CUESTIONARIO PRACTICA SMART MOTORS: Parte 1

Alumno 1 : Jezael Pérez Herrera_____

Alumno 2: Elena Cantero Molina_____

Grupo : INSE15_____

Selección de parámetros

Trabajaremos con la base de datos de señales en motores “sensorless” que encontrará en moodle.

La base de datos consta de 58500 ejemplos con 48 atributos por cada ejemplo y 11 clases (tipos de defectos)

Lo primero que vamos a hacer es utilizar un clasificador de los que hemos visto usando todos los parámetros que son 48. Usaremos una prueba con percentage Split de 66/34 por rapidez. Cuando tengamos que determinar más detalle sobre significancia entre métodos, usaremos 10-fold. Para el caso J48 ponga también los valores máximo y mínimo según el margen de confianza. Puede usar un fichero en Moodle para calcular los valores máximo y mínimo.

1. Rellene la siguiente tabla

Método	**Tasa de éxito Prueba 66/34**
0R	*8.79*
1R	*63.57*
Naive Bayes	*75.34*
J48	*98.77* *Máx : 98.93 , Min : 98.62*

2. Volvemos a la pantalla de “Preprocess” y visualizamos todos los parámetros. Cuando los valores de atributo están muy repartidos entre clases, en teoría tendremos mejores posibilidades de clasificación. De los parámetros bb, elija 6 que tengan bastante dispersión entre clases y haga una prueba usando solo esos 6. Para ello debe eliminar los otros de la base de datos. Seleccione esos 6 junto con la clase, y use las opciones “invert” y “remove”. No se olvide de no borrar la clase. Vuelva a hacer los experimentos anteriores y rellene la tabla:

Método:	Atributos seleccionados	Prueba 66/34
0R	*b19-b24*	*8.79*
1R	*b19-b24*	*11.1966*
Naive Bayes	*b19-b24*	*9.2609*
J48	*b19-b24*	*15.274* *Máx: 15.7739 , Min: 14.7740*

Explique la variación de las tasas 0R y 1R si es que existen con este método. Diga si en los demás métodos hay variación de la tasa.
Como con el método 0R no se tienen en cuenta los atributos por lo tanto es indiferente el número de atributos seleccionados y la tasa de éxito es la misma. Sin embargo, en el método 1R al tener en cuenta tanto las clases como el atributo con menor número de errores la tasa de éxito es bastante menor que anteriormente.

Como Naive Bayes y J48 hacen uso de todos los atributos y cómo sólo seleccionamos 6 en este caso, la tasa de éxito es bastante menor.

3. Si hubiéramos elegido solo los 6 primeros de los aa, tendríamos un resultado siguiente: póngalo en la tabla

Método	Prueba 66/34
0R	*8.7934*
1R	*16.6013*
Naive Bayes	*13.6601*
J48	*8.8688* *Máx: 89.1282 , Min: 88.2478*

Explique el resultado que obtiene
En este caso, el método 0R no tiene variación por lo comentado anteriormente. El método 1R y Naive Bayes aumentan ligeramente y el J48 sin embargo disminuye. Estando en las condiciones anteriores con 6 atributos, podemos observar que la diferencia de tasa de éxito cambia por la desviación típica de ambos ejemplos.

Vemos también que la selección de atributos es importante para tener un buen sistema. Si por ejemplo, podemos conformarnos con por ejemplo, un 72,5% de tasa de error, podemos ahorrarnos un montón de atributos que nos harán tener un sistema mucho más sencillo.

4. Vamos a trabajar ahora con los métodos automáticos que tengo para preprocesar o seleccionar atributos. Empezaremos con los métodos de preproceso no supervisados. Asegúrese primero de hacer un “Undo”.
   Empezaremos con unsupervised attribute discretize y ponga un bin de 20 y veamos en “visualize all” la distribución de los datos. Haga una captura de ventana.

Pruebe los resultados de Naive Bayes y J48 con esta nueva discretización. Siempre que lo considere necesario para sus conclusiones, calcule también el margen de confianza tal como se lo hemos explicado (fórmula de 1,96 * raíz cuadrada (x (100-x)/N). Para ello abra el fichero de Excel que está en moodle y ponga ahí los números correspondientes.

Método	Prueba 66/34 ±
Naive Bayes	*98.7883* ± 0.1520
J48	97.3655 ± 0.2226

Comente los resultados obtenidos comparando éstos con el sistema base y diga si las diferencias son significativas estadísticamente

Con el método J48 no vemos diferencias significativas mientras que en el Naive Bayes sí gracias a esta discretización.

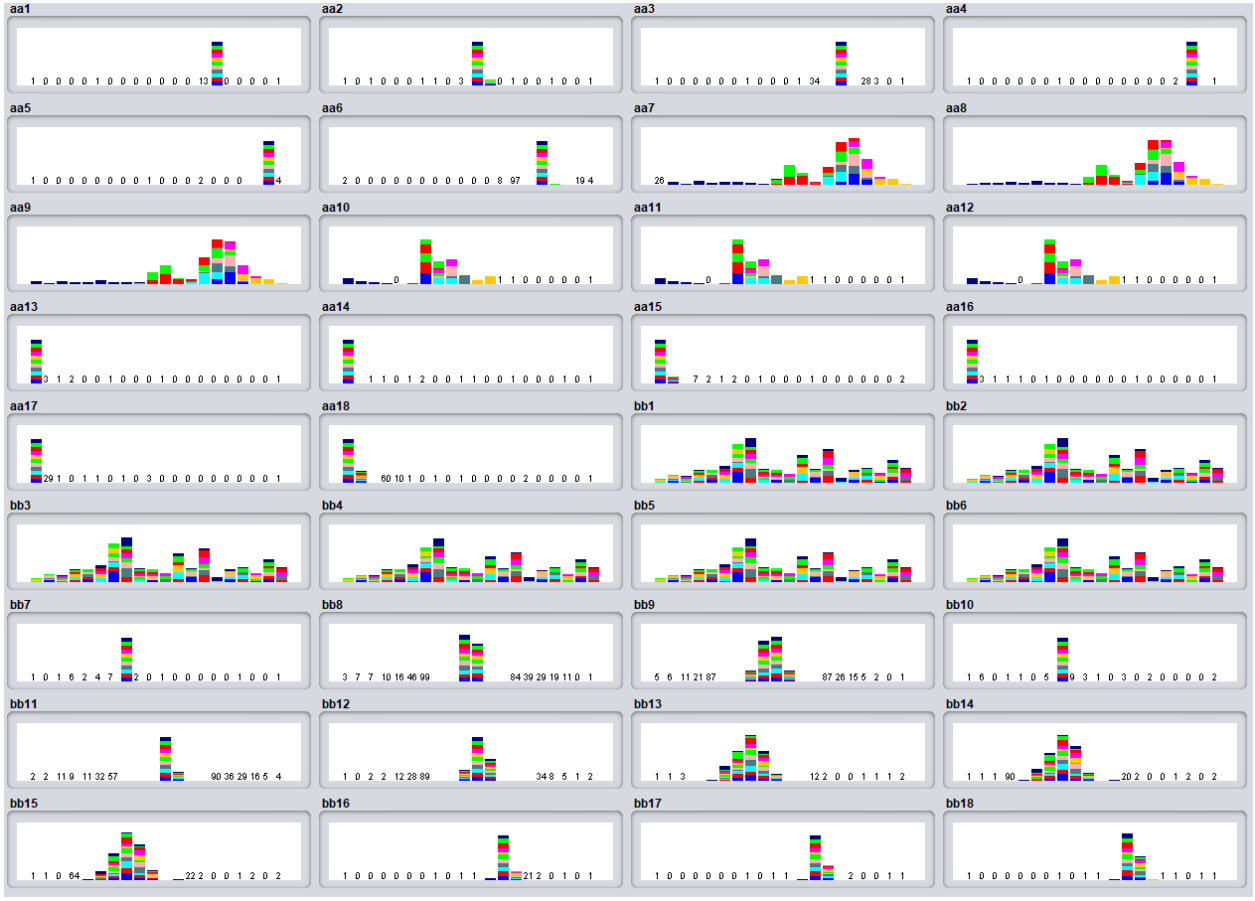
5. Si nos fijamos en los parámetros visualizados anteriormente, vemos que hay una serie de parámetros que tienen todas las clases en el mismo bin. Estos son aa1, aa3, aa4, aa5, aa6, aa13, aa14, aa16, aa17, bb7, bb10, bb19, bb20, bb22, bb23. Remueva ahora esos parámetros y realice una nueva clasificación. Ponga en la tabla los resultados correspondientes

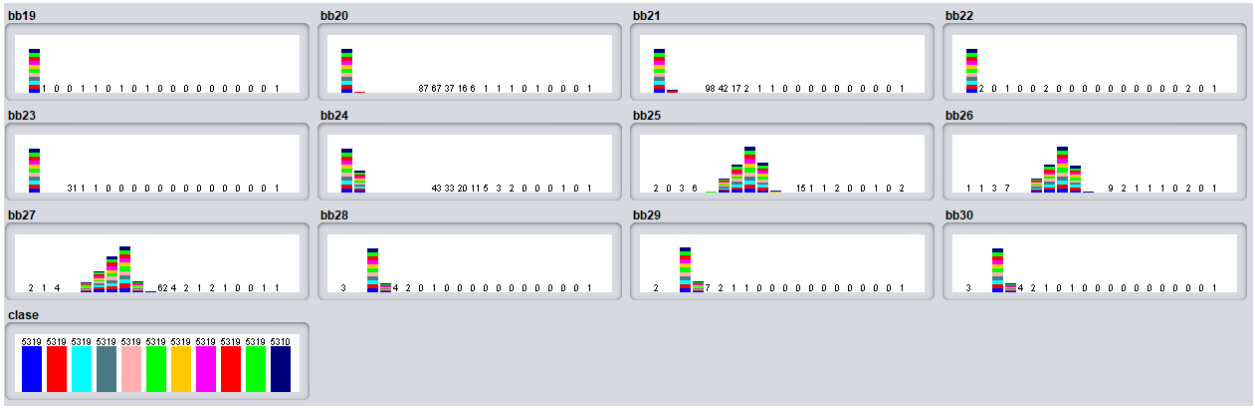
Método	Prueba 66/34
Naive Bayes	*98.4062*
J48	*97.2851*

Diga si estos resultados se diferencian significativamente de los resultados del apartado anterior. ¿Qué conclusiones extrae de esta prueba?

Se puede observar que apenas hay diferencias significativas entre los resultados de este apartado y el anterior porque tenemos el mismo bin y por lo tanto hay información redundante y así se puede eliminar y hacer un sistema más sencillo.

6. Vamos a probar ahora con “equal frequency binning”. Para ello pulse el “undo” para volver al sistema original y tendremos que pulsar ahora “Use equal frequency” a True con el mismo número de bins. Haga un Visualize all y una captura de pantalla y rellene la correspondiente tabla



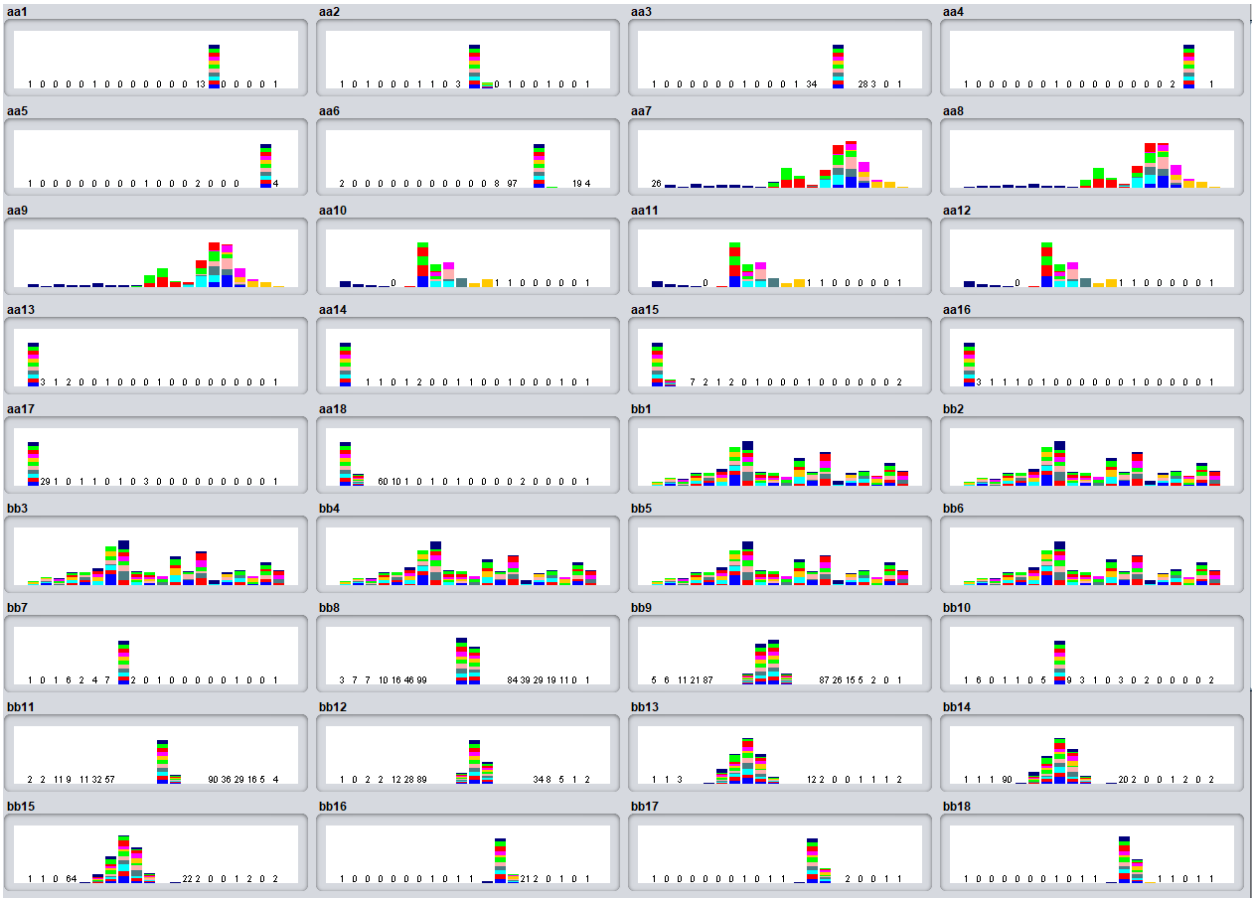


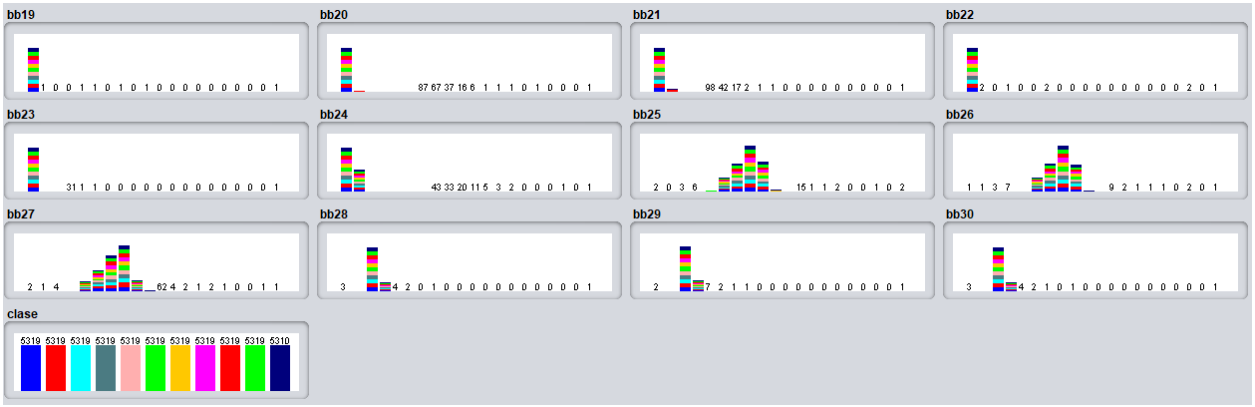
Método	Prueba 66/34
Naive Bayes	*98.7883*
J48	*97.3655*

Comente cómo es ahora la distribución de datos por bin. Explique por qué en algún caso no es plana. Comente el resultado de tasa obtenido.

Los resultados son parecidos y a veces es plana porque siempre tenemos la misma frecuencia de binning.

7. Pasemos ahora a hacer discretización supervisada. En “supervised attribute” pulse “discretize” Capture la ventana haciendo un visualize all y ejecute los algoritmos y rellene la tabla.





Método	Prueba 66/34

Naive Bayes	*98.7883*
J48	*97.3655*

Note que los bins no están equiespaciados. Explique por qué. Comente los resultados comparándolos con los anteriores

Los resultados obtenidos en este apartado y en el anterior son idénticos. Esto se debe a lo que se comenta en el siguiente apartado, ya que esta igualdad en los resultados carecía de lógica.

Sin embargo, lo que acabamos de hacer tiene un defecto, la discretización supervisada se ha fijado en las clases para determinar los mejores intervalos posibles y ha usado todos los datos para ello (no ha separado entrenamiento y test). Al hacer la clasificación y la separación en training y test el experimento no es correcto porque las etiquetas del test ya han sido vistas en la fase de discretización. Por ello, para hacerlo correctamente debe hacer "undo" para revertir la discretización e ir a la pestaña "Classify" y utilizar la opción Filtered Classifier. Dentro de esa ventana y si selecciona las opciones, ahora puede seleccionar en la pestaña "filter" el método de filtro (supervised attribute discretize) y en la pestaña "classifier" el clasificador que utilizaremos, primero Naive Bayes y después el de J 48. Rellene la tabla-

Método	**Prueba 66/34**
Naive Bayes	*98.723*
J48	*97.702*

Estos resultados son los correctos. Compare los mismos con los obtenidos en el párrafo anterior y con los obtenidos por defecto al principio del cuestionario y haga un comentario

No nos ha dado tiempo

Si no conociéramos a priori los datos de test, pero hubiéramos establecido que queremos hacer discretización entonces tenemos que aplicar el algoritmo de discretización obtenido de los datos de entrenamiento a un nuevo test. Para ello debe utilizar un comando de consola:

java weka.filters.supervised.attribute.Discretize –b –i fichero original entrenamiento –o fichero original discretizado –r fichero de test original –s fichero de test discretizado –c índice de la clase

y después calcular la tasa de clasificación usando el clasificador que queramos y usando como test en vez de "percentage Split" el "Supplied test set" que será el fichero que obtengamos discretizado de la fase anterior y que ya probamos en la primera práctica. A modo de ejemplo hemos hecho esos experimentos dividiendo el fichero original en dos partes , los primeros 35000 datos y los segundos 19500 y aplicando el discretizador. Los resultados han sido en Naive Bayes de 98,71% y en J 48 de 97,55% que son indistinguibles de los obtenidos en el apartado anterior (difieren un poco dado que

hemos usado un Split exacto 66,66, 33,33 y en el experimento anterior hemos usado 66/34). El comando que hemos usado para la discretización ha sido el siguiente:

```
java weka.filters.supervised.attribute.Discretize -b -i C:\Users\Manolo\Sensorless_drive_diagnosis_58500-random-pri35000.arff -o C:\Users\Manolo\Sensorless_drive_diagnosis_58500-random-pri35000-d.arff -c 49 -r C:\Users\Manolo\Sensorless_drive_diagnosis_58500-random-seg19500.arff  -s C:\Users\Manolo\Sensorless_drive_diagnosis_58500-random-seg19500-d.arff
```