

# Trabajo Práctico Integrador

## Minería de Datos - Grupo 1

Facundo Miglierini, Pamela Heredia, Juan Cruz Cassera Botta

### Introducción

El [dataset elegido](#) trata sobre el powerlifting, un deporte de fuerza que consiste en la realización de tres ejercicios de levantamiento de peso: la sentadilla, el press de banca y el peso muerto.

En cuanto a la forma en la que se encuentra estructurado este conjunto de datos, podemos decir que cada fila del mismo consta de la participación de un atleta en una competencia de powerlifting. Por cada participación existen atributos que almacenan información sobre las características personales de los levantadores de peso, tales como su nombre y peso corporal. Por otro lado, se registran los resultados obtenidos a nivel competitivo, junto con algunas fórmulas que ayudan a determinar el rendimiento de los competidores teniendo en cuenta su peso corporal y la carga levantada en los diferentes ejercicios. Por último, se describe la información relacionada al evento donde participan los atletas.

Al día de hoy, se recolectaron datos de 2.924.617 participaciones de diversos atletas en múltiples competencias, que van desde septiembre de 1964 hasta el mes de junio del año corriente.

Cabe destacar que si bien el dataset dispone de una gran cantidad de entradas, el mismo presenta un elevado número de valores faltantes dispersos entre los distintos atributos. Esto es así porque existen competencias que emplean modalidades diferentes y, en consecuencia, recolectan información diferente.

### Hipótesis

Para la realización de este trabajo consideramos emplear dos modelos que describan el conjunto de datos, los cuales fueron implementados mediante las técnicas de agrupamiento y reglas de asociación. Para que sea posible su realización, hemos utilizado el algoritmo K-Medias y el algoritmo Apriori, respectivamente.

Por otro lado, implementamos un conjunto de modelos predictivos que fue llevado a cabo mediante diferentes técnicas de minería de datos, tales como Naïve Bayes y árboles de clasificación. Estos modelos los utilizaremos para realizar predicciones sobre un mismo atributo de clase que detallaremos posteriormente.

Tal como indicamos previamente, uno de los métodos empleados para poder describir los datos consiste en realizar un modelo de agrupamiento. A través del mismo buscamos conocer la forma en la que se relacionan los distintos atributos del dataset para cada uno de los grupos generados. Consideramos como hipótesis el hecho de que podríamos agrupar a los atletas de

acuerdo a sus resultados obtenidos en cada levantamiento, y a partir de allí discernir un conjunto particular de características que los identifiquen, como por ejemplo el año en el que se realizó la competencia, el peso corporal de los atletas, su fuerza relativa, su sexo, entre otros.

En cuanto a las reglas de asociación, nos basamos en la idea de que existen diferentes relaciones entre los valores de los distintos atributos. Partiendo de dicha hipótesis, el objetivo de este modelo consiste en ver cómo se relacionan los atributos del dataset.

En lo que respecta a la realización del modelo predictivo, podemos decir que partimos desde la hipótesis basada en que se puede predecir el sexo de un competidor a partir de las características físicas y el rendimiento del mismo. En este caso, nuestro objetivo consiste meramente en poner a prueba diversos modelos de predicción, aplicando las distintas técnicas de minería vistas en clase para dichos fines. Posteriormente, evaluaremos la precisión con la que se realizaron las predicciones para los distintos modelos.

## Preprocesamiento

Tal como indicamos previamente, el dataset elegido para este trabajo presenta una gran cantidad de datos faltantes. Este es el principal problema a destacar, ya que el algoritmo de K-Medias que conocemos para aplicar los agrupamientos requiere que todas las entradas del conjunto de datos tengan un valor no nulo.

Por otra parte, debemos tener en cuenta que el conjunto de datos elegido presenta un tamaño considerable, lo cual provoca que todo procesamiento llevado a cabo sobre el mismo conlleve un elevado costo en términos de tiempo de ejecución. Para solventarlo, optamos por eliminar las columnas de menor importancia, las cuales suelen almacenar sustantivos propios que no aportan ningún tipo de información que enriquezca el modelo. Entre dichos atributos podemos mencionar el nombre de la persona y su nacionalidad, el nombre de la competencia, entre otros.

Además, optamos por conservar únicamente el levantamiento de mayor peso para cada ejercicio, teniendo en cuenta los tres primeros intentos dado que el cuarto no es tenido en cuenta dentro de los torneos. Por último, descartamos determinados atributos que almacenan valores utilizados por ciertas federaciones para categorizar o comparar el rendimiento de los atletas, pero que presentan muchos datos faltantes. En el anexo 2 se describe detalladamente cuáles atributos fueron eliminados.

En lo que respecta a la edad de la persona, pudimos rellenar los datos faltantes a partir del promedio del valor indicado en el atributo “Age Class”, el cual indica la categoría de edad en la que se inscribió el atleta al momento de realizar la competencia. A su vez, aplicamos la misma idea haciendo uso del atributo “BirthYearClass” para los casos en los que tampoco se encuentra indicado el valor de categoría de edad. Una vez hecho esto, conservamos únicamente el atributo de edad.

Posteriormente, eliminamos determinadas filas del conjunto de datos, ya que a nuestro criterio podrían alterar los resultados. Entre ellas, se encuentran las participaciones de los atletas que se desconoce si fueron sometidos a un test de drogas, si no fueron sometidos a un test de drogas, quedaron descalificados o no aparecieron en la competencia, entre otros. Además, decidimos borrar las filas de los atletas con valor del atributo “Sex” equivalente a “Mx” debido a la poca cantidad de entradas que contenían esta variante.

Cabe destacar que existen ocasiones en las que los atletas fallan sus levantamientos. Para representar esto, los valores de carga indicados en los mismos son almacenados con valores

negativos. Con el objetivo de disponer de todos los datos con valores positivos, optamos por dar por realizados estos levantamientos, restándole 5, 10 o 15 Kg de peso para los ejercicios “Bench”, “Squat” y “Deadlift”, respectivamente.

En lo que respecta a la fecha de la competición, decidimos conservar únicamente el año para simplificar los datos. Junto a esto, redondeamos el valor de la edad para aquellos casos en los que se encuentra cargada como un número real, de forma tal que podamos convertir el tipo del atributo a entero y disminuir el peso del dataset en consecuencia.

## Modelos Realizados

En esta sección se presentará detalladamente el trabajo realizado sobre los distintos modelos de minería de datos.

### K-Medias

#### Preprocesamiento

Para poder aplicar el algoritmo de k-medias, tuvimos que numerizar el atributo relacionado al sexo y el atributo que indica el equipamiento utilizado durante la competencia. Para numerizar el sexo empleamos numerización binaria, dado que no existe un valor que no tenga más peso sobre otro. Sin embargo, a la hora de modificar el equipamiento decidimos establecer una numerización manual, de manera que mayor sea el número utilizado cuanto menos restricciones existan en el equipamiento que puede utilizar el atleta.

Además, decidimos descartar el atributo del peso total levantado, ya que consideramos que es un valor redundante que se deriva del peso indicado en los tres ejercicios. Por lo tanto, no aporta información relevante a la hora de intentar describir distintos grupos en el conjunto de datos.

Por último, eliminamos el atributo “WeightClassKg” debido a que se trata de un atributo cualitativo que no vale la pena convertir a uno de tipo cuantitativo, ya que el peso del atleta lo podemos obtener directamente desde el atributo “BodyweightKg” que es de tipo entero.

#### Procesamiento

Una vez preprocesado el dataset, decidimos aplicar normalización sobre todos los atributos, de manera que ninguno tenga mayor influencia que el resto. A continuación, pusimos a prueba el algoritmo de k-medias provisto por Rapidminer con diferentes valores de k, evaluando sus rendimientos con el índice de Davies Bouldin. Utilizamos esta métrica en lugar de Silhouette por el elevado costo que conlleva su ejecución.

Finalizadas las pruebas de ejecución de los algoritmos, notamos que el valor de Davies Bouldin disminuye a medida que crece el valor de k. De acuerdo a este índice, el mejor agrupamiento es realizado cuando se presentan dos grupos diferentes.

## Davies Bouldin

Davies Bouldin: 0.112

Índice de Davies Bouldin obtenido con  $k = 2$

## Conclusión

Cuando el valor de  $k$  equivale a 2, los centroides tienen los siguientes valores:

Attribute	cluster_0	cluster_1
Equipment	-0.097	0.055
Age	0.033	-0.019
BodyweightKg	-0.660	0.372
Glossbrenner	-0.665	0.376
Date	0.107	-0.060
Squat	-0.940	0.530
Bench	-0.992	0.560
Deadlift	-1.024	0.578
Sex_F	1.201	-0.678
Sex_M	-1.201	0.678

Valores de los centroides para cada cluster cuando  $k = 2$

En base a estos resultados, podemos sacar algunas conclusiones. Teniendo en cuenta qué tan alejados de la media se encuentran los valores de los centroides respecto de los diferentes atributos, podemos destacar cuáles son los factores determinantes que permiten distinguir un grupo del otro.

Como podemos observar, el sexo tiene un gran impacto a la hora de caracterizar el agrupamiento. Por un lado se encuentra un grupo que presenta en su mayoría mujeres, y por el otro lado un grupo que dispone de hombres en su mayoría. Partiendo de esto, notamos también que el peso corporal del grupo donde predominan las mujeres tiende a ser considerablemente menor respecto del otro grupo. En consecuencia, los pesos levantados en todos los ejercicios también son menores.

Si analizamos de forma más detallada la diferencia existente entre los valores de los levantamientos de ambos grupos, podemos ver que la mayor diferencia existe en el ejercicio “Deadlift”. Esto quiere decir que es el movimiento que más influenciado se encuentra por el sexo y el peso corporal. A esta variante la sucede el ejercicio “Bench”, y por último se encuentra la variante “Squat”, siendo esta la menos dependiente del peso y el sexo de la persona.

Por otra parte, también podemos distinguir que la edad no es un factor determinante a la hora de determinar el rendimiento de los atletas. Tampoco lo son el equipamiento utilizado para realizar los levantamientos, ni la fecha en la que se realizó el evento. Esto sugiere que no

ocurrieron demasiados cambios en el deporte con el transcurso de los años. Sin embargo, dichos resultados pueden deberse a que las entradas corresponden en su mayoría a la última década, teniendo los levantamientos de esta época mayor influencia sobre el modelo.

A continuación, mostraremos un diagrama de coordenadas paralelas que muestra cómo se realiza el agrupamiento de acuerdo a los diferentes atributos del conjunto de datos.

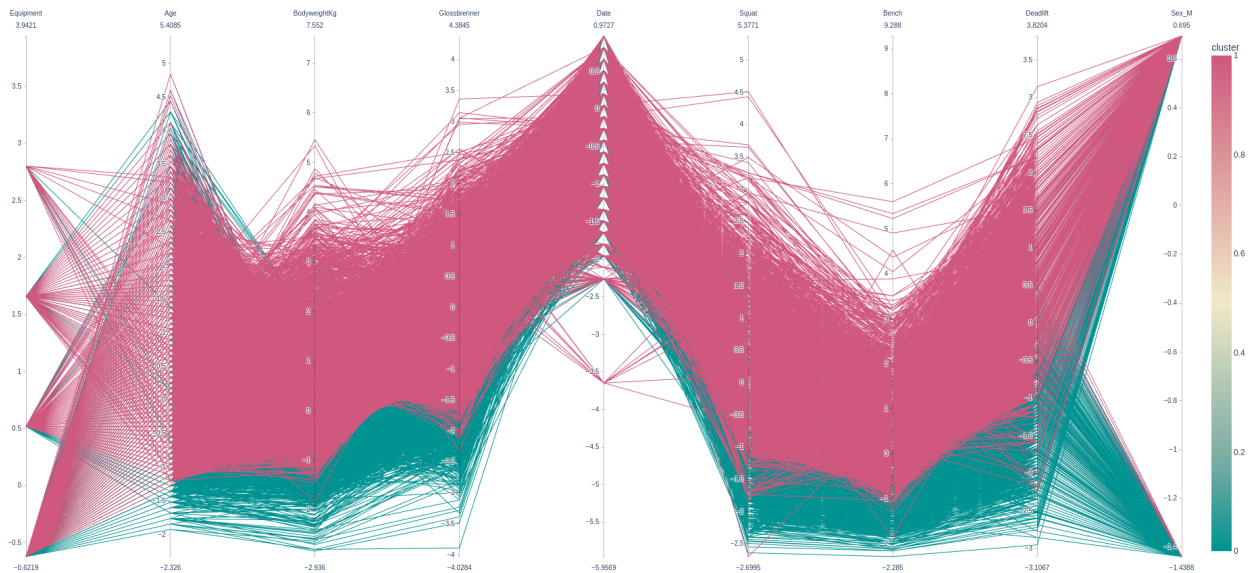


Diagrama de coordenadas paralelas que muestra la relación existente entre los valores normalizados de los atributos para cada grupo. Las líneas de color verde pertenecen al cluster 0, mientras que las líneas de color rosa pertenecen al cluster 1

En base a lo mostrado en la imagen, podemos ver que existe cierto solapamiento entre los elementos de ambos grupos respecto de los atributos seleccionados para emplear el gráfico (se utilizó únicamente el atributo de sexo “Sex\_M” para simplificar el diagrama).


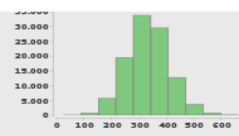

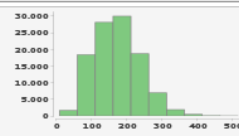

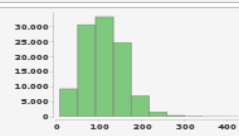


## Naïve-Bayes

### Preprocesamiento

Para generar el modelo de Naïve-Bayes eliminamos nuevos atributos, quedándonos con aquellos cuyos valores siguen una distribución normal. Por ejemplo, los atributos que indican el peso levantado en los ejercicios realizados y la edad de los lifters mantienen este tipo de distribución en sus valores.

Por otro lado, generamos un nuevo atributo nominal “Sex” para poder usarlo como etiqueta. Su valor es generado a partir de los atributos “Sex\_M” y “Sex\_F”, los cuales fueron descartados en este modelo.

Además, eliminamos el atributo “BodyweightKg”, ya que luego de diversas pruebas llegamos a la conclusión de que su presencia en el dataset no impacta de forma positiva en la performance de este modelo.

Result History		SimpleDistribution (Naive Bayes)					
Name	Type	Missing	Statistics	Filter (10 / 10 attributes): <input type="text" value="Search for Attribute"/>			
 Glossbrenner	Real	0	 Open visualizations	Min 27.320	Max 660.050	Average 333.702	
 Squat	Real	0	 Open visualizations	Min 10	Max 515	Average 173.147	
 Bench	Real	0	 Open visualizations	Min 9.100	Max 426	Average 111.438	
 Deadlift	Real	0	 Open visualizations	Min 15.900	Max 410	Average 194.221	

Distribución de los valores para los atributos “Glossbrenner”, “Squat”, “Bench” y “Deadlift”

Tras finalizar el preprocesamiento, nos quedó un dataset con los siguientes atributos: “Sex”, “Age”, “WeightClass”, “Glossbrenner” y los pesos de cada uno de los tres ejercicios realizados en la competencia.

## Procesamiento

Usamos el algoritmo Naïve-Bayes sobre un total de 531.557 ejemplos, de los cuales el 80 % fue usado para entrenamiento y el 20 % restante para testing. A la hora de implementar dicho algoritmo, tuvimos en cuenta el uso de una corrección de Laplace.

## Conclusión

Llegamos a la conclusión de que sí es posible predecir el sexo de los lifters con este modelo con una tasa de acierto relativamente alta a partir del rendimiento obtenido en la competencia, la edad y las características físicas de los diferentes atletas.

Tal como se mencionó anteriormente, pudimos notar que el atributo “BodyweightKg” por sí solo no tiene relevancia en el modelo ya que su presencia no produce mejoras en la tasa de aciertos. Sin embargo, el atributo “Glossbrenner” almacena valores que destacan la fuerza relativa de los competidores, la cual es calculada en base al peso corporal del atleta y el peso levantado durante la competencia. Esto implica que de forma indirecta el peso corporal sea tenido en cuenta en el modelo. Por lo tanto, es acertado pensar que dicho valor no aporta al mismo cuando es utilizado de forma independiente. Sin embargo, se vuelve útil cuando se lo utiliza junto con los valores relacionados a las cargas levantadas en los diferentes ejercicios.

Tras realizar un análisis de rendimiento del modelo obtenido sobre los datos de prueba, notamos que el mismo obtuvo una tasa de acierto del 94,57 %.

SimpleDistribution (Naive Bayes)			
<input checked="" type="radio"/> Table View <input type="radio"/> Plot View			
<b>accuracy: 94.57%</b>			
	true F	true M	class precision
pred. F	33106	4251	88.62%
pred. M	1520	67434	97.80%
class recall	95.61%	94.07%	

Matriz de confusión a partir de un modelo Naïve Bayes aplicado sobre un conjunto de datos de prueba

## Árbol de clasificación

### Preprocesamiento

En este caso, a diferencia del modelo Naïve-Bayes, decidimos conservar el atributo usado para representar la cantidad total de kilos levantados entre los tres ejercicios.

Por otra parte, conservamos también los atributos que representan los puntos obtenidos en cada uno de los ejercicios realizados por el lifter. Si bien la decisión de conservar dichos atributos disminuye en un 0.2 % el accuracy del modelo, se obtiene un árbol mucho más legible, con menos ramas y menor tamaño que el obtenido al eliminar estos atributos. También generamos un nuevo atributo “Sex”, con los valores “F” para femenino y “M” para masculino, para poder usarlo como etiqueta, eliminando los atributos “Sex\_F” y “Sex\_M”.

Por último, decidimos eliminar ciertos atributos que aumentaban la complejidad en el modelo e incrementaban los costos en términos de tiempo de procesamiento. Entre ellos se encuentran los atributos “Date”, “Equipment” y “WeightClassKg”. Al haber implementado modelos de prueba incluyendo dichas columnas en el dataset, la cantidad de ramas del árbol aumentaba en gran medida.

Tras finalizar el preprocesamiento, nos quedó un dataset con los siguientes atributos: “Sex” (label), “Age”, “TotalKg”, “Glossbrenner”, “Deadlift”, “Bench” y “Squat”.

### Procesamiento

Utilizamos el algoritmo C4.5 (operador W-j48 en RapidMiner) que permite trabajar con atributos numéricos. En esta ocasión optamos por separar el conjunto de datos en tres partes: datos de entrenamiento, de validación y de prueba. A dichos datasets les asignamos el 70 %, el 10 % y el 20 % del total inicial de entradas, respectivamente. A partir del conjunto de datos de validación, buscamos obtener los mejores valores de  $\alpha$  para realizar el pruning del árbol y la mejor cantidad mínima de elementos por hoja, con el objetivo de obtener un modelo simple que sostenga una elevada tasa de aciertos.

Una vez obtenidos los mejores parámetros, los cuales fueron 0.75 para  $\alpha$  y 5000 para la cantidad mínima de elementos por hoja, pusimos a prueba el árbol sobre el conjunto de datos de testing.

## Conclusión

En base al árbol resultante podemos destacar qué atributos son los más determinantes a la hora de tomar las decisiones para realizar la clasificación deseada. Esto es posible identificando los atributos que son tenidos en cuenta en aquellos nodos del árbol que se encuentran más cerca de la raíz.

En primer lugar, observamos que el peso levantado en el ejercicio “Bench” es uno de los principales criterios utilizados para determinar si el atleta es hombre o mujer. Podemos intuir que esto se debe a que los hombres suelen tener una mayor proporción de masa muscular en el tren superior, cuyos grupos musculares son fuertemente implicados en este ejercicio.

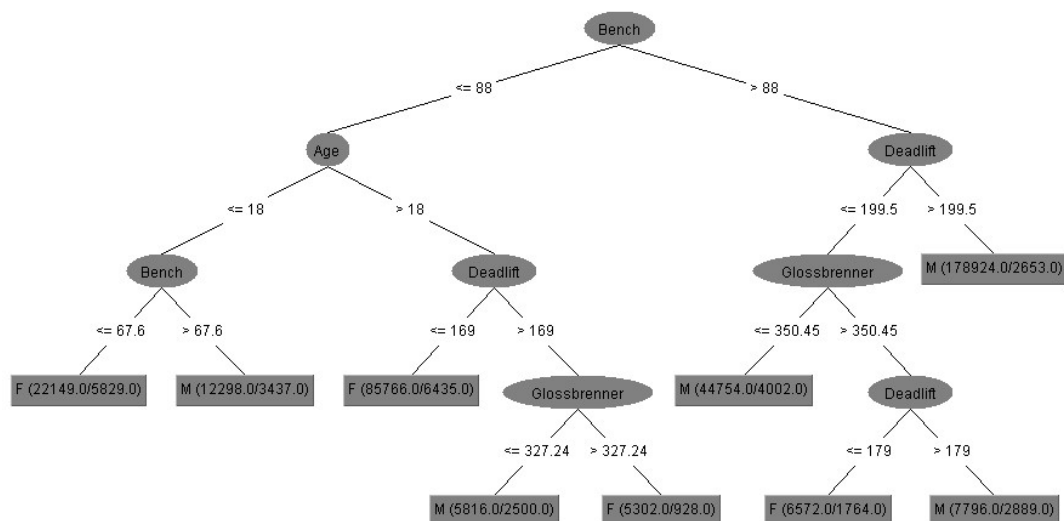
Por otro lado, un criterio ampliamente utilizado en diversas ramas del árbol se basa en el atributo “Age”, el cual es tenido en cuenta únicamente cuando el peso levantado en “Bench” no supera los 88 Kg. Consideramos que esto tiene sentido, ya que el hecho de que un atleta levante cargas relativamente bajas en este ejercicio puede deberse a que se trata de un chico que es menor de edad, o de una mujer.

Además, en el subárbol que abarca aquellos competidores que levantan más de 88 Kg en el ejercicio “Bench”, el criterio principal utilizado para decidir el sexo del atleta se basa en los resultados obtenidos en el levantamiento “Deadlift”.

Cabe destacar que en los nodos de mayor profundidad del árbol también se hace uso de los atributos “Glossbrenner”, “Bench” y “Deadlift”, cuyos valores tienen menor impacto sobre el modelo, dado que sus bifurcaciones son realizadas sobre una menor cantidad de entradas en relación a los campos explicados anteriormente.

Para finalizar, podemos decir que el árbol resultante es legible, ya que consta de 9 hojas y el tamaño del listado de reglas obtenido a partir del mismo es de 17 líneas. Si bien el modelo resultó ser bastante simple, su rendimiento es favorable, con una tasa de aciertos de 91.62 % obtenida sobre los datos de prueba.

A continuación adjuntamos una imagen del árbol obtenido en RapidMiner.



Árbol obtenido aplicando pruning



## W-J48

J48 pruned tree

-----

```
Bench <= 88
|   Age <= 18
|   |   Bench <= 67.6: F (22149.0/5829.0)
|   |   Bench > 67.6: M (12298.0/3437.0)
|   Age > 18
|   |   Deadlift <= 169: F (85766.0/6435.0)
|   |   Deadlift > 169
|   |   |   Glossbrenner <= 327.24: M (5816.0/2500.0)
|   |   |   Glossbrenner > 327.24: F (5302.0/928.0)
Bench > 88
|   Deadlift <= 199.5
|   |   Glossbrenner <= 350.45: M (44754.0/4002.0)
|   |   Glossbrenner > 350.45
|   |   |   Deadlift <= 179: F (6572.0/1764.0)
|   |   |   Deadlift > 179: M (7796.0/2889.0)
|   Deadlift > 199.5: M (178924.0/2653.0)
```

Number of Leaves : 9

Size of the tree : 17

Listado de reglas asociadas al árbol de clasificación realizado.

accuracy: 91.62%

	true M	true F	class precision
pred. M	66886	4565	93.61%
pred. F	4275	29810	87.46%
class recall	93.99%	86.72%	

Matriz de confusión a partir del modelo de árbol aplicado sobre un conjunto de datos de prueba

## Reglas de Asociación

### Preprocesamiento

Para realizar este modelo, optamos por implementar una discretización por frecuencia sobre todos los atributos. La misma fue realizada en dos intervalos: range1 y range2. Decidimos utilizar dicha cantidad de intervalos porque es el mejor valor del índice Davies Bouldin a la hora de realizar el clustering expuesto previamente con distinta cantidad de grupos ocurre cuando  $k$  equivale a 2. Los rangos obtenidos fueron los siguientes:

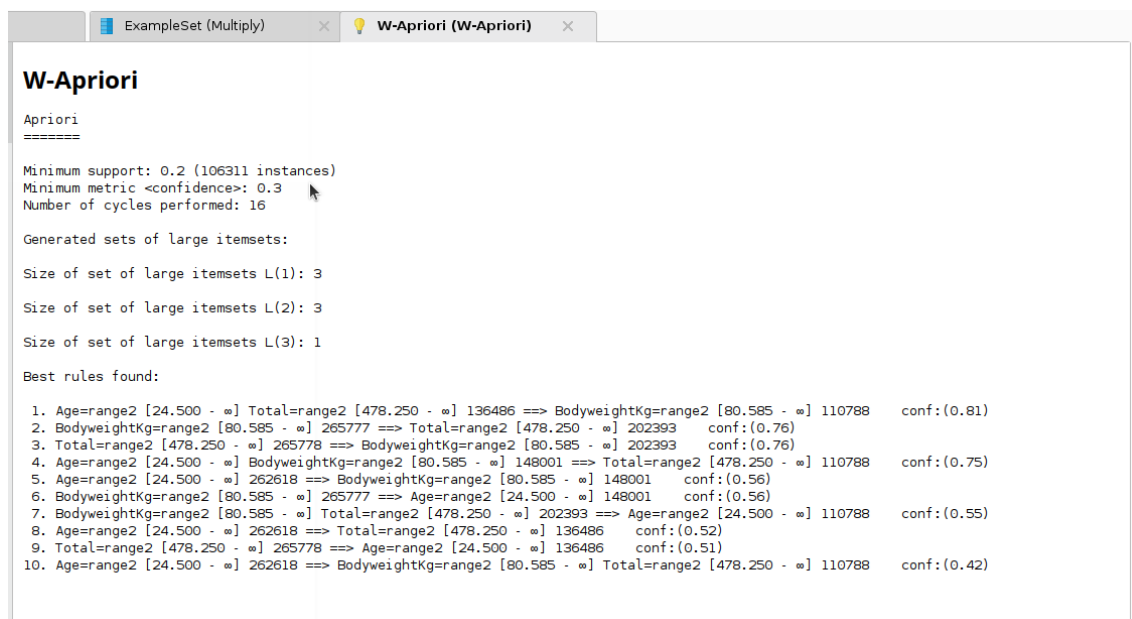
1. Age: (-inf, 24.5) [24.5, +inf)
2. Bodyweight: (-inf, 80.585) [80.585, +inf)
3. TotalKg: (-inf, 478.25) [478.25, +inf)

Además, eliminamos los atributos “Date”, “Equipment”, “Sex\_M”, “Sex\_F”, ya que no aportan información relevante, debido a que no generan nuevas reglas. También eliminamos “Bench”, “Deadlift” y “Squat”, que sólo se relacionan entre sí, generando reglas que aportan muy poca información.

Tras finalizar el preprocesamiento, nos quedó un dataset con los siguientes atributos nominales: “Age”, “BodyweightKg” y “TotalKg”.

### Procesamiento

Para realizar el procesamiento utilizamos el algoritmo Apriori, y como métrica de evaluación usamos la confianza, con un mínimo de 0.3.



```
W-Apriori

Apriori
=====

Minimum support: 0.2 (106311 instances)
Minimum metric <confidence>: 0.3
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 3
Size of set of large itemsets L(2): 3
Size of set of large itemsets L(3): 1

Best rules found:

1. Age=range2 [24.500 - ∞] Total=range2 [478.250 - ∞] 136486 ==> BodyweightKg=range2 [80.585 - ∞] 110788   conf:(0.81)
2. BodyweightKg=range2 [80.585 - ∞] 265777 ==> Total=range2 [478.250 - ∞] 202393   conf:(0.76)
3. Total=range2 [478.250 - ∞] 265778 ==> BodyweightKg=range2 [80.585 - ∞] 202393   conf:(0.76)
4. Age=range2 [24.500 - ∞] BodyweightKg=range2 [80.585 - ∞] 148001 ==> Total=range2 [478.250 - ∞] 110788   conf:(0.75)
5. Age=range2 [24.500 - ∞] 262618 ==> BodyweightKg=range2 [80.585 - ∞] 148001   conf:(0.56)
6. BodyweightKg=range2 [80.585 - ∞] 265777 ==> Age=range2 [24.500 - ∞] 148001   conf:(0.56)
7. BodyweightKg=range2 [80.585 - ∞] Total=range2 [478.250 - ∞] 202393 ==> Age=range2 [24.500 - ∞] 110788   conf:(0.55)
8. Age=range2 [24.500 - ∞] 262618 ==> Total=range2 [478.250 - ∞] 136486   conf:(0.52)
9. Total=range2 [478.250 - ∞] 265778 ==> Age=range2 [24.500 - ∞] 136486   conf:(0.51)
10. Age=range2 [24.500 - ∞] 262618 ==> BodyweightKg=range2 [80.585 - ∞] Total=range2 [478.250 - ∞] 110788   conf:(0.42)
```

Reglas de asociación obtenidas a partir del dataset

## Conclusión

A continuación mencionaremos nuestras conclusiones obtenidas en base a las reglas resultantes.

En primer lugar, la regla con mayor confianza (0.81) establece que si un lifter tiene una edad dentro del rango2 y un Total en el rango2, entonces es muy probable que tenga un “BodyweightKg” en el rango 2. Esto indica que los lifters más viejos, con mayores kilos levantados, tienden a tener un peso corporal más alto.

La segunda regla también refuerza esta idea, ya que establece que si un lifter tiene un valor de “BodyweightKg” en el rango2 es probable que tenga un “Total” en el rango2. Esto sugiere que los lifters con un peso corporal más alto tienen una mayor probabilidad de levantar más peso.

La tercera regla muestra que si un lifter tiene un valor de “Total” en el rango2, es probable que tenga el valor de peso corporal en el rango2. Esto indica que los lifters con mayores cargas totales realizadas tienden a tener un peso corporal más alto.

En general, las reglas sugieren una relación positiva entre la edad y el peso corporal de los lifters con el peso total levantado en las competencias. Parece haber una tendencia de que los lifters más viejos, con mayor peso corporal, tengan más kilos levantados. Esto puede deberse a que dichos atletas disponen de una mayor experiencia y años de entrenamiento respecto de los competidores más jóvenes.

## Conclusión general del trabajo

En este apartado vamos a analizar los resultados obtenidos a partir de los diferentes modelos, y contrastarlos con las hipótesis definidas al comienzo de este trabajo.

En cuanto a los modelos descriptivos, podemos observar que encontramos ciertas relaciones entre los atributos del conjunto de datos. Principalmente, hallamos una fuerte correlación entre el peso corporal de los atletas y la carga levantada en los ejercicios. Por otra parte, las reglas nos ayudaron a detectar nuevas relaciones que no se podían percibir fácilmente utilizando agrupamientos. Por ejemplo, pudimos notar que existe cierta relación entre la edad de los competidores y su rendimiento obtenido, donde una edad más avanzada indica mayor tiempo de preparación.

En lo que respecta a los modelos predictivos, consideramos que los resultados fueron satisfactorios debido a que fue posible realizar la predicción del sexo de los atletas con una elevada tasa de aciertos para los modelos de Naïve Bayes y de árboles, superando ambos el 90 % de accuracy sobre los datos de prueba.

En conclusión, creemos que hemos cumplido con lo propuesto en la hipótesis para todos los modelos presentados.

## Anexo 1: Descripción detallada de los atributos

El dataset original posee 41 atributos:

1. Name: Nombre del participante.
2. Sex: Sexo del participante (Masculino, Femenino, No Binario).
3. Event: Tipo de competición. Especifica cuáles de los tres ejercicios de powerlifting se realizarán.
4. Equipment: Categoría de equipamiento que el lifter utilizó.
5. Age: Edad del participante/lifter.
6. AgeClass: Categoría de edad del lifter (entre qué edad y qué edad).
7. BirthYearClass: Categoría de fecha de nacimiento del lifter.
8. Division: Opcional. División donde compite el lifter.
9. BodyweightKg: Peso del participante a la fecha de la competición.
10. WeightClassKg: Categoría de peso donde el participante compitió.
11. Squat1Kg, Squat2Kg, Squat3Kg: Indican kilos levantados en el primer, segundo y tercer intento, respectivamente, en sentadillas. El valor negativo indica intento fallido.
12. Squat4Kg: Peso levantado en el cuarto intento, en sentadillas. El puntaje obtenido es sólo para récord personal, no cuenta en la competencia.
13. Best3SquatKg: Mejor puntaje de los 3 intentos de sentadillas.
14. Bench1Kg, Bench2Kg, Bench3Kg: Indican kilos levantados en el primer, segundo y tercer intento, respectivamente, en banca. El valor negativo indica intento fallido.
15. Bench4Kg: Peso levantado en el cuarto intento, en banca. El puntaje obtenido es sólo para récord personal, no cuenta en la competencia.
16. Best3BenchKg: Mejor puntaje, de los 3 intentos, en banca.
17. Deadlift1Kg, Deadlift2Kg, Deadlift3Kg: Peso levantados en el primer, segundo y tercer intento, respectivamente, en peso muerto. El valor negativo indica intento fallido.
18. Deadlift4Kg: Cuarto intento en peso muerto. El puntaje obtenido es sólo para récord personal, no cuenta en la competencia.
19. Best3DeadliftKg: Mejor puntaje, de los 3 intentos, en peso muerto.
20. TotalKg: Suma de todos los kilos levantados.
21. Place: Resultado obtenido en la competencia.

22. Dots: Fórmula que indica el rendimiento del atleta en comparación del resultado obtenido por otros atletas.
23. Wilks: Puntos de fuerza relativa.
24. Glossbrenner: Puntos basados en Wilks, que sirven para elegir al mejor deadlifter.
25. Goodlift: Indica el rendimiento obtenido en comparación de otros atletas de su misma categoría.
26. Tested: Indica si el lifter se sometió a un test de drogas antes de competir.
27. Country: Nacionalidad del atleta.
28. State: Estado o provincia donde reside el atleta.
29. Federation: Nombre de la federación que regula la competencia.
30. ParentFederation: Nombre de la federación que gestiona la federación mencionada previamente.
31. Date: Fecha de la competencia.
32. MeetCountry: País en el que se desarrolló la competencia.
33. MeetState: Estado o provincia donde se desarrolló la competencia.
34. MeetTown: Ciudad donde se desarrolló la competencia.
35. MeetName: Nombre de la competencia.

## Anexo 2: Atributos eliminados

1. Name: Consideramos que el nombre del atleta no es relevante a la hora de establecer modelos, ya que no aporta ningún tipo de información que lo caracterice.
2. Deadlift4kg: Dado que este levantamiento lo hacen quienes quieren batir su propio récord personal y no cuenta a nivel competitivo, muy pocos atletas lo realizan. Por lo tanto, este atributo tiene demasiados valores faltantes.
3. Bench4kg: Ocurre lo mismo que con Deadlift4kg.
4. Squat4kg: Ocurre lo mismo que con Deadlift4kg.
5. State: Consideramos poco relevante la importancia del estado o provincia del atleta a la hora de realizar agrupamientos que destaquen las características físicas de los atletas. Además, para utilizar el algoritmo de K-Medias no pueden haber atributos nominales.
6. MeetState: No es relevante el estado donde se realizó la competencia.
7. MeetTown: Ocurre lo mismo que con MeetState.

8. MeetCountry: Ocurre lo mismo que con MeetState.
9. MeetName: No es importante el nombre de la competencia, no aporta información útil.
10. Dots: Es una fórmula compleja que utilizan pocas federaciones de powerlifting, por lo que genera muchos valores faltantes.
11. Wilks: Es un valor que indica resultados similares a Glossbrenner. Por lo tanto, elegimos quedarnos únicamente con este último atributo.
12. Division: No aporta información relevante porque se utiliza únicamente para proveer contexto.
13. ParentFederation: El nombre de una federación no ayuda a describir un modelo que destaque las características de los atletas.
14. Federation: Ocurre lo mismo que con ParentFederation.
15. Country: Ocurre lo mismo que con State.
16. GoodLift: Determinar si un levantamiento fue bueno o malo se basa en un cálculo utilizado solamente por una federación, por lo que dispone de muchos datos faltantes.