Universidad de Granada

Universidad de Granada

Aplicación de la minería de datos

Tratamiento de Datos sobre el dataset del beisbol en USA

Autores:
Manuel Jesús García Manday
Mario Ortega Aguayo

Master en Ingeniería Informática

1 de febrero de 2017

Índice

1. Introducción **3**

2. Dataset **3**

3. Análisis exploratorio **5**

4. Clustering **6**

5. Árboles de decisión **9**

6. Regresión **11**

7. Conclusiones **13**

1. Introducción

A través de la minería de datos es posible extraer información relevante sobre conjuntos grandes de datos aplicando una serie de técnicas y algoritmos que se encargan de explorarlos.

En este documento se mostrará la exploración de un conjunto de datos utilizando para ello algunas de las técnicas mas comunes como la clasificación, la regresión y el agrupamiento, exponiendo en cada punto del presente trabajo el tratamiento aplicado así como la interpretación de los resultados.

Son varias las herramientas existentes para explorar grandes conjuntos de datos, Knime y R Studio son de las más comunes y utilizadas en este ámbito. En esta ocasión he decidido trabajar con R Studio por ser una herramienta basada en scripts de R y tener algo de experiencia previa, lo que me ayudará a incrementarla y profundizar en su conocimiento.

2. Dataset

Tras investigar y navegar por diferentes sitios web como kaggle.com que ofrecen múltiples datasets , hemos optado por seleccionar un conjunto de datos aplicado al ámbito del deporte. En este caso el baseball, un deporte mundialmente conocido y que tiene su principal foco en los Estados Unidos desde hace muchos años.

Hemos seleccionado este dataset después de analizarlo y evaluar las diferentes variables y observaciones que presenta, viendo que sería interesante conocer y profundizar más sobre los detalles de este deporte.

Este conjunto de datos contiene información completa de los equipos de baseball de la liga de los Estados Unidos desde 1871 hasta 2015, en dichos datos se pueden ver estadísticas sobre los bateadores y pitchers de los equipos así como estadísticas los partidos jugados, los partidos ganados y perdidos, los entrenadores, las ligas, etc. La última versión de este dataset, la cual es la que hemos utilizado, se puede encontrar en la página www.seanlahman.com/baseball-archive, así como versiones anteriores y diferentes formatos del conjunto de datos.

Son muchas las variables de las que dispone este dataset, por lo que después de realizar un análisis sobre cada una de ellas viendo lo que puede aportar en el estudio y la relevancia de sus datos, hemos seleccionado un subconjunto de ellas creando un nuevo dataset sobre el que realizaremos todas las técnicas de minería de datos mencionadas anteriormente, manteniendo las mismas filas que en el original.

Para evitar que no se seleccionen variables relevanes para el nuevo dataset, hemos realizado un estudio previo de todo el conjunto de variables para eliminar las que tengas valores nulos o que no aporten nada al estudio. Para ello hemos realizado un análisis exploratorio previo donde poder ver las relaciones entre las diferentes variables del conjunto inicial.

Como se puede observar en la siguiente imagen, existen campos en el dataset que no son tan importantes para el estudio como los dos que se han comparado más abajo. En este caso las variables de "media de partidos ganadosz de "factor de 3 años para los bateadores"no van a aportar ninguna conclusión ni dato de peso en el estudio, por lo que estas dos son entre otras más las que no entraran en el conjunto final del dataset.

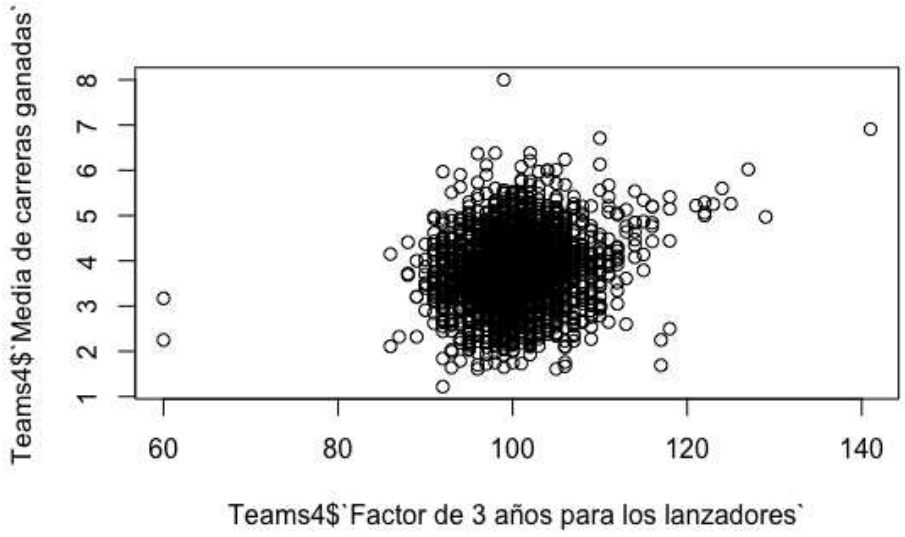

Figura 1

En la siguiente tabla se muestra un resumen de las variables que representan este nuevo conjunto de datos en función a cada equipo que ha participado en la liga de baseball entre los años 1871 y 2015, indicando su tipo y medida:

Variable	Tipo	Medida
yearID	fecha	año de fundación
lgID	texto	liga a la que pertenece
Games	número (entero)	partidos jugados
Win	número (entero)	partidos ganados
Lost	número (entero)	partidos perdidos
LgWin	booleano	si ha ganado la liga
R	número (entero)	carreras anotadas
AB	número (entero)	veces que un bateador cambia contra un pitcher
H	número (entero)	golpes de los bateadores
SO	número (entero)	strikeouts de los bateadores
SB	número (entero)	bases robadas
RA	número (entero)	carreras anotadas por el contrario
CG	número (entero)	juegos completados
SHO	número (entero)	juego terminado en el que el equipo perdedor no anota carrera
E	número (entero)	errores

Cuadro 1: Conjunto de varibales.

Como se ha podido comprobar en la anterior tabla, el conjunto de datos seleccionado contiene información sobre los aspectos meramente deportivos sucedidos durante el transcurso de todos los partidos disputados por cada equipo en el periodo indicado anteriormente, de la cual se puede obtener conclusiones curiosas e importantes como veremos en los posteriores puntos de este trabajo.

3. Análisis exploratorio

Una vez que se ha dispuesto del nuevo dataset con el subconjunto de variables del original y manteniendo el mismo número de datos (mismo número de fila), hemos realizado un primer análisis exploratorio sobre las diferentes variables para comenzar a ver la relación que existe entre los diferentes campos y obtener resultados que nos ayude mejor a comprender este conjunto de datos.

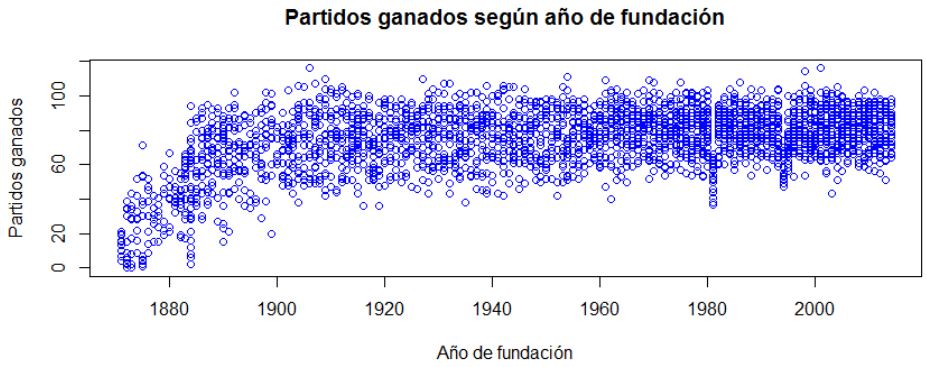

Figura 2

En este primer análisis realizado sobre el conjunto de datos podemos ver como el número de partidos ganados tiene un crecimiento a finales del siglo XIX y principios del siglo XX, esto puede ser debido a que la cantidad de equipos y partidos aumentaran durante esa época. Lo que también se puede observar es durante el periodo de finales del siglo XX y comienzos del siglo XXI son más la cantidad de puntos sobre la gráfica, lo que no interviene en la cantidad de partidos ganados ya que se mantiene en la misma línea, sino que refleja el crecimiento en número de equipos durante ese periodo, que fue la época dorada del baseball en los Estados Unidos según su historia.

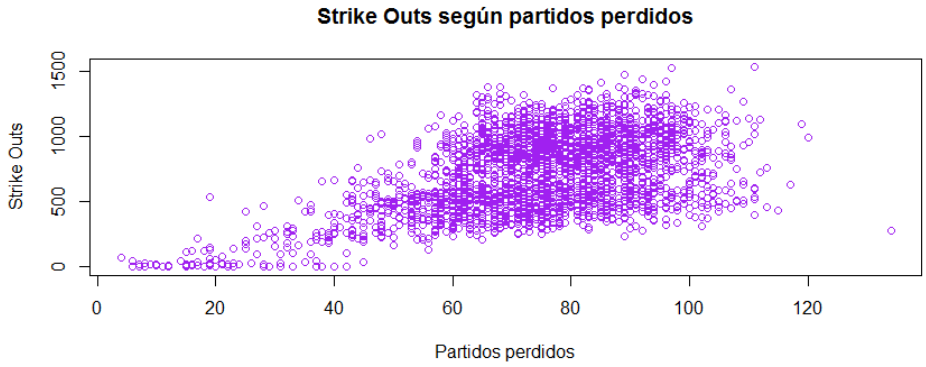

Figura 3

La Figura 3 muestra la relación existente entre el número de partidos perdidos y la cantidad de strikes outs acumulados. Como se puede observar los resultados obtenidos son totalmente razonables y casan con la lógica, ya que la mayor cantidad de strikes outs se encuentran concentrados en el rango de mayor número de partidos perdidos que va de 60 a 100. Esto se debe a que en un partido de baseball, un strike out penaliza al equipo que lo comete, ya que elimina a uno de sus bateadores.

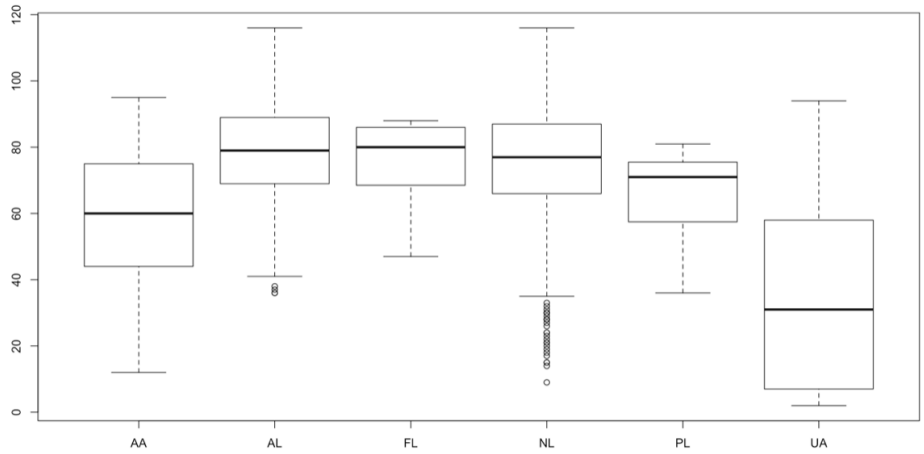

Figura 4

Para la Figura 4 se ha utilizado otro tipo de gráfico para expresar la relación entre la cantidad de partidos ganados y la liga en los que se han disputados. Observando la gráfica anterior se puede como existen 3 ligas que son las que mayor número de victorias han conseguido sus equipos, siendo la liga UA la que menor puntuación tiene. Esto puede ser debido a la calidad de los diferentes equipos en relación a sus jugadores, algo que si duda afecta en el ámbito de cualquier deporte. Otro detalle a tener en cuenta que arroja esta gráfica es la disposición de la raya horizontal negra dentro de cada caja (lo que es conocido como bigote), que como podemos observar en dos de ellas no se encuentra situadas en la mitad de la misma, sino que la separación entre la parte superior y la parte inferior que divide dicha raya horizontal es diferente, esto es debido a que la cantidad de partidos ganados en esas ligas fueron conseguidos por un número menor de equipos, es decir, el número de victorias en esas ligas estás concentradas en un menor número de equipos que en las demás ligas.

4. Clustering

Cuando tenemos un dataset con una cantidad elevada de datos, estos se muestran aparentemente complicados de organizar.

La técnica clustering consiste en clasificar los datos formando grupos o clusters de elementos, de forma que los datos dentro de cada agrupacion presenten cierto grado de homogeneidad en base a los valores adoptados sobre un conjunto de variables. Para nuestro dataset, vamos a utilizar el algoritmo kmeans de R para realizar el cluster.

Una vez que se han eliminado los datos vacíos dentro de nuestro dataset a analizar, se cogen variables que aparentemente tengan relación, asi podremos ver mejor la bondad de esta técnica. Cabe destacar que se han seleccionado para este nuevo conjunto variables que no tienen mucha relación entre ellas para ver como se comporta la bondad en este caso.

De esta manera podremos aplicar la técnica de clustering a variables que tienen una mayor bondad frente a otras que tienen menos y poder sacar conclusiones referente a las agrupaciones obtenidas.

En la siguiente tabla se muestran los campos que se han escogido del dataset con el hemos comenzado el estudio de tratamiento de datos:

Una vez con el subconjunto seleccionado, procedemos a estandarizar las variables y así tener las mismas unidades, es decir, una desviación estándar.

Variable	Medida
yearID	año de fundación
Games	partidos jugados
AB	veces que un bateador cambia contra un pitcher
SHO	juego terminado en el que el equipo perdedor no anota carrera
SB	bases robadas
hline E	errores

Cuadro 2: Conjunto de varibales para Clustering.

Kmeans es un algoritmo de partición de datasets en distintos grupos, donde se clasifican según similitudes que hay entre ellos. Hemos decidido utilizar este algoritmo ya que consideramos que es más idóneo para analizar un gran número de casos, y por otro lado no es necesario trabajar sobre una matriz distancias, sino que la hace sobre la original, y así no requiere tanto uso de procesamiento y consumo de memoria.

El algoritmo va asignando a un grupo u otro según el intervalo al que pertenezca. El número de grupos lo determinamos por el valor k, el cual estableceremos en la propia función de R.

El primer paso es determinar el número de clústers en que se va a dividir el conjunto sobre el que se va a trabajar, cuestión que no es fácil de determinar. Sin embargo, usa solución usada frecuentemente para determinar el número óptimo de clusters es el método Elbow (del inglés codo, referente a la curvatura que se forma en la gráfica), el cual implica observar un conjunto de posibles números de agrupaciones en relación con la forma en que minimizan la suma de cuadrados dentro del grupo. Vemos la gráfica a continuación y observamos como el número óptimo de clusters sería 4, ya que es el punto en que tiende a decrementar.

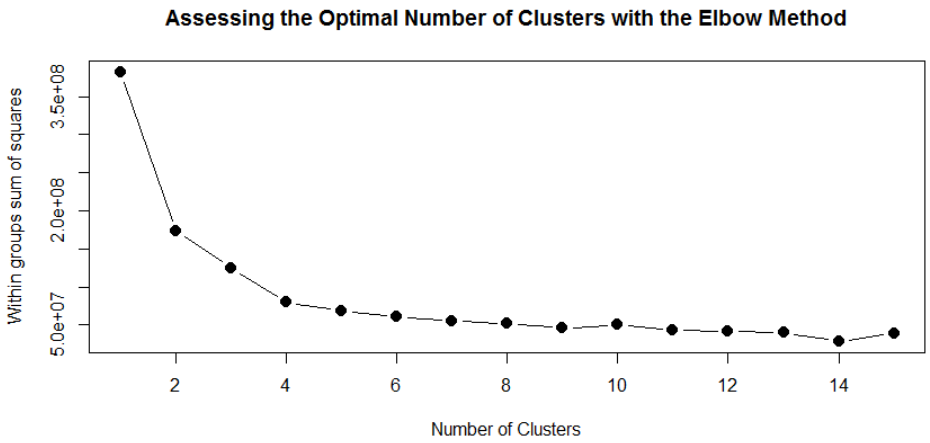

Figura 5

Para comprobar este hecho, en primer lugar vamos a hacer un clustering con 2 y 3 grupos, como vemos a continuación.

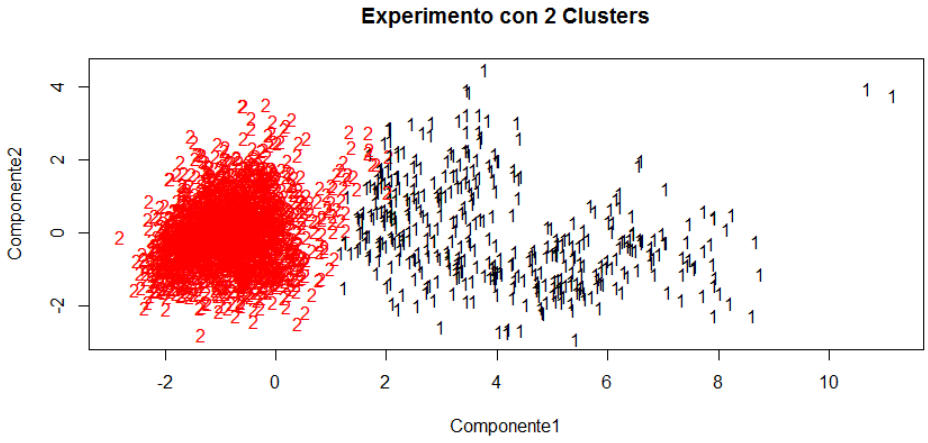

Figura 6

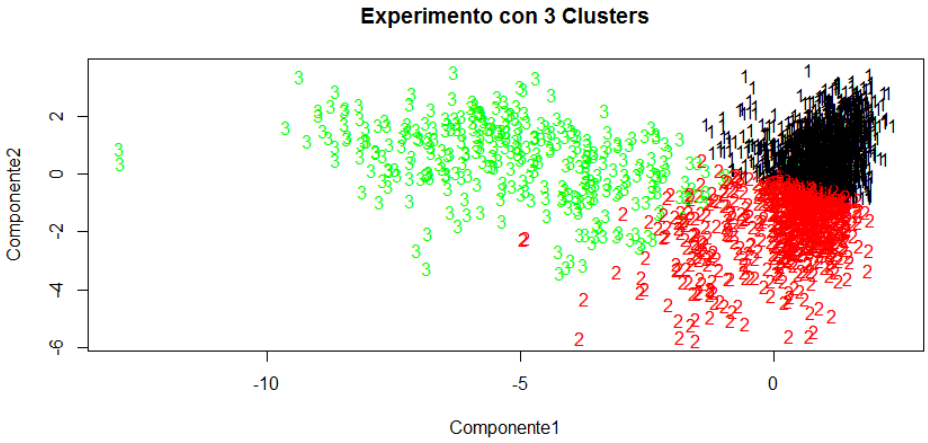

Figura 7

Como observamos, en la solución para 2 clusters se crean dos conjuntos diferenciados pero uno de ellos es bastante disperso, por lo que no podremos concluir que haya muchas similitudes entre los conjuntos.

En la solución para 3 clusters se crean igualmente dos grupos que parecen tener bastante similitud pero uno disperso y no demasiado mezclado con los grupos restantes.

A continuación vamos a ver la restante, con 4 clusters, como ya comentamos anteriormente sería el más óptimo para este conjunto de datos.

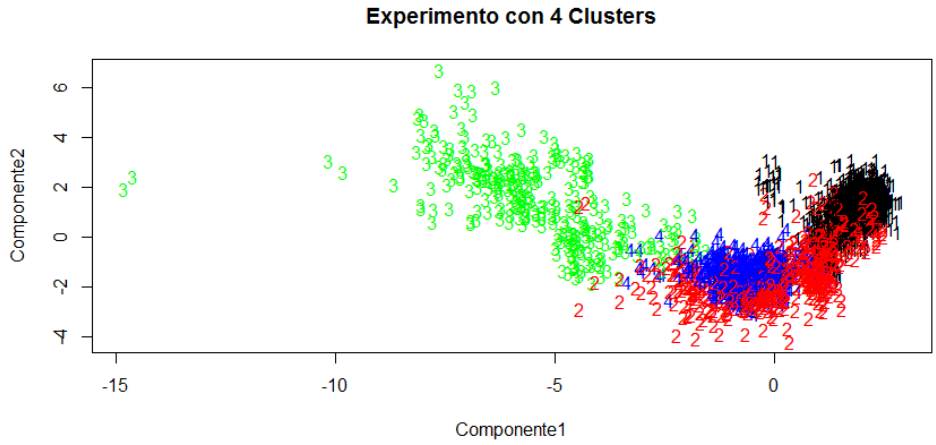

Figura 8

Como podemos ver en la gráfica, a pesar de que un conjunto se mantiene disperso, los 3 conjuntos restantes estan bastante mezclados y cercanos, por lo que entre ellos se observa bastante similitud.

5. Árboles de decisión

Mediante los árboles de decisión vamos a clasificar en diversos grupos los equipos para de este modo predecir si un equipo puede ganar la liga o no en función de los valores de las variables del modelo. Con el conjunto de datos ya cargado, lo siguiente es cargar una serie de librerías necesarias para trabajar con los árboles de decisión, en este caso utilizaremos la librería rpart.

library(rpart)
library(rpart.plot)

Ahora vamos a generar dos conjuntos más de manera aleatoria, uno para entrenamiento y otro para los test.

indice = sample(2, row(Teams2), replace=TRUE, prob=c(0.7, 0.3))
entrenaTeams = Teams2[indice==1,]
testTeams = Teams2[indice==2,]

Una vez generado los dos conjuntos de muestreo con la probabilidad indicada en el parámetro correspondiente, lo siguiente será describir el modelo de clasificación, junto con la clase y las variables que intervienen, y crear el árbol de decisión, así como la predicción.

modeloTeams = LgWin R + H + SO + RA arbolTeams = rpart(modeloTeams, data=entrenaTeams)
arbolTeams2 = rpart(modeloTeams, data=entrenaTeams, parms=list(split=ïnformation”))
plot(arbolTeams) text(arbolTeams) plot(arbolTeams2) text(arbolTeams2)
prediccion = predict(myModel, testTeams) mc = table(prediccion, entrenaTeams.LgWin)
rpart.plot(arbolTeams, type=1, extra=100, cex=.7, box.col=c(”gray99”, ”gray88”)[arbolTeams.LgWin])

El árbol de decisión obtenido al realizar la clasificación con dicho modelo es el que se muestra en las figuras de abajo.

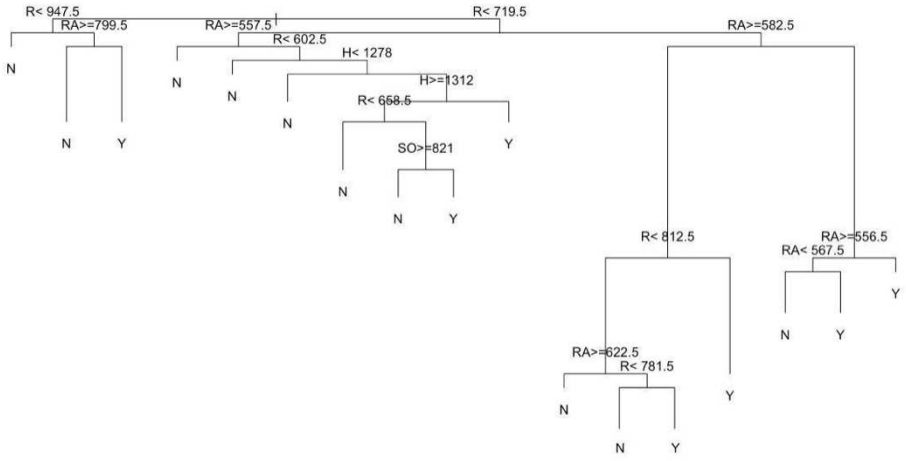

Figura 9

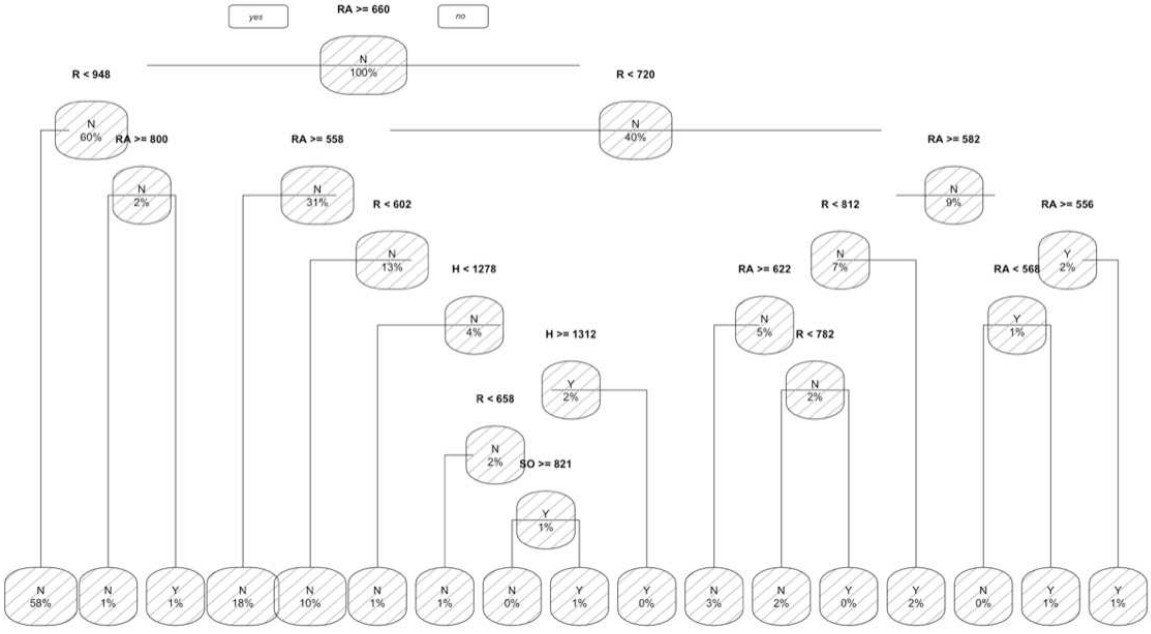

Figura 10

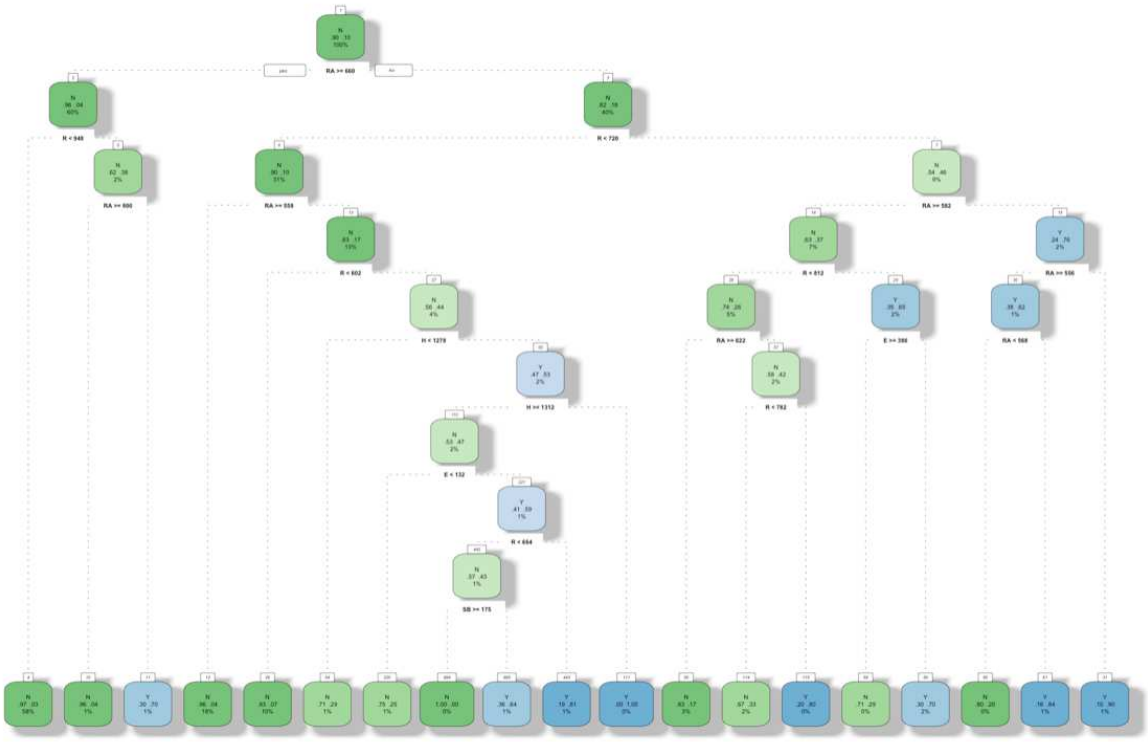

Figura 11

Como se ha podido comprobar en las imágenes anteriores, la clasificación se ha realizado en base a todas las variables del modelo, lo cual quiere decir que el modelo utilizado mantiene una buena relación entre todas ellas. Observando los resultados obtenidos vemos que el porcentaje de ganar la liga en función de esas variables es muy bajo, debido a que son muchas las observaciones sobre las que se ha realizado (más de 2700 equipos) y que sólo hay un campeón por año en cada liga, por lo que la predicción para ganar una liga determinada en función de los resultados que obtenga el equipo en base a las variables del modelo es bajo como los que muestran en las Figuras 9, 10 y 11.

6. Regresión

Para la realización del método de regresión lineal se van a comparar los siguientes campos:

Variable	Medida
W	total de partidos ganados
H	golpes de los bateadores
R	carreras anotadas

Cuadro 3: Conjunto de varibales para Regresión.

Como podemos imaginar, son campos que pueden tener una alta relación. En primer lugar, se ejecuta el método pairs, el cual nos va a relacionar pares de gráficos de dispersión y encontramos la siguiente imagen.

En una primera visualización, podemos observar que el campo H y R, en especial, van a tener una relación positiva, aunque en las otras comparaciones también encontraremos una buena relación.

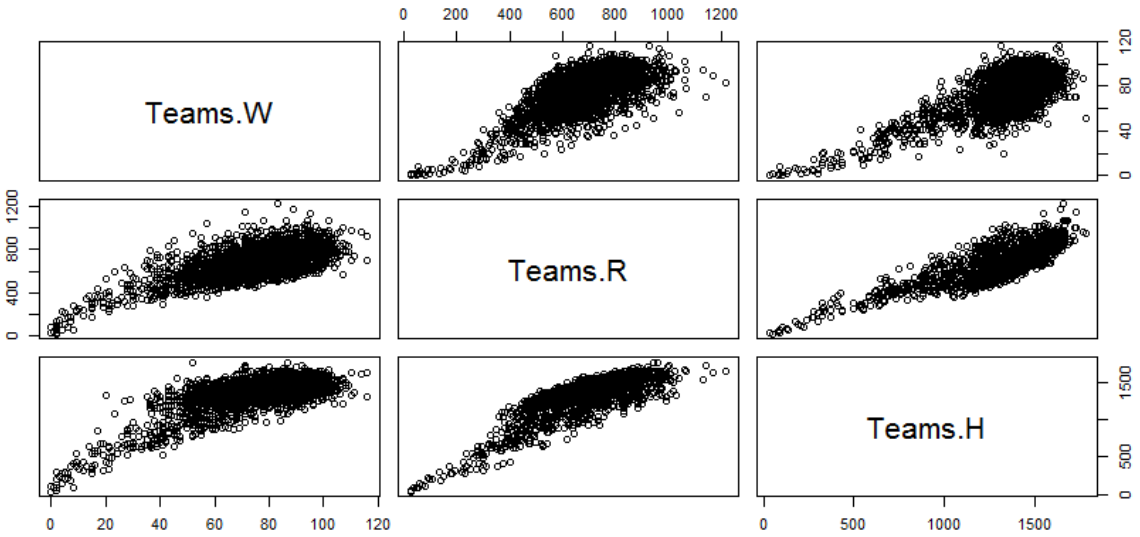

Figura 12

A continuación se va a ejecutar el método cor, y así visualizaremos una matriz que va a determinar la correlación lineal entre los campos a comparar. Este método nos muestra la siguiente matriz:

	Teams.W	Teams.R	Teams.H
Teams.W	1.0000000	0.6782768	0.7264355
Teams.R	0.6782768	1.0000000	0.8023696
Teams.H	0.7264355	0.8023696	1.0000000

Como vemos, R y H tienen una correlación positiva de 80aproximadamente. Por tanto, son buenos datos para realizar este experimento. Vamos a realizar el experimento sobre las variables de mayor correlación, es decir, los golpes de los bateadores y las carreras anotadas.

Tras esto, creamos un modelo donde relacionamos valores dependientes entre ellos, donde decimos que H (golpes de los bateadores) es dependiente de R (carreras anotadas), siendo H la variable independiente y R la dependiente.

Ahora se crea un vector con el comando lineal mode, donde definimos la ecuación, introduciendo en el vector regresión los valores del modelo lineal, a través de los parámetros modelo y el propio dataset.

Hacemos un summary para mostrar todas las características de la regresión que acabamos de hacer.

Call:
lm(formula = modelo, data = newteams2)

Residuals:
Min 1Q Median 3Q Max
-554.47 -44.02 30.90 85.55 271.61

Coefficients:
Estimate Std. Error t value Pr(¿—t—)
(Intercept) 459.10554 12.77059 35.95 ¡2e-16 *
Teams.R 1.30016 0.01837 70.80 ¡2e-16 ***

Signif. codes: 0 ‘*’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

Residual standard error: 131.9 on 2773 degrees of freedom
Multiple R-squared: 0.6438, Adjusted R-squared: 0.6437
F-statistic: 5012 on 1 and 2773 DF, p-value: ¡2.2e-16

Vemos que la variable R e Intercepto son significativos a mas del 99 por ciento, ya que lo indica ***. Por tanto, ahora nos aseguramos que podemos crear la gráfica de regresión, a la cual añadiremos la línea que indica que la regresión lineal es creciente y significativa.

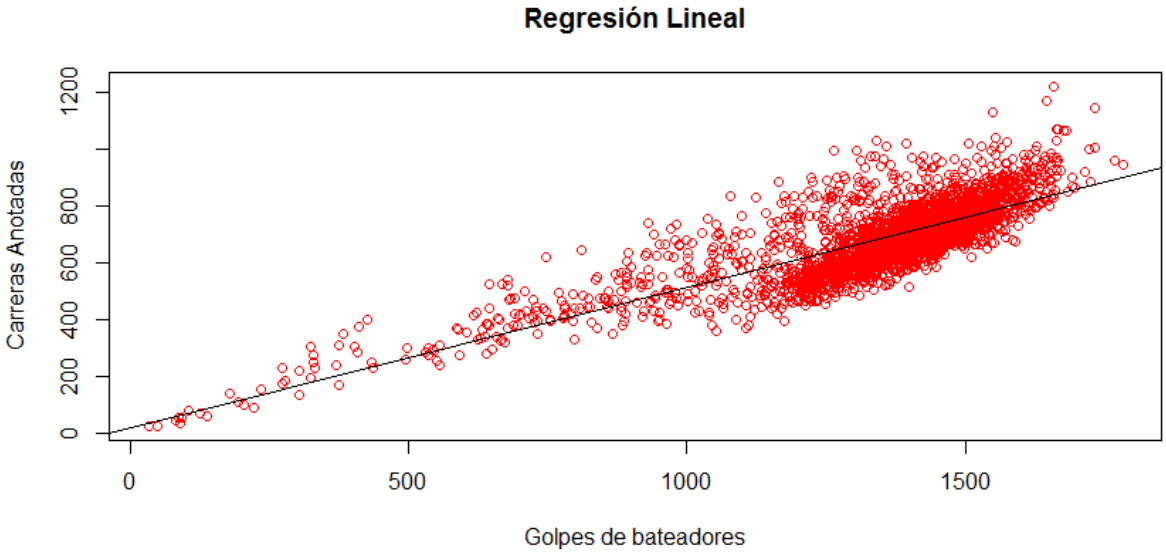

Figura 13

Vemos a través de la gráfica de la Figura 13 que es una relación con pendiente positiva, por tanto el experimento ha resultado satisfactorio.

7. Conclusiones

Hemos decido utilizar un dataset bastante extenso ya que, pensamos que a la hora de llevar este trabajo a un ambiente laboral, la cantidad de datos va a ser bastante grande, por tanto, hemos querido experimentar el abordaje de este tipo de tarea. Por otro lado, hemos decidido utilizar R, ya que, aparte de tener motivación para aprenderlo, ya que nunca antes habíamos trabajado con el, es un lenguaje robusto, potente y muy utilizado en data mining, por tanto hay un gran volumen de documentación referente al mismo.

El análisis de este dataset nos ha permitido profundizar en las técnicas de agrupamiento, clasificación y regresión para grandes volúmenes de datos, mediante los cuales hemos podido sacar importantes deducciones y detalles significativos como pueden ser los factores que se tienen que dar para que un equipo pueda ganar

la liga, que la mayoría de los partidos ganados en algunas ligas se concentran en un subconjunto más pequeño de equipos o que existen un incremento positivo en el número de carreras anotadas en función de los golpes certeros de los bateadores.

Las técnicas empleadas en este trabajo pensamos que han sido las más adecuadas ya que, como comentamos anteriormente, el dataset tiene un gran volumen. Por ello, pensamos que a través de estos métodos veremos con más claridad el análisis de los datos y obtendremos estadísticas más fiables y más optimas para este tipo de análisis.

Referencias

[1] Dataset., *http://www.seanlahman.com/baseball-archive/statistics/*

[2] Apuntes asignatura Master en Ingeniería Informática, *https://decsai.ugr.es/*