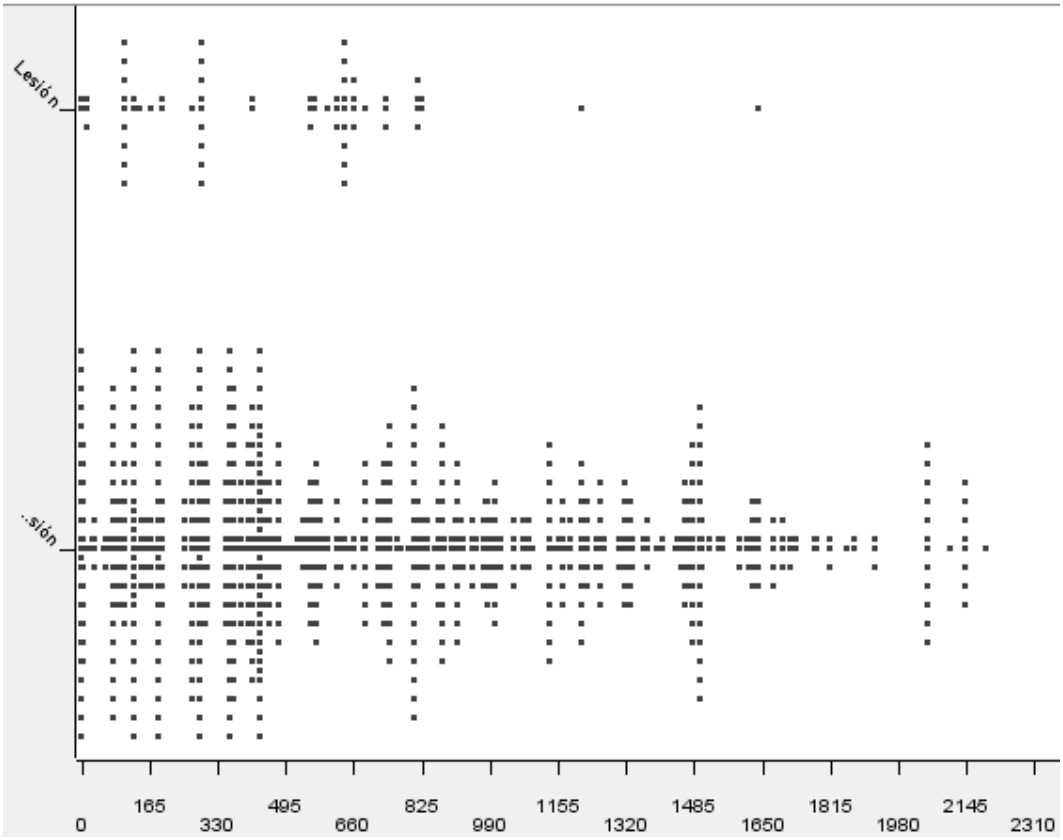


Ilustración 5.12 Scatter plot de las variables de Minutos Jugados y Lesión.

Esto nos llevó a cuestionar el porqué de este patrón. Bien, primero vamos a tomar como punto de partida las afirmaciones de diversos estudios que colocan la carga de entrenamiento como la principal responsable por las incidencias de lesiones durante o después de los mismos y así haremos un análisis un poco más profundo.

Utilizaremos el gráfico de histograma para poder analizar la relación entre 3 o más variables. Entonces, fijaremos dos variables (minutos jugados y lesión), mientras que variamos otros conjuntos de para verificar su efecto. A continuación destacamos tres análisis que apuntan cierto estándar.

En el primer análisis, reflejado en la Ilustración 5.13, elegimos las variables de GPS: distancia total recorrida (barra roja), velocidad media (barra azul) y cantidad de *sprints* (barra verde) y notamos que los jugadores lesionados (barra púrpura) que tuvieron menos de 1.000 minutos de juego (eje x), tuvieron una carga de intensidad media (eje y) similar a jugadores con una cantidad mayor de minutos jugados (eje x), pero con menor índice de lesiones (barra púrpura). En el caso de los jugadores con menor tiempo de juego (<1.000), el promedio de *sprints* (barra verde) es mayor que el promedio de aquellos con mayor tiempo de juego, revelándose un factor importantísimo en la ocurrencia de lesiones.

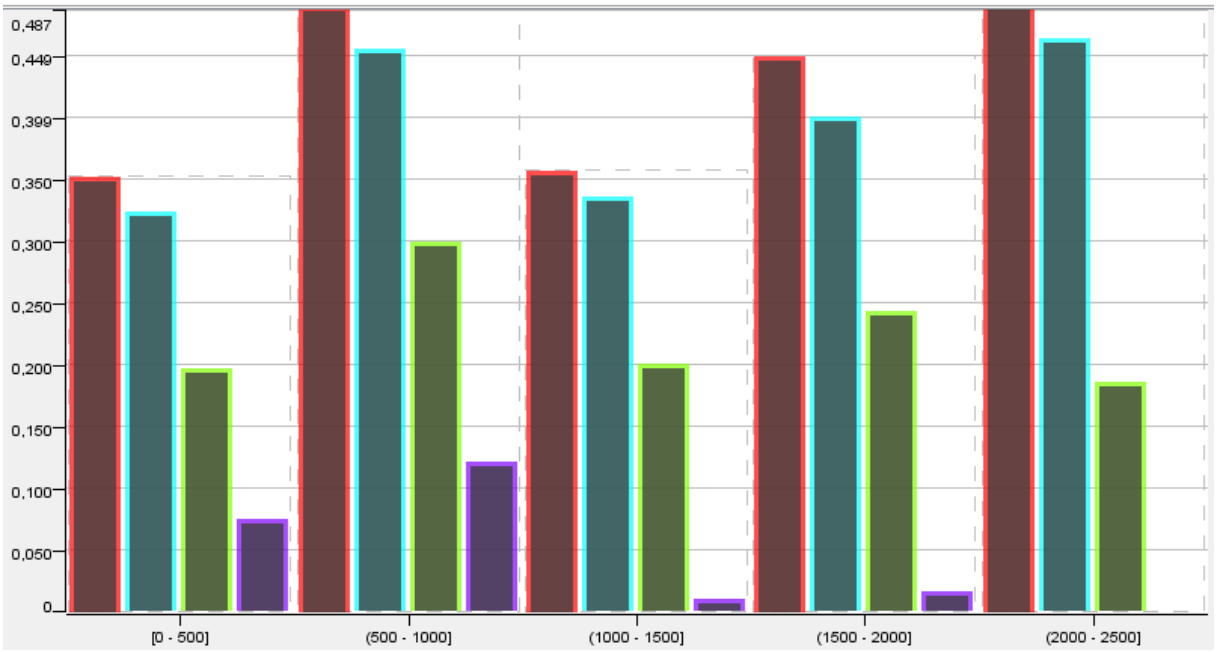


Ilustración 5.13 Frecuencia de las variables de GPS (Distancia Total, Velocidad Media y Sprints) y Lesión en relación a cantidad de minutos jugados.

Pasando al segundo análisis, tomamos como variables la carga de entrenamiento: TRIMP (*TRaining IMPulse* en inglés) y UA (unidad arbitraria).

TRIMP es una técnica que pretende integrar en un solo término o score, tanto el volumen como la intensidad de entrenamiento y es específico para entrenamiento de resistencia, ya que utiliza la frecuencia cardíaca o zonas de frecuencia cardiaca (Foster, y otros, 2001). Esta puntuación se multiplicaría por la duración en cada zona, multiplicado por el score de la fase. (1 min en la zona 1 se da score 1 (TRIMP 1 × 1), 1 min en la zona 2 se da score 2 (TRIMP 1 × 2) ....). El TRIMP score total se obtiene como la suma de los resultados de todas las zonas. Cuanto mayor el TRIMP de la actividad, mayor sería la carga fisiológica impuesta al atleta.

UA consiste en una medida utilizada para cuantificar la carga de entrenamiento y su resultado es el producto de la escala de percepción subjetiva de esfuerzo del atleta (PSE) por el volumen de entrenamiento en minutos. Así podemos determinar qué sesiones podrían estar fuera de los parámetros de entrenamiento esperados. PSE no es más que la nota proporcionada por el atleta, en una escala de 0 a 10, en relación a la dificultad de esfuerzo del entrenamiento aplicado como establecieron (Foster, Daines, Hector, Snyder, & Welsh, 1996) y se muestra en la Tabla 5.4.

Clasificación	Descriptor
0	Reposo
1	Muy fácil
2	Fácil
3	Moderado
4	Algo difícil
5	Difícil
6	-
7	Muy difícil
8	-
9	-
10	Máximo

Tabla 5.4 Escala PSE de Foster

En este segundo análisis podemos ver en la Ilustración 5.14, que tenemos una carga de entrenamiento e intensidad casi similar entre los atletas que disputaron menos de 1.000 minutos en la temporada y aquellos que disputaron entre 1.000 y 1.500 minutos, pero en carga e intensidad menores que aquellos que disputaron

más de 1.500 minutos. Esto puede sugerir que tal vez la cantidad de carga e intensidad esté poco sobredimensionada para estos atletas con menos tiempo de juego (<1.000). Sin embargo vamos a seguir con nuestro último análisis.

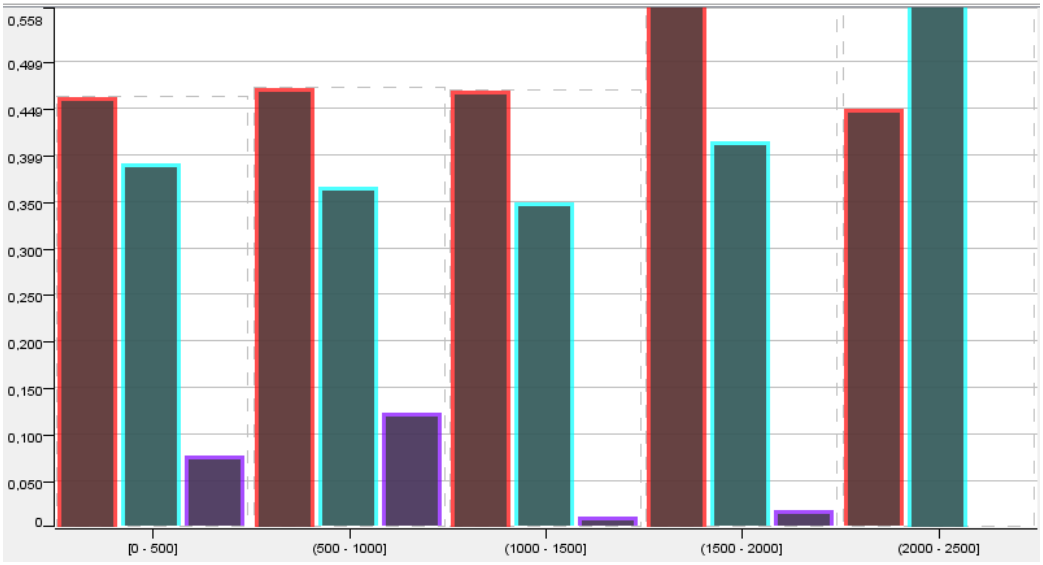


Ilustración 5.14 Frecuencia de las variables de carga de entrenamiento (TRIMP y UA) y Lesión en relación a cantidad de minutos jugados.

En este tercer y último análisis, vamos a evaluar las variables de índice de masa corporal (IMC) y de masa corporal magra (MCM). Podemos percibir en la Ilustración 5.15, que atletas con menores índices musculares y menor tiempo de partida en la temporada fueron los que presentaron mayor incidencia de lesiones.

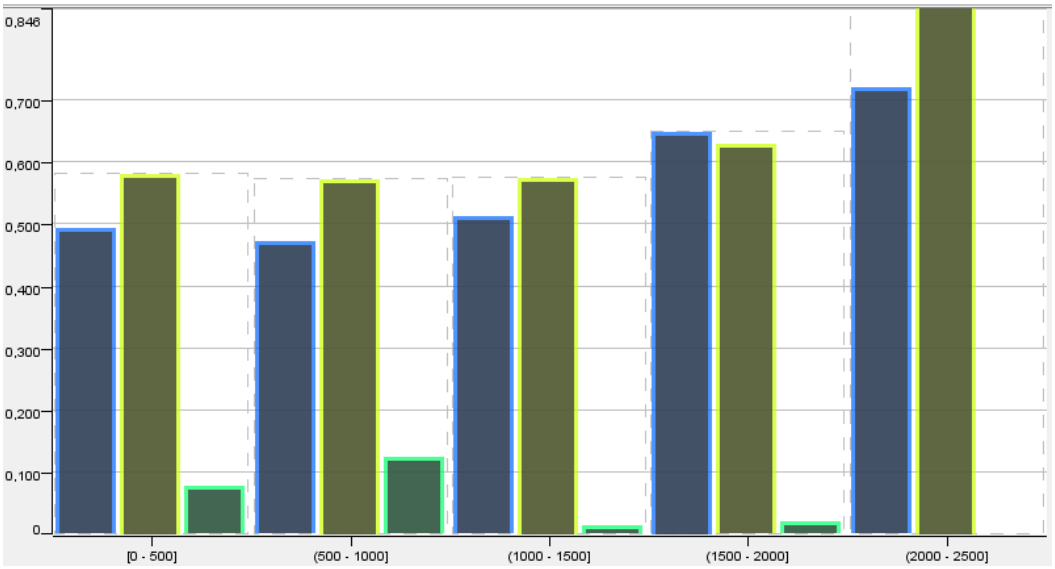


Ilustración 5.15 Frecuencia de las variables de índice de masa corporal, masa corporal magra y Lesión en relación a cantidad de minutos jugados.

Esto nos sugiere que la falta de un mejor condicionamiento muscular, es decir, trabajo de gimnasio (musculación) unido a una carga e intensidades sobreestimadas a jugadores con poco tiempo de partido en la temporada (<1.000 minutos) poseen fuerte relación con el riesgo de lesiones.

En base a esta hipótesis, pasamos a analizar el comportamiento de las medianas y cuartiles de estas variables a través del gráfico *box plot*, diagrama de caja o diagrama de extremos y cuartiles que es una herramienta gráfica para representar la variación de datos observados de una variable numérica por medio de cuartiles (Ross, 2004). En resumen, el *box plot* identifica dónde se encuentra el 50% de los valores más probables, la mediana y los valores extremos y se utiliza frecuentemente para analizar y comparar la variación de una variable entre diferentes grupos de datos (Devore, 2006).

Para la primera variable analizada (minutos jugados), en la Ilustración 5.16 vemos una pequeña variación en la mediana entre las clases, pero su dispersión y distribución son muy diferentes, como podemos percibir en la línea que conectan las medias, en los cuartiles y límite superior.

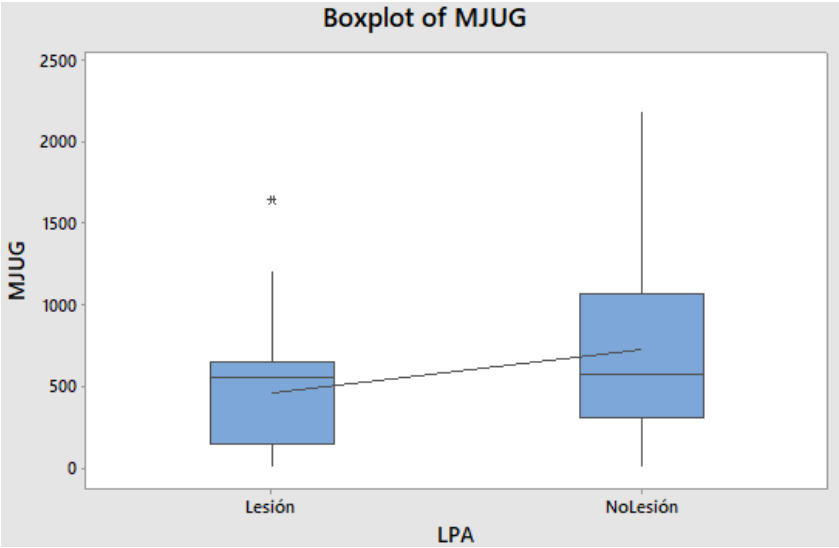


Ilustración 5.16 Box plot de las variables de Minutos Jugados y Lesión.

Para la segunda variable analizada (*sprints*), en la Ilustración 5.17 vemos una gran variación en la mediana, media, dispersión y distribución entre las clases, como podemos percibir en la línea que conectan las medias, en los cuartiles y límite superior.

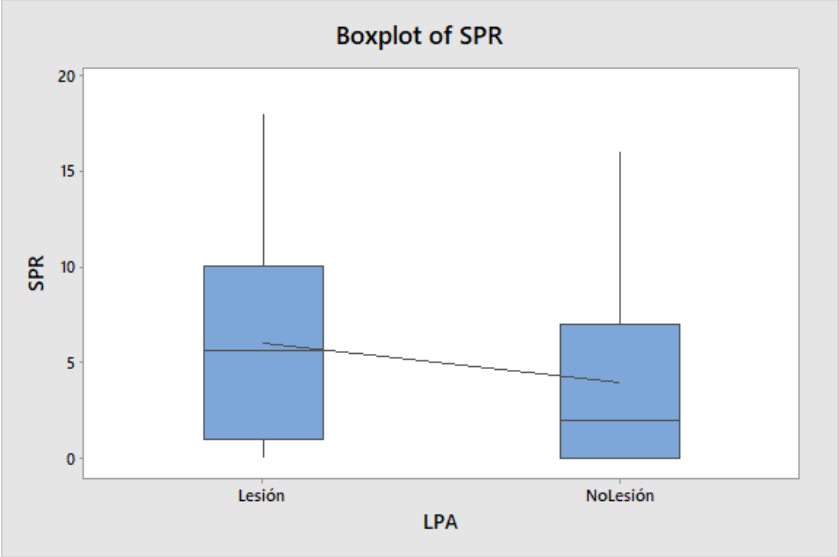


Ilustración 5.17 Box plot de las variables de Sprints y Lesión.

Para la tercera variable analizada (distancia total), en la Ilustración 5.18 vemos una gran variación en la mediana, media y dispersión, pero pequeña en la distribución entre las clases, como podemos ver en la línea que conectan las medias, en el primer cuartil y límite superior.

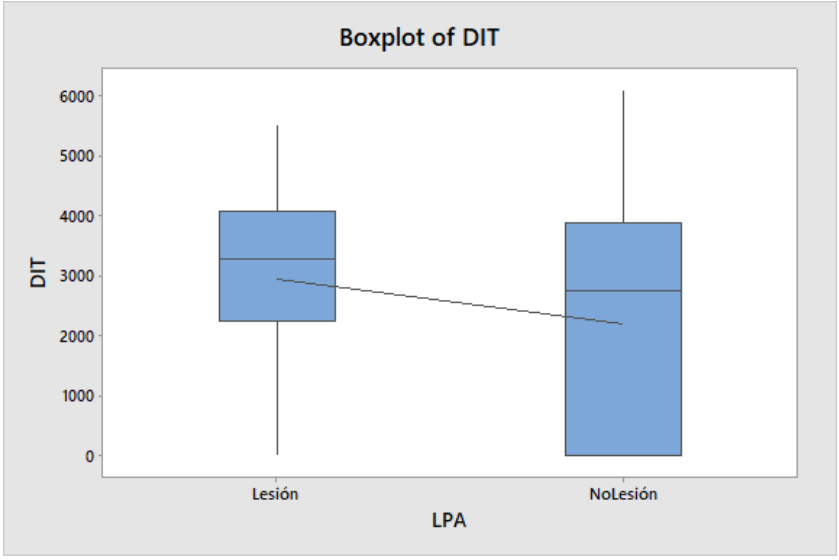


Ilustración 5.18 Box plot de las variables de Distancia Total y Lesión.

Para la cuarta variable analizada (velocidad media), en la Ilustración 5.19 vemos una gran variación en la mediana, media, dispersión y distribución entre las clases, como podemos ver en la línea que conectan las medias, en los cuartiles y límites. Podemos observar también la posible existencia de anomalías (*outliers*) que podrían o no influir en este análisis, pero como dijimos en el preprocesamiento, intentamos eliminarlos donde la media de su desviación estándar era mayor o menor que 2, por lo que estamos asumiendo estos valores que quedaron por debajo de este límite, aquí presentados.

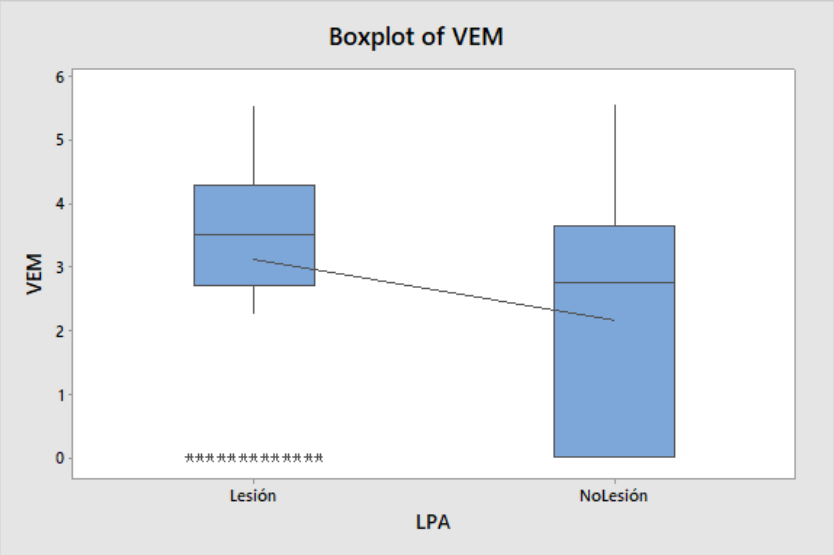


Ilustración 5.19 Box plot de las variables de Velocidad Media y Lesión.

Para la quinta variable analizada (unidad arbitraria), en la Ilustración 5.20 vemos una similitud en la mediana y tercer cuartil con pequeña variación en la media, dispersión y distribución entre las clases, como podemos ver en la línea que conectan las medias, en el primer cuartil y límite superior.

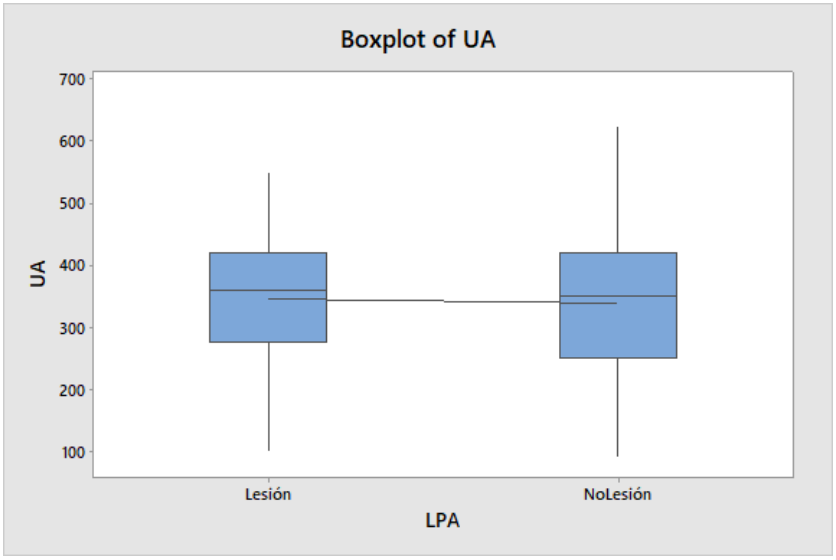


Ilustración 5.20 Box plot de las variables de Unidad Arbitraria (UA) y Lesión.

Para la sexta variable analizada (TRIMP), en la Ilustración 5.21 vemos una variación en la mediana, media, dispersión y distribución entre las clases, como podemos observar en la línea que conectan las medias, en el tercer cuartil y límite superior.

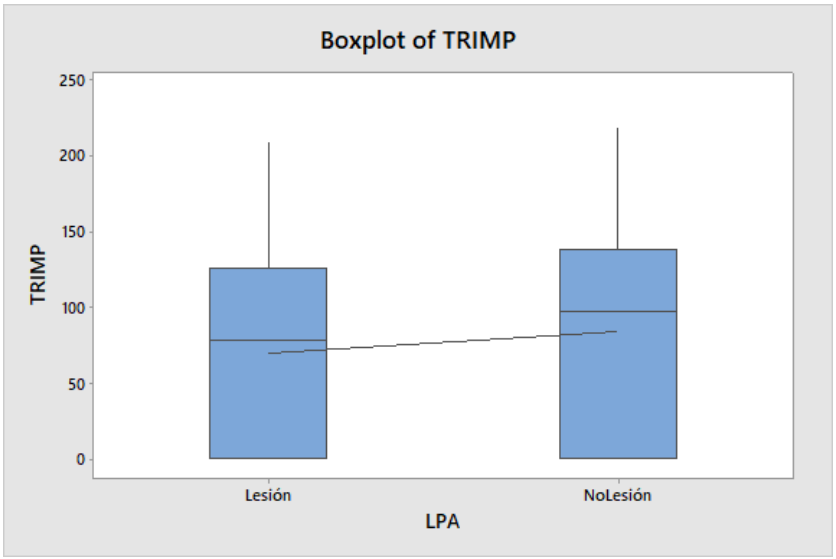


Ilustración 5.21 Box plot de las variables de TRIMP y Lesión.

Para la séptima variable analizada (IMC), en la Ilustración 5.22 vemos una pequeña variación en la mediana, pero mayores variaciones en la dispersión y

distribución entre las clases, como podemos ver en la línea que conecta las medias, en los cuartiles y límites.

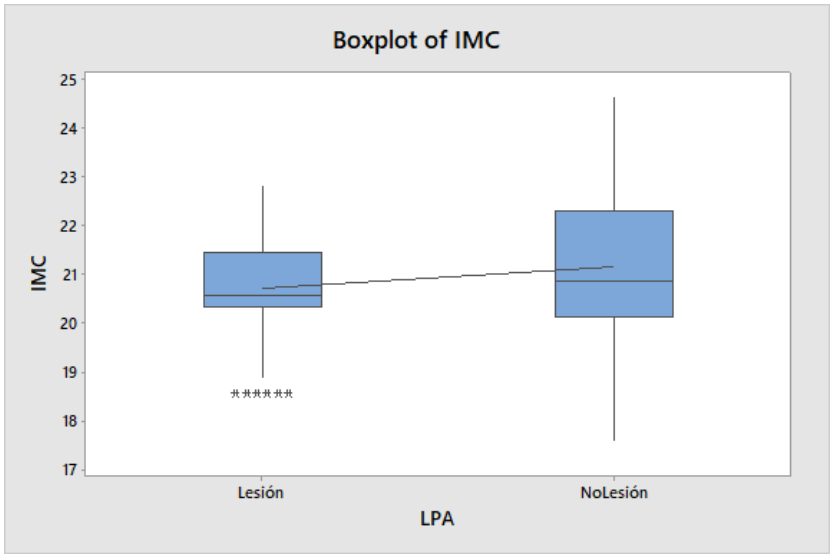


Ilustración 5.22 Box plot de las variables de índice de masa corporal (IMC) y Lesión.

Para la octava variable analizada (MCM), en la Ilustración 5.23 vemos una gran variación en la mediana, pero pequeñas variaciones en la dispersión entre las clases, como podemos observar en la línea que conecta las medias, en los cuartiles y límites.

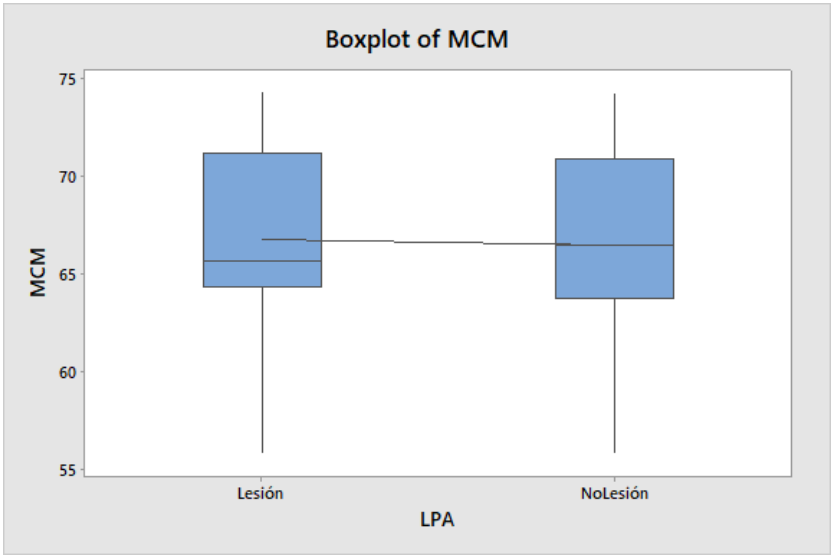


Ilustración 5.23 Box plot de las variables de masa corporal magra (MCM) y Lesión.

Además de las ocho variables citadas, se han analizado otras tres variables que nos llamaron la atención en los gráficos.

Para la novena variable analizada (potencia máxima), notamos en la Ilustración 5.24 una gran variación tanto en la mediana cuanto en la dispersión, pero pequeña en la distribución entre clases, como podemos ver en la línea que conecta sus promedios, en los cuartiles y límites. Los *outliers* presentados siguen el mismo tratamiento realizado durante el preprocesamiento explicado anteriormente cuando presentamos la variable velocidad media.

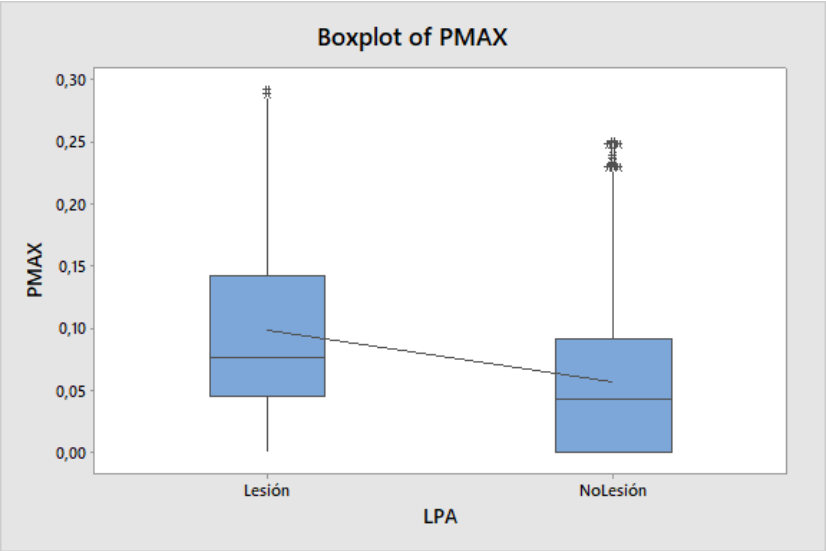


Ilustración 5.24 Box plot de las variables de potencia máxima (PMAX) y Lesión.

Para la décima variable analizada zona 2 del frecuencímetro (Z2), notamos en la Ilustración 5.25 una gran variación en la mediana y en la dispersión entre clases, como podemos ver en la línea que conecta sus promedios y en los cuartiles.

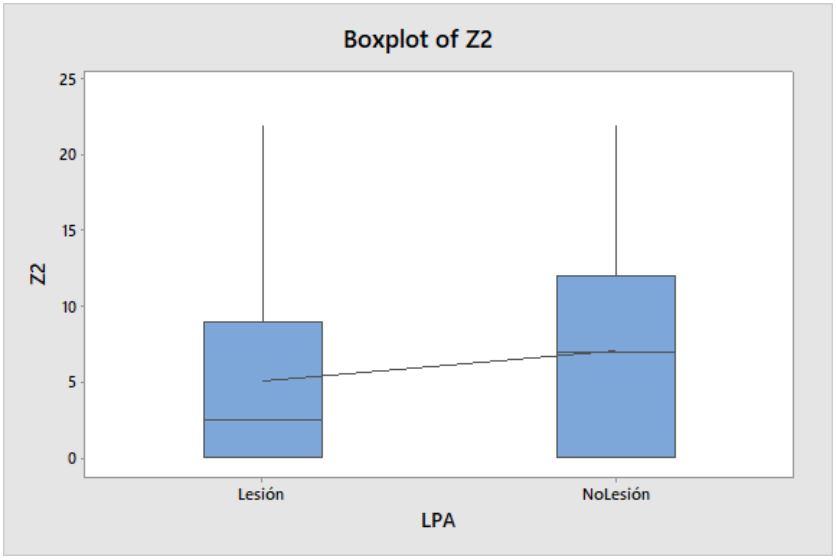


Ilustración 5.25 Box plot de las variables Z2 y Lesión.

Para la undécima variable analizada (porcentaje de grasa corporal), notamos en la Ilustración 5.26 una pequeña variación en la mediana y moderada variación en la dispersión y distribución entre clases, como podemos ver en la línea que conecta sus promedios, en los cuartiles y límites.

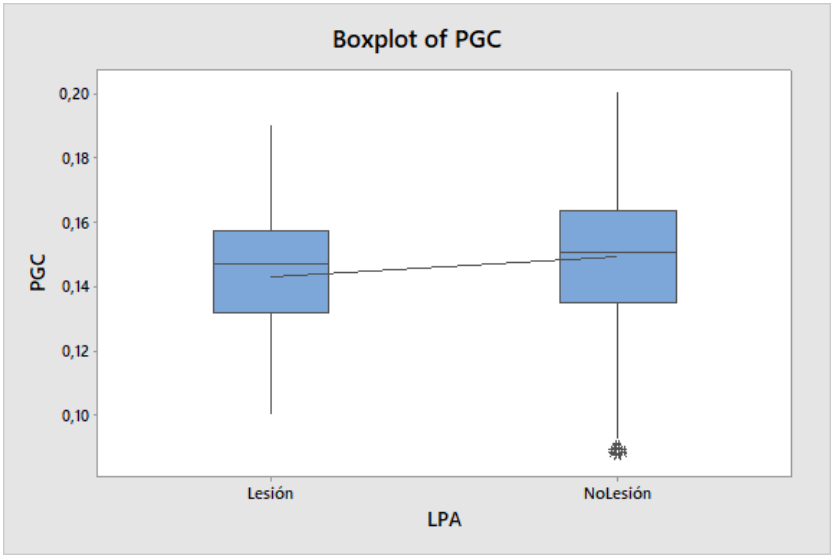


Ilustración 5.26 Box plot de las variables porcentaje de grasa corporal (PGC) y Lesión.

Como es posible percibir arriba, presentamos solamente las variables que nos han dado comportamientos muy distintos entre las clases, para poder analizar y formular algunas hipótesis. De este análisis observamos que algunas de estas

variables se destacan por presentar grandes diferencias en la mediana, como también en los cuartiles y límites de su población y esto puede ser una señal clara de que son variables importantes para la elaboración de un modelo de predicción de lesiones.

En resumen, después de analizar los diagramas de caja, podemos decir que las variables MJUG, SPR, VEM, DIT, PMAX, TRIMP, IMC, MCM, Z2 ejercen gran influencia en la clase lesión, mientras la variable UA ejerce poca influencia. Podemos también concluir de manera aislada que la incidencia de lesión ocurre:

- En atletas con poco tiempo de partido (MJUG).
- En atletas con una mayor intensidad de entrenamiento (SPR, VEM, DIT, PMAX).
- En atletas con bajo índice de masa corporal, masa muscular y porcentual de grasa corporal (IMC, MCM, PGC).
- En atletas con muy bajo tiempo de en la zona 2 del frecuencímetro que corresponde a 60-70% de la frecuencia cardiaca máxima (entrenamiento de baja intensidad - mejora la resistencia y la quema de grasa, por ejemplo: trotar).

5.1.4. Fase de modelado

En la fase anterior presentamos el análisis exploratorio mediante la visualización gráfica de los datos. Esta fase es parte inicial de un ciclo que pasa por el preprocesamiento, modelado y prueba en un proceso continuo de optimización de los resultados (ajustes de hiperparámetros e ingeniería de características), como se ve en la Ilustración 5.1, para encontrar el modelo más ajustado al objetivo del problema estudiado.

En este paso, modelado, abordamos un conjunto de algoritmos (*Bag of Training* o *Bag of Models* en inglés) de aprendizaje automático, incluyendo redes neuronales, donde elegimos los 5 algoritmos con mejores resultados para presentarlos en este estudio. Este conjunto de algoritmos fue alimentado con datos provenientes de cada técnica de reducción (RC, PCA y BFE) consolidada en un meta nodo (*Bag of Techniques* en inglés), además de recibir una nueva ronda de

datos sin aplicar ninguna reducción, que denominamos línea de base; esto se puede observar en la Ilustración 5.27 y Ilustración 5.28. El modelado y prueba recibió un conjunto particionado en un 80%, que luego se particionó nuevamente en la validación cruzada, mientras que otro 20% fue para la fase de validación que será presentada a posteriori (Ilustración 5.31 y Ilustración 5.28).

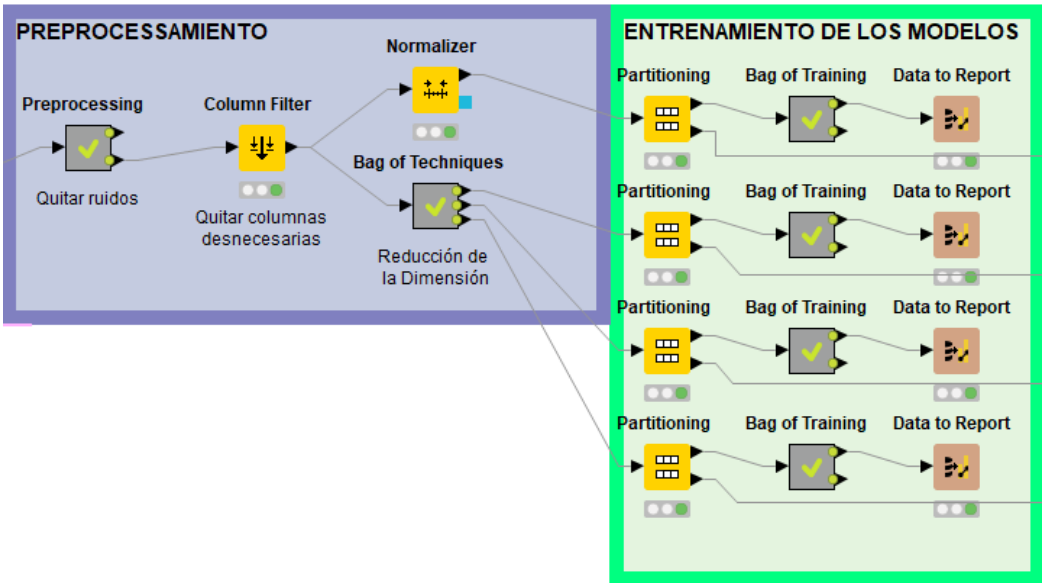


Ilustración 5.27 Diagrama representando la conexión entre preprocesamiento y modelado.

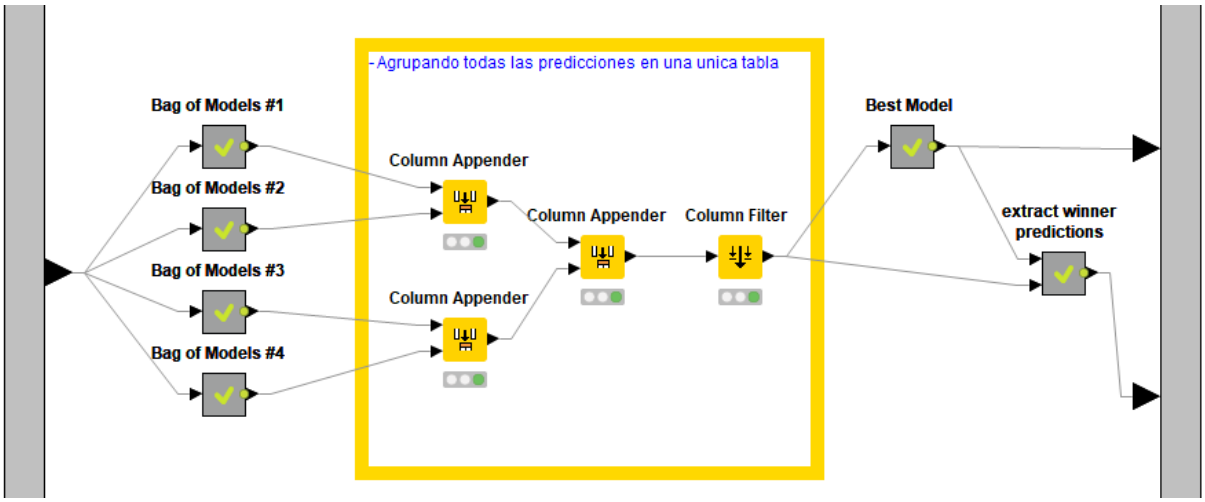


Ilustración 5.28 Diagrama representando las tareas de la meta nodo "Bag of Training".

Cada algoritmo se ejecuta a través de un método de validación cruzada (k = 10) que consiste en dividir el conjunto total de datos en k subconjuntos mutuamente

excluyentes del mismo tamaño y, a partir de esto, un subconjunto se utiliza para probar y los k-1 se utilizan para la estimación de los parámetros y se calcula la exactitud del modelo. Pero antes del inicio de este método por tratarse de un conjunto de datos con clases muy desequilibradas ($\cong 94/6$), lo que dificultaría el aprendizaje del modelo, tuvimos que equilibrarlas a través de una técnica de sobremuestreo denominada SMOTE (*Synthetic Minority Over-sampling Technique* en inglés (Chawla, Bowyer, Hall, & Kegelmeyer, 2002)) que consiste en crear muestras sintéticas de la clase más pequeña en lugar de crear copias de los datos. El algoritmo selecciona dos o más instancias similares (usando una medida de distancia) y modifica el valor de un atributo en una instancia en una cantidad aleatoria, de acuerdo con la diferencia de valores en las instancias vecinas. Después de equilibrar las clases necesitamos barajar las instancias para que el método de validación cruzada particione en un orden equilibrado las clases como se puede ver en la Ilustración 5.29 y la Ilustración 5.30.

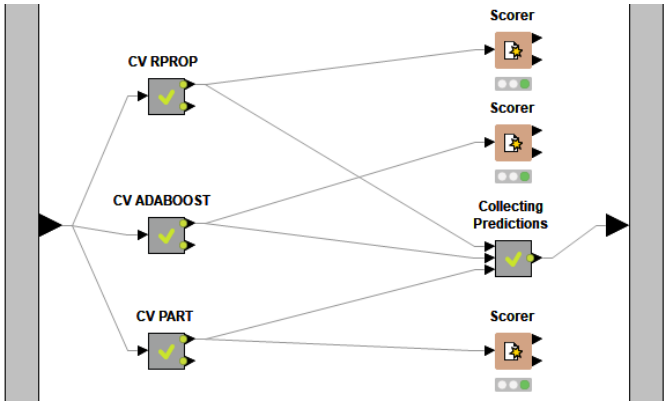


Ilustración 5.29 Diagrama representando un conjunto de algoritmos y su meta nodo validación cruzada.

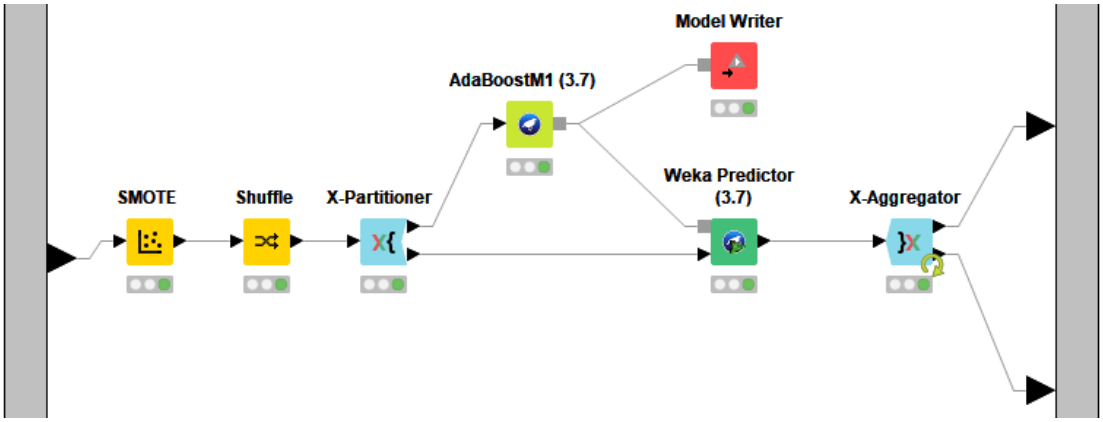


Ilustración 5.30 Diagrama representando las tareas de la meta nodo de validación cruzada.

5.1.5. Fase de validación

En esta fase recibimos el 20% del conjunto de datos particionado tras la fase de preprocesamiento y aplicamos el mismo conjunto de algoritmos de la fase de modelado, con la diferencia de que en esta fase aplicaremos el modelo aprendido de la fase de modelado y evaluamos los resultados alcanzados (Ilustración 5.31 y Ilustración 5.32) para su posterior interpretación y presentación en la fase de informes.

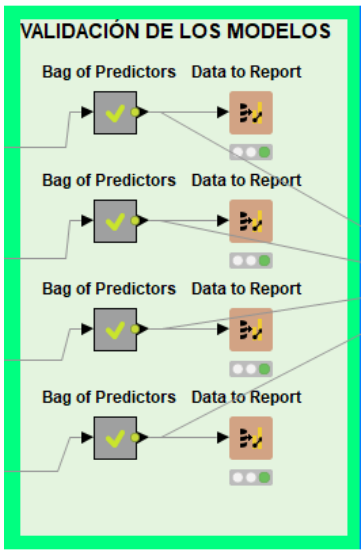


Ilustración 5.31 Diagrama de la fase de validación.

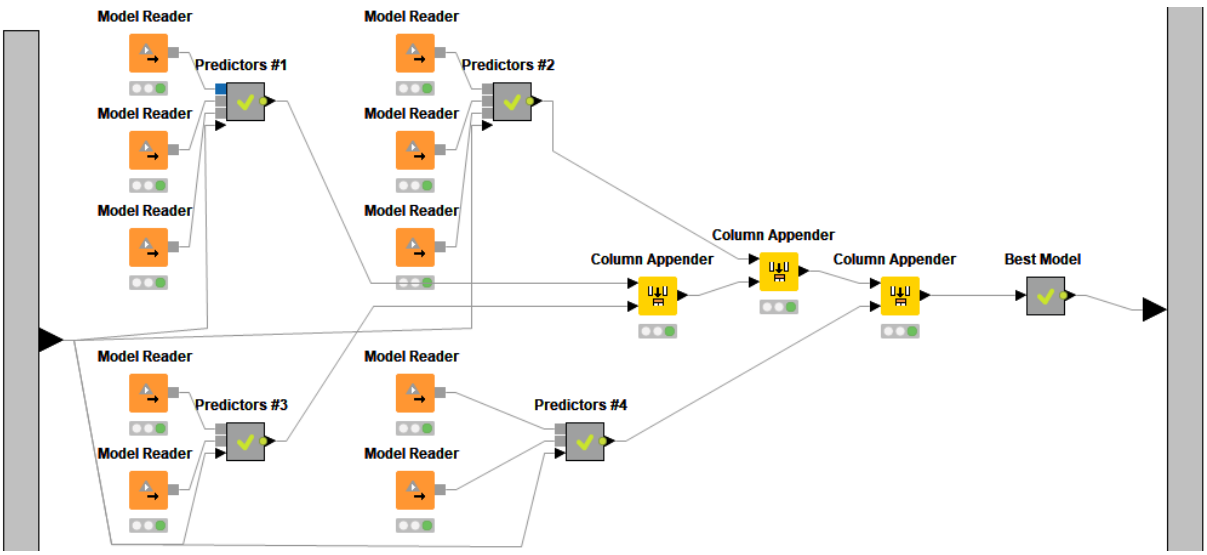


Ilustración 5.32 Diagrama con las tareas de validación (predictores).

5.1.6. Fase de informes

Desde el punto de vista de rendimiento, el mejor algoritmo es aquel capaz de generar un alto índice de aciertos en la clase positiva, es decir, prever el riesgo de lesión con alta confiabilidad llevando a reducir su ocurrencia en los entrenamientos y garantizando así número mayor de atletas a disposición de la comisión técnica para los partidos. Pero además, también debe ser capaz de generar un bajo índice de falsas alarmas (falsos positivos), es decir, prever erróneamente riesgo de lesión en atletas que, si continuan el entrenamiento sin ninguna intervención, no se lesionarían.

Hipotéticamente, si para evitar el riesgo de lesión en un atleta, la sugerencia es disminuir la carga e intensidad de entrenamiento, podríamos decir que un alto FPR haría que tuviéramos muchos atletas por debajo de sus mejores condiciones físicas, disponibles para la comisión técnica en los partidos. En contrapartida un bajo índice de aciertos en la clase positiva, llevaría a muchos atletas a contraer lesiones, impidiéndoles disputar partidos y disminuyendo así las opciones de la comisión técnica. Por lo tanto debemos procurar encontrar el mejor equilibrio en esta relación de métricas, que pasa por un proceso de decisión de toda la comisión técnica. En la óptica de este estudio, la prioridad es el bienestar del atleta, por lo que se debe dar mayor peso al índice de aciertos (recall +) que al FPR (1-especificidad) como criterio de desempate.

En la última fase de nuestra investigación, llamada de informes, recolectamos y consolidamos todos los resultados de la fase de validación presentándolos en forma de clasificación desde mejor índice MCC hasta el peor, seleccionando así los 5 mejores algoritmos (Random Forest, Adaboost, Part, Fuzzy y RProp) tanto en el enfoque A (línea de base) como en el B (Wilcoxon Mann-Whitney U) como se presenta en la Ilustración 5.33 y Ilustración 5.34.

Row ID	MCC	Recall	1-Specifity	rank
BFE_FUZZ	0.439	0.692	0.094	1
BASE_ADB	0.427	0.538	0.055	2
BFE_ADB	0.403	0.462	0.044	3
BFE_RPROP	0.34	0.846	0.238	4
BASE_FUZZ	0.328	0.462	0.072	5
PCA_ADB	0.328	0.538	0.099	6
BASE_RF	0.319	0.769	0.215	7
BASE_PART	0.316	0.462	0.077	8
RC_RPROP	0.306	0.923	0.331	9
BFE_RF	0.303	0.846	0.282	10
RC_ADB	0.294	0.385	0.061	11
PCA_RPROP	0.227	0.462	0.133	12
RC_FUZZ	0.221	0.385	0.099	13
RC_PART	0.214	0.462	0.144	14
BFE_PART	0.21	0.538	0.193	15
PCA_RF	0.197	0.615	0.26	16
PCA_FUZZ	0.197	0.385	0.116	17
RC_RF	0.163	0.615	0.309	18
BASE_RPROP	0.142	0.385	0.166	19
PCA_PART	0.137	0.308	0.122	20

Ilustración 5.33 Resultados del enfoque A – línea de base.

Row ID	MCC	Recall	1-Speci...	rank
BASE_FUZZ	0.489	0.692	0.072	1
BFE_PART	0.447	0.846	0.144	2
BFE_ADB	0.444	0.538	0.05	3
BFE_FUZZ	0.418	0.692	0.105	4
BASE_RF	0.365	0.923	0.254	5
BFE_RF	0.361	0.846	0.215	6
BASE_ADB	0.359	0.538	0.083	7
RC_ADB	0.305	0.462	0.083	8
BFE_RPROP	0.287	0.692	0.204	9
RC_FUZZ	0.235	0.462	0.127	10
PCA_PART	0.23	0.846	0.392	11
BASE_RPROP	0.226	0.692	0.276	12
PCA_RF	0.213	0.923	0.497	13
PCA_FUZZ	0.197	0.615	0.26	14
RC_PART	0.188	0.692	0.331	15
RC_RF	0.177	0.385	0.133	16
BASE_PART	0.153	0.385	0.155	17
RC_RPROP	0.149	0.615	0.331	18
PCA_ADB	0.084	0.308	0.177	19
PCA_RPROP	-0.023	0.231	0.271	20

Ilustración 5.34 Resultados del enfoque B – Wilcoxon Mann–Whitney U.

De acuerdo con estos resultados analizamos las métricas obtenidas según los enfoques que siguen:

- Enfoque A – Línea de Base, teniendo en cuenta la premisa dictada anteriormente, de que la métrica llamada "Recall +" tendrá un peso mayor que la métrica FPR (tasa de falsos positivos), nuestra elección se muestra en la Tabla 5.5 presentada abajo.

Modelo	Parámetros	Recall+	FPR	MCC
Backward Feature Elimination + Resilient Backpropagation (RPROP)	Iterations = 150 Hidden layers = 3 Hidden neurons per layer = 8	84,6%	23,8%	0,418
Backward Feature Elimination + Fuzzy (FURIA)	Folds = 3 Weight = 2 Optimizations = 2 Tnorm = standard UncovAction = standard	69,2%	9,4%	0,439
Rank Correlation (RC) + Resilient Backpropagation (RPROP)	Iterations = 150 Hidden layers = 3 Hidden neurons per layer = 8	92,3%	33,1%	0,306
Backward Feature Elimination + Random Forest	Tree depth = 9 Trees = 2000 No static random seed	84,6%	28,2%	0,303
Baseline + Random Forest	Tree depth = 9 Trees = 2000 No static random seed	76,9%	21,5%	0,319

Tabla 5.5 Los mejores parámetros del modelo para el enfoque A. Las métricas se muestran como la media +/- desviación estándar. El simbolo +,quiere decir, recall de la clase positiva.

- Enfoque B – Wilcoxon Mann-Whitney U, teniendo en cuenta la premisa dictada anteriormente, de que la métrica llamada "Recall +" tendrá un peso mayor que la métrica FPR (tasa de falsos positivos), nuestra elección se muestra en la Tabla 5.6 presentada abajo.

Modelo	Parámetros	Recall+	FPR	MCC
Backward Feature Elimination + Partial Decision Trees (PART)	Confidence = 0.95 Folds = 3 Objectives = 2	84,6%	14,4%	0,47
Baseline + Random Forest	Tree depth = 9 Trees = 2000 No static random seed	92,3%	25,4%	0,365
Backward Feature Elimination + Random Forest	Tree depth = 9 Trees = 2000 No static random seed	84,6%	21,5%	0,361
Baseline + Fuzzy (FURIA)	Folds = 3 Weight = 2 Optimizations = 2 Tnorm = standard UncovAction = standard	69,2%	7,2%	0,489
Backward Feature Elimination + Fuzzy (FURIA)	Folds = 3 Weight = 2 Optimizations = 2 Tnorm = standard UncovAction = standard	69,2%	10,5%	0,418

Tabla 5.6 Los mejores parámetros del modelo para el enfoque B. El simbolo +,quiere decir, recall de la clase positiva.

La discusión y análisis de los resultados presentados previamente se realizara en el apartado 6.

5.1.7. Fase de toma de decisiones

En esta fase, se toma la decisión respecto del mejor modelo o modelos a partir de las interpretaciones hechas en la fase anterior, que buscan desmitificar el problema a través de la identificación de sus principales síntomas y de esta forma permitir al cliente decidir cuál es el mejor tratamiento para la cura de este problema, en este caso prevenir lesiones. En esta fase cuando hablamos en cliente, de una forma práctica queremos decir, que la participación de toda la comisión técnica en el proceso es fundamental para la toma de decisión. Sin embargo, debido a que no existe la posibilidad de participación del club, tenemos que quedarnos en el campo de la hipótesis y realizar una sugerencia que será explicada en el capítulo 6.

5.1.8. Fase de implementación

La creación del modelo generalmente no es el final del proyecto. Incluso si el objetivo del modelo es aumentar el conocimiento sobre los datos, el conocimiento adquirido deberá organizarse y presentarse de manera que sea útil para el cliente (Shearer, 2000).

Por eso esta fase va mucho más allá de la simple tarea de implementar, dentro de un flujo de trabajo, los procesos ya desarrollados dentro de un modelo que comporte desde la recolección y procesamiento previo de los datos hasta la aplicación del mejor modelo de predicción y su optimización. El desafío está justamente en entender los requisitos de negocio, escenarios e indicadores de desempeño que se desean mejorar para implantar este flujo en fase de producción, con el objetivo de mantener en activa evolución el modelo aprendido, una vez que la recolección regular de nuevos datos en lotes o en línea influirá en los resultados a lo largo del tiempo, ya sin necesidad de satisfacer los criterios de aceptación del cliente.

De esta manera, el cliente debe estar participando activamente de este proceso, buscando siempre la alineación frente a sus necesidades. En esta investigación esta fase fue descartada en consecuencia de la inviabilidad de participación de toda la comisión técnica.

6. RESULTADOS Y DISCUSIÓN

El algoritmo de aprendizaje automático PART en el enfoque B (Tabla 5.6), fue el que presentó el mejor resultado con el 84,6% de aciertos en la clase positiva, identificando correctamente a los atletas que sufrieron una lesión asociada al entrenamiento con una tasa del 14,4% de tasa de falsos positivos, representando un coeficiente de 0,47 en la escala MCC (que va de -1 a +1). Esta elección se debe a los criterios adoptados en la fase de informes dentro de la sección 5.1.6 del capítulo 5. Tales criterios tienen en cuenta el índice de exactitud general MCC, buscando un equilibrio entre la métrica *Recall* y FPR. Esto quiere decir que no siempre el algoritmo con el mejor *Recall* (*Random Forest*=92,3%) será el elegido, si éste también presenta un alto FPR (*Random Forest*=25,4%), por lo que tenemos el MCC (*Random Forest*=0,37) para ayudarnos en la elección.

El algoritmo PART (MCC = 0,47, *Recall* = 84,6%, FPR = 14,4%) obtuvo una leve ventaja en el MCC en relación al algoritmo FURIA (MCC = 0,49, *Recall* = 69,2%, FPR = 7,2%), lo que influyo en nuestra decisión, optando por decidir de acuerdo con el criterio de desempate que prefiere un mayor valor de *Recall*.

Esto demuestra que este algoritmo es una buena elección, ya que es citado en diversos estudios científicos que tratan con clases altamente desequilibradas. Sadratrasoul et al., en su estudio (Sadatrasoul, Gholamian, & Shahanaghi, 2013) concluye que este algoritmo fue el que presentó los mejores resultados en todos los conjuntos de datos balanceados y desequilibrados. (Su, Ju, Liu, & Yu, 2015) en su artículo propone una mejora del algoritmo PART usando la divergencia de Kullback-Leibler o ganancia de información como criterio de división para construir árboles de decisión parciales y comprueba su robustez en presencia de clases desequilibradas, junto con la técnica de *oversampling* SMOTE, recomendando el uso de ambas al tratar con clases altamente desequilibradas.

Para que el modelo de árbol de decisión y el modelo de lista de decisiones PART sean más legibles para el ser humano, cada ruta de raíz a hoja se puede transformar en una regla *IF-THEN*. Si la condición es satisfecha, la conclusión se verifica. A partir de esta explicación, nuestro modelo fue capaz de generar 21 reglas cubriendo correctamente en promedio el 93,63% de las instancias. De entre las 21

reglas, 9 determinan la clase como positiva (lesión) tal como se presenta en la Tabla 6.1 a seguir.

Regla	Condición	Resultado	Confianza
1	IMC > 0.135231 AND MJUG > 0.265027 AND PGC <= 0.65473 AND IMC <= 0.534604 AND DIT <= 0.823272 AND DIT > 0.045382	(175.0/26.0)	87,06%
2	IMC > 0.135231 AND MJUG <= 0.059055 AND IMC <= 0.503201	(132.0/12.0)	91,7%
3	MJUG > 0.219945 AND IMC > 0.534604 AND PGC <= 0.667158 AND MJUG > 0.287341	(118.0/2.0)	98,33%
4	IMC <= 0.582436 AND IMC > 0.135231 AND PGC <= 0.415271 AND MJUG <= 0.293372 AND DIT <= 0.670099 AND PGC > 0.29094 AND MJUG <= 0.139149 AND DIT > 0.346335	(64.0)	100%
5	IMC <= 0.582436 AND IMC > 0.135231 AND IMC <= 0.309899 AND MJUG <= 0.260218 AND IMC > 0.180555 AND DIT <= 0.493185	(54.0)	100%
6	IMC <= 0.582436 AND IMC > 0.135231 AND MJUG <= 0.323829 AND MJUG > 0.234716	(100.0/33.0)	75,2%
7	MJUG <= 0.077998	(20.0/1.0)	95,24%
8	MJUG > 0.08561 AND MJUG > 0.122923 AND PGC > 0.46455 AND MJUG > 0.133489	(22.0/1.0)	95,65%
9	MJUG > 0.08561 AND MJUG <= 0.122923	(21.0)	100%

Tabla 6.1 Reglas extraídas de la clase positiva (lesión).

Los números en paréntesis en la columna de resultados indican el número de ejemplos en las hojas. El número de ejemplos erróneamente clasificados también se daría, en este caso después de una barra (/).

Entre las variables enumeradas destacan las siguientes en orden ocurrencia:

- MJUG que aparece en todas las reglas.
- IMC que aparece en 6 de las 9 reglas.
- PGC que aparece en 4 de las 9 reglas.
- DIT que aparece en 3 de las 9 reglas.

Entre las reglas enumeradas, solamente las 3 primeras poseen una cobertura del 59,5% de las instancias con 91,4% de confianza.

Por último, nuestras hipótesis planteadas durante la fase de análisis exploratorio de datos en el capítulo 5 confirman, para este estudio, la fuerte correlación entre lesión y las variables MJUG, IMC, PGC y DIT, reforzando la idea de que atletas con pocos minutos jugados, bajo IMC y PGC y una sobrecarga de entrenamientos son susceptibles de lesionarse. Por eso, el resultado de este estudio recomienda un aumento en el trabajo en gimnasio (musculación) junto con una buena dieta para la ganancia de masa muscular y una reducción en la carga de entrenamiento para atletas con menor tiempo de juego.

7. CONCLUSIONES

De acuerdo con los resultados de este TFM, se puede afirmar que el uso de aprendizaje automático puede ser útil como herramienta de apoyo a la decisión de la comisión técnica dentro de un programa de prevención de lesiones.

También podemos afirmar que se recomienda fuertemente el empleo de las pruebas de hipótesis para la selección previa de características, antes de la aplicación de las técnicas de reducción de la dimensión, como se mostró en el capítulo 5 y en los resultados de los dos enfoques (A y B) expuestos en las ilustraciones 5.33 y 5.34.

Este estudio hace las siguientes contribuciones: En primer lugar, proporciona una estructura general para predecir lesiones mediante la agregación semanal de datos GPS. En segundo lugar, compara dos enfoques diferentes: el primero utiliza todas las variables disponibles y el segundo sólo las variables que rechazaron la

hipótesis nula. En tercer lugar, comparan varios métodos de aprendizaje automático como forma de construir un modelo predictivo.

La principal contribución es la utilización de la prueba de hipótesis de Wilcoxon Mann-Whitney U y la reducción de la dimensionalidad a través de la técnica de eliminación de características hacia atrás, junto con el algoritmo PART, para manejar la clasificación y su aplicación al problema actual. El modelo BFE + PART no sólo funciona suficientemente bien para demostrar que la tarea es factible y con un rendimiento óptimo de ejecución, sino que también permite extraer las condiciones (reglas) que pueden llevar a la lesión. Este recurso es particularmente útil para el profesional de deportes que busca entender cómo el programa de entrenamiento puede causar lesiones. Desde que las unidades GPS se introdujeron recientemente en el fútbol, este estudio proporciona una herramienta útil para utilizar los datos que se recopilan, mientras que establece una referencia para futuros estudios.

Por otro lado cabe también resaltar que este modelo tiene ciertas limitaciones:

- La población utilizada sólo contiene datos de jugadores de una liga concreta. No se disponen de los datos de todos los jugadores de las ligas profesionales en el mundo. Por lo tanto, el modelo obtenido de este estudio puede ser considerado aplicable solamente a esta muestra.
- Las variables recopiladas corresponden a un dispositivo GPS concreto. Sabemos que cada dispositivo GPS proporciona sus variables, no habiendo un patrón universal establecido por un órgano regulador. De la misma forma, cada comisión técnica trabaja con sus variables en la periodización semanal de entrenamiento, no utilizando siempre todas las que el dispositivo ofrece. Por lo tanto, podríamos alcanzar mejores resultados recogiendo y analizando un mayor número de variables fisiológicas, físicas, químicas, psicológicas, entre otras, como señalan diversos estudios en este área.

En contrapartida, se están realizando esfuerzos por conocer los datos de la población y tener un estándar universal en la recolección de datos (variables) de los dispositivos EPTS, como en el caso de la FIFA, tal como se ha explicado en los

capítulos 1 y 2. Esto quiere decir que en el futuro, a partir de estos datos, será posible obtener un modelo activo y evolutivo que pueda servir de herramienta de soporte a la decisión en tiempo real para cualquier comisión técnica en el mundo.

7.1. Trabajos futuros

Hay varias direcciones que se pueden explorar en el futuro. Esta sección resume algunas sugerencias generales.

En primer lugar, se puede explorar una gama más amplia de modelos para la investigación que la descrita en este TFM. Aunque este TFM examinó la idoneidad de varios modelos, hay más métodos que podrían utilizarse. Obviamente, una comparación completa de cada algoritmo de clasificación, regresión, aprendizaje profundo o reforzado estaba más allá del alcance de este TFM. La elección de los algoritmos utilizados estuvo motivada por una serie de factores, como la idoneidad de un algoritmo para un problema particular y su éxito en problemas previos con características similares. El aprendizaje automático, sin embargo, es un área de investigación muy activa con nuevos algoritmos que se publican continuamente. Por lo tanto, todavía hay terreno fértil para la investigación sobre qué métodos son los mejores para un problema en particular.

En segundo lugar, la investigación futura también podría indicar datos adicionales que podrían ser utilizados por los modelos actuales. La recopilación de datos en el fútbol es problemática y difícil, como se indicó anteriormente. Lo que este TFM logró en cada investigación fue ilustrar que incluso con datos limitados, es posible construir modelos predictivos para lesiones en el fútbol. Sin embargo, el rendimiento de todos los modelos podría beneficiarse enormemente de datos adicionales. Más específicamente, todos los estudios se beneficiarían al incluir datos de más clubes.

En tercer lugar, los estudios requieren replicación, a fin de comprender el grado en que los resultados son generalizables. Estos estudios particulares se realizaron en colaboración con un club de la primera división brasileña. Sin embargo, sigue siendo la cuestión de si estos estudios pueden generalizarse bien a otras

situaciones, ya que puede haber complicaciones, variando el estilo de juego, la constitución física o la nutrición.

Por último, quizá en un futuro podamos llegar a una solución de implementación práctica y global, siguiendo ejemplos de otras dos empresas (Kitman Labs, 2018) y (TopSportsLab, 2018) que ya poseen hoy oferta de *software* disponible en el mercado e ir más allá de sus funcionalidades, presentando cuadros de mando como se muestra en las ideas recogidas en la siguiente ilustración.

Ilustración 7.1 Esbozo de una solución de monitoreo y predicción de lesiones en tiempo real.

Bibliografía

Abdi, H., & Williams, L. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2 (4)*, págs. 433–459.

Abney, S. (2007). *Semisupervised Learning for Computational Linguistics.* Chapman & Hall/CRC.

Abotel, K. (2015). *The Crippling Cost Of Sports Injuries.* Obtenido de Forbes: https://www.forbes.com/sites/sap/2015/08/11/the-crippling-cost-of-sports-injuries/#7835bc2f4d1f

Ahlawat, N., Gautam, A., & Sharma, N. (2014). Use of Logic Gates to Make Edge Avoider Robot. *International Journal of Information & Computation Technology, 4 (6)*, pág. 630.

Alvarez, E. (08 de Marzo de 2017). *FIFA envisions a future where players wear in-game fitness trackers.* Obtenido de Engadget: https://www.engadget.com/2017/08/03/fifa-epts-wearable-technology/

Andrienko, N., & Andrienko, G. (2005). *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach.* New York: Springer-Verlag.

Bishop, C. (2008). *Pattern Recognition and Machine Learning.* New York: Springer-Verlag.

Boughorbel, S., Jarray, F., & M., E.-A. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE, 12 (6)*, pág. 4.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Int. Res. 16 (1)*, págs. 321-357.

Cohen, W. W. (1995). Fast effective rule induction. *Proceedings of the Twelfth International Conference on Machine Learning* (págs. 115-123). Morgan Kaufmann.

Cox, M., & Ellsworth, D. (1997). *Application-Controlled Demand Paging for Out-of-Core Visualization.* IEEE Computer Society Press.

Dallinga, J. M., Benjaminse, A. M., & Lemmink, K. A. (2012). Which screening tools can predict injury to the lower extremities in team sports? *Sports Medicine, 42 (9)* , págs. 791-815.

Devore, J. L. (2006). *Estatística e Probabilidade para Engenharia e Ciências.* Cengage Learning, págs. 35-38.

Dhar, V. (2013). Data Science and Prediction. *Communications of the ACM, 56 (12)*, pág. 64.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Unsupervised Learning and Clustering. En *Pattern Classification, (2nd ed.).* Wiley.

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research, 15 (1)*, págs. 3133-3181.

FIFA. (03 de Mayo de 2007). *Prevention of injuries.* Obtenido de FIFA: http://www.fifa.com/development/news/y=2007/m=5/news=prevention-injuries-513864.html

FIFA. (09 de Octubre de 2015). *FIFA and IFAB to develop global standard for electronic performance and tracking systems.* Obtenido de FIFA: http://www.fifa.com/about-fifa/news/y=2015/m=10/news=fifa-and-ifab-to-develop-global-standard-for-electronic-performance-an-2709918.html

FIFA. (6 de Abril de 2018). *JUNIOR FOOTBALL RESEARCH & DATA MANAGER*. Obtenido de GlobalSportsJobs: https://www.globalsportsjobs.com/job/964902/junior-football-research-and-data-manager/?TrackID=66024&utm_source=jbe&utm_medium=email&utm_campaign=DateUnknown&BatchID=1938

Flach, P. A. (2012). *Machine Learning, The Art and Science of Algorithms that Make Sense of Data.*

Foster, C., Daines, E., Hector, L., Snyder, A., & Welsh, R. (1996). Athletic performance in relation to training load. *Wisconsin medical journal, 95 (6)*, págs. 370-374.

Foster, C., Florhaug, J., Franklin, J., Gottschall, L., Hrovatin, L., Parker, S., . . . Dodge, C. (2001). A New Approach to Monitoring Exercise Training. *Journal of Strength and Conditioning Research, 15 (1)*, págs. 109-115.

Frank, E., & Wittenx, I. H. (1998). Generating Accurate Rule Sets Without Global Optimization. *Fifteenth International Conference on Machine Learning* (págs. 144-151). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Gabbett, T. (2016). The training-injury prevention paradox: should athletes be training smarter and harder? *British Journal of Sports Medicine, 50 (5)*, 273-280.

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* New York: Springer Texts in Statistics.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Feedforward Networks. En *Deep Learning* (pág. 196). The MIT Press.

Hägglund, M., & Waldén, M. (de 2016). *Epidemiology of football injuries.* Obtenido de https://www.researchgate.net/publication/303312362_Epidemiology_of_football_injuries

Hayashi, C. (1998). What is Data Science ? Fundamental Concepts and a Heuristic Example. En C. Hayashi, K. Yajima, H. Bock, N. Ohsumi, Y. Tanaka, & B. Y., *Studies in Classification, Data Analysis, and Knowledge Organization* (págs. 40-51). Tokyo: Springer.

Ho, T. K. (1995). Random Decision Forests . *ICDAR '95 Proceedings of the 3rd International Conference on Document Analysis and Recognition* (págs. 278–282). Montreal: IEEE Computer Society.

HO, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell., 20 (8)*, págs. 832-844.

Huehn, J. C., & Huellermeier, E. (2009). *FURIA: An Algorithm for Unordered Fuzzy Rule Induction.* Data Mining and Knowledge Discovery.

Jarrell, S. B. (1994). *Basic Statistics (Special pre-publication ed.).* Dubuque: Wm. C. Brown Pub, pág. 492.

Johnson, A., & Ivarsson, A. (2011). Psychological predictors of sport injuries among junior soccer players. *Scandinavian Journal of Medicine and Science in Sports, 21 (1)*, 129–136.

Junge, A. (2000). The Influence of Psychological Factors on Sports Injuries. *The American Journal of Sports Medicine, 28 (5)*, págs. 10-15.

Junge, A; Dvorak, J; Rösch, D. (2000). Psychological and Sport-Specific Characteristics of Football Players. *American Journal of Sports Medicine, 28 supplement 5, S-22 - S-28*.

Kampakis, S. (2016). *Predictive modeling of football.* University College London.

Kitman Labs. (17 de Abril de 2018). Obtenido de Kitman Labs: https://www.kitmanlabs.com/our-system/

Knime. (5 de Deciembre de 2015). Seven Techniques for Data Dimensionality Reduction.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification. *Proceedings of the 2007 conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies* (págs. 3-24). Amsterdam: IOS Press .

Lantz, B. (2013). *Machine Learning with R.* Packt Publishing Ltd.

Laux, P., & et al. (2015). Recovery–stress balance and injury risk in professional football players. *Journal of Sports Sciences, 33 (20)*, 2140–2148.

Leek, J. (12 de Deciembre de 2013). *The key word in "Data Science" is not Data, it is Science.* Obtenido de Simply Statistics: https://goo.gl/UdEtxo

Lewis, M. (2004). *Moneyball: The Art of Winning an Unfair Game.* W. W. Norton & Company.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook.* New York: Springer.

Mannila, H. (1996). Data mining: machine learning, statistics, and databases. *International Conference on Scientific and Statistical Database Management* (págs. 2-9). Washington DC: IEEE Computer Society.

Miller, B. (. (2011). *Moneyball [Motion Picture].*

Minitab Inc. (2016). A comparison of the Pearson and Spearman correlation methods.

Mueller-Wohlfahrt, H.-W., & et al. (18 de Octubre de 2012). Terminology and classification of muscle injuries in sport: The Munich consensus statement. *British Journal of Sports Medicine, 47 (6)*, págs. 342-350.

Nielsen, A. B., & Yde, J. (1989). Epidemiology and traumatology of injuries in soccer. *The American Journal of Sports Medicine, 17 (6)*, págs. 803-807.

QSports. (2013). Obtenido de QStarz: http://www.qstarz.com/Products/Software%20Products/QSports-F.htm

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Francisco: Morgan Kaufmann Publisher.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research, 4 (1)*, págs. 77-90.

Riedmiller, M., & Braun, H. (1992). Rprop - A Fast Adaptive Learning Algorithm. *Proceedings of the International Symposium on Computer and Information Science VII*, (págs. 586-591).

Rivest, R. L. (1987). Learning decision lists. *Machine Learning, 2 (3)*, págs. 229–246.

Ross, S. (2004). *Introduction to Probability and Statistics ofr Engineers and Scientists (3rd ed.).* Elsevier, pág. 27.

Rossi, A. (10 de Abril de 2017). *PREDICTIVE MODELS IN SPORT SCIENCE: MULTI-DIMENSIONAL ANALYSIS OF FOOTBALL TRAINING AND INJURY PREDICTION.* Università degli Studi di Milano.

Sadatrasoul, S., Gholamian, M. R., & Shahanaghi, K. (2013). Inducing Valuable Rules from Imbalanced Data: The Case of an Iranian Bank Export Loans. *International Journal of Information, Security and Systems Management, 2 (1)*, págs. 130-135.

SAS. (29 de Marzo de 2016). *Machine Learning: What it is and why it matters.* Obtenido de SAS: https://www.sas.com/it_it/insights/analytics/machine-learning.html

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, págs. 85–117.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing, 5 (4)*, págs. 13-22.

Su, C., Ju, S., Liu, Y., & Yu, Z. (2015). Improving PART algorithm with K-L divergence for imbalanced classification. *Intelligent Data Analysis, 19 (5)*, págs. 1035-1048.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction.* Cambridge: MIT Press.

Swain, D., Abernathy, K., Smith, C., Lee, S., & Bunn, S. (1994). Target heart rates for the development of cardiorespiratory. *Medicine and Science in Sports and Exercise, 26 (1)*, págs. 112-116.

Talukder, H., T, V., Foster, G., Hu, C., Huerta, J., Kumar, A., . . . Simpson, S. (2016). Preventing in-game injuries for NBA players. *Sports Analytics Conference.* Boston.

TopSportsLab. (17 de Abril de 2018). Obtenido de TopSportsLab: https://www.topsportslab.com/software/injury-prevention.php

Tsai, C. F., Hsu, Y. F., & Yen, D. C. (2014). *A comparative study of classifier ensembles for bankruptcy prediction.* Applied Soft Computing 24, 977–984.

Uth, N., Sørensen, H., Overgaard, K., & Pedersen, P. (2005). Estimation of VO2max from the ratio between HRmax and HRrest--the Heart Rate Ratio Method. *European Journal of Applied Physiology, 93 (4)*, págs. 508-509.

Utts, J. M. (2005). *Seeing Through Statistics (3rd ed.).* Thomson Brooks/Cole, págs. 166-167.

Van Gerven, M., & Bohte, S. (2017). Artificial Neural Networks as Models of Neural Information Processing. *Frontiers in Computational Neuroscience, 11 (114)*, págs. 1-2.

Walker, J., Halliday, D., & Resnick, R. (2012). *Fundamentos de Física, 1.* Rio de Janeiro: LTC, pág. 340.

Wernick, Y., Brankov, Y., & Strother, S. C. (2010). Machine Learning in Medical Imaging. *IEEE Signal Processing Magazine, 27 (4)*, págs. 25–38.

Zell, A., Mache, N., Hübner, R., Mamier, G., Vogt, M., Schmalzl, M., & Herrmann, K.-U. (1994). SNNS (Stuttgart Neural Network Simulator). En *Skrzypek J. (eds) Neural Network Simulation Environments. The Kluwer International Series in Engineering and Computer Science, 254.* Boston, MA, USA: Springer.