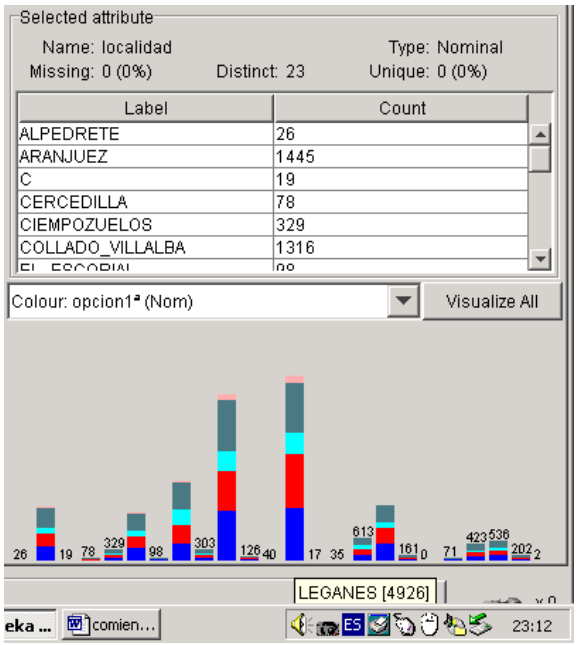
SOLUCION

1. Características de los datos y filtros

Una vez cargados los datos, aparece un cuadro resumen, *Current relation*, con el nombre de la relación que se indica en el fichero (en la línea @relation del fichero arff), el número de instancias y el número de atributos.

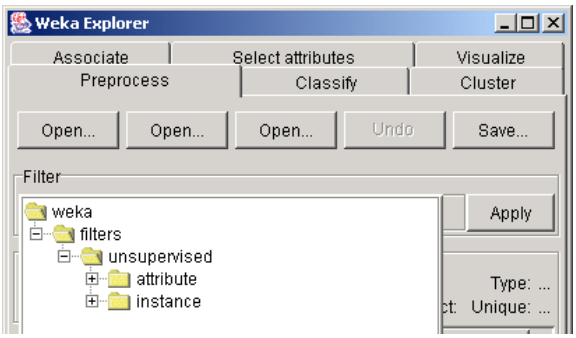
Además, en la parte inferior se presenta gráficamente el histograma con los valores que toma el atributo. Si es simbólico, la distribución de frecuencia de los valores, si es numérico, un histograma con intervalos uniformes. A modo de ejemplo, a continuación mostramos el histograma por localidades, indicando con colores la distribuciones por opciones elegidas.



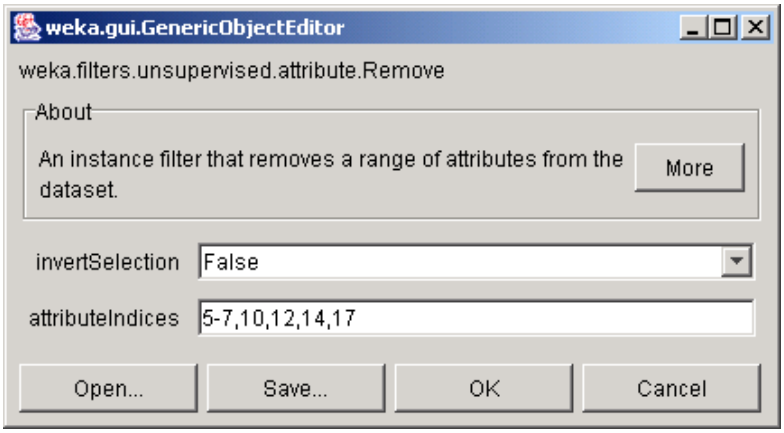
Se ha seleccionado la columna de la localidad de Leganés, la que tiene más instancias, y donde puede verse que la proporción de las opciones científicas (1 y 2) es superior a otras localidades, como Getafe, la segunda localidad en número de alumnos presentados.

En cuanto al filtrado para la preparación de los datos, WEKA tiene integrados filtros que permiten realizar manipulaciones sobre los datos en dos niveles: atributos e instancias. Las operaciones de filtrado pueden aplicarse “en cascada”, de manera que cada filtro toma como entrada el conjunto de datos resultante de haber aplicado un filtro anterior.

Para aplicar un filtro a los datos, se selecciona con el botón **Choose** de **Filter**, desplegándose el árbol con todos los que están integrados.



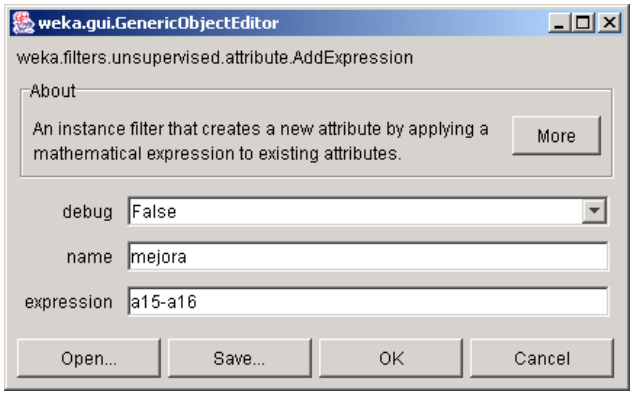
Vamos a utilizar el filtro de atributos "*Remove*", que permite eliminar una serie de atributos del conjunto de entrada. Como primer filtro de selección, vamos a eliminar de los atributos de entrada todas las calificaciones parciales de la prueba y la calificación final, quedando como únicas calificaciones la nota de bachillerato y la calificación de la prueba. Por tanto tenemos que seleccionar los índices 5,6,7,10,12,14 y 17, indicándolo en el cuadro de configuración del filtro *Remove*:



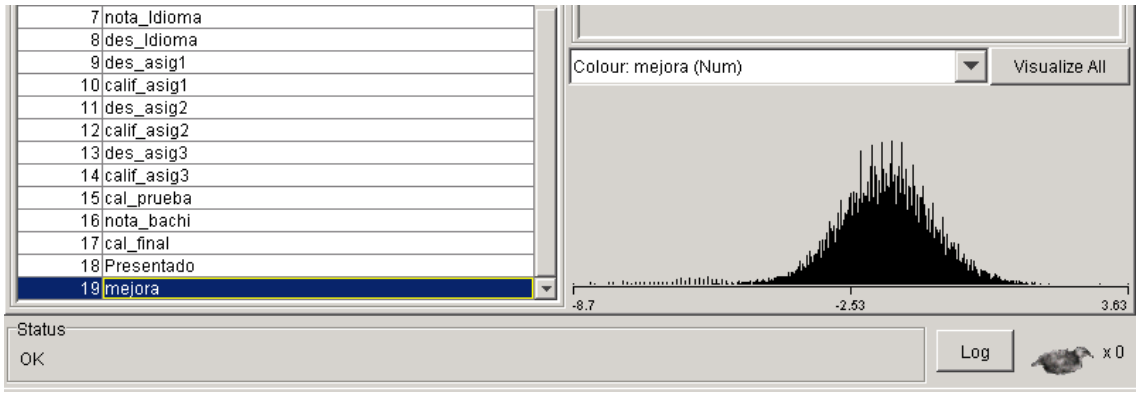
Como puede verse, en el conjunto de atributos a eliminar se pueden poner series de valores contiguos delimitados por guión (5-7) o valores sueltos entre comas (10,12,14,17). Además, puede usarse "first" y "last" para indicar el primer y último atributo, respectivamente.Una vez configurado, al accionar el botón **Apply** del área de filtros se modifica el conjunto de datos (se filtra) y se genera una relación transformada.

Filtros de añadir expresiones

Muchas veces es interesante incluir nuevos atributos resultantes de aplicar expresiones a los existentes, lo que puede traer información de interés o formular cuestiones interesantes sobre los datos. Por ejemplo, vamos a añadir como atributo de interés la "mejora" sobre la nota de bachillerato, lo que puede servir para calificar el "éxito" en la prueba. Seleccionamos el filtro de atributos **AddExpression**, configurado para obtener la diferencia entre los atributos calificación en la prueba, y nota de bachillerato, en las posiciones15 y 16:



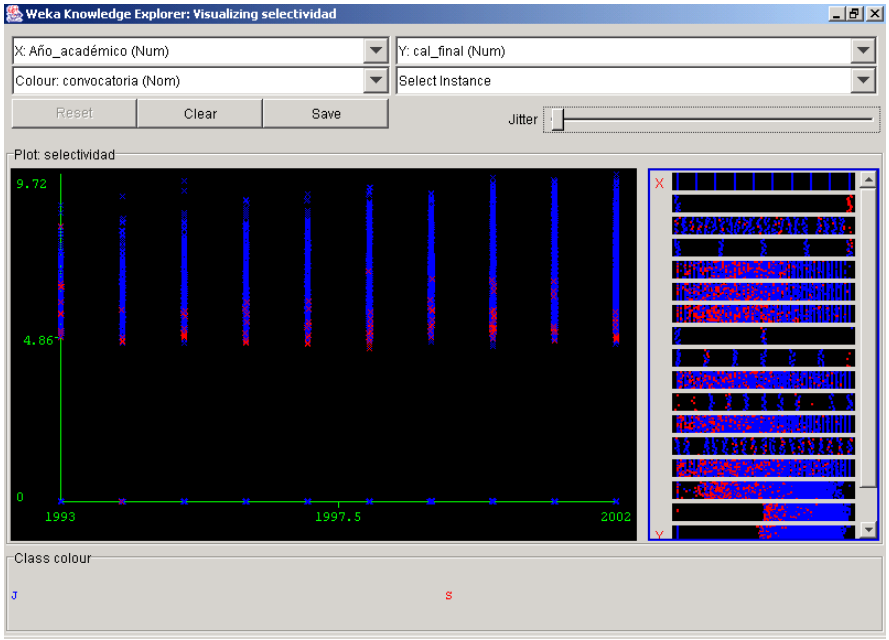
después de aplicarlo aparece este atributo en la relación, sería el número 19, con el histograma indicado en la figura:



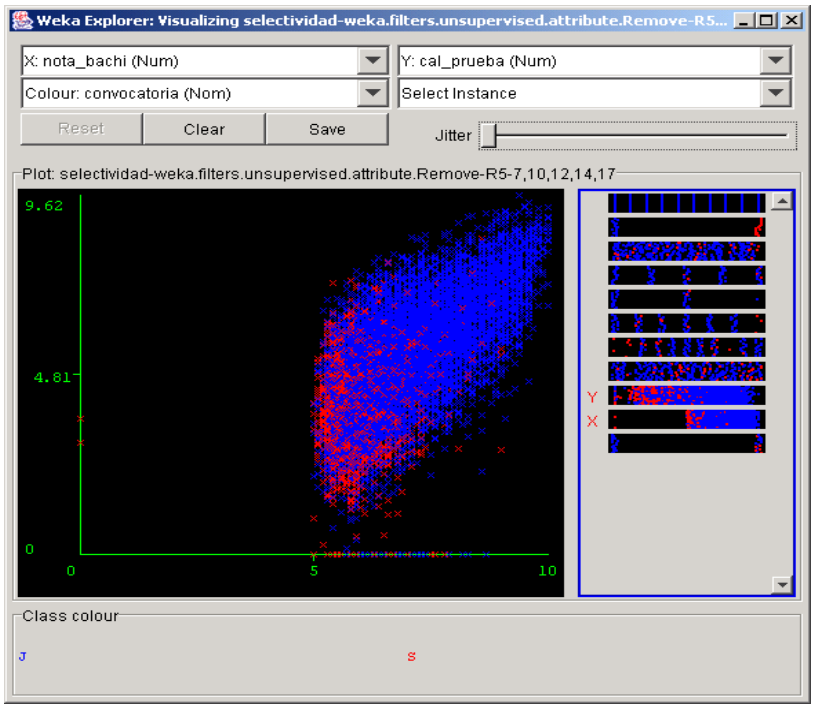
2. Visualización

Una de las primeras etapas del análisis de datos puede ser el mero análisis visual de éstos, en ocasiones de gran utilidad para desvelar relaciones de interés utilizando nuestra capacidad para comprender imágenes. La herramienta de visualización de WEKA permite presentar gráficas 2D que relacionen pares de atributos, con la opción de utilizar además los colores para añadir información de un tercer atributo.

Las instancias se pueden visualizar en gráficas 2D que relacionen pares de atributos. Al seleccionar la opción **Visualize** del *Explorer* aparecen todas los pares posibles de atributos en las coordenadas horizontal y vertical. La idea es que se selecciona la gráfica deseada para verla en detalle en una ventana nueva. En nuestro caso, aparecerán todas las combinaciones posibles de atributos. Como primer ejemplo vamos a visualizar el rango de calificaciones finales de los alumnos a lo largo de los años, poniendo la convocatoria (junio o septiembre) como color de la gráfica.

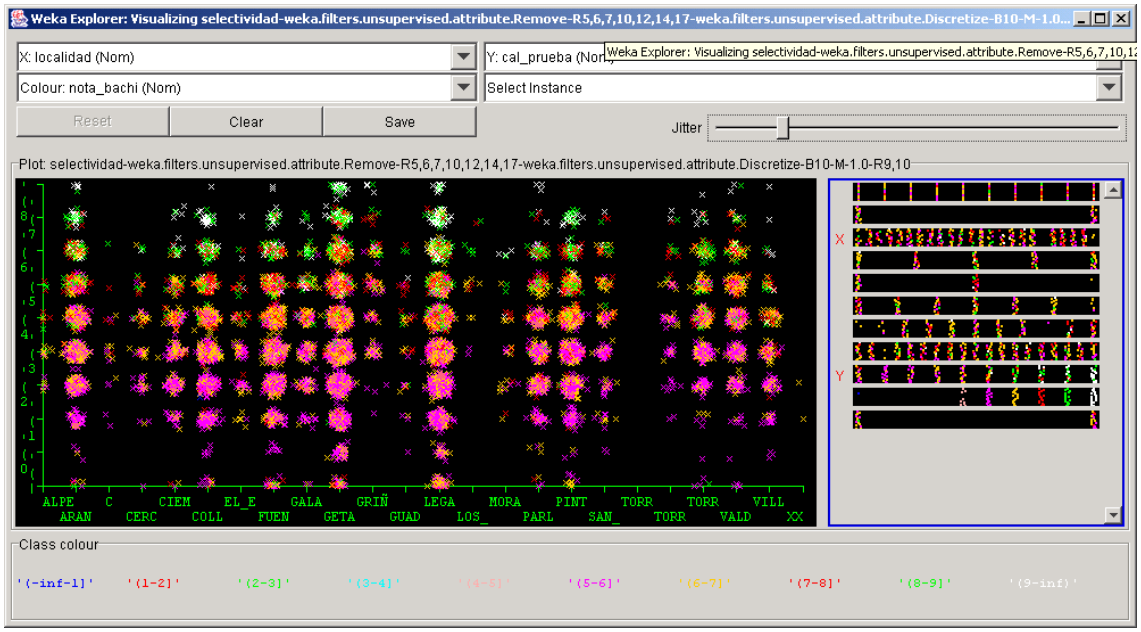


Vamos a visualizar ahora dos variables cuya relación es de gran interés, la calificación de la prueba en función de la nota de bachillerato, y tomando como color la convocatoria (junio o septiembre).



en esta gráfica podemos apreciar la relación entre ambas magnitudes, que si bien no es directa al menos define una cierta tendencia creciente, y como la convocatoria está bastante relacionada con ambas calificaciones.

Se sugiere preparar el siguiente gráfico, que relaciona la calificación obtenida en la prueba con la localidad de origen y la nota de bachillerato, estando las calificaciones discretizadas en intervalos de amplitud 2



Aquí el color trae más información, pues indica en cada intervalo de calificaciones de la prueba, la calificación en bachillerato, lo que permite ilustrar la "satisfacción" con la calificación en la prueba o resultados no esperados, además distribuido por localidades.

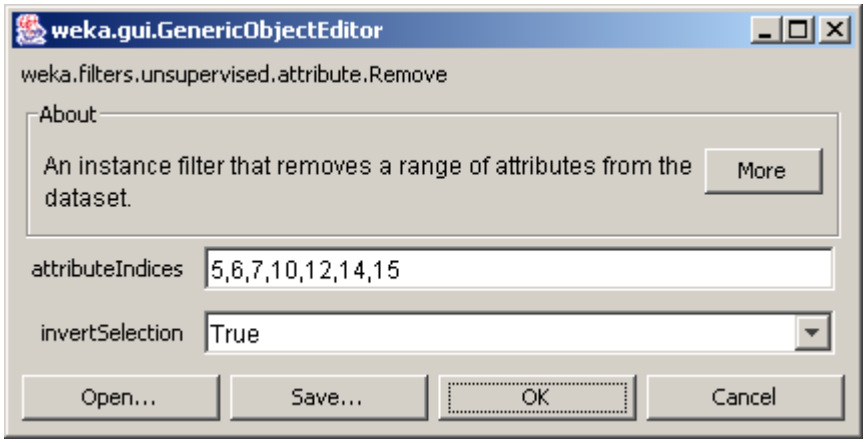
3. Predicción numérica

La predicción numérica se define en WEKA como un caso particular de clasificación, en el que la clase es un valor numérico. Los algoritmos de predicción numéricas aparecen mayoritariamente en el apartado *classifiers->functions*, aunque también en *classifiers->trees*.

Vamos a ilustrar algoritmos de predicción numérica en WEKA con dos tipos de problemas. Por un lado, "descubrir" relaciones deterministas que aparecen entre variables conocidas, como calificación en la prueba con respecto a las parciales y la calificación final con respecto a la prueba y bachillerato, y buscar otros modelos de mayor posible interés.

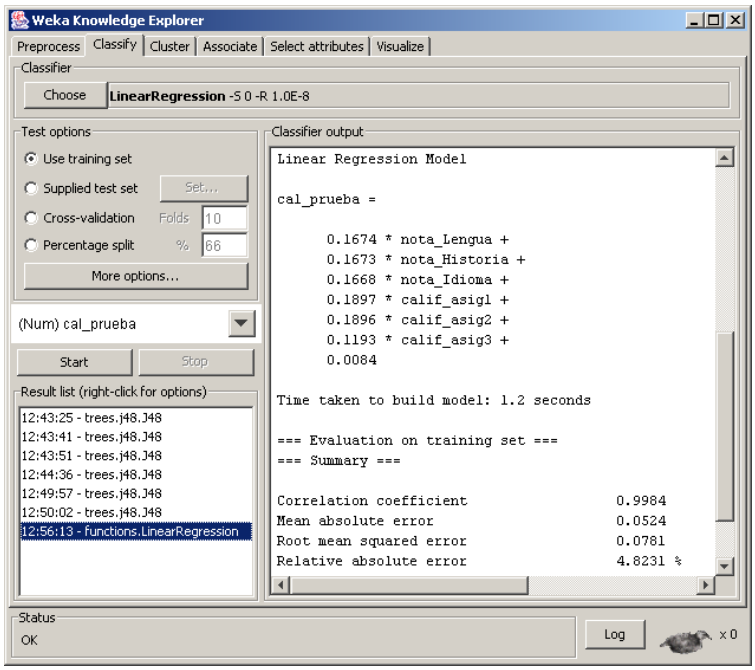
Relación entre calificación final y parciales

Seleccionamos los atributos con las 6 calificaciones parciales y la calificación en la prueba:



Vamos a aplicar el modelo de predicción más popular: regresión simple, que construye un modelo lineal del atributo clase a partir de los atributos de entrada: **functions->LinearRegresion**

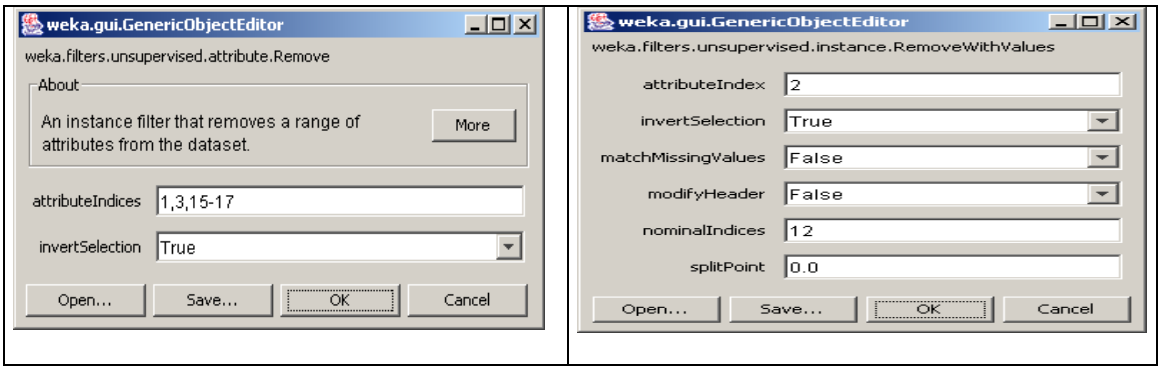
Como resultado, aparece la relación con los pesos relativos de las pruebas parciales sobre la calificación de la prueba:



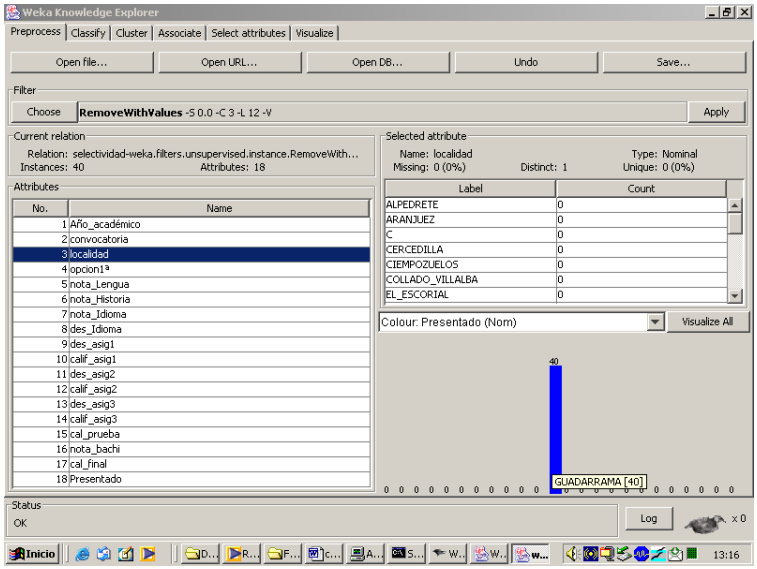
Hay que observar que en los problemas de predicción la evaluación cambia, apareciendo ahora el coeficiente de correlación y los errores medio y medio cuadrático, en términos absolutos y relativos. En este caso el coeficiente de correlación es de 0.998, lo que indica que la relación es de una precisión muy notable.

Si aplicamos ahora esta función a la relación entre calificación final con calificación en la prueba y nota de bachillerato (filtro que selecciona únicamente los atributos 15-17), podemos determinar la relación entre estas variables: qué peso se lleva la calificación de bachillerato y de la prueba en la nota final. Vamos a hacerlo primero con los alumnos de una población pequeña, de Guadarrama (posición 12 del atributo localidad). Aplicamos los filtros

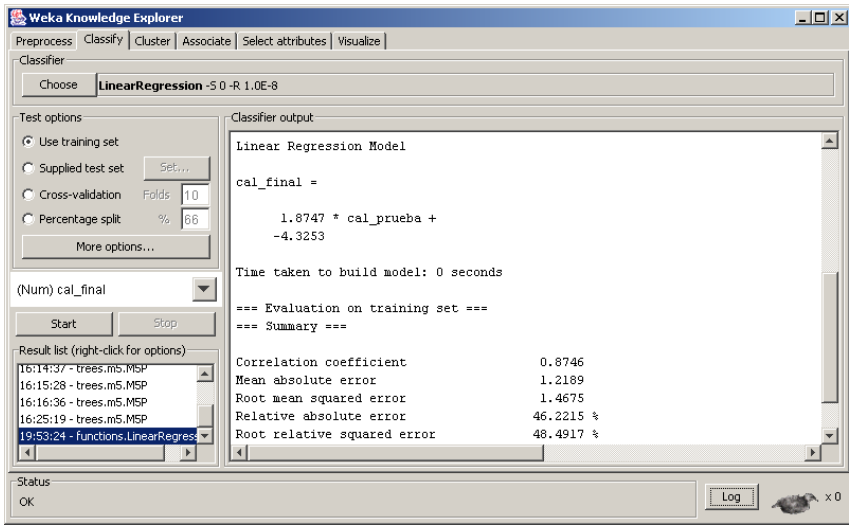
correspondientes para tener únicamente estos alumnos, y los atributos de calificaciones de la prueba, bachillerato y final:



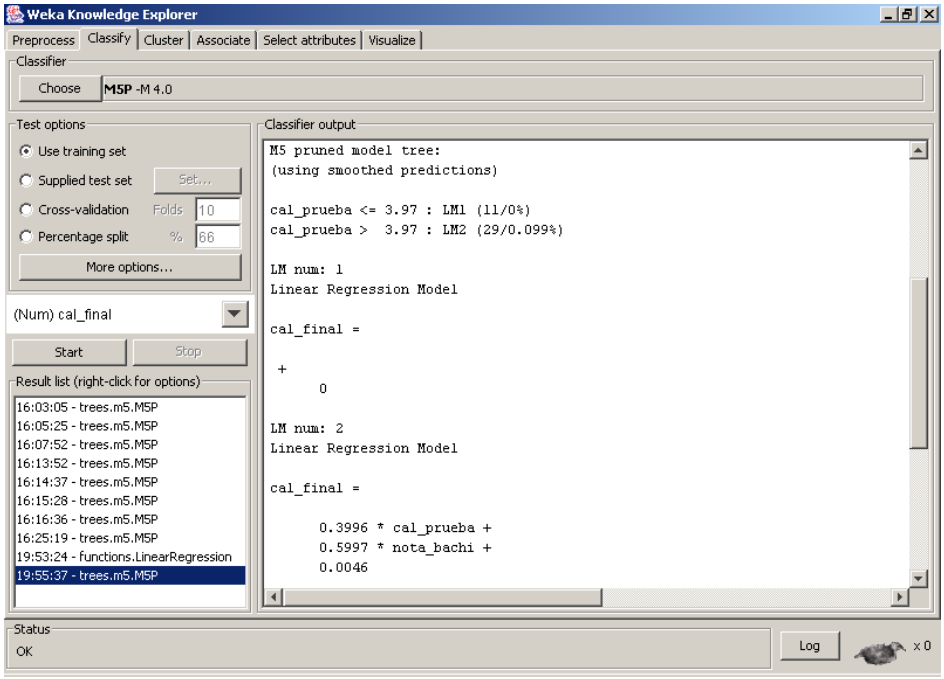
llegamos a 40 instancias:



si aplicáramos regresión lineal como en el ejemplo anterior, obtenemos el siguiente resultado:



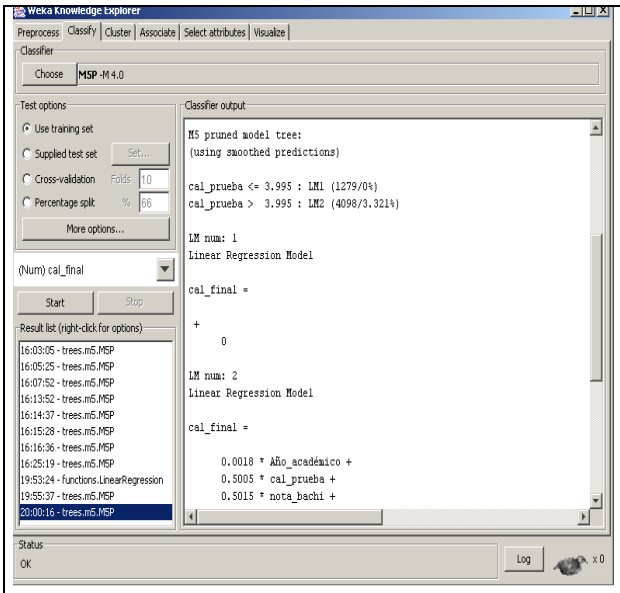
el resultado deja bastante que desear porque la relación no es lineal. Para solventarlo podemos aplicar el algoritmo M5P, seleccionado en WEKA como **trees->m5->M5P,** que lleva a cabo una regresión por tramos, con cada tramo determinado a partir de un árbol de regresión. Llegamos al siguiente resultado:



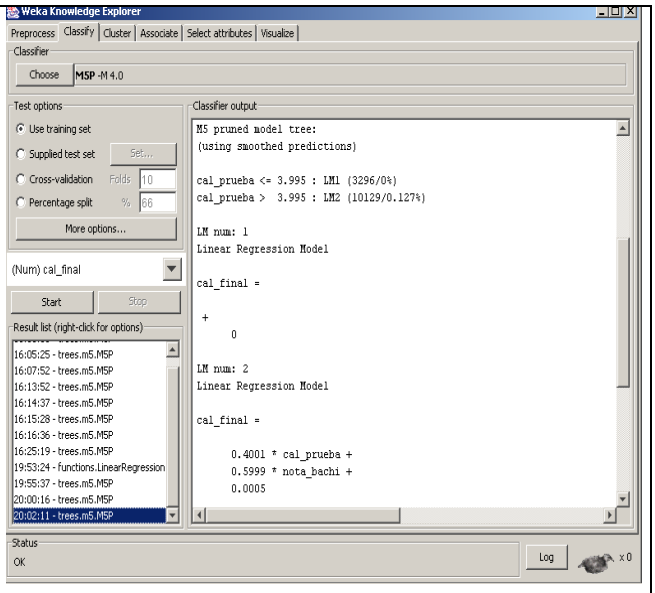
que es prácticamente la relación exacta utilizada en la actualidad: 60% nota de bachillerato y 40% de la prueba, siempre que se supere en ésta un valor mínimo de 4 puntos.

Si aplicamos este algoritmo a otros centros no siempre obtenemos este resultado, por una razón: hasta 1998 se ponderaba al 50%, y a partir de 1999 se comenzó con la ponderación anterior. Verifíquese aplicando este algoritmo sobre datos filtrados que contengan alumnos de antes de 1998 y de 1999 en adelante. En este caso, el algoritmo M5P no tiene capacidad para construir el modelo correcto, debido a la ligera diferencia en los resultados al cambiar la

forma de ponderación. Los árboles obtenidos en ambos casos se incluyen a continuación:



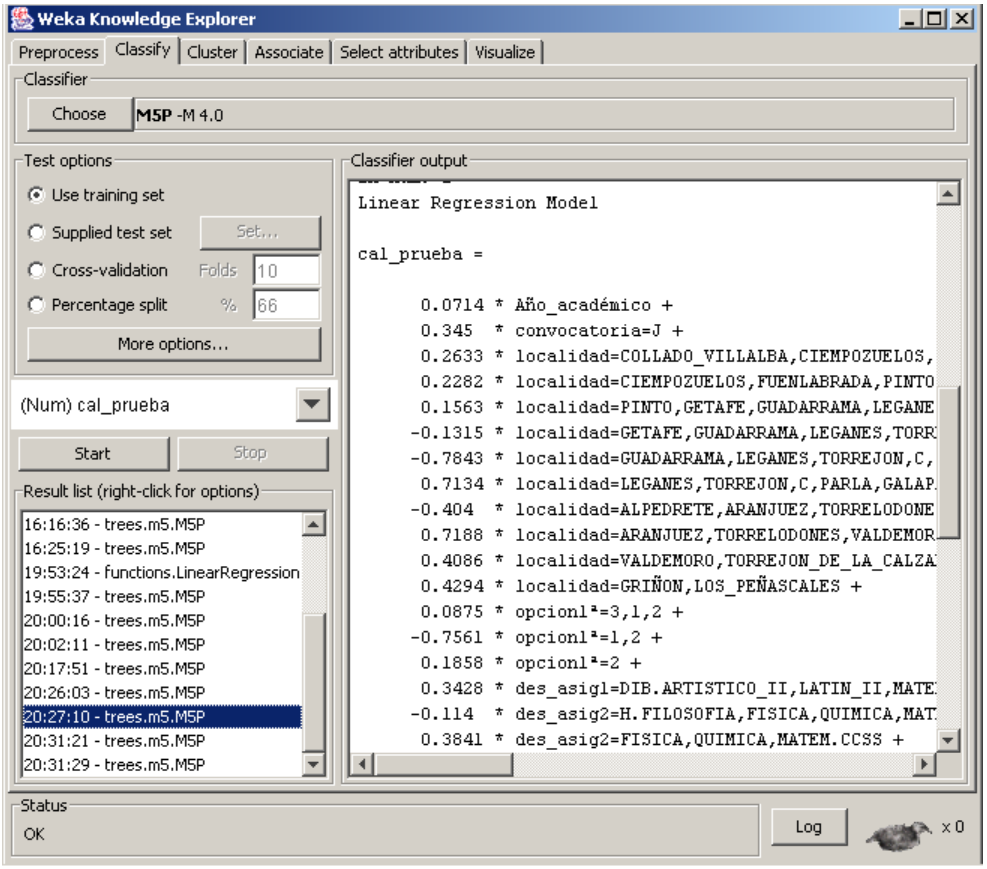
hasta 1998



de 1999 en adelante

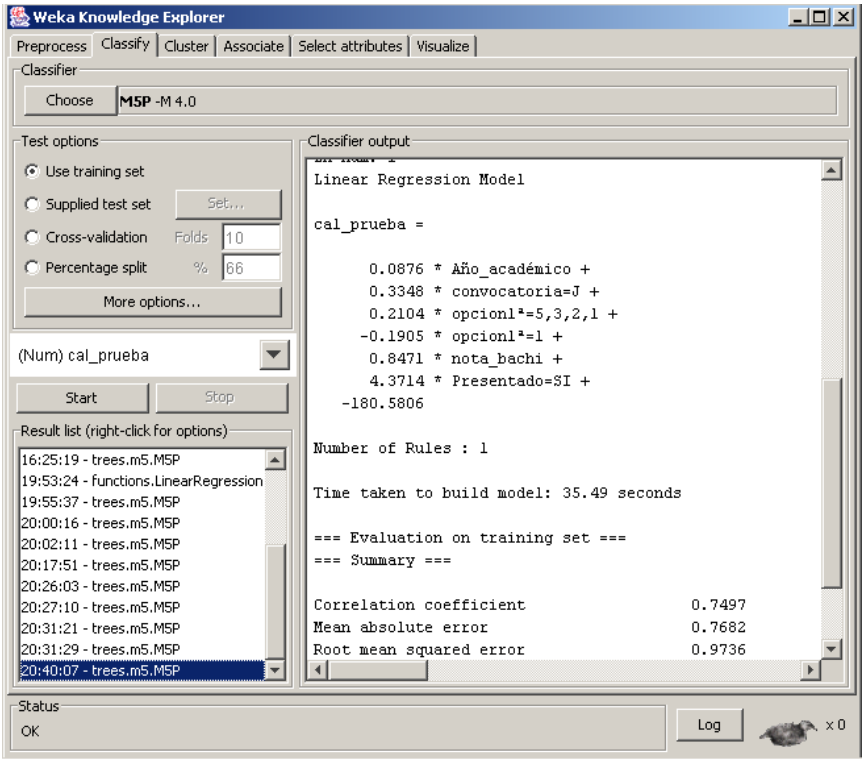
Predicción de la calificación

Vamos a aplicar ahora este modelo para intentar construir un modelo aplicación más interesante, o, al menos, analizar tendencias de interés. Se trata de intentar predecir la calificación final a partir de los atributos de entrada, los mismos que utilizamos para el problema de clasificar los alumnos que aprueban la prueba. Si aplicamos el algoritmo sobre el conjunto completo llegamos al siguiente modelo:

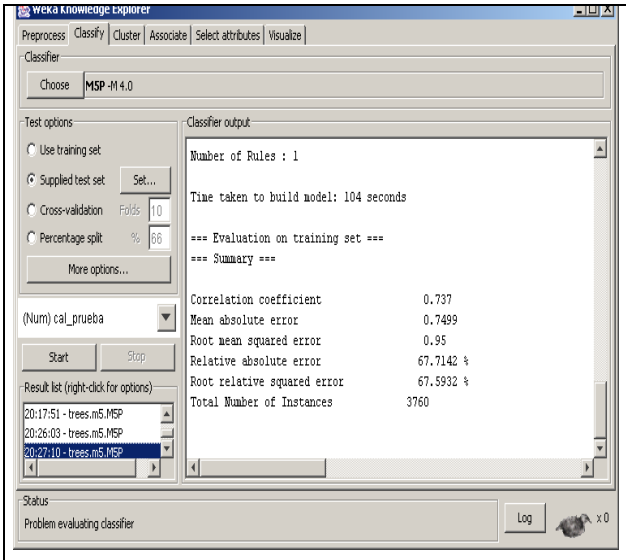


obsérvese cómo trata el algoritmo los atributos nominales para incluirlos en la regresión: ordena los valores según el valor de la magnitud a predecir (en el caso de localidad, desde Collado hasta Los Peñascales y en el de opción, ordenadas como 4º, 5º, 3º, 2º, 1º), y va tomando variables binarias resultado de dividir en diferentes puntos, determinando su peso en la función. En esta función lo que más pesa es la convocatoria, después la nota de bachillerato, y después entran en juego la localidad, asignaturas optativas, y opción, con un modelo muy complejo.

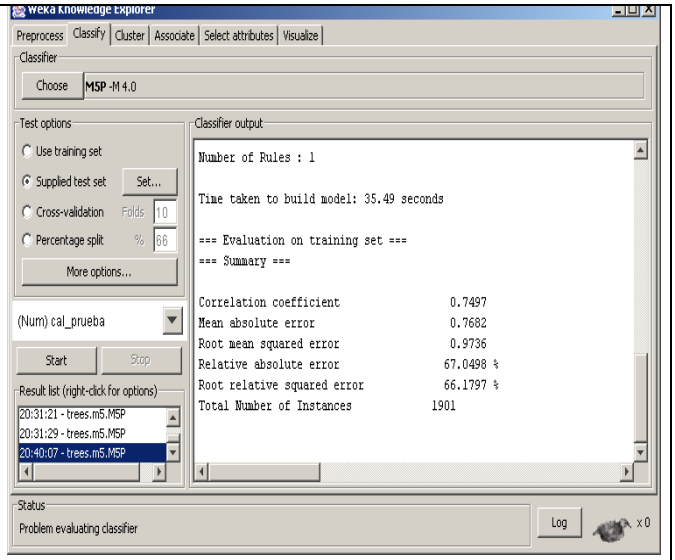
Si simplificamos el conjunto de atributos de entrada, y nos quedamos únicamente con el año, opción, nota de bachillerato, y convocatoria, llegamos a:



este modelo es mucho más manejable. Compare los errores de predicción con ambos casos:



modelo extenso

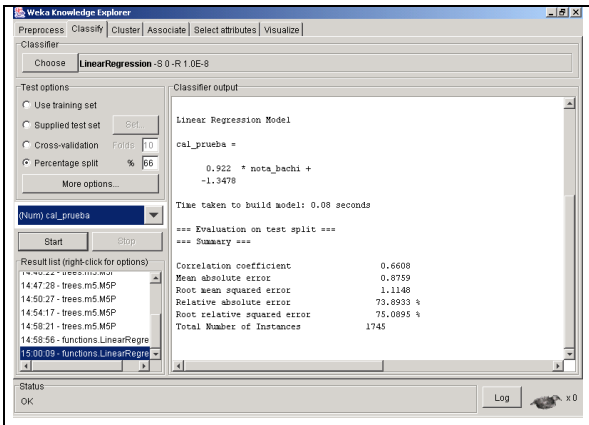


modelo simplificado

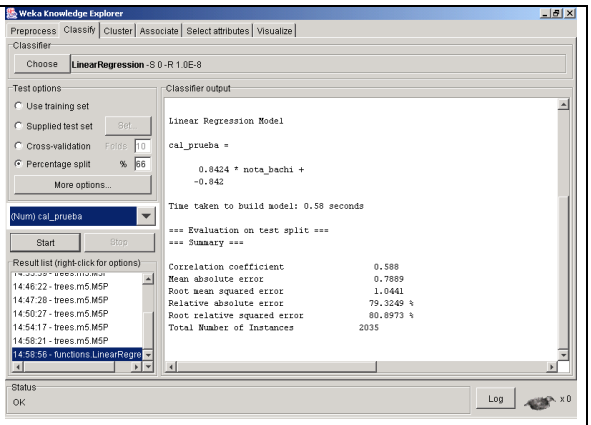
Correlación entre nota de bachillerato y calificación en prueba

Finalmente, es interesante a veces hacer un modelo únicamente entre dos variables para ver el grado de correlación entre ambas. Continuando con nuestro interés por las relaciones entre calificación en prueba y calificación en bachillerato, vamos a ver las diferencias por opción. Para ello filtraremos por un lado los alumnos de opción 1 y los de opción 4. A continuación dejamos

únicamente los atributos calificación en prueba y nota de bachillerato, para analizar la correlación de los modelos para cada caso.



alumnos opción 1º



alumnos opción 4º

podemos concluir que para estas dos opciones el grado de relación entre las variables sí es significativamente diferente, los alumnos que cursan la opción 1º tienen una relación más "lineal" entre ambas calificaciones que los procedentes de la opción 4º