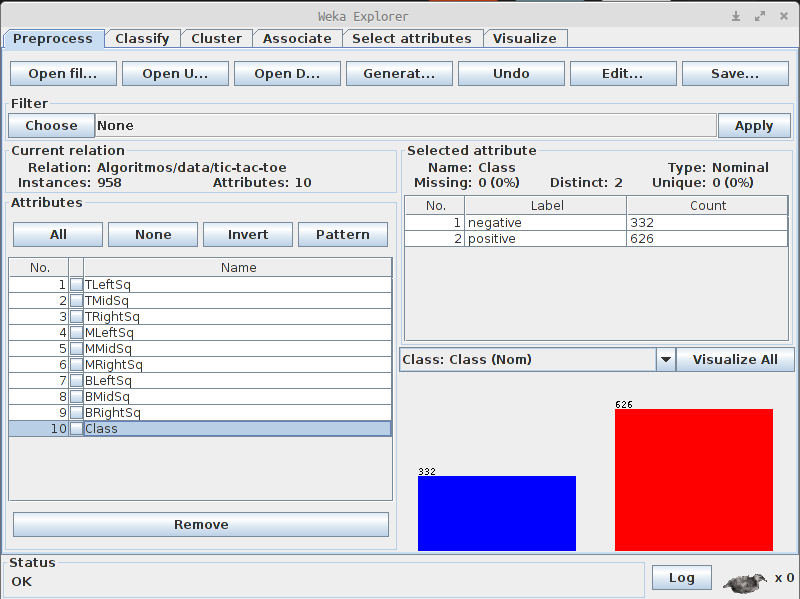
**Utilizando la herramienta Weka, realizar una clasificación NaiveBayes para los conjuntos de datos estudiados en la práctica 1:**

Antes de poder analizar los ficheros con Weka ha sido necesaria la conversión del formato en el que se encontraban los datos, al formato *\*.arrf*.

Una vez convertidos los ficheros ya podemos analizarlos con Weka. Para ello, cargamos el fichero *tic-tac-toe.arff* y observamos:



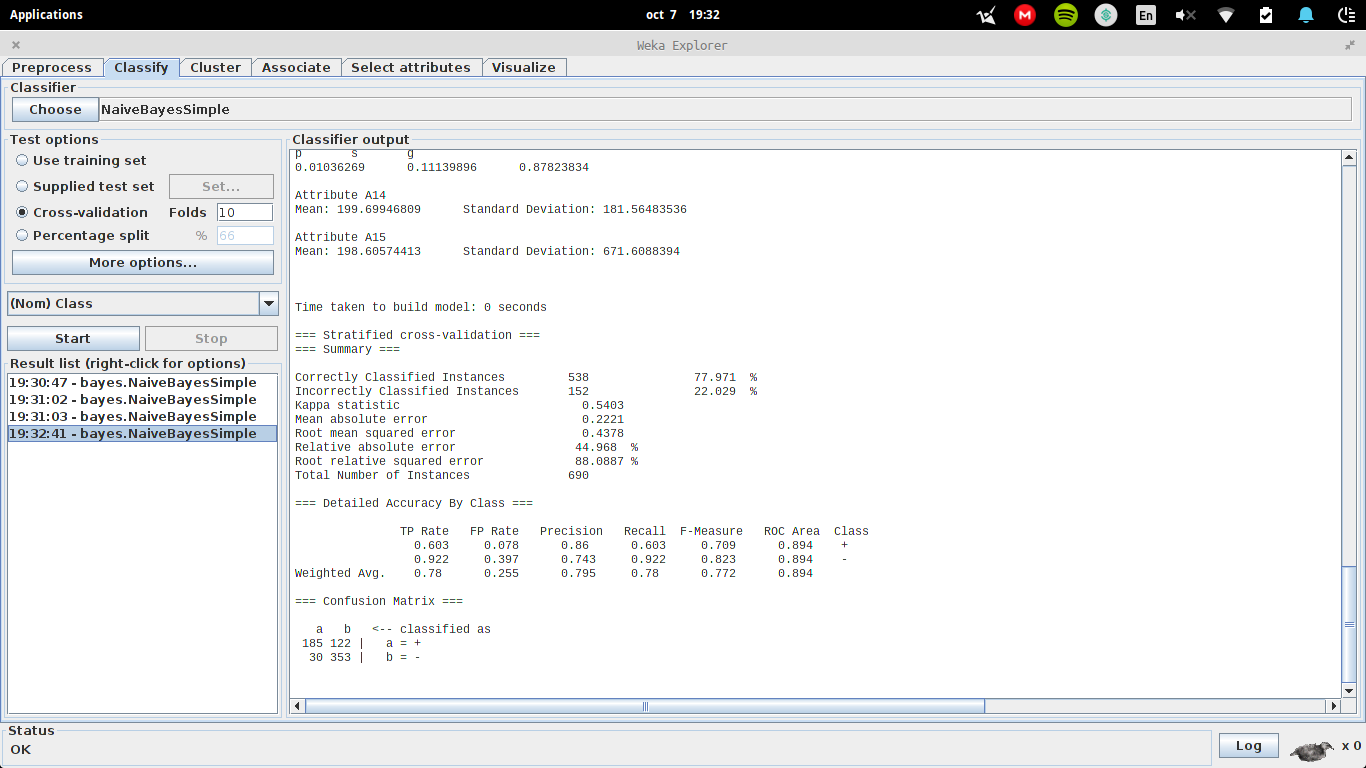
Nuestro clasificador a Priori contaba: { negative = 254, positive = 464} . ¿A qué se debe esta diferencia? Nuestro clasificador a Priori ignora al leer el fichero todas las filas que contengan alguna "?". En cambio Weka, sólo ignora estas filas cuando necesita utilizar ese dato. Para clasificar a Priori no necesitamos saber la media exacta del atributo TLeftSq por ejemplo. Si en ese atributo hay alguna "?", Weka elimina la fila al necesitar el valor, nuestra implementación al cargar el fichero.

Independientemente de esta diferencia de implementación (que podría ser significativa) observamos que el error cometido por Weka es 34.6555% y por nuestra implementación 35.3760 %.

Una vez comentado esto, procedemos a clasificar con NaiveBayes los datos.

Crx

Incluimos una captura de pantalla de Weka, y a continuación comentamos los resultados:



Observamos que ha clasificado mal un 22% de las muestras. Actualmente no podemos comparar este clasficador por el implementado en la Práctica 1 ya que no llegamos a implementarlo.

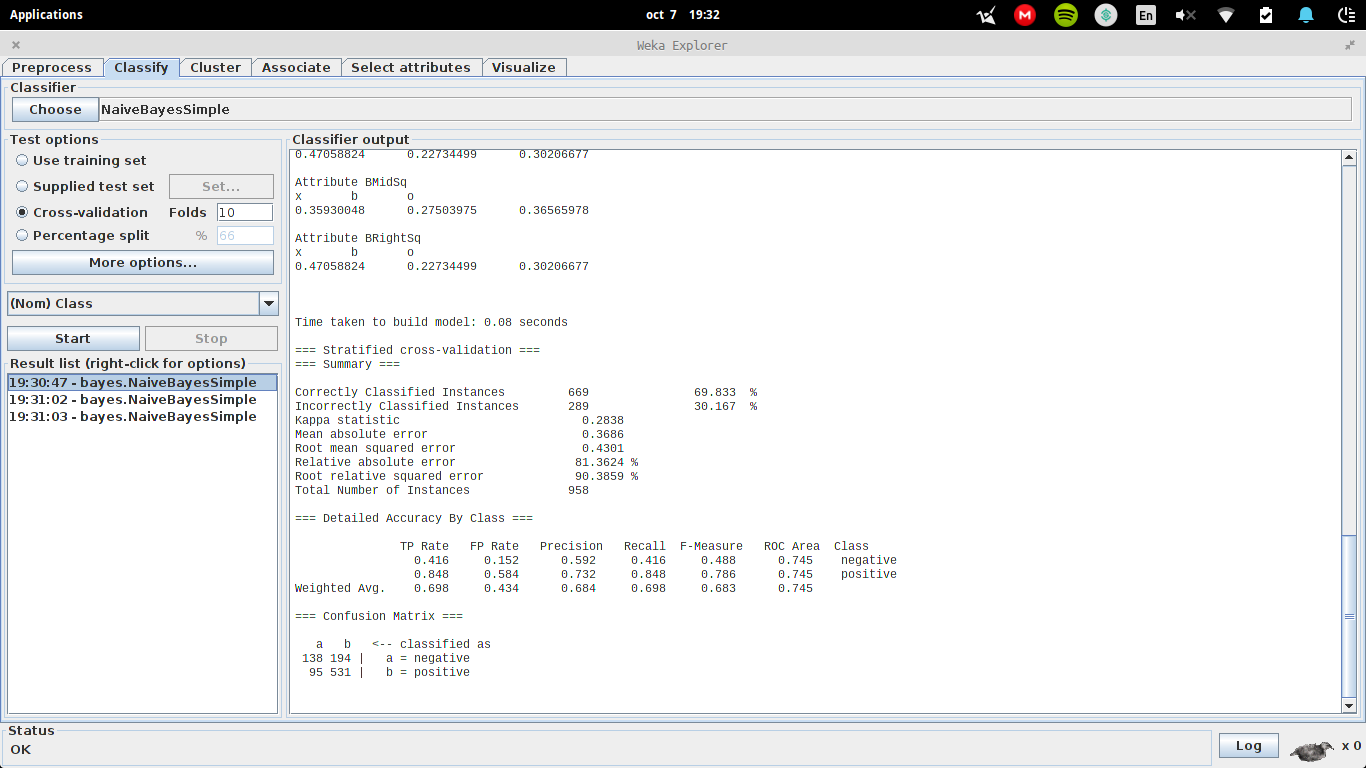
Comentamos la matriz de confusión, que es una manera de representar el error muy visual:

|  |  |  |
| --- | --- | --- |
| Clasificado como --> | a | b |
| siendo a = + | 185 | 122 |
| siendo b = - | 30 | 353 |

Es curioso que, siendo una muestra de clase "-", es clasificada correctamente en un porcentaje mucho mayor (92%) que siendo "+" (60%). Concluimos que es más fácil reconocer los "-" que los "+".

Tic-tac-toe

Incluimos una captura de pantalla de Weka, y a continuación comentamos los resultados:



Observamos que ha clasificado mal un 30% de las muestras. Este porcentaje es demasiado elevado como para poder confiar en este clasificador. Esperamos que con los algoritmos que vayamos aprendiendo a lo largo del curso se pueda clasificar con un porcentaje menor este conjunto de datos.

Actualmente no podemos comparar este clasficador por el implementado en la Práctica 1 ya que no llegamos a implementarlo.

Comentamos la matriz de confusión, que es una manera muy visual de representar el error:

|  |  |  |
| --- | --- | --- |
| Clasificado como --> | a | b |
| a = negative | 138 | 194 |
| b = positive | 95 | 531 |

En este caso, observamos que la tasa de error de dar como "positive" un elemento "negative" es superior al 50% (58%), en cambio dar como "negative" un elemento "positive" es del 15%, un porcentaje mucho más aceptable.