Proyecto en base de datos Angus

María Inés Fariello

1 Descripción del problema

La mejora genética ha interesado al hombre desde le principio de la domesticación, presionando selectivamente hacia los fenotipos más convenientes. Para ello, generalmente se basó en observar los fenotipos y elegir los animales que creía más convenientes para mejorar las características de su interés.

En los últimos tiempos se ha podido incorporar información genética y genómica para hacer dicha selección. Utilizando las técnicas de reconocimiento de patrones se puden tomar básicamente dos caminos. La selección genómica se aplica cuando se tiene información sobre muchos marcadores (en este caso hablamos de miles) se puede hacer calsificación, tratando de predecir basándose en un conjunto de entrenamiento nuevos individuos. El problema de esta técnica es que para clasificar a un nuevo individuo es que se necestia conseguir la misma cantidad de datos que tenían los individuos en el conjunto de entrenamiento. Y que en general en este tipo de métodos la clasificación se mira como "caja negra", ya que no se mira cuáles variables influyen en la clasificación en general. La otra técnica es la selección asistida por marcadores (MAS). Esta consiste en hacer selección de variables para ver cuáles son los genes o SNPs que son necesarios para hacer la clasificiación. Una vez optimizado el método y establecido las clases, para clasificar un nuevo individuo se necesita solamente evaluarlo en las variables seleccionadas. Esto por un lado reduce los costos, y puede ayudar además a ver qué genes están involucrados en determinados procesos biológicos.

1.1 Objetivo

En este trabajo se buscará qué variables genotípicas son responsables de las variaciones de algunos de los fenotipos medidos.

1.2 Descripción de la base de datos

La base de datos de Angus es muy amplia en cuanto a la cantidad y diversidad de variables medidas. Además es compleja ya que presenta por un lado dependencia entre las posibles variables de respuesta y por otro presenta dos tipos de variables explicativas, variables con efectios ambientales, como por ejemplo la edad, el sexo o el departamento en el que fue criado el animal y por otro, variables genotípicas. Los fentoipos son el producto de la interacción entre el genotipo y el ambiente, por lo tanto se deben

tomar en cuenta ambos tipos de variables. En algunos casos los efectos de las variables ambientales pueden ser mayores que los genéticos, por lo que no se deben descartar a lo largo del análisis.

1.3 Variables Fenotípicas

Las variables fenotípicas se dividen en dos grupos. El primero se refiere a las características que determinan la cantidad de carne producida: el peso y el largo canal y la compacidad que surge de devidir el peso entre el largo de la canal (Figura 1). El segundo son las características relacionadas con la calidad de la carne: pH, medidas sobre la luminocidad y el color (L, a y b prom), la pérdida de agua por cocción al día y a los 10 días de maduración medida como el peso de la carne antes de cocinarse y luego de cocinarse, la terneza y la cantidad de lípidos (Figura 2).

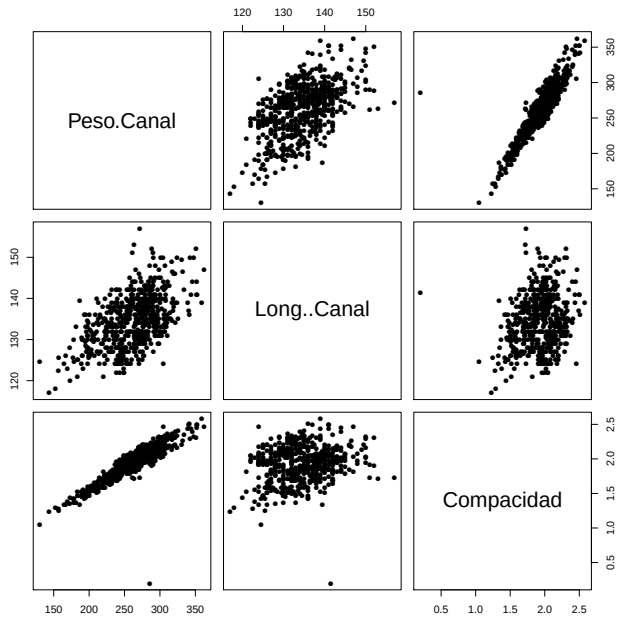

Figure 1: Características referidas a la cantidad de carne producida

1.4 Variables Ambientales

Las variables ambientales, son aquellas pueden influir en el fenotipo, pero que no forman parte de la información genética del individuo. En este caso las variables ambientales presentes en la base de datos son:

- Planta y fecha de faena
- Número de tropa

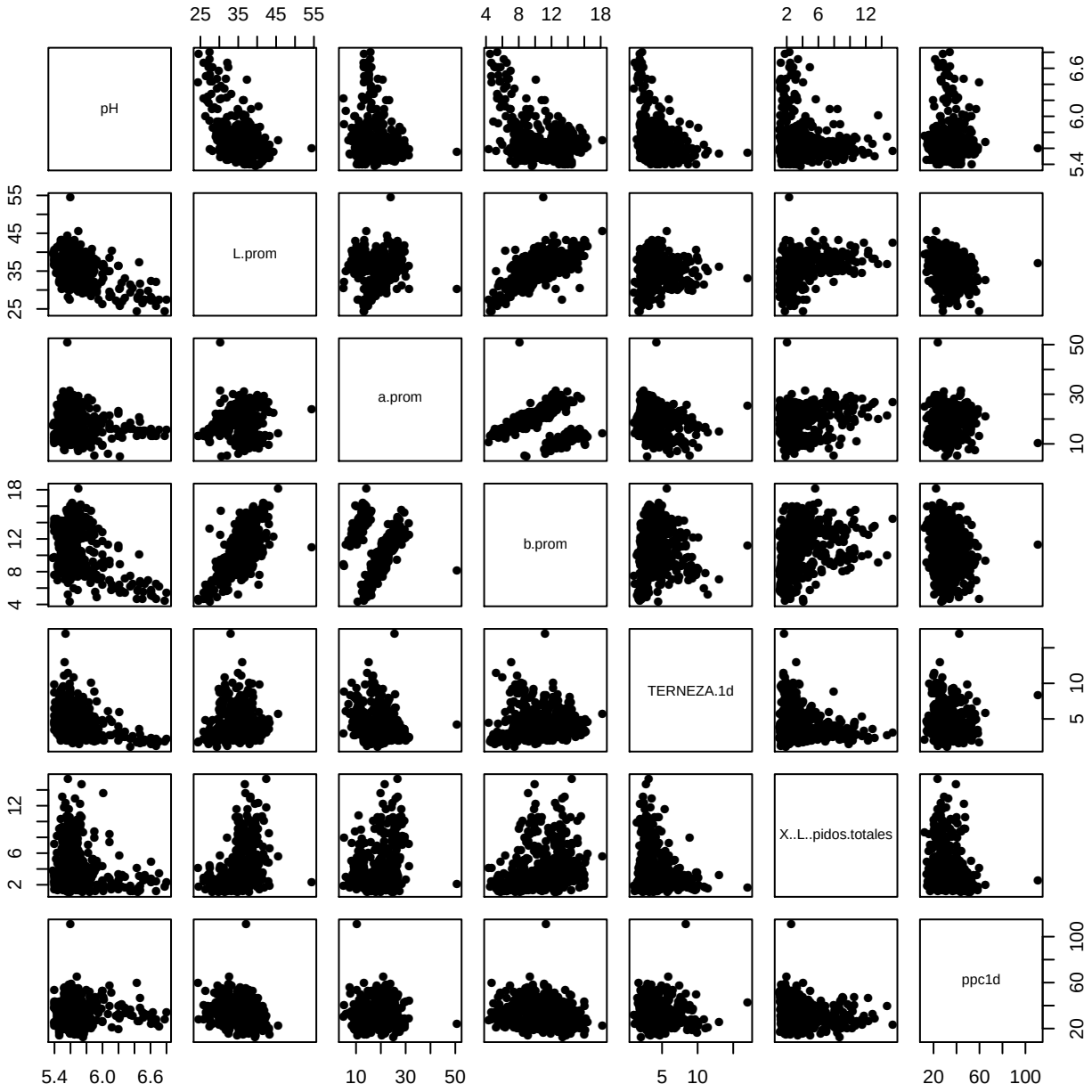

Figure 2: Características referidas a la calidad de carne producida

- Sexo (si bien es genética, no tenemos información genética de los cromosomas sexuales)

- Edad
- Alimentación
- Productor y Departamento

1.5 Variables Genotípicas

En la base de datos hay 91 SNP. Los SNPs se nombran con el nombre del gen y en el caso que haya información de varios SNPs presentes en el mismo gen, se agrega un número desde uno a la cantidad de SNPs presentes dentro del mismo gen al final de su nombre. Dada esta nomenclatura, el gen que aparece como GH1 probablemente pertenezca al grupo de SNPs presentes en el egen GHR 1, ya que los SNPs de este grupo se llaman GHR2, 3, y 4 y no hay un grupo de SNPs llamados GH.

Cuando se trabaja con gentipos se debe verificar que no existe una estructuración genética entre las distintas tropas, ya que se pueden esto puede llevar a confusión de efectos genéticos entre genotipos y fenotipos. Para ver esto se puede hacer un análisis de componentes principales (PCA) a partir de distancias genéticas entre individuos y verificar que estos no formen clusters que se asocien con alguna característica en particular, como ser productor o departamento.

2 Análisis de Datos

Para algunos de los fenotipos medidos se buscarán genes que estén implicados en su variación. Ya que los fenotipos son continuos, se construirán clases para obtener variables discretas, se hará selección de variables y se evaluará el porcentaje de aciertos de la clasificación. Para no confundir efectos ambientales con genéticos se analizarán los grupos establecidos en la sección 1.4.

En el caso que sea posible aplicar los métodos de filtro y wrapper en cada uno de los grupos se contrastarán los resultados obtenidos entre los mismos. Idealemente los genes seleccionados deberían ser los mismos sin importar el grupo. Además se podrían usar los distintos grupos como train y test, ya que si los genes realemente están relacionados con un fenotipo, no debería importar el grupo. Para hacer esto se debe verificar que la distribución de los fenotipos es la misma entre grupos.

2.1 Limpieza de datos

En una primera instancia se retiraron SNPs que tuvieran más de un 10% de datos faltantes, reduciendo el conjunto de 91 a 52 SNPs.

Construcción de matriz de distancias de individuos: Cada uno de los genotipos de los individuos se representan con 0: si el individuo tiene dos copias de uno de los dos aleleos, 1 si tiene una copia de cada uno y 2 si tiene dos copias del otro alelo. Esto quiere decir que cada individuo está representado genéticamente con una "tira" de 52 0s, 1s y 2s,

en el caso que no tenga datos faltantes. Para construir la matriz de distancias los datos se estandarizan primero, para que todos los SNPs tengan la misma media y tomar la información de cada SNP con el mismo peso, sin ponderar. Una vez estandarizados, la matriz de distancias se puede construir multiplicando la matriz de genotipos por su traspuesta. Este procedimiento no admite datos faltantes, por lo tanto los genotipos faltantes se sustituten por la media del SNP.

Una vez construída la matriz de distancias, se realiza la PCA usando la función `princomp` de R.

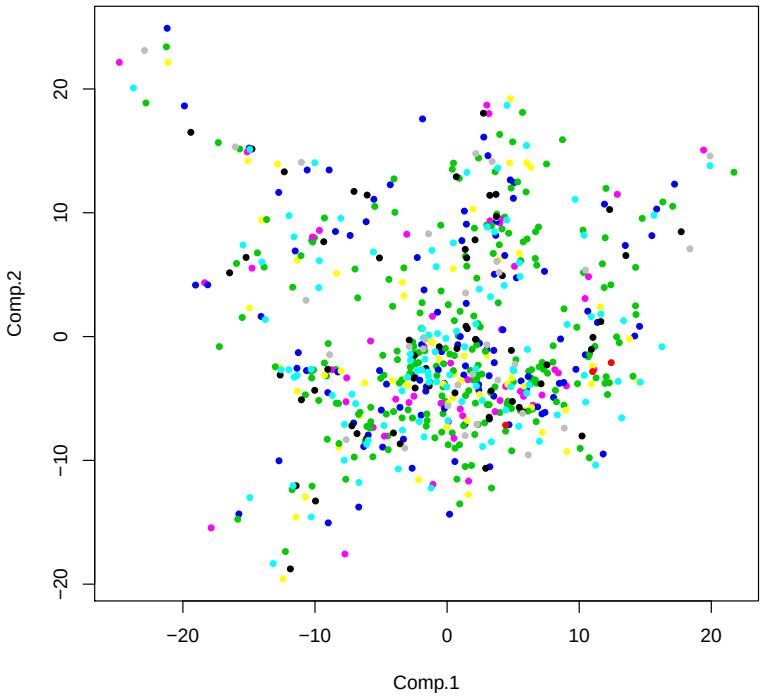

Figure 3: Componentes principales de la matriz de distancias genotípicas de los individuos. Los puntos están coloreados según el productor que porporcionó el individuo.

EN la gráfica del PCA se observa que hay algunos individuos que se separan del grupo principal, pero no se ve una estrucutración obvia. En particular no hay estructura respecto a los productores.

Para tener en cuenta variables abmientales que podrían en influir en el resultado, se agruparon los individuos según su edad (1, 2 o 3 años), sexo (M o F) y tipo de alimentación (pasto o feedlot), resultando en 4 grupos de individuos. Las características de estos grupos y la cantidad de individuos se presentan en la tabla siguiente:

Si miramos los 90 SNPs y sacáramos los individuos que tuvieran solamente como máximo un cuarto de los genotipos faltantes, se debaerían descartar 63 individuos. Este

Grupo	edad	sexo	alimentación	cantidad
1	3	2	feedlot	281
2	3	2	pasto	228
3	1	1	pasto	153
4	1	2	pasto	43

filtro no se hizo en un principio, porque hay que sacar algunos SNPs con muchos datos faltantes, y esto puede depender del fenotipo que se esté midiendo, ya que al filtrar por los individuos que contengan información del fenotipos, el total de SNPs sobre los que se trabaja varía. De todas maneras los datos faltantes son un aspecto a tener en cuenta a lo largo de todo el trabajo.

2.2 pH

Se eligió el pH como primer fenotipo a estudiar, ya que la división en clases discretas no es arbitraria. Según el INAC (http://www.inac.gub.uy, algunas definiciones prácticas) pHs mayores a 5,8 son perfjudiciales sobre la calidad y la duración de la carne. Por lo tanto los pHs menores o iguales a 5,8 se etiquetaron como *correcto* y los mayores a 5,8 como *alto*.

El problema principal de ésta característica es que las clases son muy desbalanceadas. En el grupo 1, de 281 individuos, tenemos información de pH de 156 individuos, de los cuales 149 individuos tienen un pH correcto y solamente 7, pH alto. En el grupo 2 este desbalance es de 148 a 62, por lo que si bien sigue siendo desbalanceado, el desbalance es menor. En el grupo 3 es de 135 a 18 y en el 4, 2 a 43. Este es un primer indicio que probalemente la componente ambiental tiene un fuerte efecto sobre el valor del pH (Figura 4). Se estudiará este probelma usando solamente el grupo 2. De esta manera se pueden descartar algunos efectos ambientales.

Si bien el objetivo es seleccionar algunos SNPs mediante métodos de filtrado y wrappers, para luego clasificar a los individuos a partir de los SNPs seleccionados, se deben usar los métodos de clasificación con todos los SNPs para poder comparar y saber si los SNPs con útiles o no. Los métodos de selección de variables siempre seleccionan por lo menos una variable, lo que indica que es probable que estos métodos seleccionen falsos positivos. Por lo tanto, luego de la selección se debe evaluar la capacidad de clasificación del conjunto de variables seleccionado.

El método tradicional para relacionar genes con fenotipos es construir un modelo lineal con los genotipos como variables dependientes y el fenotipo (sin estratificar en clases) como variable de respuesta y ver cuáles SNPs son significativos. En general, estos modelos asumen que los efectos de los genes son aditivos. En algunos casos se incluyen otro tipo de interacciones, pero no es la práctica usual. El clasificador SVM puede ser de gran ayuda en este tipo de problemas, ya que no se deben especificar interacciones a priori. El único problema que presenta SVM es que se deben encontrar los parámetros que hagan mejor funcionar el modelo, lo que es costoso desde el punto de vista del tiempo.

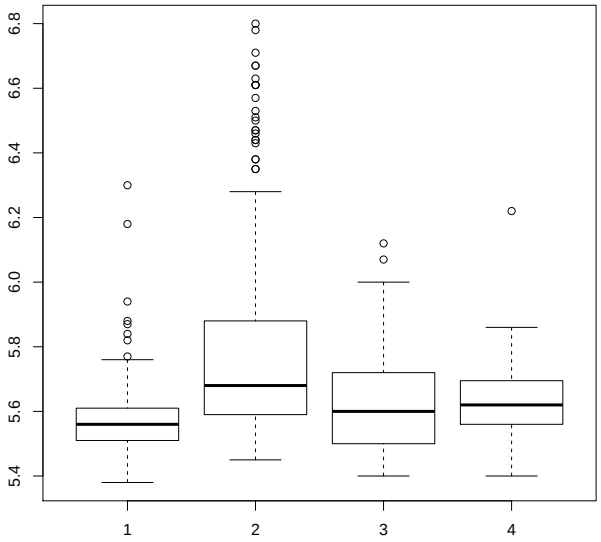

Figure 4: Boxplot del pH dentro de los diferentes grupos

Para tener una idea exploratoria de la capacidad de clasificación de los datos, se clasificaron los datos usando lo métrodos C4.5 y random forest. Estos métdos tienen un buen poder de clasificiación sin necesidad de optimizar los parámetros. El único problema es que cuando hay demasiadas variables, pierden interpretabilidad, pero con 3 o 4 variables, son muy fáciles de interpretar. Estos métodos nos permitirán ver rápidamente si los subconjuntos de variables seleccionadas pueden mejorar la clasificación o no.

Balanceo de clases: Cuando las clases son muy desbalanceadas, es probable que los métodos tomen menos en cuenta a los individuos que están en la clase más chica. Por lo tanto, se aplicarán todas las técnicas tanto a clases desbalanceadas (las originales) como a clases balanceadas. Para aplicar clases balanceadas, se utilizaron las técnicas Resample y SMOTE. La diferencia entre Resample y SMOTE, es que Resample muestrea con reptición de las muestras originales y SMOTE crea nuevos individuos. Por lo tanto, Resample no agrega información pero obliga a los métodos a tomar más en cuenta la clase minoritaria al momento de minimizar los errores de clasificación, ya que cuando las clases están desbalanceadas los errores de la clase más chica le cuestan menos, mientras que aplicando re-sample, los errores de clasificar cuestan los mismo en las dos clases. SMOTE crea individuos sintéticos de la clase minoritaria, por lo tanto puede agregar un poco más de información, ya que crea nuevos individuos, distintos a los otros de la clase. Estos métodos, si bien ayudan a clasificar, pueden agregar falsos positivos para explicar la clase minoritaria, ya que algunas variables que pueden aparecer como importantes,

pueden estar sobre repreesentadas por efecto del re-muestreo solamente. El filtro Resample se aplicó con la opción `biasToUniformClass` = 1, para obtener clases balanceadas: 99 instancias para correcto y 111 para alto. En el caso de SMOTE se aplicó con los parámetros por defecto, obteniendo 148 y 124 instancias respectivamente.

Selección de variables: Los métdos utilizados para seleccionar subconjuntos de variables fueron `CfsSubsetEval`, `GainRaioAttributeEval` e `InfoGainAttributeEval`. El primero evalúa subconjuntos de atributos considerando la habilidad predictiva de cada variable, pero teniendo en cuenta el grado de redundancia entre las variables del subconjunto, por lo tanto intenta seleccionar variables que estén bien correlacionadas con las clases pero que mantengan una baja correlación entre ellas. Los otras dos métodos miden la ganancia de información comparando la inforación que brinda usar cada variable con la información obtenida si no se usaraa. La diferencia entre ambas es que el ratio re-escala por la información brindada por la variable. Estos métodos miden la utilidad de cada variable por separado, por lo tanto puede haber redundancia entre las variables seleccionadas.

2.2.1 Resultados

Los errores de clasificación de los métodos aplicados a los distintos juegos de datos se presentan en la tabla 1. El balanceo de clases permite disminuir los errores de clasificación en ambos métodos (C4.5 y Random Forest), por lo que parecería que puede ser más útil para la selección de variables que utilizar clases desbalanceadas. El método de balanceo de clases que mejoró más la clasificación fue re-sample.

	J48	Random forest
Clases Desbalanceadas	39.54	35.71
Balanceadas (Resample)	19.52	16.19
Balanceadas (SMOTE)	28.68	24.26

Table 1: Porcentaje de error de clasificaión utilizando dsitintas técnicas de árboles, para un juego de datos desbalanceados y dos balanceados.

SMOTE además de tener un peor desempeño en la clasificación (Tabla 1), agrega valores en las variables que no existen en el conjunto original y lo hace solamente para la clase minoritaria (Figura 5). Estos nuevos valores de variables pueden tener una gran influencia en los métodods de selección de varibales, ya que proporcionan mucha información porque los individuos que tienen esos valores pertenecen solamente a la clase minoritaria. Por lo tanto las variables seleccionadas clasificarían bien a los individuos creados por SMOTE.

En lo que sigue se utilizarán las clases balanceadas creadas por re-sample o clases desbalanceadas.

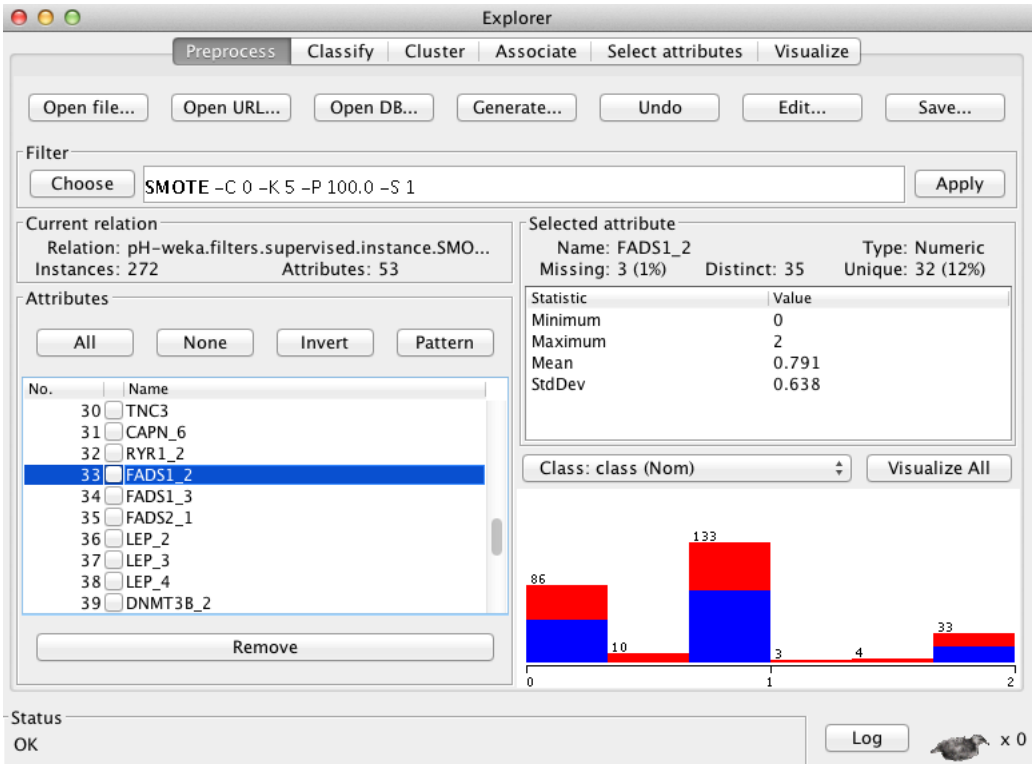

Figure 5: Distribución de la variable FADS1 2. Los colores representan la proporción de individuos que contienen cada valor que puede tomar la variable. Los valores que no son enteros son los provenientes de los individuos creados sintéticamente por SMOTE

Selección de variables Utilizando clases balanceadas por el métdo re-sample y los distintos métodos de selección de variables, se obtuvieron las siguientes variables (Tabla 2). Las variables 8, 14, 27 y 43 están siempre presentes, sin importar el método utilizado.

Método	Variables
BestFirst+CfsSubsetEval	8,14,27,43
BestFirst+CfsSubsetEval (10 X-val)	8, 27, 43 (9), 45 (7), 21 (7)
InfoGainAttributeEval+Ranker	27, 8, 9, 45, 14, 43, 47, 21
GainRatioAttributeEval+Ranker	27, 45, 47, 21, 43, 8, 9, 14

Table 2: Números de las variables seleccionadas según el métdo utlizado (10 X-val significa 10 fold cross validation)

Para saber si las variables seleccionadas son útiles, se volvió clasificar usando árboles como en la parte anterior, pero usando distintos subconjuntos de variables obtenidos usando los métodos de selección de variables. Los subconjuntos utilizados fueron la unión de todas las variables seleccionadas por los tres métodos, la intersección, y la intersección sin la clase 14, ya que aparece como última cuando se utiliza el método

`GainRatioAttributeEval`.

Usando las variables seleccionadas para construír los árboles para clasificar el conjunto con clases balanceadas, ambos métodos clasificaron bastante peor. Mientras que con todas las variables el error de cerca del 20%, con el conjunto reducido de las mismas el error pasa a más de un 30% para todos los subconjuntos de variables posibles (Tabla 3). Aplicando la clasificación al conjunto de las clases desbalanceadas, hay una mejora en el error de clasificación, pero que no es para nada útil, ya que surge de clasificar todos los individuos como *correcto* (Tabka 4).

Variables	J48	Random forest
8,9,14,21,27,43,45	33.8	31.9
8,14, 27, 43	32.38	35.23
8, 27, 43	36.28	36.67

Table 3: Error de clasificación usando clases balanceadas, para los métodos C4.5 y Random forest utlizando distintos subgrupos de variables

Variables	J48	Random forest
8,9,14,21,27,43,45	29.52	38.09
8,14, 27, 43	29.52	31.90
8, 27, 43	29.52	30.00

Table 4: Error de clasificación usando clases desbalanceadas, para los métodos C4.5 y Random forest utlizando distintos subgrupos de variables

Dado el mal rendimiento de los métodos de clasificación usando un conjunto de varaiables reducidos, se hará el mismo procedimieto, pero usando clases balanceadas. Además, sabemos que en el caso anterior existe riesgo de falsos positivos debido a posibles sesgos de muestreo. De esta manera es posible comparar con las variables obtenidas utilizando clases balanceadas.

Las variables obtenidas fueron las mismas que en la parte anterior (Tabla 5), excepto por la variable 43. La razón de esta diferencia es que por azar la variable 43 quedó relacionada solamente un genotipo (de los 3 posibles) con la variable pH alto, por lo que aparentaba ser informativa, pero fue solamente un efecto del re-muestreo, por lo que la variable 43 es un falso positivo.

Se agrega la variable 9, que había aparecido como candidata en la selección de variables con clases balanceadas, pero no ocupaba los primeros puestos.

Los resultados de la clasificación usando las variables obtenidas al usar toda la muestra, dieron el mismo resultado que en la tabla 4 y el error de clasificación se se debe a la clasificación de todas las muestras como con pH *correcto*. Para ver a qué se debía este reslutado se construyeron los haplotipos usando las variables 8,9,14 y 27. Sacando a los haplotipos con menos de 5 observaciones en la población, se obtuvo la siguiente tabla (6):

Método	Variables
BestFirst+CfsSubsetEval	8,14,27
BestFirst+CfsSubsetEval (10-X val)	8, 14, 27
InfoGainAttributeEval+Ranker	14, 8, 27, 9
GainRatioAttributeEval+Ranker	8,27,9,14

Table 5: Números de las variables seleccionadas según el métdo utlizado (10 X-val significa 10 fold cross validation)

haps	alto	correcto
0222	1	12
1022	1	6
1112	1	11
1121	0	5
1122	22	37
2012	3	12
2021	0	5
2022	32	46

Table 6: Cantidad de haplotipos por clase

Para realizar un test de χ^2 se tomaron en cuenta solamente los haplotipos 0222, 1112, 1122, 2012 y 2022 de la tabla de contingencia, ya que sino quedan demasiadas celdas con menos de 5 observaciones. El p-valor obtenido es 0.02756. Por lo que aparentemente habría una influencia de estos haplotipos en el pH de los animales, pero que probablemente interactúen con otros efectos, tanto ambientales, como genéticos no presentes en la muestra.

En la tesis de licenciatura en biología de Signe Haakonson [**?**], observaron que la tropa 6 tenía valores de pH más elevados que el resto de las tropas. Por lo tanto se hizo un test de χ^2 cuadrado para los mismos haplotipos, utilizando las tropas en la tabla de contingencia, en vez de los fenotipos. La tabla obtenida es la Tabla 7 y el p-valor 1.5810^{-07}, por lo tanto la asociación del haplotipo es mucho más fuerte con las tropas que con los fenotipos (con clases discretas, al menos).

Estos resultados apuntan a que si bien estos genes podrían estar relacionados con el pH, se encuntran en mayor frecuencia en la tropa 6. Por lo que se hace difícil separar el efecto tropa y la relación con el pH (Figura **??**).

Dado que no se había tomado en cuenta la tropa en la selección de variables anterior se hizo una nueva selección de variables, incluyendo las mismas variables genotípicas, pero además variables ambientales y tomando todos los datos. Las variables ambientales incluídas fueron tropa, planta y sexo. Luego de realizada la selección de variables usando los métodos anteriores se obtinenen las variables tropa y el SNP scap 3.

Este SNP es heterocigoto para 28 individuos y homocigoto para un alelo en 609 individuos. La particularidad es que para 26 de los 28 individuos el pH es inferior a 5.8,

tropa	3	5	6	9
0222	1	0	4	8
1022	2	4	1	0
1112	4	5	1	2
1121	1	4	0	0
1122	3	15	33	8
2012	7	3	4	1
2021	1	3	1	0
2022	8	15	47	8

Table 7: Cantidad de haplotipos por tropa

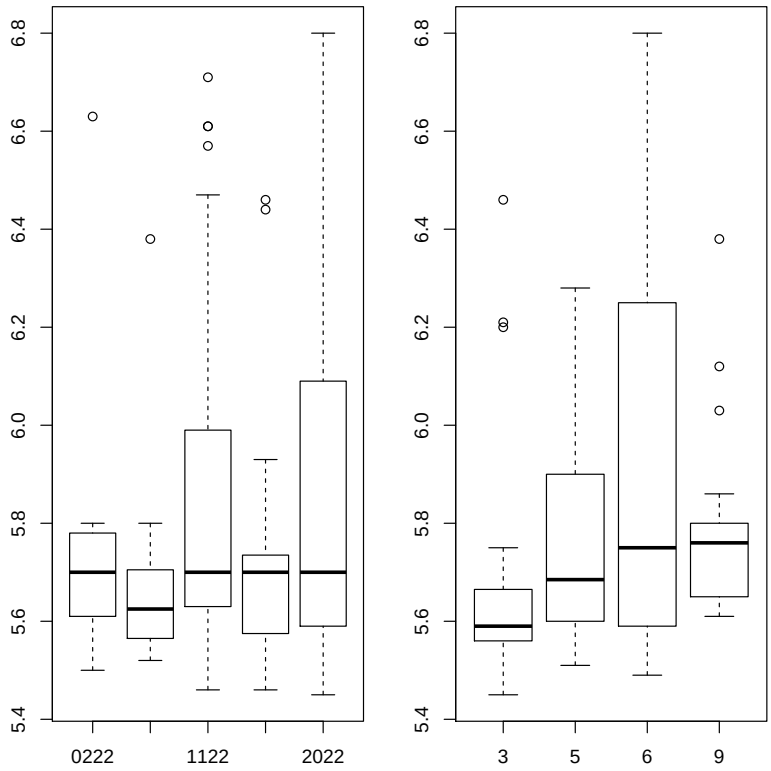

Figure 6: Boxplots relacionando el pH con los haplotipos (izquierda) y las tropas (derecha)

por lo que scap 3 podría estar relacionado con los valores de pH en la carne.

Se intentó utilizar el método `WrrapperSubsetEval`. Este método se basa en buscar subconjuntos de variables con buena predicción basadas en clasificación con SVM. Se intentó hacer pero los valores encontrados para el costo y el gamma del kernel no eran estables. Para ver por qué no estaba funcionando el SVM se hizo un LDA para tener

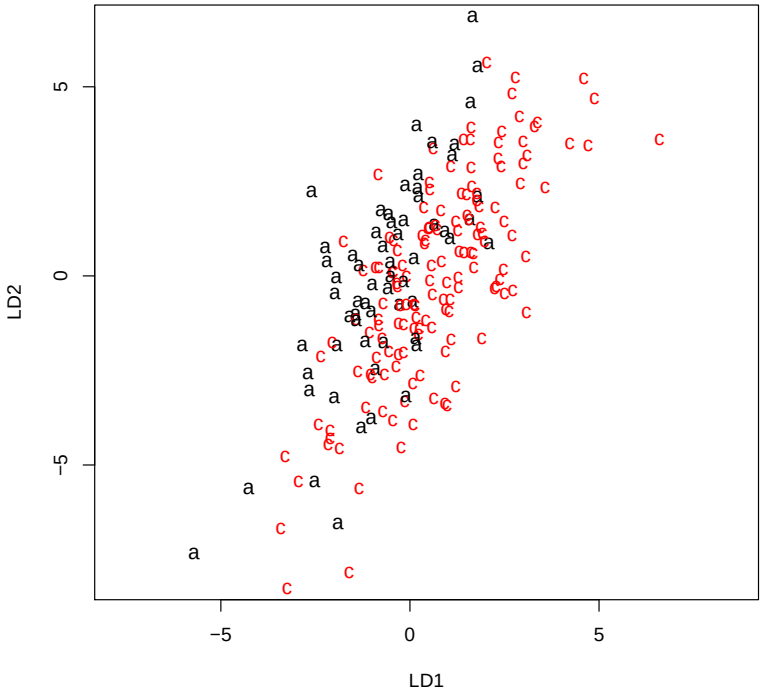

Figure 7: Análisis Discriminante Lineal para los individuos del grupo 2 y las clases *correcto* y *alto*.

una idea de la separabilidad de los datos.

En la figura 7 se oberva la gran superposición que existe entre ambas clases, lo que hace difícil que cualquier método de clasificación funcione.

2.2.2 Discusión (pH)

Mirando los SNPs seleccionados por los distintos métodos, se observó que en todos los casos los genotipos que están presentes en los individuos con pH alto, también los están los que tienen pH correcto. Por esta razón, los árboles no pueden clasificar bien, ya que no es posible tener nodos puros para la clase pH alto.

No parece haber sido una buena idea dividir por clases para intentar clasificar. Para ello hay dos soluciones posibles: utilizar más clases o no dividir en clases y usar métodos de clasificación que sean compatibles con variables reales, en vez de discretas, como por ejemplo regresión SVM. Si bien la cantidad de posibilidades es menor, Weka permite hacerlo.

Además en la tesis de Haakonsson el único SNP que se relaciona con el pH es el IGF2 que no está presente en la base de datos con la que se trabajó, por lo tanto no es posible

comparar ambos resultados. En dicho trabajo también llegan a la conclusión que hay un efecto tropa fuerte. El pH también puede ser influenciado por el manejo de la tropa antes de la faena, lo que en estos datos se resume como un efecto tropa, por lo tanto es lógico que esta variable aparezca dentro de la selección de variables.

Sobre la manera de construir las variables balanceadas y la manera que las construye SMOTE en particular, se podría arreglar probablemente poniendo el atributo class y especificando las clases en luger de poner numeric, para impedir que los nuevos individuos tengan variables con fracciones.

En dicha tesis las otras asociaciones buscadas se hicieron para los fenotipos relacionados con el color de la carne, L, a y b y la retención de agua, evaluada a través de la pérdida (de agua) por cocción, con un día y 10 de maduración. A continuación se intentará enocntrar SNPs relacionado con el b promedio (uno de los parámetros que mide el color) sin discretizar el fenotipo.

2.3 b promedio

El fenotipo para el cual se encontraron más SNPs relacionados en la tesis de Haakansson fuel el b promedio, que es una medida de color y se relaciona directamente con lo que miran los compradores de carne. Por lo tanto es otro factor importante de la calidad de la carne.

Al retirar los individuos que no tenían medido el fenotipo, nos quedamos con 445 animales distribuídos en 7 tropas. Para realizar selección de variables en este caso la división del fenotipo en clases discretas no es obvia. En la figura **??**, vemos que si además miramos el a promedio (que es la otra medida de color), los individuos se clusterizan en dos grupos diferentes. En un cluster hay individuos del grupo 3 y 4 y en el otro cluster individuos de los grupos 1, 2 y 3. El grupo 3 se separa en dos clusters porque contiene individuos que si bien están en el mismo grupo comparten tropas con animales de otros grupos. Por lo que podemos conluir que la agrupación es más bien por tropas. En el boxplot de la distribución de b según tropas (Fig. 9), vemos que hay un claro efecto tropa. Al momento de discretizar esto sería un problema, ya que si discretizármaos deberíamos lograr encontrar una manera de definir clases equivalentes entre tropas corrigiendo el efecto tropa. Por esta razón se decidió usar b sin discretizar e incluir la tropa en variables explicativas, como para considerar su efecto.

Para hacer la selección de variables se usaron los métodos `CfsSubsetEval` que es independiente del método de clasificiación y `WrapperSubsetEval` junto a la función `SMOreg` para seleccionar las variables en función del resultado de la regresión por SVM. Los resultados se encuentran en el apéndice al final del trabajo.

Las variables incuídas para el análisis además de las genéticas fueron: planta, tropa, sexo y alimentación.

Como en la tesis se utilizó solamente un subconjunto de los datos, se hizo selección de variables usando todos los SNPs disponibles, y los SNPs que fueron utilizados en la tesis.

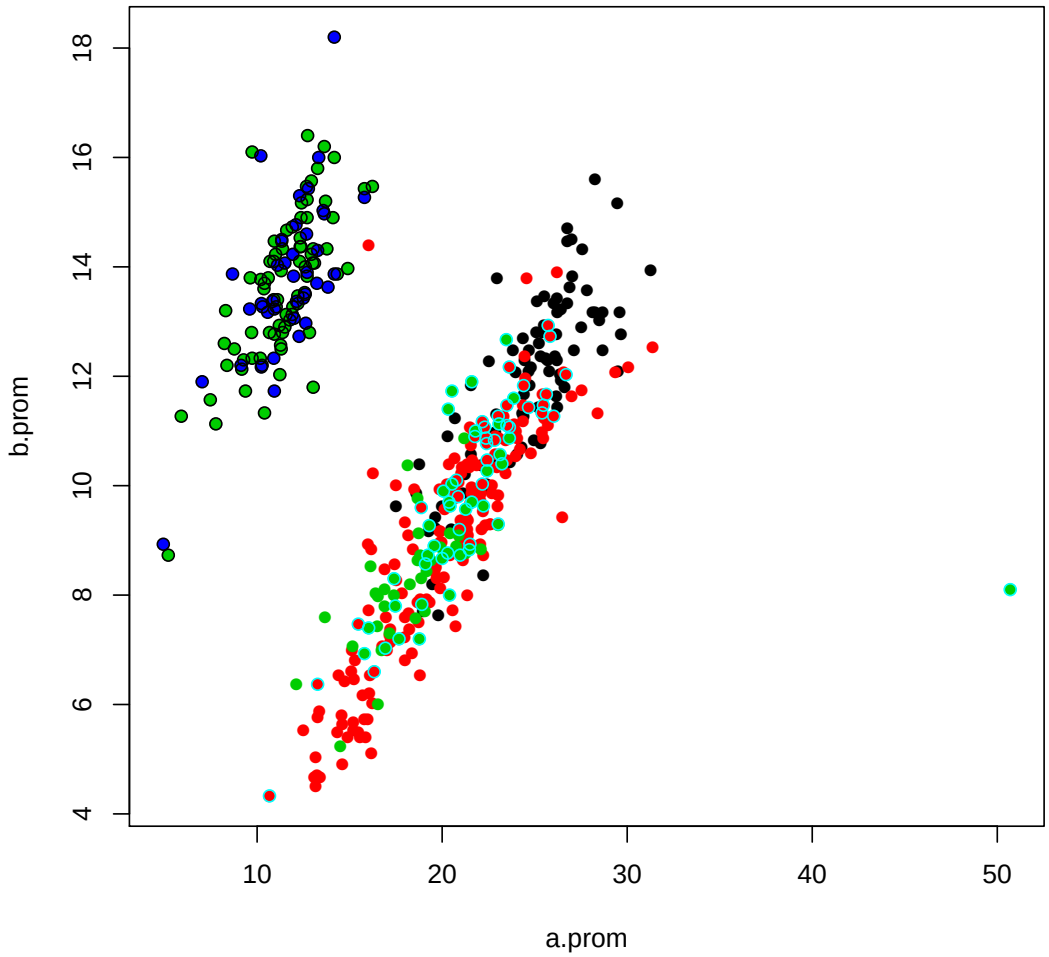

Figure 8: a y b promedio para cada individuo. Los colores de los puntos representan al grupo que pertenecen los individuos (1,2, 3 o 4) y los puntos que tienen además color exterior pertenecen a la misma tropa pero a grupos diferentes

2.3.1 Resultados (b promedio)

Variables ambientales Sin importar el método, las variables ambientales tropa y alimentación salieron siempre en primer logar, indicando que es el efecto más fuerte, y lo es más aún que el efecto genético. El efecto planta y sexo fueron también elegidos por algunos métdos.

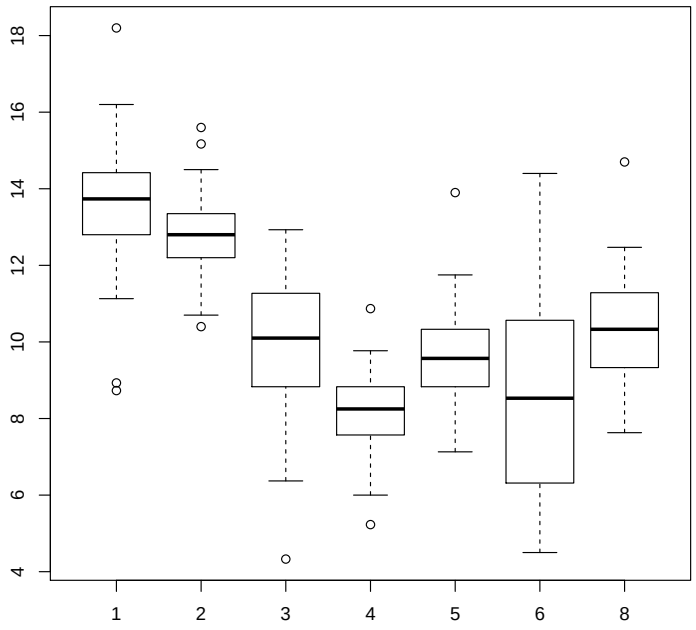

Figure 9: Distribución de b promedio por tropa

Variables genéticas En el informe para color b, aparece SCD 5 como significativo. Este SNP fue genotipado solamente en 81 animales, por lo tanto no se encuentra dentro de los SNPs testeados. Pero en nuestro análisis aparece SCD6, que es un SNP en el mismo gen. El otro que aparece en la tesis es CAPN1 3, en nuestros datos aparecen solamente una vez CAPN.316, CAPN.6. (Los resultados obtenidos para esta parte se encuentran en el apéndice).

2.3.2 Discusión (b promedio)

Nuevamente el efecto tropa en este fenotipo es muy importante. No obstante se obtuvieron algunos genes que podrían tener algún sentido. En un futuro se discutirá con los investigadores que porporcionaron la base de datos para profundizar sobre el sentido de éstos genes.

2.4 Software

Se utilizaron los softwares R y weka.

3 Conclusión

La información contenida en la base de datos parece no ser suficiente para predecir los fenotipos (al menos los testeados en este trabajo). Por un lado puede ser que la cantidad de datos no sea suficiente, pero por otros se debaría trabajar con una base de datos limpia y validada ya que la cantidad de datos faltantes es mucha. Esta falta se da en todos los tipos de variables, hay animales que no tienen ciertos fenotipos registrados, y hay muhos genes a los que les falta un porcentaje muy grande de datos.

Por otro lado, los efectos ambientales en los fenotipos medidos son en general mayores a los genéticos. Para seguir con este trabajo, se deberá discutir con los expertos cómo puede hacerse para corregir estos efectos, por ejemplo, si tendría sentido hacer una normalización entre clases o no y si pueden discretizarse o no. La discretización ayuda a comprender más fácilmente los errores cometidos por los clasificadores, por lo que es más fácil sacar conclusiones. Cuando se trabaja con variables continuas si bien se pueden comparar los errores entre los distintos métodos, es más difícil ver de dónde vienen.

4 Apéndice

Método y atributos seleccionados:

Evaluator: weka.attributeSelection.CfsSubsetEval, Search:weka.attributeSelection.BestFirst -D 2 -N 5
10(100 %) 2 Tropa
5(50 %) 4 Alimentacion
8(80 %) 12 CAST2
4(40 %) 21 GH1
6(60 %) 32 SCD6
10(100 %) 39 RYR1.2
10(100 %) 40 FADS1.2
6(60 %) 49 CACNA2D3
10(100 %) 54 CRH.2

Oreden sin X-val: Tropa, Alimentacion, CAST2, GH1, SCD6, RYR1.2, FADS1.2, CACNA2D3, CRH.2. (Total number of subsets evaluated: 464, Merit of best subset found: 0.65)

Evaluator: weka.attributeSelection.ClassifierSubsetEval -B weka.classifiers.trees.M5P -T -H "Click to set hold out or test instances" – -M 4.0 Search:weka.attributeSelection.BestFirst -D 2 -N 5

Total number of subsets evaluated: 2087 Merit of best subset found: 1.053

Planta, Tropa, SST2, PPARGCIA1, PPARA4, CAST1, CAST2, CAST3, CAST4, IGF1.3, IGF1.5, GH1, PRNP1, PRNP2, LPL1, SCD4, CAPN.316, CAPN.6, CACNA2D1, CACNA2D3, DNAJA1.1, PRKAG3.1, DGAT.1, MB.1

Mejores parámetros regresión SVM C=20, Gamma=0.1.

Correlation coefficient 0.9775 Mean absolute error 0.1577 Root mean squared error 0.5662 Relative absolute error 7.13 % Root relative squared error 21.2487 % Total Number of Instances 441

Usando estos parámetros para el wrapper: Evaluator: weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.functions.SMOreg -F 5 -T 0.01 -R 1 – -C 20.0 -N 0 -I "weka.classifiers.functions.supportVec -L 0.001 -W 1 -P 1.0E-12 -T 0.001 -V" -K "weka.classifiers.functions.supportVector.RBFKernel -D -C 250007 -G 0.1" Search:weka.attributeSelection.BestFirst -D 2 -N 5

Tropa, sexo, Alimentacion, CAST1, FADS1 2, DNAJA1 1, POMC, PRKAG3 1,

Evaluator: weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.functions.SMOreg -F 5 -T 0.01 -R 1 – -C 20.0 -N 0 -I "weka.classifiers.functions.supportVector.RegSMOImproved -L 0.001 -W 1 -P 1.0E-12 -T 0.001 -V" -K "weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.1" Search:weka.attributeSelection.BestFirst -D 1 -N 5 X validation

3(100 %) 2 Tropa
3(100 %) 3 sexo
3(100 %) 4 Alimentacion
2(67 %) 7 PPARD4

2(67 %) 18 IGF1 3
2(67 %) 45 LEP 3
1(33 %) 8 PPARD7
1(33 %) 22 PRNP1
1(33 %) 27 INSIG1 2
1(33 %) 38 CAPN 6
1(33 %) 41 FADS1 3
1(33 %) 44 LEP 2
1(33 %) 46 DNMT3B 2
1(33 %) 50 DNAJA1 1
1(33 %) 58 MB 1

Restringiendo al subset de genes utilizados en la tesis (no están todos los SNPs en mi base de datos.)

Evaluator: weka.attributeSelection.CfsSubsetEval Search:weka.attributeSelection.BestFirst -D 2 -N 5

10(100 %) 2 Tropa
9(90 %) 4 Alimentacion
4(40 %) 9 CAST1
6(60 %) 10 CAST2
4(40 %) 15 IGF1 3
6(60 %) 18 GH1
7(70 %) 23 SCD6
0(0 %) 7 PPARD7

Sin X-val: Total number of subsets evaluated: 217 Merit of best subset found: 0.625: Tropa, Alimentacion, PPARD7, CAST2, GH1, SCD6.

SVMreg:

Coordinates: [2.7, -1.5] Values: 2.7 (X coordinate), 0.03162277660168379 (Y coordinate) Correlation coefficient 0.7417

Evaluator: weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.functions.SMOreg -F 5 -T 0.01 -R 1 – -C 2.7 -N 0 -I "weka.classifiers.functions.supportVector.RegSMOImproved -L 0.001 -W 1 -P 1.0E-12 -T 0.001 -V" -K "weka.classifiers.functions.supportVector.RBFKernel -C 250007 -G 0.0316227766016837" Search:weka.attributeSelection.BestFirst -D 1 -N 5 Relation: Bprom-weka.filters.unsupervised.attribute.Remove-R5,10,17,22-27,39-41,43-55,57-58

3(100 %) 1 Planta
3(100 %) 2 Tropa
3(100 %) 3 sexo
3(100 %) 4 Alimentacion
3(100 %) 29 CAPN 6
2(67 %) 30 TNC4
2(67 %) 13 CAST5

2(67 %) 18 GH1
1(33%) 5 PPARGCIA1
1(33 %) 6 PPARD4
1(33 %) 8 PPARA4
1(33 %) 9 CAST1
1(33 %) 10 CAST2
1(33 %) 14 IGF1 1
1(33 %) 16 IGF1 5
1(33 %) 17 IGF1 7
1(33 %) 23 SCD6
1(33 %) 28 CAPN530