LOS JUEGOS OLIMPICOS: ¿PODEMOS PREDECIR EL NUMERO DE OROS QUE VA A CONSEGUIR UN PAIS?

Diego Del Río y Clara Esteban

I. INTRODUCCION

Los Juegos Olímpicos son el mayor evento deportivo internacional multidisciplinario en el que participan atletas de diversas partes del mundo. Son considerados la principal competición del mundo deportivo, en la cual participan mas de doscientas naciones. Existen los Juegos Olímpicos de verano, que se celebran cada cuatro años, y los Juegos Olímpicos de invierno los cuales se celebran cada dos años.

Siempre se cree que los países con una buena situación socioeconómica o mayor numero de población, serán mas propensos a ganar mas medallas en los juegos olímpicos. Incluso se pueden encontrar artículos como el publicado por el diario económico *Expansión* bajo el titulo "*Como influye el PIB en el medallero de los JJOO"* el cual respalda esta idea con las siguientes palabras:

"*En los países con un desarrollo medio o alto se suele cumplir que, cuanto mayor es su PIB, mejor es su puntuación en los Juegos Olímpicos. Puede deberse a que en ellos un alto PIB suele traducirse en una alta recaudación de impuestos, lo que a su vez implica mayores presupuestos para los gobiernos. Estos países, al tener cubiertas las necesidades mas básicas, pueden permitirse dedicar una parte de su gasto a becas y ayudas para entrenar deportistas de elite."*

Parece lógico pensar que una buena situación socioeconómica permitirá a las naciones dedicar una mayor parte de sus presupuestos al deporte, además cuanto mas elevada sea la población del país, mayor cantidad de atletas de prestigio podremos encontrar en el y por consiguiente un mayor número de atletas será enviado a representar al país a los juegos.

Hemos pensado que las medallas ganadas por un país ya sean de oro, plata o bronce no pueden depender tan solo del PIB, sino que tiene que haber otra serie de factores los cuales influyan en el medallero que consiga cada país y que sean independientes del PIB del país. Por consiguiente, queremos ver si realmente son solo los países mas desarrollados los cuales obtienen mejores resultados en los juegos o si por el contrario también pueden obtener grandes resultados países menos desarrollados.
En este análisis trataremos de desarrollar un modelo que pueda predecir los países que mas medallas de oro van a ganar con la mayor precisión posible.

II. CONJUNTO DE DATOS Y CARACTERISTICAS

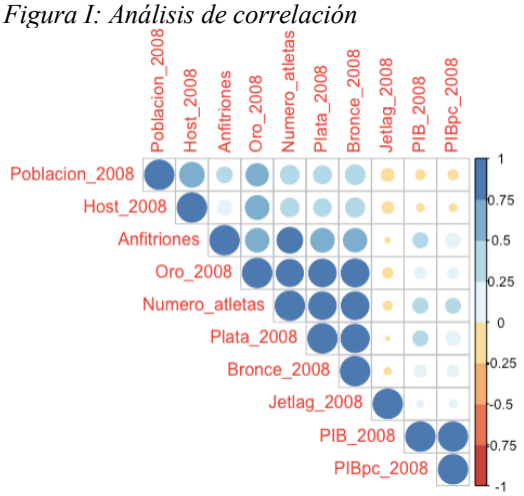
Para este proyecto hemos utilizado una BBDD de resultados históricos de los juegos olímpicos cuyos datos hemos extraído de *Kaggle.* Además, hemos añadido otras variables a la base de datos que nos parece que son interesantes y que influyen en los resultados.

Algunas de las variables que podemos encontrar son Oro_2008 (Nº de medallas de oro ganadas), Host_2008 (si el país es o no anfitrión), PIBpc_2008 (el PIB per cápita del país), etc. El numero de elementos de nuestra BBDD es de 112 que son los países que participaron en los JJOO.

Hemos iniciado nuestro análisis tratando de conocer los datos con los que estamos trabajando. Para ello, hemos realizado un análisis de la correlación entre las variables de nuestra base de datos (*véase la Figura I*). Prácticamente todas las variables mantienen una relación lineal positiva y muchas de ellas

con bastante correlación. El que muchas de las variables tengan una relación lineal positiva es lógico, ya que, por ejemplo, un país a mayor numero de habitantes mas posibilidades hay de tener atletas prestigiosos y mas posibilidades por tanto de ganar una medalla. La única variable que parece tener una relación lineal negativa es la variable Jetlag_2008, lo cual también puede ser lógico ya que a mayor numero de horas de diferencia entre el país de origen del atleta y el país anfitrión peor van a descansar los jugadores y por tanto se entiende que van a tener una peor actuación en los JJOO.

Figura I: Análisis de correlación

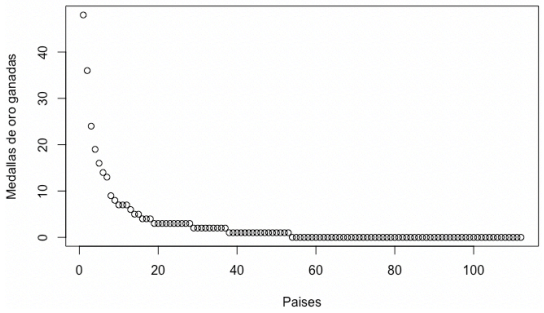


Las altas relaciones lineales o no lineales de las variables pueden deberse o bien porque ocurren simultáneamente o bien porque tienen alguna relación causal. Por tanto, una alta correlación no es un buen indicio de selección de variables. Atendiendo a nuestro grafico de correlación, observamos que muchas variables mantienen correlaciones altas con otras variables. Sin embargo, existen otras variables como pueden ser Jetlag_2008 (Nº de horas de diferencia entre el país y el país anfitrión), PIBpc_2008 (PIB per cápita del país) que tienen correlaciones muy bajas con el resto de las variables. Sabiendo que son variables que conocemos antes del comienzo de los Juegos Olímpicos, creemos que podrían tener una alta capacidad predictiva y por tanto ser buenas variables objeto.

Además, hemos seleccionado la variable Poblacion_2008 (Nº de habitantes del país) y Numero_atletas (Nº de atletas que cada delegación presenta a los JJOO) las cuales a pesar de mantener altos grados de correlación con el resto de variables, pueden ser buenas variables explicativas ya que son datos que se conocen antes de los juegos y los cuales son constantes, por ejemplo si un país dice que va a presentar X deportistas, si uno de esos deportistas se lesiona lo sustituyen por el siguiente que este en la lista de deportistas no convocados.

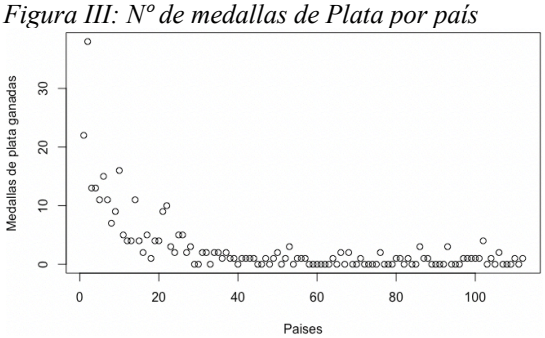
Observando el grafico de medallas de oro ganadas *(véase la Figura II),* a partir del país 14 se puede apreciar como el numero de medallas de oro ganadas por cada país baja drásticamente hasta alcanzar las 0 medallas de oro ganadas.

Figura II: Nº de medallas de Oro por país



Como podemos ver en la siguiente Figura *(véase Figura III)* la cual muestra el número de medallas de plata ganadas por cada país, vemos como los países que más medallas de oro han ganado son también los que han obtenido un mayor numero de medallas de plata. Se puede observar como a partir del país numero 22 el numero de medallas de plata ganada por los países se estanca.

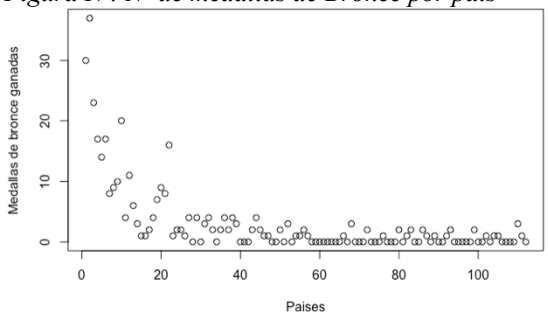
Figura III: Nº de medallas de Plata por país



Por último, hemos realizado un grafico que muestra el número de medallas de bronce ganadas por cada país, en el podemos observar también como los países que han obtenido mejores resultados en los juegos olímpicos, es decir los que han obtenido más

medallas de oro y plata, son los que obtuvieron más medallas de bronce. *(véase Figura IV)*

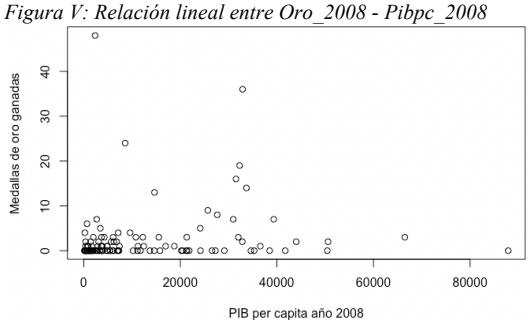
Figura IV: Nº de medallas de Bronce por país



Hemos realizado un modelo de regresión lineal múltiple Oro_2008 ~ PIBpc_2008 + Poblacion_2008 + Anfitriones + Numero_atletas + Jetlag_2008, según este modelo únicamente son significativas las variables Población_2008 y Numero_atletas atendiendo a los p-valores.

Según este modelo el coeficiente de variación de PIBpc_2008 es negativo (-1.260e-05). Sin embargo, el grafico de nube de puntos entre estas dos variables nos muestra lo contrario, que pese alguna que otra excepción (como es el caso de China, el cual con un PIB per cápita bajo obtiene grandes resultados) los países que tienen un PIB per cápita superior consiguen un mayor numero de medallas de oro que los que están por debajo, siempre existen países que pese a tener un PIB per cápita elevado no saben aprovecharlo para mejorar en deportes, podríamos decir que son anomalias.

Figura V: Relación lineal entre Oro_2008 - Pibpc_2008



Esto nos quiere decir que no es muy sensato medir el impacto de una variable en otra mediante un modelo de regresión lineal múltiple. Por lo tanto, se entiende que seria mas sensato el medir el impacto de las distintas variables sobre Oro_2008 mediante un modelo de regresión lineal simple ya que de esta forma se estima mejor el efecto global de cada una de las variables sobre el numero de medallas de oro ganadas.

Los resultados obtenidos mediante el modelo de regresión simple son: 5.498e-05 para PIBpc_2008, 12.379 para Anfitriones, 0.046 para Numero_atletas, -0.2127 para Jetlag_2008 y 2.239e-08 para Población_2008. El mayor cambio ha sido que la variable Anfitriones ha pasado de ser una variable que no es significativa en el modelo de regresión lineal múltiple a una variable que es significativa en el modelo de regresión lineal simple. Otro cambio notable ha sido en la variable Jetlag_2008, la cual paso de tener un impacto positivo mediante el análisis por modelo de regresión lineal múltiple a tener un impacto negativo mediante el modelo simple.

Hemos estudiado si podíamos hacer un análisis de PCA con las variables que nos interesa, pero, observamos que en el grafico de la correlación PIBpc_2008, Jetlag_2008, Numero_atletas, Anfitriones y Poblacion_2008 no mantienen un alto grado de correlación. Este análisis solo tiene sentido si los datos están muy correlacionados, así seria sensato el extraer el componente común de la información conjunta de cada variable y de este modo reducir la dimensión del conjunto. Pero como nuestra BBDD tampoco es tan grande, no tiene mucho sentido si quiera el tratar de hacer este análisis.

Si quisiésemos predecir el numero de medallas de plata o de bronce que va a obtener un país realizaríamos el mismo proceso, solo que no saldrían las mismas variables significativas que para conseguir oros.

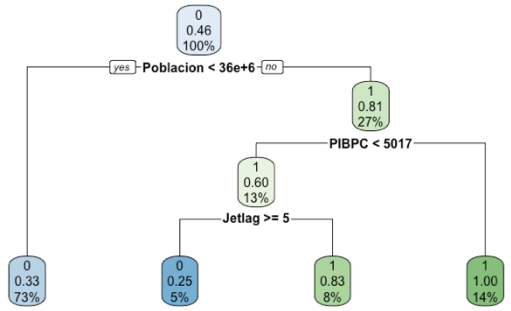
III. METODOS

Ahora que ya tenemos nuestro modelo predictivo definido, decidimos usar este método para conocer su capacidad predictiva. Partimos de nuestro conjunto de datos del modelo.

Decidimos construir un modelo de árbol de clasificación para la variable cualitativa Ganadores oro 2008, en la cual aparecen

reflejados con un 1 los países que ganaron alguna medalla de oro y con un 0 los que no ganaron ninguna.
Para esto hemos empleado el parámetro de complejidad ("cp") el cual se utiliza para controlar el tamaño del árbol de decisión y para seleccionar el tamaño óptimo del árbol. Cuanto mas pequeño sea este parámetro mas se extenderá nuestro árbol.
Con esto decidimos que nuestro parámetro de complejidad optimo seria 0.01.
Tras esto nos salió el siguiente árbol, hay que tener en cuenta que al estar cogiendo muestras aleatorias los modelos con menor error de predicción que filtramos van variando.

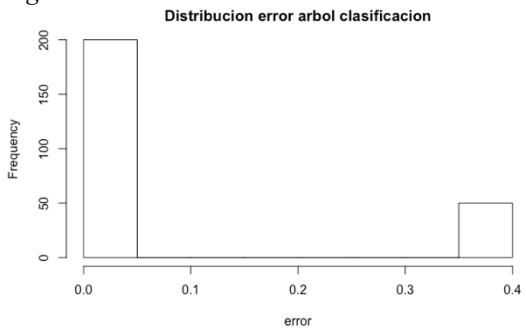
Figura VI: Diagrama de Árbol de decisión.



El primer nodo de nuestro árbol tiene el 100% de la muestra (que tiene un 46% de eventos iguales a 1), si nos quedáramos en ese nodo y no hiciésemos nada mas, deberíamos predecir que un país escogido al azar de nuestra población no va a ganar ninguna medalla de oro, ya que el 54% de los países no lo hacen. Si vamos al siguiente nodo si un país tiene menos de 36000000 habitantes entonces predecimos que un país no va a ganar medalla de oro, el 73% de la muestra son países con menos de 36millones de habitantes y de estos el 33% han ganan medalla de oro. Si la población del país resulta ser superior a los 36millones entonces el árbol se bifurca. Si el PIB per cápita del país es superior a 5017$ entonces predecimos que el país va ha ganar alguna medalla, el 14% de la muestra son países con mas de 5017$ de PIB per cápita. Si por el contrario el PIB per cápita es inferior entonces la cosa se bifurca. Si el cambio horario del país con respecto a el lugar donde se realizan los juegos es superior o igual a 5 horas entonces predecimos que el país no va a ganar ninguna medalla de oro, el 5% de la muestra tiene mas de 5 horas de diferencia y de estos el 25% no ha obtenido ninguna medalla de oro. Por el contrario, si el cambio horario es inferior a las 5 horas predecimos que el país va a ganar alguna medalla de oro, el 8% de la muestra tienen menos de 5 horas de diferencia y de estos el 83% ha ganado medalla de oro.

Aplicando la validación cruzada a nuestro modelo optimo, obtenemos la siguiente distribución. *(véase Figura VII)*

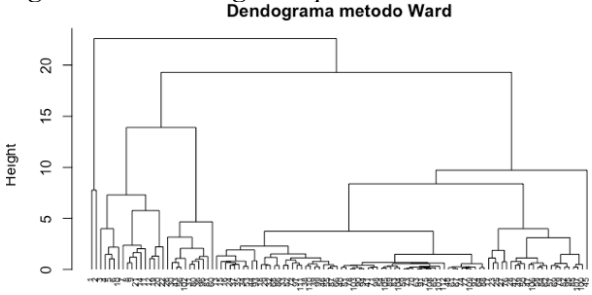
Figura VII: Distribución de errores árbol



Con la distribución de errores que nos ha dado, nos dice que nuestro diagrama de árbol falla bastante poco, pero de vez en cuando se desvía 0.4 decimas.

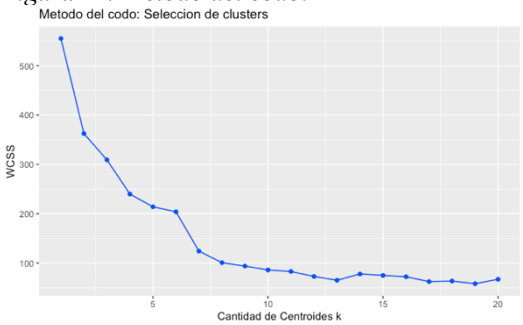
Una vez aquí, decidimos hacer un análisis descriptivo de nuestras variables, y así saber si podemos mejorar nuestro modelo. Para esto tratamos de averiguar si podíamos hacer grupos con los países mediante alguna relación que encontremos entre ellos. Para esto utilizaremos los dendogramas. Este método nos permite definir similitudes entre los distintos elementos (los países) a través de una medida de distancia entre los mismos. Para realizar el siguiente dendograma hemos utilizado el método "Ward", el cual forma grupos de una manera que minimiza la perdida asociada con cada grupo.*(véase Figura VIII)*

Figura VIII: Dendograma por método Ward



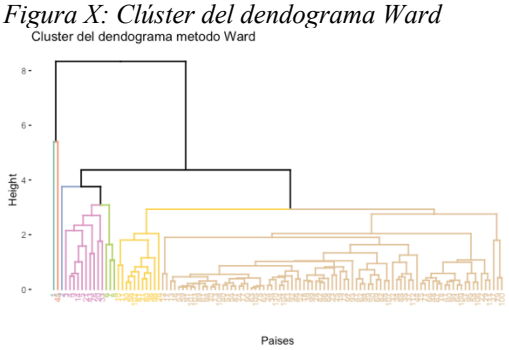
Una vez tenemos los países agrupados por similitudes, nos preguntamos cuantos grupos de países deberíamos hacer. Para esto realizamos el método del codo, el cual nos dirá el numero de k optimo. Consiste en enfrentar la suma de todas las distancias de las observaciones con sus respectivos centroides con el nº de clusters formados. A mayor numero de clusters, menor distancia total. El k a partir del cual veamos que el decremento de la distancia total se vuelva irrelevante, será el adecuado para formar clusters.

Figura IX: Método del codo.



Observando el grafico *(Figura IX)* vemos que el numero de clúster va a ser 7 ya que a partir de 7 se vuelve irrelevante el decremento. Una vez sabemos esto pasamos a representar el clúster del dendograma.

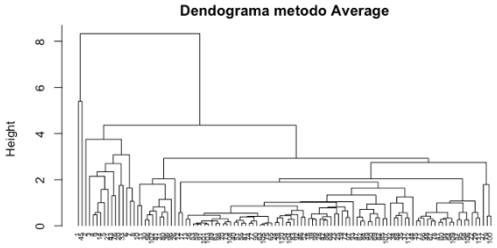
Figura X: Clúster del dendograma Ward



Como vemos en el clúster del dendograma (*Figura IX)*, se han hecho 7 grupos, por ejemplo, uno de ellos contiene el país número 1 y otro al país número 2 que se corresponden con China y USA respectivamente, esto puede tener sentido ya que ambos países tienen una gran población, presentan una gran cantidad de atletas a los juegos y ambos ganan una gran cantidad de medallas de oro, teniendo esto en cuenta parece lógico que los agrupe en grupos diferentes al resto.

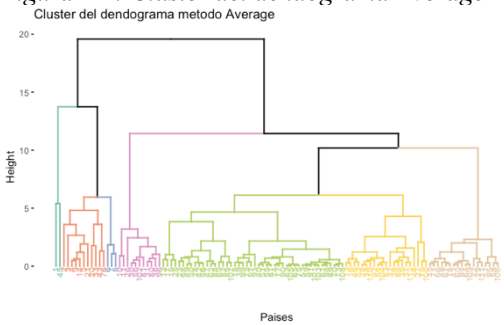
Tras esto decidimos realizar otro dendograma, pero en este caso mediante el método "Average", el cual agrupa en función del promedio de las distancias de los componentes de un conglomerado con respecto al promedio de las distancias de otro grupo. *(véase Figura XI)*

Figura XI: Dendograma por método Average



Tras esto al igual que hicimos anteriormente agrupamos por grupos de países.

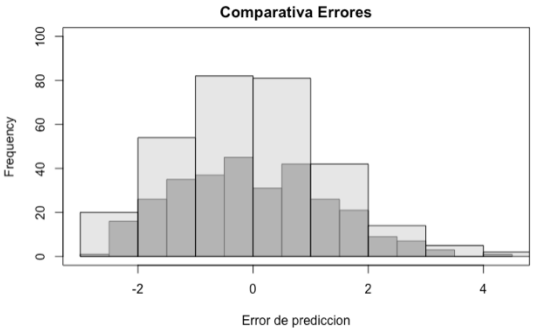
Figura XII: Clúster del dendograma Average



Como se observa en el grafico se ven los siete grupos que realiza. Entre estos tres grupos vemos que otra vez agrupa a China y USA.

Tras esto podemos sacar la distribución de los errores de predicción de ambos dendogramas. *(véase Figura XIII)*

Figura XIII: Distribución de errores de dendogramas



Observando el grafico de la distribución de errores vemos como falla mucho mas el dendograma de Ward que el dendograma de Average. Podemos ver como para ambos el

error esta mas concentrado en el lado izquierdo.

Otro método para hacer grupos de clusters es el método de los K-means que sigue un criterio especifico de cercanía entre las observaciones de dos variables, se diferencia de la técnica anterior en que en esta solo empleamos dos variables y en la anterior empleamos todas las variables seleccionadas para el estudio. K-means es un método de agrupamiento, que tiene como objetivo la partición de un conjunto n de observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio (centroide) es mas cercano. A medida que vayamos aumentando el numero de clusters que queremos formar, la distancia de cada observación con su centroide se ira reduciendo hasta llegar a 0, en el caso de tener tantos clusters como observaciones.
Hemos desarrollado una serie de clusters para predecir el numero de medallas de oro.

Figura XIV: Cluster PIB per cápita - Población

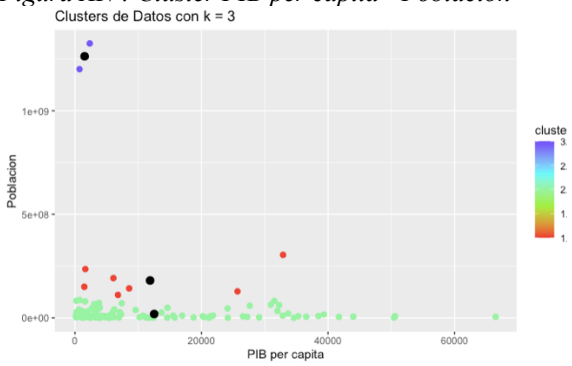


Figura XV: Cluster NºAtletas - Población

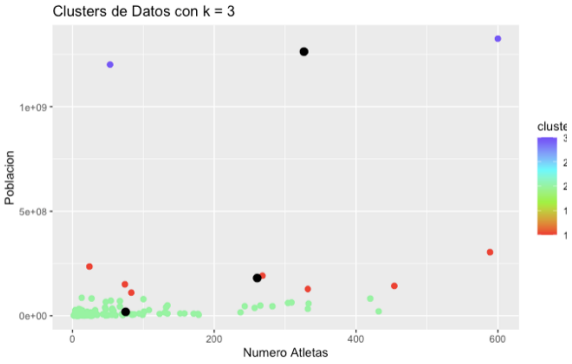


Figura XVI: Cluster NºAtletas – PIB per cápita

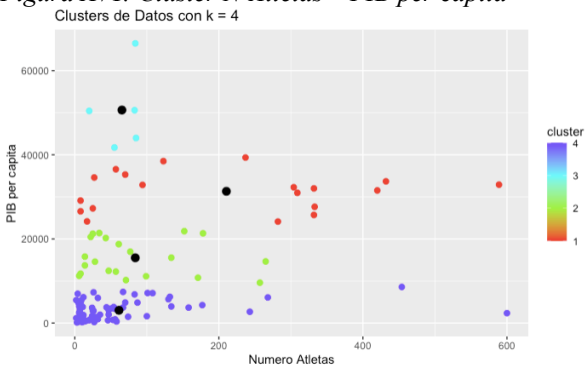
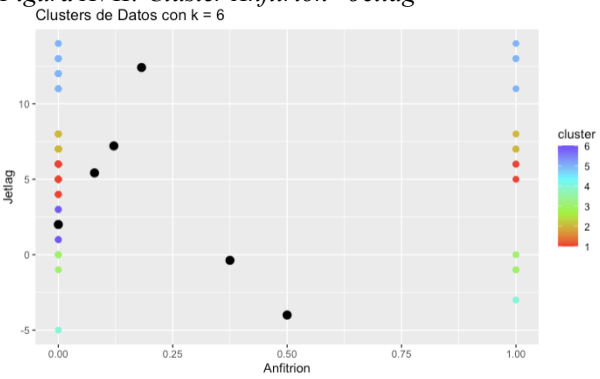
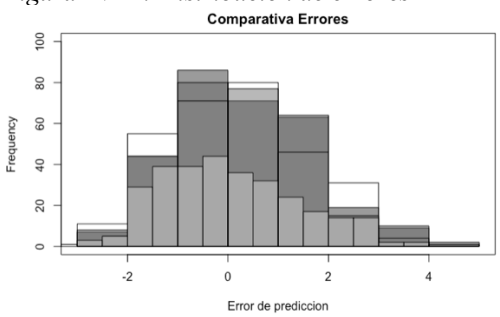


Figura XVII: Cluster Anfitrion - Jetlag



Todos los clusters formados son bastante poco complejos, creemos que es debido a la poca variedad que presentan las observaciones. Los arboles de regresión formados son prácticamente iguales, no añaden novedades notorias. Por ultimo hemos comparado, como hicimos con otros métodos, las distribuciones de los errores de predicción de estos últimos arboles, que formamos con estos pares de varibales, mediante la validación cruzada. Como podemos ver el la Figura XVIII, todas las distribuciones son bastante parejas a excepción de la que resulta del par PIBpc - NºAtletas. A pesar de eso podemos ver una cierta semajanza a las distribuciones que formaban los arboles construidos a partir de dendogramas

Figura XVIII: Distribución de errores



Con esto, llegamos a la conclusión de que los mejores modelos predictivos que hemos conseguido son los modelos de arboles predictivos donde previamente hemos agrupado las variables bien mediante dendogramas o siguiendo el criterio k-mean, ya que la distribución de errores de ambos es similar.