Estudio matemático de lesiones musculares

Trabajo de fin de grado

Curso 2020/21

UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS

Grado en Ingeniería Matemática

Alumno: **Pablo Martín Sierra**

Tutora: **Ana Carpio**

Madrid, 2 de Septiembre de 2021

Resumen

El objetivo de este estudio surge ante la necesidad de anticiparse a las lesiones que puedan surgir dentro del ámbito profesional de competición de baloncesto, viendo cómo una serie de características físicas y funciones desempeñadas por un jugador dentro de la cancha de baloncesto afectan a la posibilidad del tipo de lesión que pueda tener, dónde podría tenerla y en qué nivel de gravedad. Estudiar esta anticipación desde el punto de vista matemático será el objetivo fundamental del trabajo.

Palabras clave: machine learning, cluster, agrupamiento, predicción, regresión, clasificación, NBA, lesiones, variable cuantitativa, variable cualitativa

Abstract

The objective of this study arises from the need to anticipate injuries that may arise within the professional basketball competitive environment, identifying how a series of physical characteristics and roles by a player on the basketball court affects the possibility of the type of injury he may have, where he may have it and at what level of severity. Studying this anticipation from the mathematical point of view will be the fundamental objective of the work.

Key words: machine learning, clustering, grouping, prediction, regression, classification, NBA, injuries, quantitative variable, qualitative variable.

Índice general

1. Introducción **6**
1.1. Objeto de estudio 6
1.2. Metodología de aprendizaje automático 6
1.3. Base de datos 8
1.4. Estructura del trabajo 8

2. Aprendizaje no supervisado **9**
2.1. k-medias (k-means) 9
2.1.1. Silhouette 10
2.2. Agrupación jerárquica (hierarchical clustering) 12

3. Aprendizaje supervisado **14**
3.1. Regresión Logística 14
3.2. SVM 18
3.3. Naive Bayes 19
3.4. Curva ROC 20

4. Aplicación a lesiones deportivas **23**
4.1. Base de datos 23
4.2. Exploración 30
4.3. Variables Dummy 35
4.4. Análisis e interpretación 36
4.4.1. Resultados k-medias 36
4.4.2. Resultados Agrupamiento Jerárquico 37
4.4.3. Resultados Regresión Logística, SVM y Naive Bayes 39

5. Conclusiones **49**

Bibliografía

Capítulo 1

Introducción

1.1. Objeto de estudio

El mundo de la alta competición siempre busca optimizar los resultados de los campeonatos y ligas profesionales, siendo un valor fundamental tener a los deportistas en un perfecto estado de forma y rendimiento. El cuerpo humano no es perfectamente simétrico y pequeños hábitos hacen que se formen descomponsaciones musculares que derivan en lesión. También pueden producirse por factores que no pueden controlarse, como recibir un fuerte impacto de otro competidor o una mala caída.
Una vez ocurrida la lesión, se buscará, desde todos los factores que puedan controlarse, que no vuelva a ocurrir o que no le suceda a nadie más. En este trabajo se tienen como objetivos a cumplir:
- Generar una agrupación de la información depurada para obtener relaciones entre variables.
- Obtener un modelo de regresión para hallar una probabilidad de lesión articular de un jugador en función de ciertas cualidades físicas.
- Obtener un modelo de regresión para hallar una probabilidad de lesión de gravedad de un jugador en función de ciertas cualidades físicas.
- Obtener modelos de regresión para hallar la probabilidad de lesión en una parte del cuerpo concreta.

1.2. Metodología de aprendizaje automático

El machine learning es un conjunto de técnicas de resolución de problemas mediante la clasificación o la predicción. Los algoritmos aprenden de los datos introducidos y luego utilizan este conocimiento para sacar conclusiones de nuevos datos. Es una rama dentro de la inteligencia artificial que utiliza algoritmos matemáticos que permiten a las máquinas aprender y entrenarse para enfrentarse a los nuevos datos y ajustarlos a las necesidades de los problemas. La metodología de resolución de estos problemas sigue el siguiente esquema de trabajo:

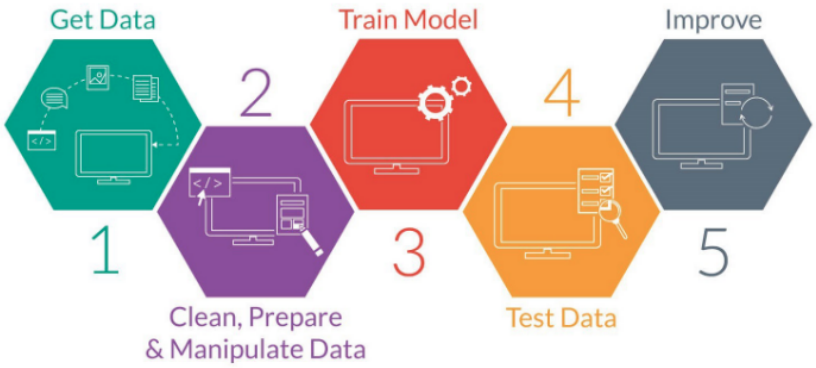


Figura 1.1: Modelo Machine Learning [1]

Se obtiene una base de datos inicial, se manipula la información con el objetivo de depurar los datos para unos resultados correctos y acordes con el estudio, se entrena el modelo con una parte de la base, se testea el modelo con los datos restantes de la base (se realiza este proceso para evitar casos de sobreentrenamiento de la máquina, que es igual de perjudicial para el resultado que si la máquina tiene un bajo número de observaciones para el entrenamiento) y se exploran resultados, buscando mejoras en el modelo en caso de necesitarlas.
En este trabajo se abordan dos modelos de aprendizaje: supervisado y no supervisado, donde el primero se utilizará para predecir una clasificación de un dato, ya que tiene una variable respuesta. En este trabajo la variable respuesta será binaria, con el objetivo de ver si pertenece o no a un grupo de clasificación. El segundo tipo de aprendizaje se utilizar para agrupar la información y obtener relaciones entre las variables donde, en este caso, no hay variables respuesta.

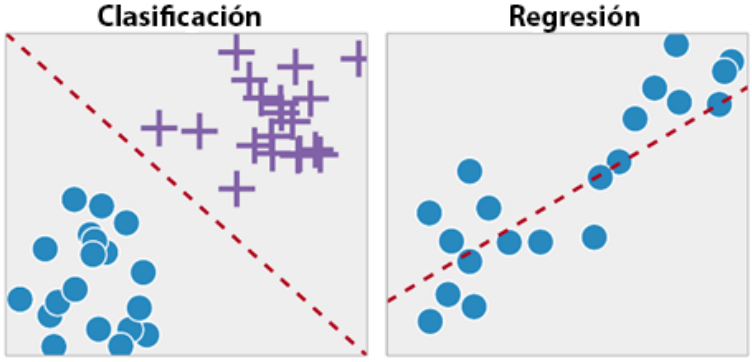


Figura 1.2: Agrupación frente a predicción [2]

1.3. Base de datos

Las Bases de Datos juegan un papel importante en la mayoría de las áreas de la sociedad actual, como la medicina, el deporte, la banca, todo tipo de empresas como las encargas de ventas de productos, marketing, etc. Una buena base de datos permite almacenar grandes volúmenes de información acerca del sector, donde el correcto tratamiento del dato debe estar en una forma que sirva para administrar, planear, controlar y tomar decisiones que ayuden, lo mejor posible, al contexto en el que se encuentre.

1.4. Estructura del trabajo

El presentente estudio tiene el siguiente esquema de trabajo:

- Explicación teórica de los modelos matemáticos utilizados.
- Construcción de una base de datos para el correcto tratamiento de los datos.
- Aplicación de estos modelos a las lesiones deportivas, en el caso concreto a jugadores de la NBA.
- Conclusiones del estudio.

Capítulo 2

Aprendizaje no supervisado

Actualmente, el procesado de información se realiza en masa, en cantidades que el ser humano no puede abarcar, por tanto, se han creado las herramientas necesarias para que las máquinas la procesen. Estas herramientas dotan a la máquina de información que aprende para poder realizar una clasificación o un análisis a futuro de una forma automatizada, dividiendo el aprendizaje de la máquina en dos grandes bloques: aprendizaje supervisado y aprendizaje no supervisado. Las técnicas de análisis de información mediante el aprendizaje no supervisado permiten explorar la base de datos sin conocer previamente un posible resultado o una clasificación, en otras palabras, se desconoce la variable respuesta del conjunto de variables que van a estudiarse, permitiendo desarrollar modelos de aprendizaje automático sin conocer a dónde llevará el análisis.

Para abordar este análisis, se estudia por separado las variables cuantitativas y cualitativas y luego se agrupan para buscar relaciones entre ellas. En primer lugar se estudiarán las variables cuantitativas, utilizando los métodos k-means y agrupación jerárquica.

2.1. k-medias (k-means)

K-means [3] es el algoritmo de aprendizaje no supervisado de uso más frecuente, siendo también, unos de los primeros algoritmos desarrollados, que consiste en la elección aleatoria de un punto inicial (centroide) desde el que se irán midiendo las distancias a los puntos vecinos, generando agrupaciones de datos. El resultado esperado es que haga agrupaciones óptimas en base a las características de dato, utilizando la minimización de la distancia cuadrática:

$$\min_{\mathrm{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

donde x_j es cada una de las observaciones conocidas del vector de observaciones $(x_1 \cdots, x_n)$, μ_i corresponde a cada centroide y $S = S_k \cdots, S_k$ es cada de los grupos a los que pertenece el centroide.

El algoritmo funciona mediante los siquientes pasos [4]:

Paso 0: se especifica el número de clusters y se eligen aleatoriamente los centroides.

Paso 1: cada dato x_j se va asignando a su centroide más cercano.

Paso 2: se actualiza la posición del centroide de cada grupo tomando como nuevo centroide la posición del promedio de los datos pertenecientes a dicho cluster. El centroide se convierte en el punto medio de las observaciones del cluster.

Paso 3: se repiten pasos 1 y 2 hasta que dejen de agruparse observaciones a clusters, se alcance una distancia mínima o se cumpla el número máximo de iteraciones.

Se visualiza el procedimiento del algoritmo:

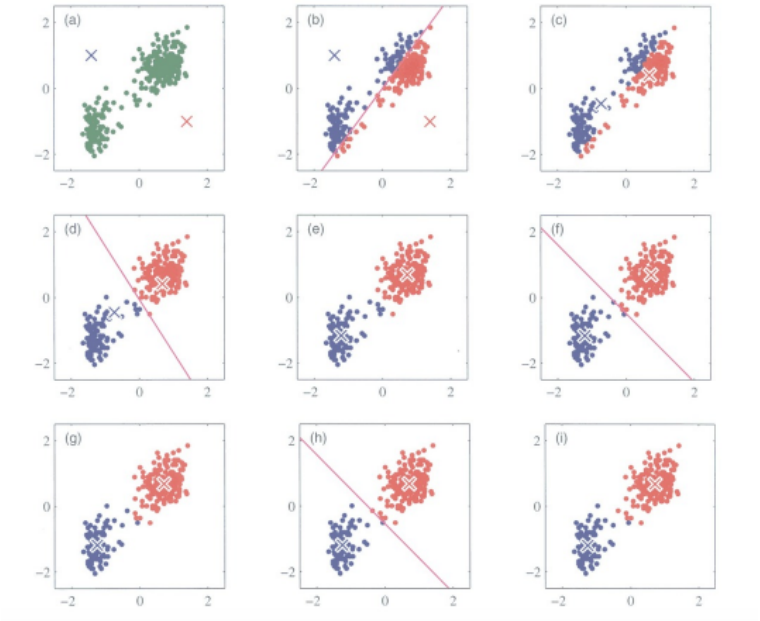


Figura 2.1: Evolución k-means [5]

2.1.1. Silhouette

Silhouette [6] [7] es un método de interpretación y validación de la coherencia de la elección de clusters para métodos de aprendizaje no supervisado.

Para abordar este algoritmo, en primer lugar, se realiza una agrupación mediante un método de clustering, en este caso k-means.

Sea $i \in C_i$ tal que i es un punto perteneciente al cluster C_i, se tiene:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

siendo $a(i)$ la distancia media entre i y todos los demás puntos de datos en el mismo cluster, donde $d(i, j)$ es la distancia entre los puntos de datos i y j en el cluster C_i (se divide por $|C_i| - 1$ porque no se incluye la distancia $d(i, i)$ en la suma). Se puede interpretar $a(i)$ como

una medida de lo bien que i está asignada a su cluster (cuanto más pequeño es el valor, mejor es la asignación).

Luego se define la diferencia media del punto i a algún cluster C como la media de la distancia desde i a todos los puntos en C (donde $C \neq C_i$).

Se define ahora

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

la menor distancia media de i a todos los puntos de cualquier otro cluster, del cual i no es miembro. Se dice que el cluster con esta diferencia media más pequeña es el cluster óptimo para i porque es el siguiente cluster que mejor se ajusta al punto i.

Ahora se define el valor del coeficiente de Silhoutte para un elemento del cluster.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$si |C_i| > 1$

y

$s(i) = 0$ si $|C_i| = 1$

Se puede ver, entonces, que los valores del coeficiente de Silhouette estarán comprendidos entre -1 y 1, siendo indicador de mayor calidad cuanto más próximo a 1 esté.

Para comparar métodos de Silhouette se aplica el siguiente indicador, que es el valor medio de los coeficientes:

$$SC = \max_k \tilde{s}(k)$$

donde $\tilde{s}(k)$ representa la media $s(i)$ sobre todos los datos de todo el conjunto de datos para un número específico de clusters k

Una representación gráfica del método para 3 cluster es la siguiente:

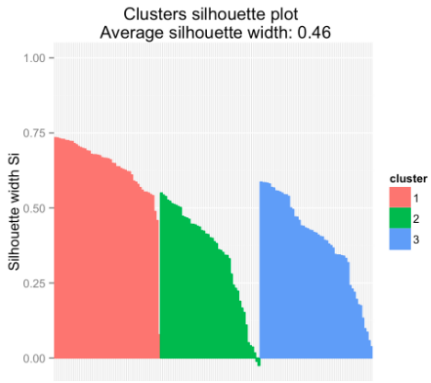


Figura 2.2: Silhouette [8]

2.2. Agrupación jerárquica (hierarchical clustering)

Otra forma de clasificación es mediante el Agrupamiento jerárquico, Clustering Jerárquico o Hierarchical Clustering [9] [10], es un método de aprendizaje no supervisado para agrupar datos en clusters. El algortimo de cluster jerárquico agrupa los datos basándose en la distancia entre cada uno y buscando que los datos que están dentro de un cluster sean los más similares entre sí que, como en k-means, utiliza la distancia euclídea para calcular la distancia entre elementos de un grupo:

$$d_E(P,Q) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_n-q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$

Esta distancia no es única, ya que pueden utlizarse distancias como la de Manhattan $\|a-b\|_1 = \sum_i |a_i - b_i|$, la distancia máxima $\|a-b\|_\infty = \max_i |a_i - b_i|$, la de Mahalanobis $\sqrt{(a-b)^\top S^{-1}(a-b)}$ donde S es la matriz de covarianza, entre otras. Posteriormente, se utilizarán distancias entre los conjuntos de datos, como la distancia mínima entre elementos de cada cluster $\min\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B}\,\}$, la distancia máxima entre elementos de cada cluster: $\max\{\, d(x,y) : x \in \mathcal{A},\, y \in \mathcal{B}\,\}$ y la distancia media entre elementos de cada cluster: $\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|}\sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{B}} d(x,y)$, siendo $\mathcal{A}$ y $\mathcal{B}$ el cluster correspondiente.

En este método, los resultados se visualizan mediante un Dendrograma, que consiste en un diagrama de árbol, que organiza los datos en subcategorías que reprsentan una agrupación por sí misma, partiendo de cluster de una única observación hasta conseguir los cluster óptimos para el problema. Este esquema de agrupación de datos permite una interpretación visual que muestra directamente la relación entre los datos, como puede verse en la siguiente imagen:

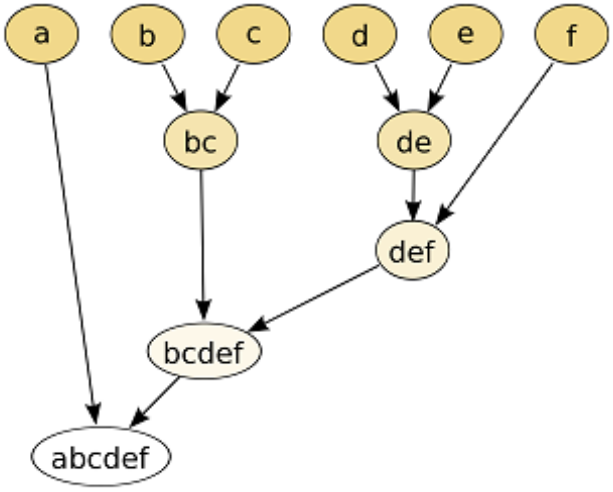


Figura 2.3: Agrupamiento Jerárquico [11]

Una vez visualizado, se puede ver cómo se agrupa en conjuntos en base a las relaciones entre observaciones. Este tipo de agrupamiento se conoce como aglomerativo, es decir, como paso inicial del algoritmo se considera que cada observación es un cluster en sí mismo, pero de un sólo elemento. Posteriormente, un cluster se junta con otro y asś sucesivamente hasta que se consigue un cierto número de cluster en los que se agrupan los datos de entrada. Se describen los pasos del algoritmo con el siguiente orden [12]:

Paso 0: Se forman tantos cluster como observaciones haya y se selcciona una distancia.
Paso 1: Se agrupan los cluster con mayor similitud, calculado con la distancia.
Paso 2: Se repite Paso 1 hasta que se completa el agrupamiento de las observaciones.

La ventaja principal de este algoritmo frente al k-means es que no se necesita seleccionar previamente un número de cluster, si no que puede obtenerse directamente del modelo.

Capítulo 3

Aprendizaje supervisado

A difrencia del aprendizaje no supervisado, estas herramientas necesitan de una variable respuesta, ya que tienen como objetivo determinar la probabilidad de que una observación se clasifique dentro de la variable respuesta. Los problemas que solucionan éstas técnicas de predicción son de ocurrencias de sucesos, viendo si el suceso de la variable respuesta ocurrirá.

3.1. Regresión Logística

Una vez visto que existe una agrupación de los los datos, se buscan diferentes predicciones mediante la regresión logística [13]. Este tipo de análisis de regresión se utiliza para predecir, con cierta probabilidad, la clasificación de una variable cualitativa en función de una serie de variables indpendientes, conocidas como predictoras, es decir, en función de unos valores ya conocidos, obtendremos la probabilidad de que unos nuevos valores se clasifiquen en una clase de la variable respuesta. La utilidad de este modelo se produce cuando la variable respuesta cumple el supuesto de 'ocurre' o 'no ocurre'. Por ejemplo:

- Una persona se contagia o no de un virus.
- Una persona comprará cierto producto o no.
- Una moneda se devaluará o no.

Lógicamente, los valores de la variable respuesta estarán dentro del intervalo $[0,1] \in \mathbb{R}$.

El origen de este método se remonta a la creación de la función logística para el estudio del crecimiento poblacional [14] en el siglo XIX, extendiendo su uso hasta finales de 1950, donde se modelizó la regresión logística simple.

En este trabajo se va a utilizar la regresión logísitica tanto con una variable respuesta binaria como con una multinomial.

La ecuación de la recta de regresión con n variables predictoras viene expresada como:

$$\mathrm{Y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n.$$

La función sobre la que se apoyará el ajuste de predicción, denominada *logit*, se define como:

$$f(t) = \frac{1}{1 + e^{-t}}$$

El objetivo es juntar la ecuación y el modelo para obtener la probabilidad de clasificación. Sustituyendo en la función se obtiene:

$$Y = \frac{1}{1 + e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}.$$

Ahora despejando para obtener la fórmula de la ecuación se llega a la siguiente regresión:

$$\ln\left(\frac{Y}{1-Y}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n.$$

Vista la anterior ecuación, se selecciona una nueva variable "p"para hallar la probabilidad de que ocurra un suceso. Primero se hace el cambio oportuno:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n.$$

Esta ecuación tiene los siguientes parámetros:

- Variable dependiente conocida como logaritmo de razón de monomios, donde "prepresenta la probabilidad de que ocurra el suceso y "1-p"la probabilidad de que no ocurra.

-Parámetros beta a estimar, cuyo ajuste se realiza a través del estimador de máxima verosimilitud. [15]

-Variables depenedientes conocidas.

Ahora, mediante operaciones elementales, se llega a la fórmula de la probabilidad:

$$p = \frac{e^{(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}.$$

¿Por qué utilizar este modelo frente a la regresión lineal al uso, que también permite una predicción y tiene una modelización más sencilla? Para responder a esta pregunta, se ilustran dos sucesos:

Comparación de la Edad y los años en la NBA:

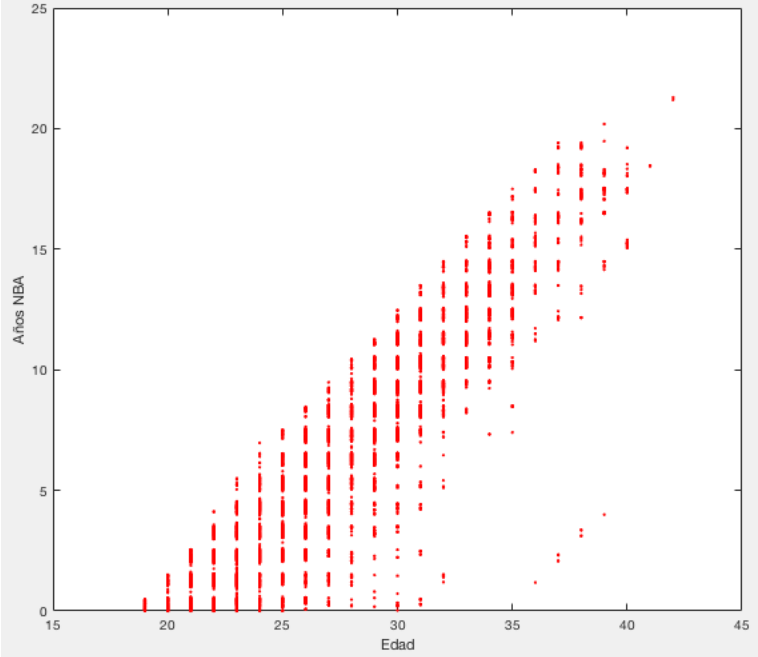


Figura 3.1: Edad vs Años NBA

No habría problema en utilizar un modelo de regresión lineal para clasificar estas dos variables objeto de estudio, sin embargo, al tener variables cualitativas como la posción en la que desempeña su papel cada jugador, se visualiza el siguiente gráfico comparativo entre la Edad del jugador y su Posición:

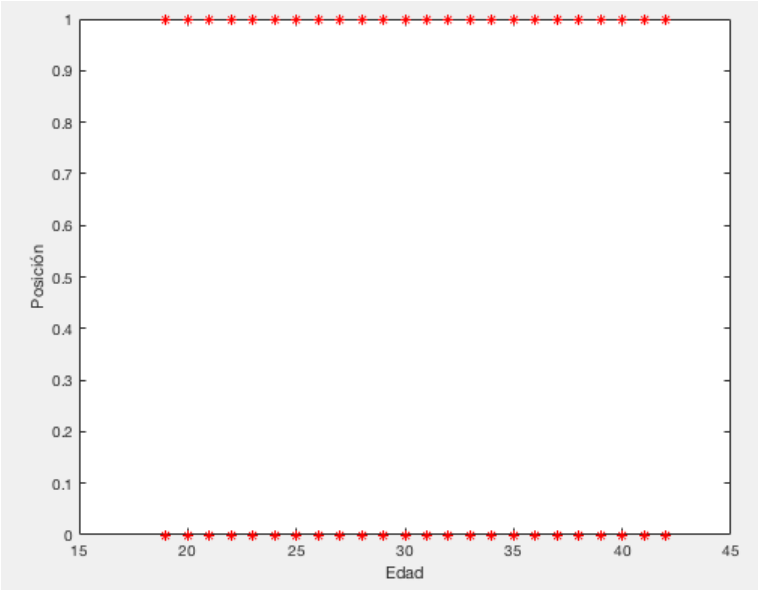


Figura 3.2: Edad vs Posición

Como se observa, ajustar una recta de regresión lineal es inviable en esta situación. Ahora, viendo el comportamiento de la función *logit*, se entiende su uso:

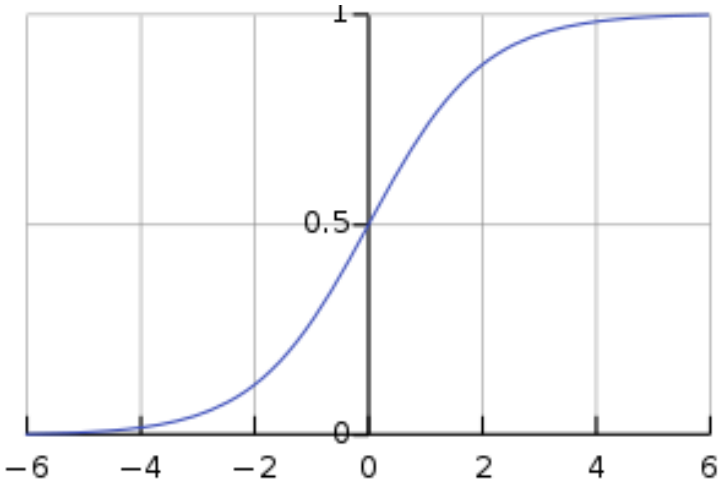


Figura 3.3: Curva Logit [16]

Esta curva irá enviando al eje x, al igual que en una proyección, los nuevos valores que se vayan agregando al conjunto de estudio.

Junto con los siguientes modelos, formarán los clasificadores de este trabajo.

3.2. SVM

Para abordar el problema de clasificación y optimizarlo interesa aplicar más de una técnica, comparando la calidad del ajuste con los resultados de las otras técnicas. Este método de clasificación surge de los laboratorios de NOKIA [17] durante el siglo XX, buscando una clasificación óptima de la información, formando una agrupación de algoritmos de aprendizaje supervisado, obteniendo como resultado una clasificación binaria [10].

El fundamento del algorimto es el siguiente: Realizar una partición del conjunto de puntos (entendidos a partir de ahora como vectores) pertenecientes al espacio n-dimensional objeto de estudio mediante hiperplanos, viendo mediante el margen máximo entre ambos vectores y el hiperplano dónde se clasifica mejor el vector. Matemáticamente hablando, se tiene la ecuación del hiperplano [18]:

$$a_1x_1 + a_2x_2 + ... + a_nx_n = b$$

entendida como la siguiente función:

$$f(x) =< a, x > +b = \sum_{i=1}^{n} a_i x_i + b$$

donde, a partir del resultado de la función, se tendrá que $f(x) \geq 0$ será considerado como resultado positivo, y negativo en caso contrario. Se utiliza la función signo $h(x) = sgn(f(x))$ que verifica que x = 1 si es mayor o igual que 0, y x = -1 si x menor que 0, siendo de utilidad para clasificar el vector en una u otra clase, positiva o negativa, de la partición del espacio, siendo x el vector a clasificar. Una representación visual de este método permitirá entenderlo mejor.

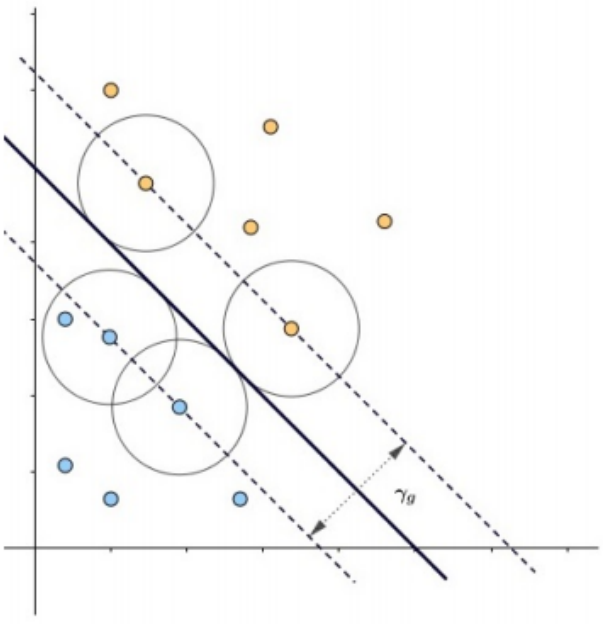


Figura 3.4: SVM [18]

En el gráfico, ejemplarizando $\mathbb{R}^2$, se ven los vectores azules y amarillos, representando la clasificación de éstos en dos clases distintas mediante los colores. La recta negra representa el hiperplano, que divide al espacio en dos partes, y la función γ_g representando el margen, que es la anchura máxima de la región paralela al hiperplano que no tiene puntos de datos interiores, definida como:

$$\gamma_g = \min_{+}(d(a, b; x_+)) - \max_{-}(d(a, b; x_-))$$

siendo x_+ los vectores de la clase positiva y x_- vectores de la clase negativa.

Al trabajar en una situación de ajuste no lineal, este problema de clasificación binaria no tiene un hiperplano simple como criterio de separación útil. Para este problema, hay una variante del enfoque matemático que conserva casi toda la simplicidad de un hiperplano de separación SVM, que se consigue a través de las funciones Kernel. Estas funciones asignan un valor de salida numérico real correspondiente a dos vectores de entrenamiento. Ejemplos de funciones Kernel:

Sigmoide: $K(x, y) = \tanh(\beta_0 x^{\mathsf{T}} y + \beta_1)$

Polinómica: $K(x, y) = (x^{\mathsf{T}} y + 1)^{\rho}$

Gaussiana: $K(x, y) = \exp(-\frac{\|x - y\|^2}{2\sigma^2})$

3.3. Naive Bayes

Este método de clasificación, como su propio nombre indica, surge del teorema de Bayes. Utilizar la probabilidad condicionada a un suceso será la base del clasificador, aunque, a pesar de ser sencillo conceptual y computacionalmente, los resultados demuestran que es un algoritmo robusto. La idea del clasificador es la siguiente:

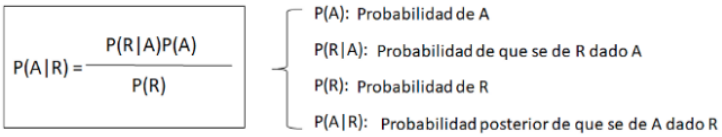


Figura 3.5: Bayes

Una vez visto este concepto, se amplía a un conjunto de sucesos independientes entre sí y se desarrolla hasta llegar al clasificador Naive Bayes [19] [20]. Generalizando el teorema se obtiene:

$$P(C|F_1, \ldots, F_n) = \frac{P(C)\ P(F_1, \ldots, F_n|C)}{P(F_1, \ldots, F_n)}.$$

A partir de este punto se inicia el clasificador, centrándose el estudio en el numerador, ya que el denominador no está condicionando a C. El numerador es equivalente a una probabilidad compuesta $P(C, F_1, \ldots, F_n)$, que puede, a partir de la definición de probabilidad condicionada, reescribirse de las siguientes formas:

$$\begin{aligned} P(C, F_1, \ldots, F_n) &= P(C)\ P(F_1, \ldots, F_n|C) = P(C)\ P(F_1|C)\ P(F_2, \ldots, F_n|C, F_1) \\ &= P(C)\ P(F_1|C)\ P(F_2|C, F_1)\ P(F_3, \ldots, F_n|C, F_1, F_2) \\ &= P(C)\ P(F_1|C)\ P(F_2|C, F_1)\ P(F_3|C, F_1, F_2)\ P(F_4, \ldots, F_n|C, F_1, F_2, F_3) \end{aligned} \tag{3.1}$$

y así sucesivamente.
Ahora se aplica la hipótesis de independencia de sucesos. Se asume que cada F_i es independiente de cualquier otra F_j para $j \neq i$ cuando están condicionadas a C. El resultado inmediato es $P(F_i|C, F_j) = P(F_i|C)$, donde, expresando la probabilidad condicionada como se ha visto antes se obtiene

$$P(C, F_1, \ldots, F_n) = P(C)\ P(F_1|C)\ P(F_2|C)\ P(F_3|C)\ \cdots = P(C) \prod_{i=1}^{n} P(F_i|C).$$

Una vez vistos estos supuestos, el numerador puede expresarse de la siguiente forma:

$$P(C, F_1, \ldots, F_n) = \frac{1}{Z} P(C) \prod_{i=1}^{n} P(F_i|C),$$

donde Z es un factor que depende únicamente de los valores F_i conocidos.

Este clasificador es iterativo [21], ya que va comparando probabilidades de que 'ocurra' en función del valor de la variable respuesta, condicionado a las observaciones. Cuando la probabilidad de que 'ocurre' es mayor que 'no ocurre', se clasifica como 'ocurre'.

3.4. Curva ROC

Dentro de los estudios estadísticos se busca optimizar la estimación de los valores obtenidos, pudiendo alcanzar, así, el resultado más preciso posible. La búsqueda de esa estimación se realiza mediante la curva ROC (Receiver Operating Characteristic o Característica Operativa del Receptor) [22] [10], que permite visualizar los resultados en distintos ámbitos. En este caso, la calidad de los ajustes estadísticos en un estudio de clasificadores. Para caracterizar el comportamiento de la sensibilidad (razón de Verdaderos Positivos) frente a la especificidad (razón de Verdaderos Negativos) se utiliza la siguiente tabla:

	Negativos	Positivos	Total
Negativos	Verdadero negativo (VN)	Falso positivo (FP)	Total negativos
Positivos	Falso negativo (FN)	Verdadero positivo (VP)	Total positivos
Total	Negativos predichos	Positivos predichos	Total observaciones

Figura 3.6: Sensibilidad vs Especificidad

Cuando se realiza una clasificación con resultado binario se tienen en cuenta los escenarios posibles mostrados en la tabla, explicando que los resultados se etiquetan como positivos (p) o negativos (n) [23]. Hay cuatro posibles resultados a partir de un clasificador binario como el propuesto. Si el resultado de una exploración es p y el valor dado es también p, entonces se conoce como un Verdadero Positivo (VP); sin embargo si el valor real es n entonces se conoce como un Falso Positivo (FP). De igual modo, tenemos un Verdadero Negativo (VN) cuando tanto la exploración como el valor dado son n, y un Falso Negativo (FN) cuando el resultado de la predicción es n pero el valor real es p.

Una vez entendida la sensibilidad vs la especificidad se genera una curva ROC para interpretar gráficamente los resultados, cuya visualización tendrá ciertas características a tener en cuenta para entender la calidad del modelo.

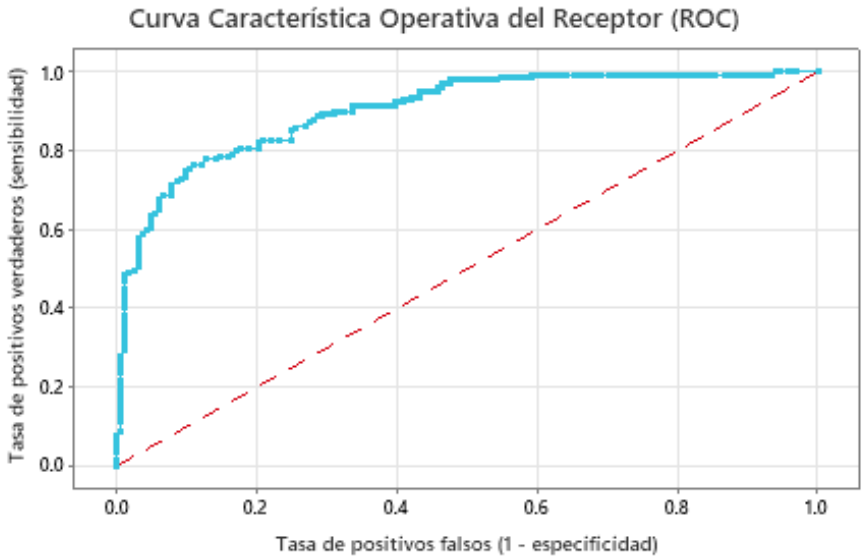


Figura 3.7: Curva ROC [24]

Por una parte, las curvas que queden por debajo de la recta, incluida ésta, serán clasificaciones aleatorias y no habrá calidad en el ajuste. Aparte de visualizarlo con la recta, numéricamente se ve el área que abarca la curva (AUC), considerándose un ajuste de mala calidad (aleatorio) AUC = 0.5 o menor, y va mejorando hasta la perfección, que sería AUC = 1.

Capítulo 4

Aplicación a lesiones deportivas

4.1. Base de datos

Realizar una exploración de una base de datos completa y ajustada a las necesidades del estudio será fundamental para encontrar la información de forma precisa, pudiendo extraer unas conclusiones lo más acertadas posible. Para llevar a cabo este objetivo, he construido la base de datos que voy a utilizar a partir de la base de datos original extraída de Kaggle [25]. La base de datos original tiene las siguientes variables: Equipo, Regreso, Jugadores, Lesiones, Fecha. Para ampliar los resultados de la exploración y obtener resultados más completos he añadido las siguientes variables: Edad, Altura, Peso, AnosNba, Color, Posición, Lateralidad. Se puede observar que tenemos dos tipos de variables objeto de estudio:

Cualitativas:
Equipo: A qué equipo pertenece el jugador lesionado.
Regreso: Regreso a la pista del jugador.
Jugadores: Nombre de los jugadores lesionados.
Lesiones: Qué tipo de lesión es.
Fecha: Cuándo se ha lesionado y cuándo ha regresado el jugador a la pista.
Color: Color de piel del jugador, que puede ser blanco, negro, moreno o asiático.
Posicion: En qué posición juega, que puede ser una o dos: Base(B), Escolta(ES), Alero(AL), Pivot(P), Ala-Pivot(AP), Escolta/Base(ES/B), Escolta/Alero(ES/AL), Alero/Ala-Pivot(AL/AP), Pivot/Ala-Pivot(P/AP).
Lateralidad: Indica si el jugador es diestro(1) o zurdo(0).
Lesion previa: Si ha tenido una lesión anteriormente.
Lesion repetida: Si ha tenido ya esa lesión.
LocalizacionParte: Parte del cuerpo donde se lesiona el juagador
Impacto: Gravedad de la lesión, creado mediante (DTD), (DNP), (out for season), (out indefinitely).

Cualitativas:
Edad: Edad que tiene el jugador.
Altura: Altura del jugador, expresada en centímetros.
Peso: Peso del jugador, expresado en kilos.
AnosNba: Período de tiempo en la NBA.

Tendremos 14 variables y 9410 observaciones. Para construir la base de datos, en primer lugar se ha utilizado la base de lesionados de Kaggle [25], de donde se ha sacado el jugador, la fecha de lesión, la vuelta a la pista del jugador y el tipo de lesión. En segundo lugar, se ha extraído la información sobre la fecha de nacimiento (que se utiliza en la variable de edad, haciendo la diferencia entre la fecha actual y la fecha de nacimiento) y peso (en libras), de la página Basketball Reference [26]". En tercer lugar, los datos de altura, la posción en la que juega y el periodo de tiempo de juego en la NBA del jugador, los he sacado de Hispanos NBA. [27]

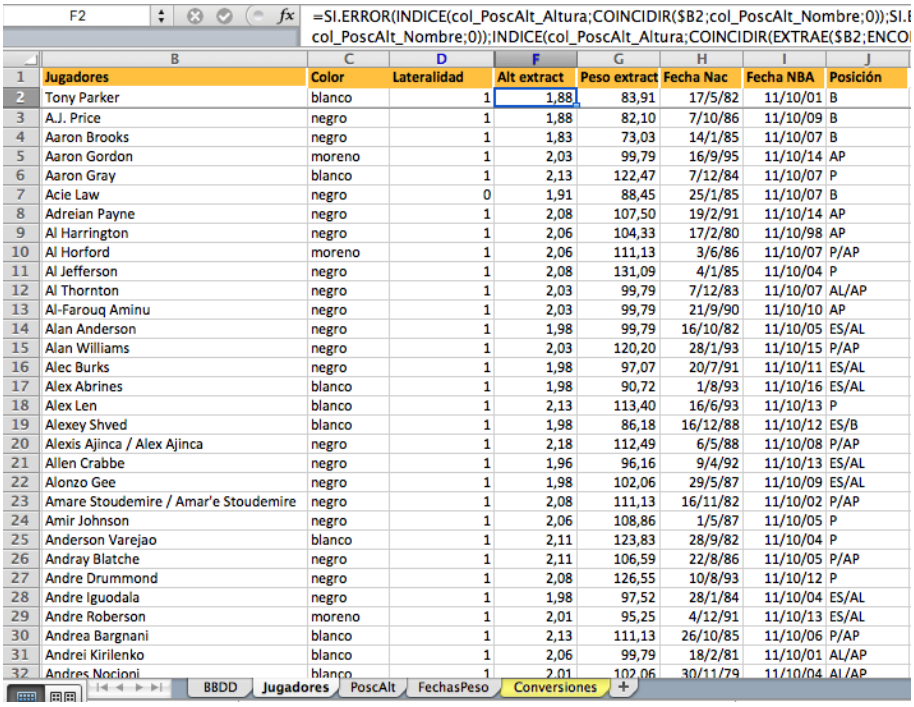
Para poder generar la base de datos con todas las variables, he utilizado un fichero de Excel, donde se ha agrupado toda la información en una única tabla, creando el fichero en un archivo CSV que se utilizará más adelante para el análisis. El objetivo del uso de Excel es poder automatizar el proceso de rellenar la base de datos con fórmulas para evitar ir jugador a jugador rellenando todo a mano y uno a uno.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Equipo	Regreso	Jugadores	Lesiones	Fecha	Edad	Altura	Peso	Años nba	Color	Posicion	Lateralidad	Lesion previa
2	Pacers		A.J. Price	strep throat (DNP)	24/1/12	25	1,88	82	2,29	negro	B	1	no
3	Pacers		A.J. Price	strep throat (DNP)	25/1/12	25	1,88	82	2,29	negro	B	1	si
4	Wizards		A.J. Price	fractured right hand (DNP)	11/12/12	26	1,88	82	3,17	negro	B	1	si
5	Wizards		A.J. Price	groin injury (DNP)	13/3/13	26	1,88	82	3,42	negro	B	1	si
6	Wizards		A.J. Price	sore right groin (DNP)	15/3/13	26	1,88	82	3,43	negro	B	1	si
7	Wizards		A.J. Price	sore right knee (DNP)	16/3/13	26	1,88	82	3,43	negro	B	1	si
8	Kings		Aaron Brooks	sore left ankle (DNP)	30/12/12	27	1,83	73	5,22	negro	B	1	no
9	Rockets		Aaron Brooks	tendinitis in knee (DNP)	15/1/14	29	1,83	73	6,26	negro	B	1	si
10	Bulls	Aaron Brooks	Aaron Brooks	returned to lineup	2/12/15	30	1,83	73	8,14	negro	B	1	si
11	Bulls		Aaron Brooks	strained left hamstring (DTD)	5/12/15	30	1,83	73	8,15	negro	B	1	si
12	Pacers		Aaron Brooks	sore knee (DTD)	26/10/16	31	1,83	73	9,04	negro	B	1	si
13	Magic		Aaron Gordon	fractured left foot (DTD)	16/11/14	19	2,03	100	0,10	moreno	AP	1	no
14	Magic		Aaron Gordon	ankle injury (DTD)	8/12/15	20	2,03	100	1,16	moreno	AP	1	si
15	Magic	Aaron Gordon	Aaron Gordon	returned to lineup	9/12/15	20	2,03	100	1,16	moreno	AP	1	si
16	Magic		Aaron Gordon	concussion (DTD)	7/4/16	20	2,03	100	1,49	moreno	AP	1	si
17	Magic		Aaron Gordon	bone bruise in right foot (DTD)	8/2/17	21	2,03	100	2,33	moreno	AP	1	si
18	Magic	Aaron Gordon	Aaron Gordon	returned to lineup	13/2/17	21	2,03	100	2,34	moreno	AP	1	si
19	Magic		Aaron Gordon	concussion (DTD)	9/12/17	22	2,03	100	3,16	moreno	AP	1	si
20	Magic		Aaron Gordon	strained right calf (DTD)	16/12/17	22	2,03	100	3,18	moreno	AP	1	si
21	Magic		Aaron Gordon	strained left hip flexor (DTD)	30/1/18	22	2,03	100	3,30	moreno	AP	1	si
22	Magic	Aaron Gordon	Aaron Gordon	returned to lineup	22/2/18	22	2,03	100	3,37	moreno	AP	1	si
23	Magic		Aaron Gordon	concussion (DTD)	9/3/18	22	2,03	100	3,41	moreno	AP	1	si
24	Magic	Aaron Gordon	Aaron Gordon	returned to lineup	19/3/18	22	2,03	100	3,44	moreno	AP	1	si
25	Magic		Aaron Gordon	sore right calf (DTD)	6/4/18	22	2,03	100	3,49	moreno	AP	1	si
26	Magic	Aaron Gordon	Aaron Gordon	returned to lineup	8/4/18	22	2,03	100	3,49	moreno	AP	1	si
27	Raptors		Aaron Gray	flu (DNP)	6/1/13	28	2,13	122	5,24	blanco	P	1	no
28	Kings		Aaron Gray	DNP	11/12/13	29	2,13	122	6,17	blanco	P	1	si
29	Kings		Aaron Gray	illness (DNP)	31/3/14	29	2,13	122	6,47	blanco	P	1	si
30	Kings		Aaron Gray	stomach illness (DNP)	2/4/14	29	2,13	122	6,48	blanco	P	1	si
31	Kings		Aaron Gray	illness (DNP)	4/4/14	29	2,13	122	6,48	blanco	P	1	si
32	Warriors		Acie Law	surgery on right wrist to remove bone spur (out for se	7/4/11	26	1,91	88	3,49	negro	B	0	no
33	Timberwolves		Adreian Payne	thrombocytopenia (blood disorder) (DTD)	3/2/17	25	2,08	108	2,32	negro	AP	1	no

BBDD | Jugadores | PoscAlt | FechasPeso | Conversiones

Figura 4.1: Base de Datos

La idea para completar la base de datos mediante el documento de Excel, ha sido generar una ficha para cada jugador con la información objeto de estudio de sus características, siendo en este caso las siguientes: Jugador, Color, Lateralidad, Altura, Peso, Fecha de nacimiento, Fecha de incorporación en la NBA y Posición. Una vez generada la ficha, será más sencillo exportar la información a la base de datos con toda la información



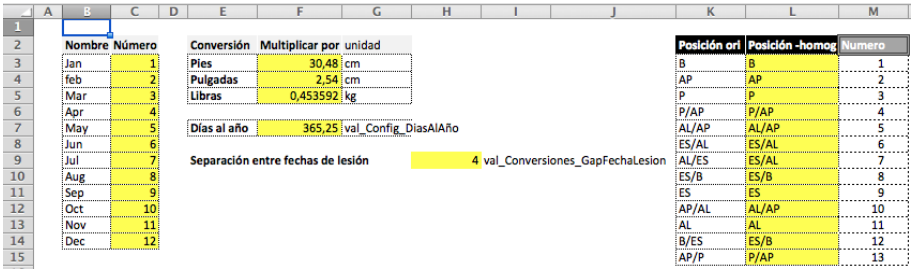
F2 =SI.ERROR(INDICE(col_PoscAlt_Altura;COINCIDIR($B2;col_PoscAlt_Nombre;0));SI.E
col_PoscAlt_Nombre;0));INDICE(col_PoscAlt_Altura;COINCIDIR(EXTRAE($B2;ENCO

	B	C	D	F	G	H	I	J
1	Jugadores	Color	Lateralidad	Alt extract	Peso extract	Fecha Nac	Fecha NBA	Posición
2	Tony Parker	blanco	1	1,88	83,91	17/5/82	11/10/01	B
3	A.J. Price	negro	1	1,88	82,10	7/10/86	11/10/09	B
4	Aaron Brooks	negro	1	1,83	73,03	14/1/85	11/10/07	B
5	Aaron Gordon	moreno	1	2,03	99,79	16/9/95	11/10/14	AP
6	Aaron Gray	blanco	1	2,13	122,47	7/12/84	11/10/07	P
7	Acie Law	negro	0	1,91	88,45	25/1/85	11/10/07	B
8	Adreian Payne	negro	1	2,08	107,50	19/2/91	11/10/14	AP
9	Al Harrington	negro	1	2,06	104,33	17/2/80	11/10/98	AP
10	Al Horford	moreno	1	2,06	111,13	3/6/86	11/10/07	P/AP
11	Al Jefferson	negro	1	2,08	131,09	4/1/85	11/10/04	P
12	Al Thornton	negro	1	2,03	99,79	7/12/83	11/10/07	AL/AP
13	Al-Farouq Aminu	negro	1	2,03	99,79	21/9/90	11/10/10	AP
14	Alan Anderson	negro	1	1,98	99,79	16/10/82	11/10/05	ES/AL
15	Alan Williams	negro	1	2,03	120,20	28/1/93	11/10/15	P/AP
16	Alec Burks	negro	1	1,98	97,07	20/7/91	11/10/11	ES/AL
17	Alex Abrines	blanco	1	1,98	90,72	1/8/93	11/10/16	ES/AL
18	Alex Len	blanco	1	2,13	113,40	16/6/93	11/10/13	P
19	Alexey Shved	blanco	1	1,98	86,18	16/12/88	11/10/12	ES/B
20	Alexis Ajinca / Alex Ajinca	negro	1	2,18	112,49	6/5/88	11/10/08	P/AP
21	Allen Crabbe	negro	1	1,96	96,16	9/4/92	11/10/13	ES/AL
22	Alonzo Gee	negro	1	1,98	102,06	29/5/87	11/10/09	ES/AL
23	Amare Stoudemire / Amar'e Stoudemire	negro	1	2,08	111,13	16/11/82	11/10/02	P/AP
24	Amir Johnson	negro	1	2,06	108,86	1/5/87	11/10/05	P
25	Anderson Varejao	blanco	1	2,11	123,83	28/9/82	11/10/04	P
26	Andray Blatche	negro	1	2,11	106,59	22/8/86	11/10/05	P/AP
27	Andre Drummond	negro	1	2,08	126,55	10/8/93	11/10/12	P
28	Andre Iguodala	negro	1	1,98	97,52	28/1/84	11/10/04	ES/AL
29	Andre Roberson	moreno	1	2,01	95,25	4/12/91	11/10/13	ES/AL
30	Andrea Bargnani	blanco	1	2,13	111,13	26/10/85	11/10/06	P/AP
31	Andrei Kirilenko	blanco	1	2,06	99,79	18/2/81	11/10/01	AL/AP
32	Andres Nocioni	blanco	1	2,01	102,06	30/11/79	11/10/04	AL/AP

BBDD | Jugadores | PoscAlt | FechasPeso | Conversiones | +

Figura 4.2: Ficha de Jugadores

El contenido de las fichas proviene de las fuentes anteriormente mencionadas, pero hay que depurar la información y adaptarla a las unidades que se estudiarán. En este caso, la información de las fichas viene de dos ventanas distintas dentro del documento Excel: 'PoscAlt' y 'FechasPeso'. También puede verse la ventana de 'Conversiones', pero ésta se utiliza para convertir los valores de las magnitudes americanas en las europeas, homogeneizar las posiciones de los jugadores, asociar los meses del año con un número y dar un valor constante a la separación entre fechas.
Para las conversiones y fórmulas voy a utilizar los siguientes recursos:
Factores de conversión de unidades:
- val_Pound2Kg: valor de conversión de libras a kilos.
- val_Inch2cm: valor de conversión de pulgadas a centímetros.
- val_Pies2cm: valor de conversión de pies a centímetros.
- val_Config_DiasAlAño: valor de conversión de tiempo con 365,25.



Nombre	Número
Jan	1
feb	2
Mar	3
Apr	4
May	5
Jun	6
Jul	7
Aug	8
Sep	9
Oct	10
Nov	11
Dec	12

Conversión	Multiplicar por	unidad
Pies	30,48	cm
Pulgadas	2,54	cm
Libras	0,453592	kg

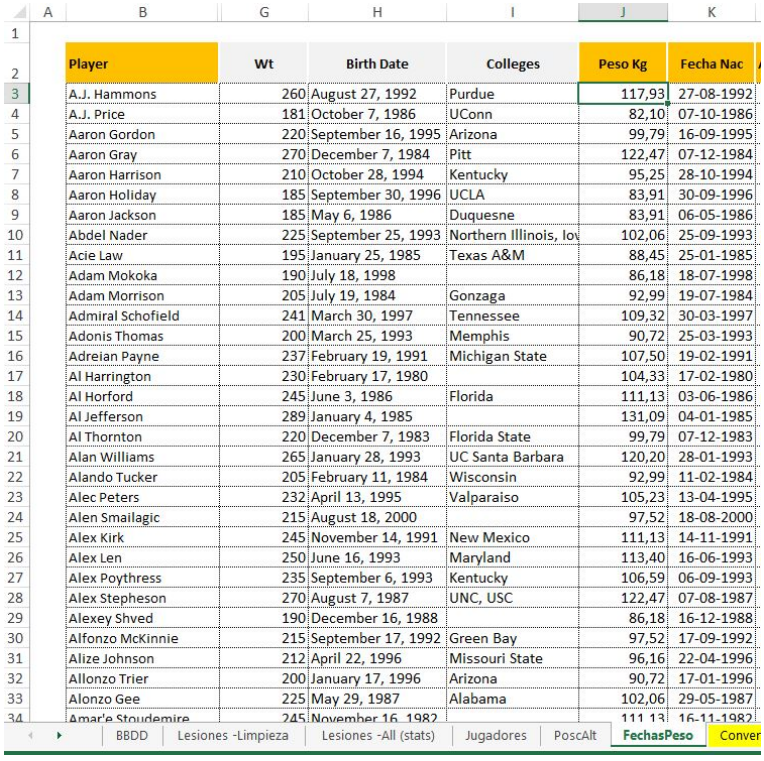
Días al año	365,25	val_Config_DiasAlAño
Separación entre fechas de lesión	4	val_Conversiones_GapFechaLesion

Posición ori	Posición -homog	Numero
B	B	1
AP	AP	2
P	P	3
P/AP	P/AP	4
AL/AP	AL/AP	5
ES/AL	ES/AL	6
AL/ES	ES/AL	7
ES/B	ES/B	8
ES	ES	9
AP/AL	AL/AP	10
AL	AL	11
B/ES	ES/B	12
AP/P	P/AP	13

Figura 4.3: Conversiones

Para listas de información:
- lst_Config-Mes3LetrasEN: Lista con las tres primeras letras de los meses en inglés.
- lst_Config-MesNum: Lista numérica con los meses del año.
- lst_Config-PosHomog: Lista de las posiciones homogeneizadas de los jugadores.
- lst_Config-PosOrig: Lista con las posiciones de los jugadores según se ha recibido de la fuente.

La pestaña de FechasPeso contiene, para cada jugador, el peso y la fecha de nacimiento:



Player	Wt	Birth Date	Colleges	Peso Kg	Fecha Nac
A.J. Hammons	260	August 27, 1992	Purdue	117,93	27-08-1992
A.J. Price	181	October 7, 1986	UConn	82,10	07-10-1986
Aaron Gordon	220	September 16, 1995	Arizona	99,79	16-09-1995
Aaron Gray	270	December 7, 1984	Pitt	122,47	07-12-1984
Aaron Harrison	210	October 28, 1994	Kentucky	95,25	28-10-1994
Aaron Holiday	185	September 30, 1996	UCLA	83,91	30-09-1996
Aaron Jackson	185	May 6, 1986	Duquesne	83,91	06-05-1986
Abdel Nader	225	September 25, 1993	Northern Illinois, Iov	102,06	25-09-1993
Acie Law	195	January 25, 1985	Texas A&M	88,45	25-01-1985
Adam Mokoka	190	July 18, 1998		86,18	18-07-1998
Adam Morrison	205	July 19, 1984	Gonzaga	92,99	19-07-1984
Admiral Schofield	241	March 30, 1997	Tennessee	109,32	30-03-1997
Adonis Thomas	200	March 25, 1993	Memphis	90,72	25-03-1993
Adreian Payne	237	February 19, 1991	Michigan State	107,50	19-02-1991
Al Harrington	230	February 17, 1980		104,33	17-02-1980
Al Horford	245	June 3, 1986	Florida	111,13	03-06-1986
Al Jefferson	289	January 4, 1985		131,09	04-01-1985
Al Thornton	220	December 7, 1983	Florida State	99,79	07-12-1983
Alan Williams	265	January 28, 1993	UC Santa Barbara	120,20	28-01-1993
Alando Tucker	205	February 11, 1984	Wisconsin	92,99	11-02-1984
Alec Peters	232	April 13, 1995	Valparaiso	105,23	13-04-1995
Alen Smailagic	215	August 18, 2000		97,52	18-08-2000
Alex Kirk	245	November 14, 1991	New Mexico	111,13	14-11-1991
Alex Len	250	June 16, 1993	Maryland	113,40	16-06-1993
Alex Poythress	235	September 6, 1993	Kentucky	106,59	06-09-1993
Alex Stepheson	270	August 7, 1987	UNC, USC	122,47	07-08-1987
Alexey Shved	190	December 16, 1988		86,18	16-12-1988
Alfonzo McKinnie	215	September 17, 1992	Green Bay	97,52	17-09-1992
Alize Johnson	212	April 22, 1996	Missouri State	96,16	22-04-1996
Allonzo Trier	200	January 17, 1996	Arizona	90,72	17-01-1996
Alonzo Gee	225	May 29, 1987	Alabama	102,06	29-05-1987
Amar'e Stoudemire	245	November 16, 1982		111,13	16-11-1982

Figura 4.4: FechasPeso

En primer lugar, se convierten los pesos que están en libras a kilos. En segundo lugar, se construye la fecha de nacimiento con el formato que interesa, ya que el extraído de la fuente no puede utilizarse como fecha y se necesitará para luego sacar la edad del jugador. Para ello, se convierte el formato de fecha original, el de la columna 'Birth Date', al de la 'Fecha Nac', crando la siguiente función:

```
=SI.ERROR(FECHA(DERECHA(H11;4);INDICE(lst_Config_MesNum;COINCIDIR(EXTRAE(H11;1;3);lst_Config_Mes3LetrasEN;0));EXTRAE(H11;
ENCONTRAR(" ";H11)+1;2));H11)
```

Figura 4.5: Función 'extraer fecha'

El formato de fecha que se construye es el utilizado en la función FECHA, implementada en Excel, que tiene primero el año, luego el mes y luego el día, luego la fórmula está orientada a este formato.

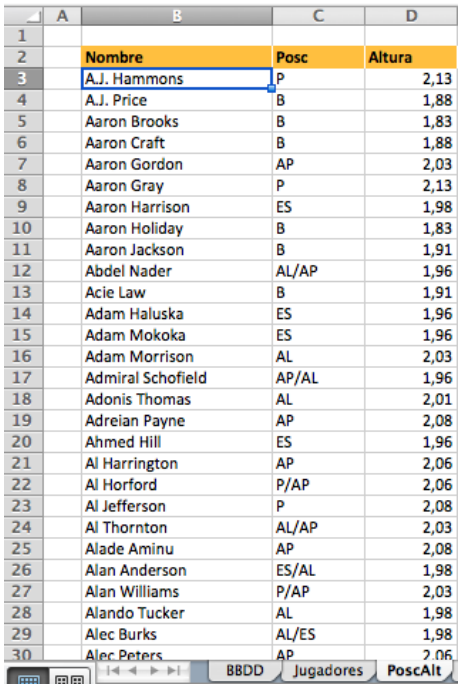
Las funciones implementadas ya por Excel que se utilizan son las siguientes:

SI.ERROR: Permite buscar y controlar el error de la fórmula. Tiene dos argumentos, el primero es la fórmula donde busca el error y el segundo el valor que devuelve si hay error. Si no hay error en el análisis, devuelve la fórmula donde ha evaluado error. Se utiliza como primer argumento la fórmula, en este caso la de extraer fecha, y como segundo la celda correspondiente a la fecha inicial. Esta función permite encontrar más fácilmente dónde puede estar el error en caso de haberlo.
FECHA: Toma tres valores numéricos y los convierte en una fecha interpretable por Excel. Se introducen tres argumentos, primero año, luego mes y luego día y se reordena en día, mes y año.
DERECHA: Selecciona los últimos 'n' caracteres de la cadena que se indique. El primer argumento es la cadena y el segundo el número de caracteres. Se utiliza para extraer el año de nacimiento.
INDICE: Devuelve el valor de una tabla. Tiene tres argumentos, la matriz en la que busca, la fila y la columna. En este caso serán dos argumentos porque la matriz es una lista. Se utiliza para extraer el mes, haciendo una correspondencia entre las tres primeras letras del nombre del mes en inglés y el valor numérico del mes. La función devuelve el número del mes.
COINCIDIR: Busca un elemento en un intervalo de celdas y devuelve la posición del elemento. Tiene tres argumentos, el primero es el valor buscado, el segundo es la matriz en la que se busca, que en este caso es una lista, y el tercero es cómo se comporta la coincidencia. En este caso es 0 porque el valor buscado puede estar en cualquier orden. Se utiliza para extraer la posición en la que coinciden las tres primeras cifras del nombre en inglés con la lista ordenada de las tres primeras letras de los meses ordenados en inglés.
EXTRAE: Devuelve un número específico de caracteres de una cadena de texto. Tiene tres argumentos, siendo el primero el texto, el segundo la posición inicial y el tercero el número de caracteres. Se utiliza para sacar las tres primeras letras del nombre del mes inglés. También se utiliza para extraer el día de nacimiento.

ENCONTRAR: Busca una cadena de texto dentro de otra cadena de texto y devuelve el número de la posición inicial en la que se encuentra. Se usan dos argumentos, el espacio en blanco y la celda correspondiente. Se utiliza para que devuelva la posición del espacio en blanco más una posición para que empiece a seleccionar la primera cifra del día de nacimiento y luego con la función EXTRAE se devuelve el día entero.

Una vez creada esta fórmula, se empieza en la primera celda de la columna 'Fecha Nac'y se arrastra hacia abajo para que se rellenen todas las celdas creando, así, la columna entera de fechas de nacimiento.

Ahora se pasa a la pestaña de Jugadores, donde se crea la ficha de cada jugador, incluyendo todas las variables relacionadas con él. Antes de comenzar se presentó el siguiente problema: dentro de los nombres de los jugadores, algunos se habían cambiado el nombre y las fuentes de información de donde se sacó la información de las nuevas variables a veces salía un nombre y otras veces otro. Para solucionarlo, dentro de la fórmula de Excel se ha incluido que, en caso de encontrar un error producido por la separación de los nombres, se distingan dos casos, uno en el que se lee el nombre empezando desde la primera posición del primer carácter del nombre hasta el último carácter (que coincide con la posición de la barra menos el espacio vacío entre la barra y el último carácter) y si da error, que empiece a leer desde la barra de separación y un espacio hasta el final del último carácter, siendo la posición de éste la denotada por LARGO, que aprovechando que da la longitud de la celda correspondiente podemos leer hasta el extremo de la celda.
Para seguir alimentando la ficha de los jugadores, me apoyo de la ventana 'PoscAlt', que contiene las posiciones homogeneizadas de los jugadores y su altura.



	A	B	C	D
1				
2		**Nombre**	**Posc**	**Altura**
3		A.J. Hammons	P	2,13
4		A.J. Price	B	1,88
5		Aaron Brooks	B	1,83
6		Aaron Craft	B	1,88
7		Aaron Gordon	AP	2,03
8		Aaron Gray	P	2,13
9		Aaron Harrison	ES	1,98
10		Aaron Holiday	B	1,83
11		Aaron Jackson	B	1,91
12		Abdel Nader	AL/AP	1,96
13		Acie Law	B	1,91
14		Adam Haluska	ES	1,96
15		Adam Mokoka	ES	1,96
16		Adam Morrison	AL	2,03
17		Admiral Schofield	AP/AL	1,96
18		Adonis Thomas	AL	2,01
19		Adreian Payne	AP	2,08
20		Ahmed Hill	ES	1,96
21		Al Harrington	AP	2,06
22		Al Horford	P/AP	2,06
23		Al Jefferson	P	2,08
24		Al Thornton	AL/AP	2,03
25		Alade Aminu	AP	2,08
26		Alan Anderson	ES/AL	1,98
27		Alan Williams	P/AP	2,03
28		Alando Tucker	AL	1,98
29		Alec Burks	AL/ES	1,98
30		Alec Peters	AP	2,06

BBDD | Jugadores | PoscAlt

Figura 4.6: Posición y Altura

Una vez vistas las pestañas y unas primeras fórmulas utilizadas con EXCEL, completar el resto de variables necesarias ha sido análogo a este proceso.

Ahora se realiza la depuración de la base para tener la base de datos lo más limpia posible de cara al estudio de aprendizaje automático.

Un elemento fundamental, en cualquier proyecto de analítica de datos y posterior preparación de modelos predictivos, es el asegurar la calidad del dato para evitar manejar datos que lleven a modelados y conclusiones erróneos. En este caso la información elemental para analizar son las lesiones (previamente se preparó y unificó la información de los jugadores). El objetivo de esta fase es conseguir una base de datos limpia y normalizada partiendo de los datos fuente para su posterior procesado en MATLAB. Se ha decido usar Excel para esta fase por las facilidades que posee para realizar estas tareas tanto de una forma manual como automática.

Un paso fundamental, en una primera revisión, es la identificación, depuración y unificación de la terminología del proyecto, sea por sinónimos o equivalencias sea por ortografía, para que sea homogénea en todo el contenido.

Unos ejemplos:

- Achilles injury
- Achilles tendon injury
- Achilles tendinitis
- Achilles

Son todas la misma lesión y se unificarán como 'Achilles', que si se dejara como en la

fuente original se procesaría como cuatro lesiones diferentes.
Para poder llegar a cabo esta depuración y obtención de la base de datos de trabajo se siguieron los siguientes pasos en Excel sobre la lista de lesiones original:
1. Quitar duplicados
2. Ordenar alfabéticamente
3. Primera revisión visual de los contenidos para identificar ejemplos como el anterior de Achilles.
4. Extracción de todo el vocabulario contenido en la descripción de las lesiones
5. Segunda revisión visual para identificar y eliminar signos de puntuación intercalados, simplificar los casos de singular/plural, identificar y corregir la falta, duplicidad u orden incorrecto de las letras.
Por ejemplo:
. Acchilles frente a Achilles
. arthroscopic frente a arthrosopic
7. Clasificación de cada uno de los elementos del vocabulario, es decir si es una parte del cuerpo concreta, el tipo de lesión,...
Por ejemplo:
. Achilles: Parte del cuerpo
. Tendinitis: Tipo de lesión
. Left: Lateralidad

Una vez hechos estos pasos se ha conseguido:
1. Unificación de terminología/equivalencias.
2. Clasificación de las lesiones con calidad, consistente y más fácilmente procesable.
. Parte del cuerpo: muñeca, tobillo, ...
. Tipo: Tejido blando (músculo, tendón), articular o no.
. Lateralidad: izquierdo, derecho.
. Tipo de lesión: rotura, esguince, ...
. Impacto: DNP, DTP, ...
Y así queda la información lista para su procesado.

Una vez terminada la depuración se prodece a su exploración y análisis. La variable Lateralidad no ha sido incluida en el estudio, ya que el 91 % de los jugadores es diestro.

4.2. Exploración

Para estudiar las variables cualitativas, en primer lugar se visualizan mediante histogramas [28], viendo qué cantidad de información contiene cada variable, agrupándola en los diferentes subconjuntos que contiene.
Se puede ver el resultado para el número de lesiones dentro de cada equipo en la siguiente figura:

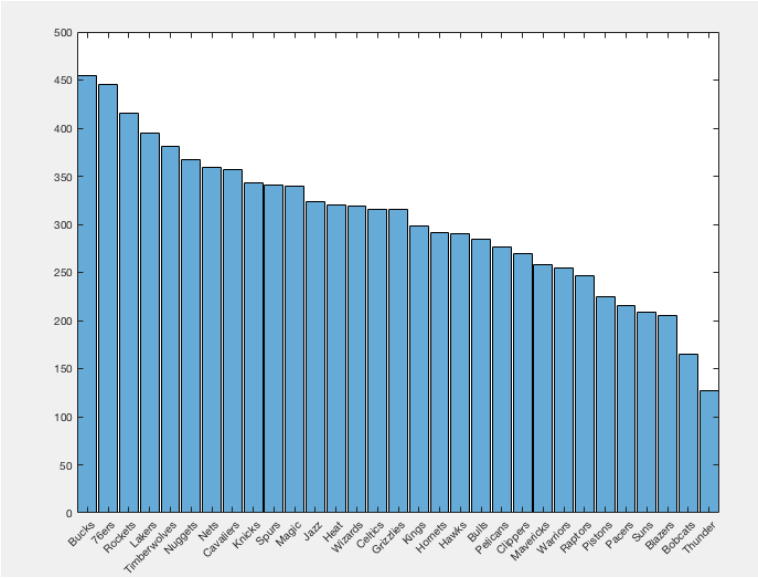


Figura 4.7: Histograma Equipos

En un primer vistazo se puede ver que los Bucks son el equipo con mayor número de lesionados. Habrá que ver qué tipo de lesiones y las características de cada jugador.
De los jugadores lesionados, vamos a ver en qué posiciones juegan.

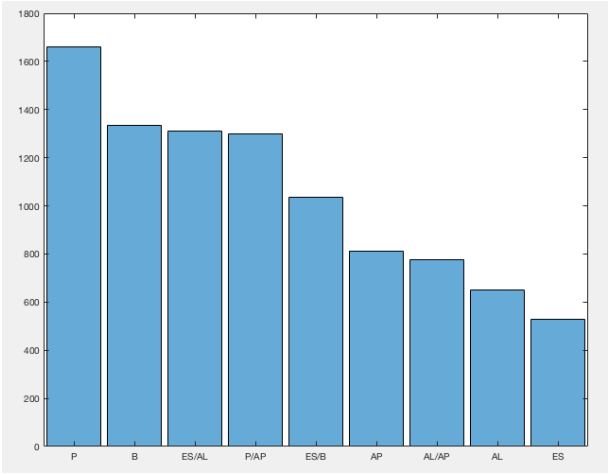


Figura 4.8: Histograma Posiciones

Podemos ver que Pívot y Base son las posiciones que tienen el mayor número de lesionados. Buscando alguna característica más, podemos ver qué color de piel tiene el mayor número de lesionados.

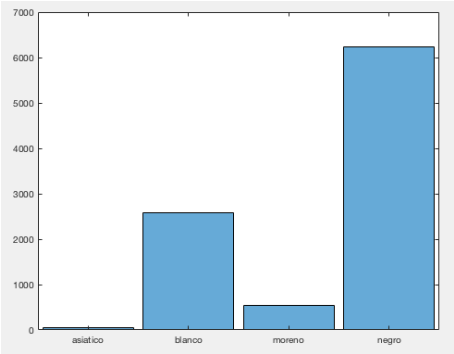


Figura 4.9: Histograma Color

Vemos que la mayoría de lesionados son negros, que tiene bastante sentido, ya que la mayoría de los jugadores de la NBA son de este color. Para un estudio numérico agrupo la información en tablas de contingencia, que hacen un resumen numérico de la información contenida en la variable.

Ahora se muestran los jugadores que más se han lesionado a lo largo de este periodo de tiempo, mostrando un top 30 de lesionados.

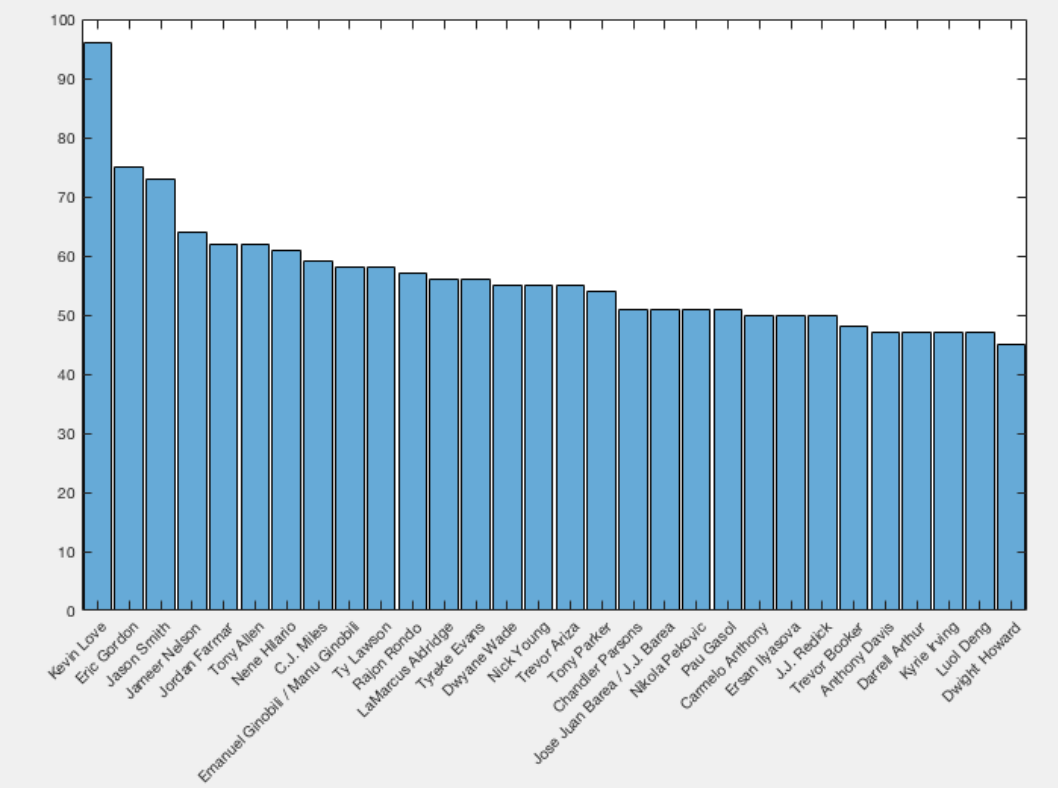


Figura 4.10: Histograma jugadores

En primer lugar, los jugadores y las posciones coinciden, luego efectivamente el papel que juega cada profesional sobre la pista es relevante a la hora de lesionarse. Reforzaremos esta teoría más adelante. También veremos si existe alguna relación más que condicione la salud de los jugadores.

Ahora vamos a visualizar, dentro de las lesiones que tenemos, las 30 que más se repiten:

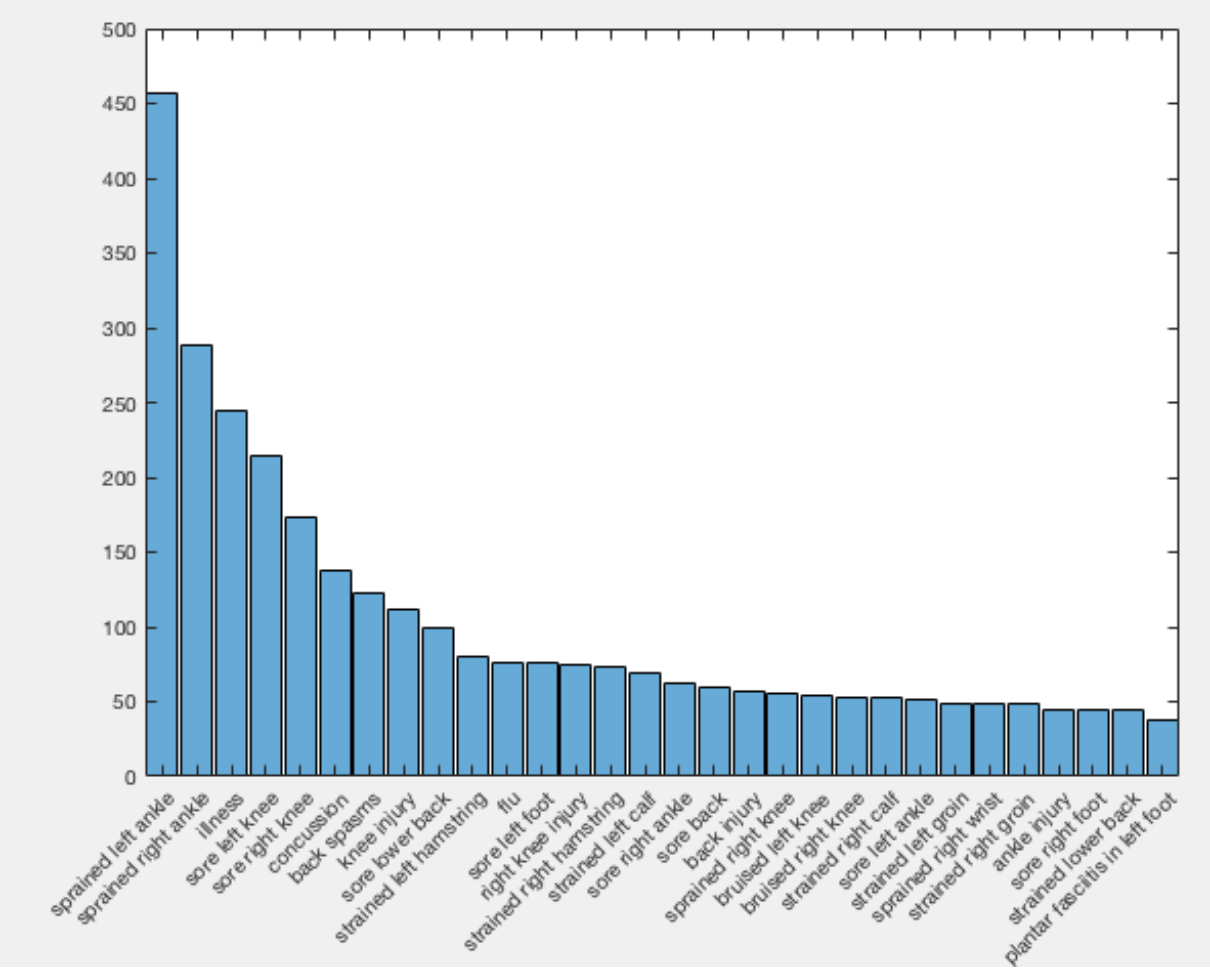


Figura 4.11: Histograma lesiones

En una primera observación, puede verse que las torceduras de tobillo son la lesión más habitual, siendo más común en el izquierdo que en el derecho, pero en este caso prácticamente el total de los jugadores son diestros, cargando el peso en el lado derecho, por tanto, dando más uso al lado derecho del cuerpo y teniendo mayor probabilidad de lesión. Por otra parte, si entendemos que una de las dos partes del cuerpo se desarrolla más por su uso, tendrá más musculatura y menor riesgo de lesión. Para seguir explorando la información de las lesiones, se ven las palabras se repiten más [29].

Figura 4.12: Nube de Palabras

Podemos ver que las palabras que más se repiten son Izquierdo, Derecho, Rodilla, Dolor o Molestia (sore) y Esguince o Torcedura (sprained), seguidas de Tobillo (ankle), Distensión (strained)... En un primer vistazo habíamos visto que, a priori, los problemas de tobillo eran los que más teníamos, algo bastante lógico, pero la rodilla ha pasado a ser el mayor problema al que se enfrentan los jugadores. También podemos ver que las lesiones de articulaciones son las más comunes. Vemos si efectivamente se confirma, haciendo una comparación entre lesiones musculares, de tejidos blandos y óseas (naranja) y de articulaciones (azul).

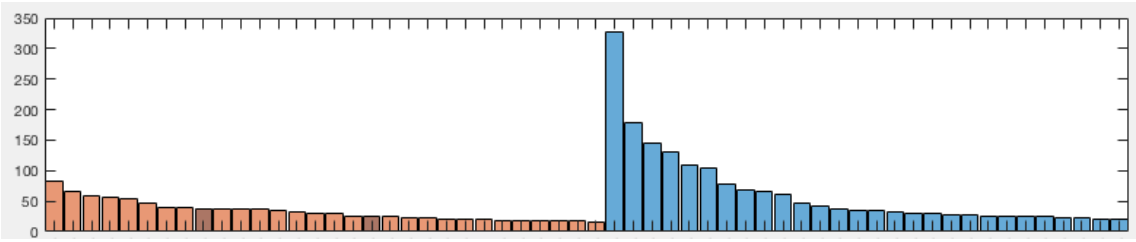


Figura 4.13: Musculares vs Articulares

Se observa que tendrá que hacerse una mayor preparación física trabajando correctamente la musculatura para evitar el sufrimiento articular.

En el apartado 4.1. se ha podido observar que al lado de cada lesión aparece (DTD), (DNP), (out for season) y (out indefinitely). Veamos su utilidad. - DNP: 'Did not play', refiriéndose a que no jugó ese día por el problema que tuviere. Son lesiones leves, ocasionadas por problemas puntuales o que no requieren de más de unos pocos días.

- DTD: 'Day to day', donde se va viendo día a día la evolución de la lesión. Estas lesiones, a parte de incluir algunas anteriores, engloban problemas que pueden durar semanas, siendo mś tediosas que las anteriores.

- Out indefinitely: sin jugar indefinidamente, puede ser un periodo breve o puede ser determinante para que el jugador termine la temporada. El rango es amplio e indica una lesión más grave que las indicadas por DNP o DTD.

- Out for season: Se termina la temporada del jugador, son las lesiones más graves, ya que el jugador automáticamente abandona la cancha hasta la siguiente temporada o se termina su carrera deportiva.

Viendo estos detalles, podemos ver que la gravedad de las lesiones es distinta. Veamos las cifras:
DNP: 4197
DTD: 2212
Out indefinitely: 425
Out for season: 491

A partir de estas etiquetas, se generará el modelo de regresión que indicará la gravedad de la lesión.

4.3. Variables Dummy

Ahora que se ha realizado una primera exploración, se procede a la profundización de los resultados donde, para estudiar el tipo de variable cualitativa en un modelo de predicción, se tiene que transformar en una variable numérica, conocida como variable ficticia o 'Dummy'.

Para ello, se genera una matriz de 1 y 0, donde será 1 si el elemento dentro de la variable pertenece a una de las categorías que tiene y 0 si no. Se crean a partir de las categorías dentro de la variable, tomando una como referencia. [30]
Por ejemplo, suponemos que tenemos la varibale 'Articulacion', que consta de tres categorías: rodilla, tobillo, hombro. Queremos estudiarlo mediante una regresión y adaptamos las categorías para estudiarlo numéricamente, obteniendo la tabla siguiente:

TipoLesion	D1	D2
'Rodilla'	0	0
'Tobillo'	1	0
'Hombro'	0	1

Figura 4.14: Dummy

Llamamos x_2 a la variable 'Articulación' que contiene las categorías que queremos estudiar. La variable x_2 se ha convertido en una matriz de 1 y 0 y ha generado 2 variables ficticias o 'dummy': D_1 y D_2.

Ahora, suponiendo un ejemplo de ecuación de regresión cualquiera:
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Pasaría a convertirse en:
$Y = \beta_0 + \beta_1 x_1 + \beta_2 D_1 + \beta_3 D_2$, donde D_1 y D_2 pueden ser ambas 0 para clasificar como rodilla, D_1 ser 1 y D_2 ser 0 para tobillo y D_1 ser 0 y D_2 ser 1 para hombro.

4.4. Análisis e interpretación

4.4.1. Resultados k-medias

Como se ha descrito anteriormente, Silhoutte permite ver la agrupación de los clusters así como utilizar esta visualización para la selección de éstos y un valor estimado de la suma de distancias totales que se han utilizado en k-means. Para aplicar Silhoutte hay que aplicar primero k-means. Una vez visualizada la agrupación en clusters se puede aproximar una agrupación óptima, pero la visualización no es un criterio definitivo, hay que reforzarlo mediante la suma total de distancias, viendo que no sea ni muy grande ni muy pequeña.

Para poder aplicar el método correctamente se ha hecho k-means como si el conjunto de variables cuantitativas se particionara en 2,3,4,5 y 6 clusters, cogiendo el resultado óptimo, y se ha aplicado Silhouette a cada posibilidad de partición del conjunto, se han visualizado los posibles conjuntos y se han visto las sumas de las distancias, pudiendo así elegir 3 como número óptimo de clusters.

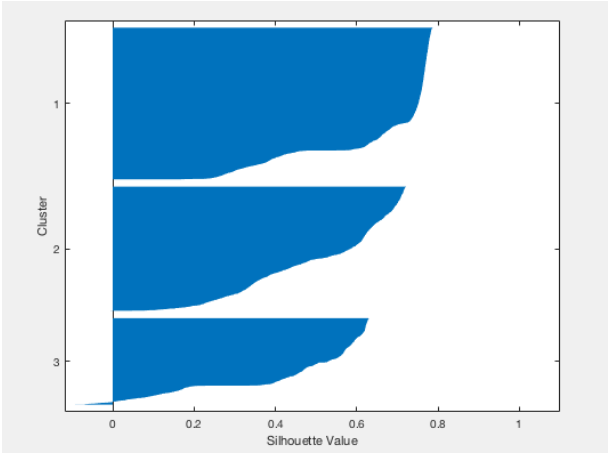


Figura 4.15: Clusters

Esta diferenciación entre agrupaciones de datos muestra lo cerca que está cada dato de un cluster respecto al dato del cluster vecino. La mayoría de los datos pueden agruparse y diferenciarse del vecino, aunque no son valores muy altos (considerados muy altos por encima de 0.8), sí que se puede realizar una agrupación, indicando una correcta aglomeración de la información. Por otra parte, aparecen unos pocos negativos, lo que indica que algunos datos no se separan en los clusters.

4.4.2. Resultados Agrupamiento Jerárquico

Los resultados del agrupamiento jerárquico son mucho más visuales, pudiendo ver el número de clusters y visualizar las observaciones. En este caso, para interpretarlo, vemos en el eje y la numeración correspondiente a la similitud entre las observaciones, que indicará la existencia del cluster, mientras que en el eje x aparecen las observaciones del problema

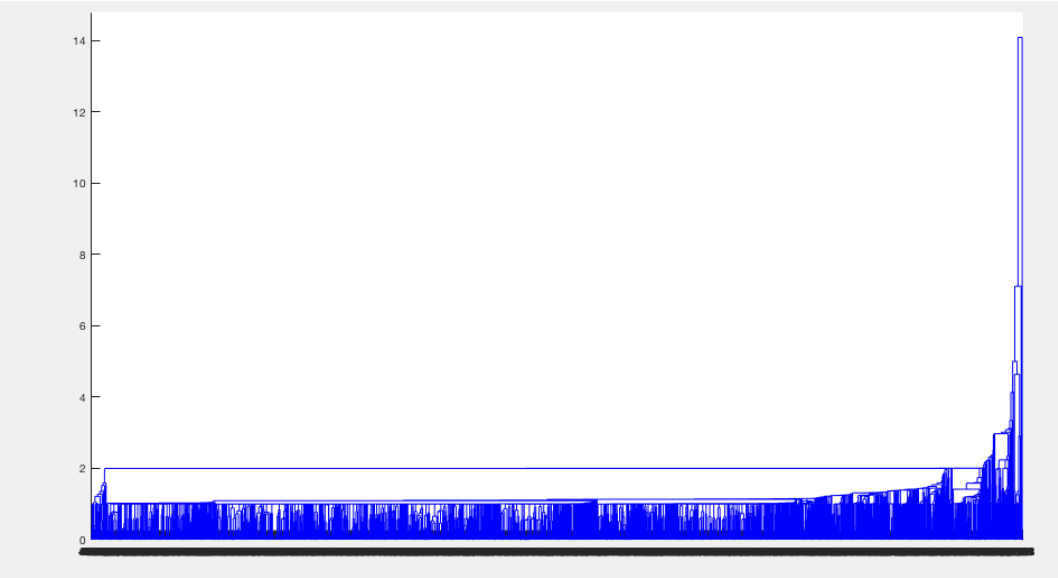


Figura 4.16: Dendrograma

A simple vista, se ve un pico a la derecha que parece corresponder a un error, pero no lo es, sino que se trata de un grupo concreto pequeño de jugadores con mucha altura, mucho peso, edad avanzada y muchos años de experiencia en la NBA. Para ver mejor el resultado, se comprimen las observaciones a 30 y así puede verse con mayor claridad el agrupamiento en cluster.

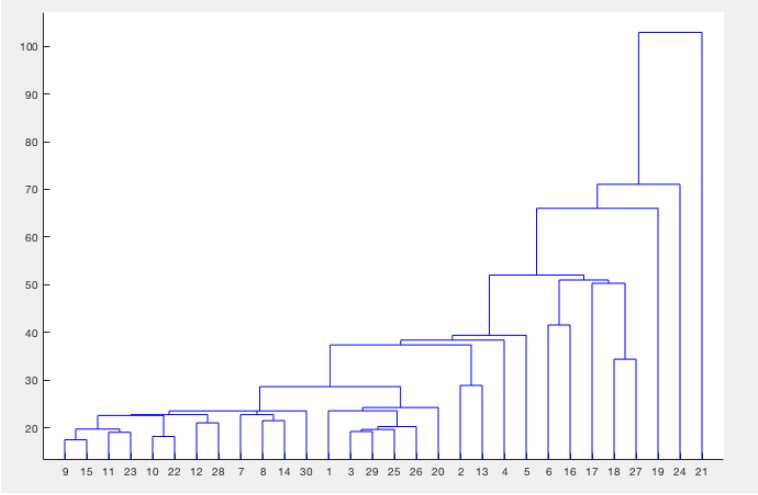


Figura 4.17: Dendrograma agrupado

Una vez vistos estos resultados y viendo que es posible la relación entre las variables, se pasa a los métodos de clasificación.

4.4.3. Resultados Regresión Logística, SVM y Naive Bayes

Ahora que ya se tienen los modelos explicados, se aplican a la base de datos construida. En primer lugar, se va a visualizar la curva ROC generada por cada uno de los modelos clasificados, buscando, así, el modelo que mejor se ajuste a la predicción de los nuevos datos. A parte de la visualización, se dará un valor del área, AUC.
Las visualizaciones van a ser varias, ya que se aplica la comparación en cada una de las respuestas a las cuestiones objeto de análisis. Siguiendo el orden de estudio, primero se analizan las articulares.

Articulares

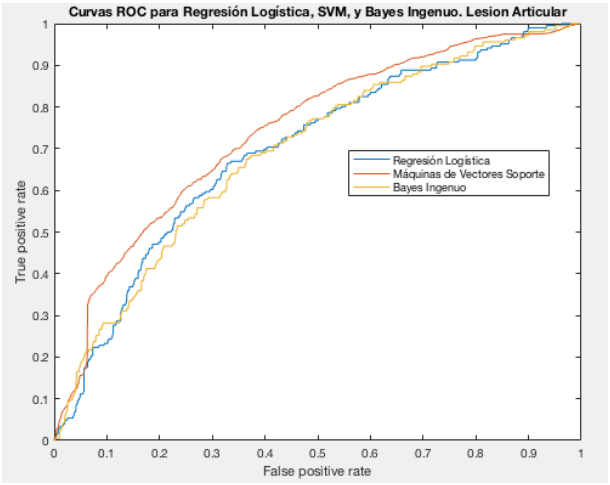


Figura 4.18: ROC Articular

Con los AUC correspondientes:

AUC = 0.6968
AUCsvm = 0.7417
AUCnb = 0.6926

Vemos cómo, en este caso, SVM es el mejor aproximador. Al trabajar con modelos no lineales, SVM interactúa con pares de observaciones y una función Kernel, luego para ejemplificar cómo interactúan las variables dentro de los modelos, se muestra mediante la estimación de coeficientes de la regresión logística.
Ahora, con los resultados de los coeficientes, se construye la probabilidad de que una lesión sea articular mediante la regresión, creando la fórmula:

$$p = \frac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5D_1+\beta_6D_2+\beta_7D_3+\beta_8D_4+\beta_9D_5+\beta_{10}D_6+\beta_{11}D_7\beta_{12}D_8+\beta_{13}D_9+\beta_{14}D_{10}+\beta_{15}D_{11})}}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5D_1+\beta_6D_2+\beta_7D_3+\beta_8D_4+\beta_9D_5+\beta_{10}D_6+\beta_{11}D_7\beta_{12}D_8+\beta_{13}D_9+\beta_{14}D_{10}+\beta_{15}D_{11})}}$$

Las cuatro primeras x_i de la regresión corresponden a las variables Edad, Altura, Peso y Años en la NBA y las D_i son las variables Dummy de las cualitativas de Posición del jugador y Lesión Previa, donde las nueve primeras corresponden a cada una de las posiciones y las dos últimas a si han tenido o no una lesión previa.

Los valores de los coeficientes β_i se muestran en la siguiente tabla:

	Estimate
(Intercept)	-0.78171
x1	-0.01484
x2	0.030142
x3	0.013007
x4	-0.024455
x5	0
x6	0.22388
x7	-0.19535
x8	0.17501
x9	0.026827
x10	0.203
x11	0.16951
x12	-0.2374
x13	-0.048358
x14	0
x15	0.14074

Figura 4.19: Coeficientes de regresión articular (β_i)

Los coeficientes figuran en la tabla como x_i, pero corresponden a los β_i de la ecuación de regresión de este trabajo. Se tiene (Intercep) como el término independiente.

Gravedad

Ahora se comparan los resultados de los ajustes para la gravedad de la lesión:

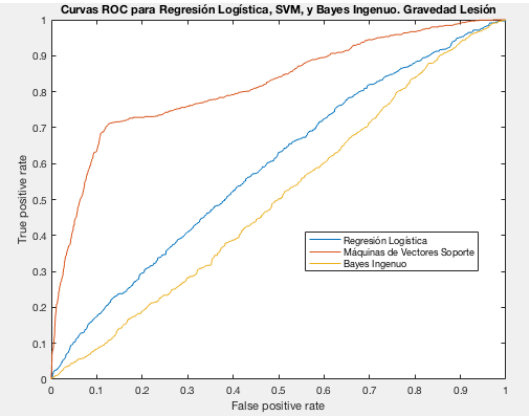


Figura 4.20: ROC Gravedad

Aqui se aprecia disparidad en los resultados, viendo cómo Naive Bayes es malo, ya que clasifica prácticamente aleatoriamente. Se fundamenta con los AUC correspondientes:

AUC = 0.5912
AUCsvm = 0.8177
AUCnb = 0.5045

Viendo que, efectivamente, Naive Bayes no es un buen clasificador para este modelo.
Ahora, se aplica lo mismo para ver la probabilidad de que un jugador vaya a tener una lesión grave:

$$p = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 D_1+\beta_6 D_2+\beta_7 D_3+\beta_8 D_4+\beta_9 D_5+\beta_{10} D_6+\beta_{11} D_7\beta_{12} D_8+\beta_{13} D_9+\beta_{14} D_{10}+\beta_{15} D_{11})}}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 D_1+\beta_6 D_2+\beta_7 D_3+\beta_8 D_4+\beta_9 D_5+\beta_{10} D_6+\beta_{11} D_7\beta_{12} D_8+\beta_{13} D_9+\beta_{14} D_{10}+\beta_{15} D_{11})}}$$

donde los coeficientes estimados son:

Estimated Coefficients:

	Estimate
(Intercept)	-5.4196
x1	-0.039172
x2	2.0531
x3	0.0098563
x4	-0.017746
x5	0
x6	-0.19703
x7	0.34929
x8	0.70911
x9	0.17423
x10	0.37264
x11	0.3087
x12	-0.10613
x13	-0.41308
x14	0
x15	-0.15032

Figura 4.21: Coeficientes de regresión gravedad (β_i)

Para ser más específicos a la hora de realizar la predicción e ir más a detalle, se calculan las probabilidades de sufrir una lesión de rodilla, tobillo, hombro, pie e isquiotibial. Se han elegido estas cinco debido a que son las más recurrentes que suceden en el cuerpo.

Rodilla

Se determinalas ROC de los modelos y se genera la ecuación con los coeficientes.

Curvas ROC

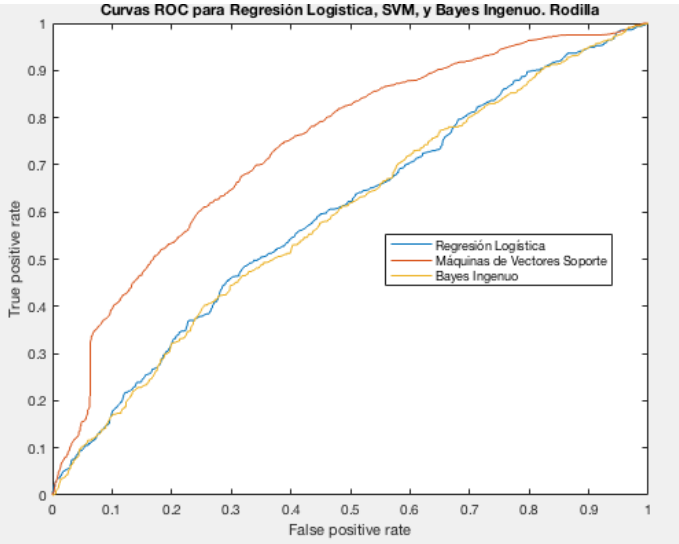


Figura 4.22: ROC rodilla

Con los AUC correspondientes:

AUC = 0.5975
AUCsvm = 0.7177
AUCnb = 0.5916

Vemos cómo SVM sigue siendo el mejor modelo de ajuste.

Se genera la fórmula con los coeficientes:

$$p = \frac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5D_1+\beta_6D_2+\beta_7D_3+\beta_8D_4+\beta_9D_5+\beta_{10}D_6+\beta_{11}D_7\beta_{12}D_8+\beta_{13}D_9+\beta_{14}D_{10}+\beta_{15}D_{11})}}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5D_1+\beta_6D_2+\beta_7D_3+\beta_8D_4+\beta_9D_5+\beta_{10}D_6+\beta_{11}D_7\beta_{12}D_8+\beta_{13}D_9+\beta_{14}D_{10}+\beta_{15}D_{11})}}$$

Coeficientes:

	Estimate
(Intercept)	-3.4844
x1	-0.039563
x2	0.10939
x3	0.024592
x4	0.040297
x5	0
x6	0.34269
x7	-0.15663
x8	0.24645
x9	-0.52034
x10	0.20949
x11	-0.18373
x12	-0.61311
x13	-0.092755
x14	0
x15	0.32291

Figura 4.23: Coeficientes de regresión rodilla (β_i)

Tobillo

Se determinalas ROC de los modelos y se genera la ecuación con los coeficientes.

Curvas ROC

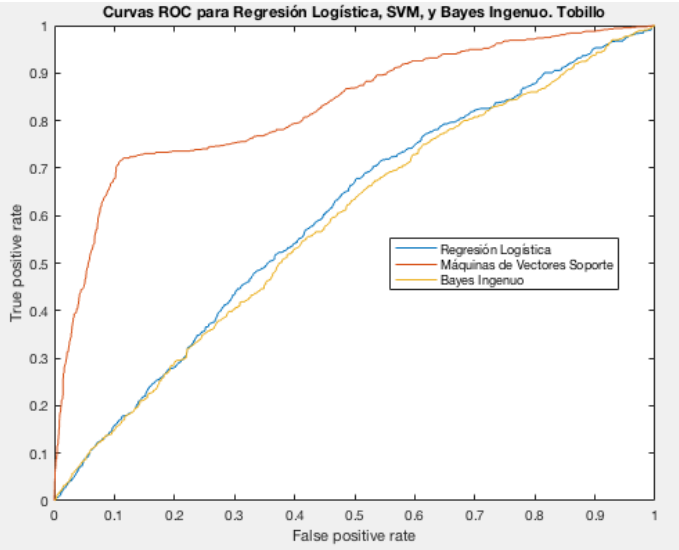


Figura 4.24: ROC tobillo

Con los AUC correspondientes:

AUC = 0.5995

AUCsvm = 0.8294
AUCnb = 0.5853

Vemos cómo SVM sigue siendo el mejor modelo de ajuste.

Se genera la fórmula con los coeficientes:

$$p = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 D_1+\beta_6 D_2+\beta_7 D_3+\beta_8 D_4+\beta_9 D_5+\beta_{10} D_6+\beta_{11} D_7\beta_{12} D_8+\beta_{13} D_9+\beta_{14} D_{10}+\beta_{15} D_{11})}}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 D_1+\beta_6 D_2+\beta_7 D_3+\beta_8 D_4+\beta_9 D_5+\beta_{10} D_6+\beta_{11} D_7\beta_{12} D_8+\beta_{13} D_9+\beta_{14} D_{10}+\beta_{15} D_{11})}}$$

Coeficientes:

Estimated Coefficients:

	Estimate
(Intercept)	−2.8033
x1	0.0053457
x2	0.48755
x3	0.0087387
x4	−0.079747
x5	0
x6	−0.34651
x7	−0.58312
x8	0.018347
x9	−0.25123
x10	−0.19089
x11	−0.028726
x12	−0.24208
x13	−0.20864
x14	0
x15	−0.1123

Figura 4.25: Coeficientes de regresión tobillo (β_i)

Hombro

Se determinalas ROC de los modelos y se genera la ecuación con los coeficientes.

Curvas ROC

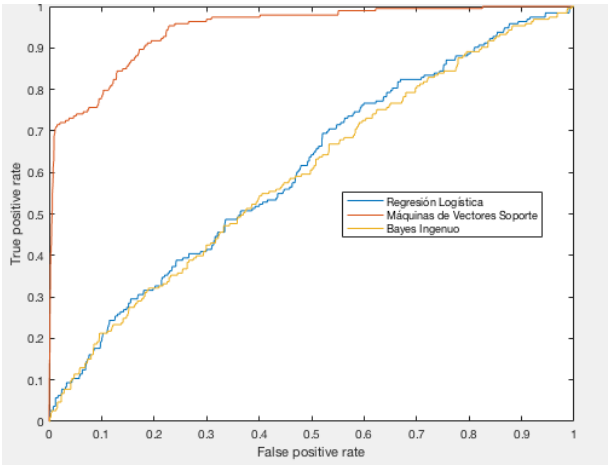


Figura 4.26: ROC hombro

Con los AUC correspondientes:

AUC = 0.6075
AUCsvm = 0.9443
AUCnb = 0.5954

Vemos cómo SVM sigue siendo el mejor modelo de ajuste.

Se genera la fórmula con los coeficientes:

$$p = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 D_1+\beta_6 D_2+\beta_7 D_3+\beta_8 D_4+\beta_9 D_5+\beta_{10} D_6+\beta_{11} D_7\beta_{12} D_8+\beta_{13} D_9+\beta_{14} D_{10}+\beta_{15} D_{11})}}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 D_1+\beta_6 D_2+\beta_7 D_3+\beta_8 D_4+\beta_9 D_5+\beta_{10} D_6+\beta_{11} D_7\beta_{12} D_8+\beta_{13} D_9+\beta_{14} D_{10}+\beta_{15} D_{11})}}$$

Coeficientes:

Estimated Coefficients:

	Estimate
(Intercept)	-5.4171
x1	-0.0086956
x2	2.5412
x3	-0.03032
x4	-0.023404
x5	0
x6	0.23672
x7	0.3466
x8	-0.054842
x9	0.96848
x10	-0.23387
x11	0.36971
x12	0.56214
x13	0.38734
x14	0
x15	0.30896

Figura 4.27: Coeficientes de regresión hombro (β_i)

Pie

Se determinalas ROC de los modelos y se genera la ecuación con los coeficientes.

Curvas ROC

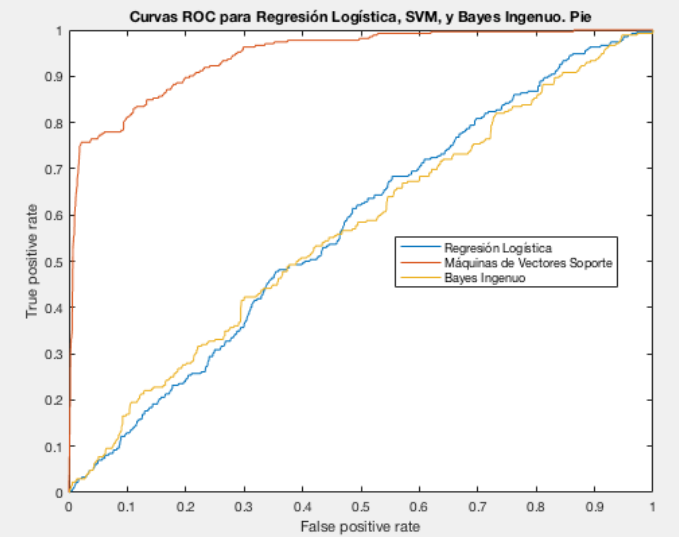


Figura 4.28: ROC pie

Con los AUC correspondientes:

AUC = 0.5721
AUCsvm = 0.9434
AUCnb = 0.5916

Vemos cómo SVM sigue siendo el mejor modelo de ajuste.

Se genera la fórmula con los coeficientes:

$$p = \frac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5D_1+\beta_6D_2+\beta_7D_3+\beta_8D_4+\beta_9D_5+\beta_{10}D_6+\beta_{11}D_7\beta_{12}D_8+\beta_{13}D_9+\beta_{14}D_{10}+\beta_{15}D_{11})}}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5D_1+\beta_6D_2+\beta_7D_3+\beta_8D_4+\beta_9D_5+\beta_{10}D_6+\beta_{11}D_7\beta_{12}D_8+\beta_{13}D_9+\beta_{14}D_{10}+\beta_{15}D_{11})}}$$

Coeficientes:

Estimated Coefficients:

	Estimate
(Intercept)	-5.4196
x1	-0.039172
x2	2.0531
x3	0.0098563
x4	-0.017746
x5	0
x6	-0.19703
x7	0.34929
x8	0.70911
x9	0.17423
x10	0.37264
x11	0.3087
x12	-0.10613
x13	-0.41308
x14	0
x15	-0.15032

Figura 4.29: Coeficientes de regresión pie (β_i)

Isquiotibial

Se determinalas ROC de los modelos y se genera la ecuación con los coeficientes.

Curvas ROC

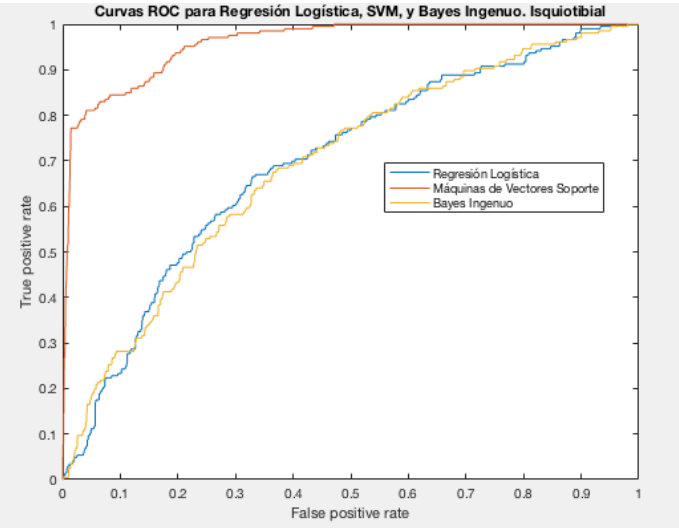


Figura 4.30: ROC isquiotibial

Con los AUC correspondientes:

AUC = 0.6969
AUCsvm = 0.9585
AUCnb = 0.6926

Vemos cómo SVM sigue siendo el mejor modelo de ajuste.

Se genera la fórmula con los coeficientes:

$$p = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 D_1+\beta_6 D_2+\beta_7 D_3+\beta_8 D_4+\beta_9 D_5+\beta_{10} D_6+\beta_{11} D_7\beta_{12} D_8+\beta_{13} D_9+\beta_{14} D_{10}+\beta_{15} D_{11})}}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4+\beta_5 D_1+\beta_6 D_2+\beta_7 D_3+\beta_8 D_4+\beta_9 D_5+\beta_{10} D_6+\beta_{11} D_7\beta_{12} D_8+\beta_{13} D_9+\beta_{14} D_{10}+\beta_{15} D_{11})}}$$

Coeficientes:

Estimated Coefficients:

	Estimate
(Intercept)	-5.8558
x1	0.035917
x2	1.0388
x3	-0.005237
x4	0.016591
x5	0
x6	-0.41017
x7	-1.242
x8	0.068647
x9	0.39276
x10	-0.31296
x11	0.6407
x12	-1.0392
x13	-1.4294
x14	0
x15	0.43474

Figura 4.31: Coeficientes de regresión isquiotibial (β_i)

Capítulo 5

Conclusiones

Los objetivos de este análisis eran hallar la existencia de una relación entre las variables cuantitativas de edad, peso, altura y años de juego en la NBA y modelos de predicción que, a partir de las características físicas de un jugador, permitan anticipar qué tipo de lesión va a desarrollar en un futuro compitiendo en su actividad deportiva habitual.

Las conclusiones son las siguientes:

- A pesar de haber una dominancia casi absoluta de jugadores diestros, la mitad de las lesiones ocurren en la parte izquierda del cuerpo.

- La variedad de lesiones que pueden ocurrir jugando al baloncesto es muy amplia, pero va a haber una tendencia hacia las lesiones articulares, en concreto de rodilla.

- El papel que desempeña un jugador sobre la cancha es un factor importante para determinar el tipo de lesión que tendrá.

- Existe una relación entre las variables cuantitativas, permitiendo la agrupación de datos en clusters.

- Es posible ajustar un modelo de regresión logística para hallar una predicción sobre el tipo de lesión que puede producirse a futuro, bien sea una lesión articular, una lesión grave o que sea en una parte concreta del cuerpo, aunque no sea un ajuste de alta calidad.

Bibliografía

[1] D. L. Betancur, "Introducción al machine learning," Consultado: 23/08/2021 https://greensqa.com/introduccion-al-machine-learning/.

[2] Varios, "Tipos de problemas en machine learning," Consultado: 23/08/2021 http://epicalsoft.blogspot.com/2018/11/azure-machine-learning-algoritmos-de.html.

[3] MathWorks, "Introducción a k-means clustering," Consultado: 23/08/2021 https://es.mathworks.com/help/stats/k-means-clustering.htmlbq_679x-19.

[4] G. Galiano-Casas and E. García-Gonzalo, "El algoritmo k-means aplicado a clasificación y procesamiento de imágenes," Consultado: 23/08/2021 https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html.

[5] D. Dupal, "K-means clustering algorithm explained," Consultado: 23/08/2021 http://dendroid.sk/2011/05/09/k-means-clustering/.

[6] Contribuciones, "Silhouette (clustering)," Consultado: 23/08/2021 https://es.wikipedia.org/wiki/Silhouette_(clustering).

[7] L. Kaufman and P. J. Rousseeuw, *Finding groups in data : An introduction to cluster analysis*, 1st ed. New York NY: Wiley-Interscience, 1990.

[8] Varios, "Clustering validation statistics: 4 vital things everyone should know - unsupervised machine learning," Consultado: 23/08/2021 http://www.sthda.com/english/wiki/wiki.php?id_contents=7952.

[9] MathWorks, "Clustering jerárquico," Consultado: 23/08/2021 https://es.mathworks.com/help/stats/hierarchical-clustering.htmlbq_6_ia.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 1st ed. New York NY: Springer Science+Business Media, 2017.

[11] Contribuciones, "Agrupamiento jerárquico," Consultado: 23/08/2021 https://es.wikipedia.org/wiki/Agrupamiento_jerárquico.

[12] Duk2, "Algoritmos de data mining para agrupar datos â clustering jerárquico," Consultado: 23/08/2021 https://estrategiastrading.com/clustering-jerarquico/.

[13] J. A. Rodrigo, "Regresión logística simple y múltiple," Consultado: 23/08/2021 https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple.

[14] J. Cramer, "The origins and development of the logit model," Consultado: 23/08/2021 http://www.cambridge.org/resources/0521815886/1208_default.pdf.

[15] Contribuciones, "Máxima verosimilitud," Consultado: 23/08/2021 https://es.wikipedia.org/wiki/Máxima_verosimilitud.

[16] Varios, "Regresión logística," Consultado: 23/08/2021 https://es.wikipedia.org/wiki/Regresión_logística.

[17] Contribuciones, "Máquinas de vectores de soporte (svm)," Consultado: 23/08/2021 https://es.wikipedia.org/wiki/Máquinas_de_vectores_de_soporte.

[18] A. L. Díaz, "Support vector machine (svm)," Consultado: 23/08/2021 https://idus.us.es/bitstream/handle/11441/77547/López %20 Díaz %20Ana %20TFG.pdf?sequence=1.

[19] Contribuciones, "Clasificador bayesiano ingenuo," Consultado: 23/08/2021 https://es.wikipedia.org/wiki/Clasificador_bayesiano_ingenuo.

[20] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, 2nd ed. New York NY: Springer Science+Business Media, 2021.

[21] F. Cardellino, "Cómo funcionan los clasificadores naive bayes: con ejemplos de código de python," Consultado: https://www.freecodecamp.org/espanol/news/como-funcionan-los-clasificadores-naive-bayes-con-ejemplos-de-codigo-de-python/.

[22] MathWorks, "Curva roc," Consultado: 23/08/2021 https://es.mathworks.com/discovery/roc-curve.html.

[23] Contribuciones, "Curva roc," Consultado: 23/08/2021 https://es.wikipedia.org/wiki/Curva_ROC.

[24] Minitab, "Ejemplo de clasificación random forests," Consultado: 23/08/2021 https://support.minitab.com/es-mx/minitab/20/help-and-how-to/statistical-modeling/predictive-analytics/how-to/random-forests-classification/before-you-start/example/.

[25] R. Hopkins, "Dataset," Consultado: 23/08/2021 https://www.kaggle.com/ghopkins/nba-injuries-2010-2018.

[26] Contribuciones, "Dataset," Consultado: 23/08/2021 "https://www.basketball-reference.com/players/".

[27] M. Mediacom, "Jugadores nba," Consultado: 23/08/2021 https://www.hispanosnba.com/jugadores/nba-todos/index.

[28] MathWorks, "Plot categorical data (histogramas)," Consultado: 23/08/2021 https://es.mathworks.com/help/matlab/matlab_prog/plot-categorical-data.html.

[29] ——, "Mapa de palabras (wordcloud)," Consultado: 23/08/2021 https://es.mathworks.com/help/matlab/ref/wordcloud.html.

[30] HospitalRamónyCajal, "Variables indicadoras('dummy')," Consultado: http://www.hrc.es/bioest/Reglin_12.html.