**PRACTICA 1**

**Tras haber llevado a cabo las tareas previas disponemos de información sobre cómo llevar a cabo distintas estrategias de evaluación de un clasificador, sabemos evaluarlo considerando costes, nos hemos familiarizado con la base de datos del problema y tenemos una idea sobre qué clasificadores de los utilizados hasta ahora ( zeroR, oneR, logística y análisis lineal discriminante) funcionan mejor según unas u otras consideraciones de evaluación .**

**Además:**

* Tenemos guardadas : una versión de los datos con los atributos **propiedad de la vivienda, estado civil, tipo de trabajo,y estatus del préstamo** como atributos **nominales**. Tenemos también una versión de los datos con los atributos propiedad-vivienda, estado-civil y tipo-trabajo **binarizados** y en la que hemos eliminado los valores perdidos ( missing o NA )
* Guardamos, también, una versión en csv de estas dos bases de datos para utilizarla con R. (Puede hacerse usando el botón “sabe” del panel de preproceso de weka.)

**Para empezar:**

Nos familiarizamos primero con el uso del **experimenter**: con discricochesR10.arff y con una versión arff de bankdata estudiando el comportamiento de Naivebayes,Logistica,J48,JRip y otros.

No sólo consideramos las cuestiones técnicas del procedimiento de uso. También “pensamos” en la posible ventaja que pueda tener su uso sobre el del explorer.

**Tras esto, nos planteamos de cara a la práctica:**

**Con weka:**

Además de lo ya realizado hasta aquí.

Usar el archivo original y el archivo binarizado con el experimenter para analizar y discutir el rendimiento de los clasificadores:

Naivebayes, logística, J48, JRip, Hoeffding Tree.

( dos bases de datos diferentes pero sobre la misma realidad y, al menos , 5 clasificadores )

Estudiar el comportamiento de, **al menos**, indicadores como % correctos, Kappa, TPR, FPR,Precisión, Medida F, área bajo la curva ROC (y quizás otros ) .No se trata de “probar por probar” debe hacerse siendo consciente de cuáles se usan y qué miden y **debe considerarse el matiz** que aporta cada uno de ellos , **especialmente** **en aquellos casos** en los que el análisis lleva a **comparaciones de diferente**

Considerar( comparar y anotar ) los resultados obtenidos ( de cara al informe final) . Comparar también, con el LDA, ya hecho con anterioridad, sin olvidar que áquel sólo se aplicó a los datos binarizados( con la otra base de datos no funcionará).

Cómo llevar a cabo esta última comparación debe decidirlo ( comentarlo y justificarlo) el analista ( estudiante)

**Con R:** Usar en todos los análisis los datos originales y los binarizados .

Llevar a cabo un análisis discriminante ( con el no binarizado no se podrá y habrá que prescindir en el análisis de las variables nominales, por ejemplo o bien trabajar con ellas en el formato númerico que tenían en el archivo primigenio de formato excel), una clasificación con el árbol rpart, y una MLP( red neuronal)

**Informe:**

Con todo el trabajo hecho realizar un informe en el que como objetivo final se recomiende una o varias estrategias para utilizar uno o varias clasificadores para el problema del creditscore.

No hay que dejar de tener en mente que se anda buscando un clasificador “ de futuro” que funcione bien “en adelante” para identificar los buenos y malos pagadores de los préstamos.

1.- Respecto a las tareas previas

El informe debe encajar las tareas previas, su realización , resultados y comentarios comparativos en el marco de una primera aproximación hacia el objetivo buscado.

Así ha de entenderse el sentido de cota mínima de éxito que supone la aplicación de los procedimientos zeroR o oneR. De la misma forma cobra sentido el llevar a cabo varias alternativas de evaluación, la consideración de distintos indicadores y la evaluación sensible al coste.

Es importante argumentar las razones por las que se prefieren unas u otras opciones de evaluación, unas u otras opciones de distintas estructuras de coste según los distintos tipos de error, o por qué unos u otros clasificadores de los inicialmente considerados son preferibles según qué medidas de calidad consideremos.

En cambio, listar una y otra vez los resultados de predicción de cada capa de test para cada clasificador es **absurdo**.

Facilitar una corta-pega de los resultados globales de clasificación sin comentario alguno es, por un lado, **insuficiente,** y, por otro lado, en algunas ocasiones, **puede no ser necesario.**

En cambio, huelga decirlo, facilitar y comentar el “modelo” puede ser fundamental ( aunque no siempre). Al menos, en aquellos casos en los que la clasificación resulte especialmente reseñable por su calidad o necesaria para la comparación.

2.-Respecto a las tareas de comparación con el experimenter

Debe quedar clara cada una de **las ventajas competitivas** que pueda tener cada clasificador, **con respecto a** qué otros clasificadores y **según qué** medidas de calidad.

Se debe especificar los criterios usados en los experimentos y lo fundamental de los resultados obtenidos en la forma que se considere pero no se trata obviamente de pegar los resultados de pantalla ( ni siquiera de repetirlos verbalmente tras haber sido pegados)

3.-Respecto a los análisis llevados a cabo con R .

Los resultados, incluyendo modelo, validación, indicadores de calidad y las representaciones más ilustrativas **deben ser comentados** adecuadamente. La comparación, en su caso, con lo ya obtenido con weka o/y otros métodos, siempre debe afrontarse y, si es relevante , reseñarse, comentarse y explicarse.

**Y finalmente** toda la información debe estructurarse hacia el objetivo final de aportar un clasificador (o unos pocos de ellos, en todo caso) como el más adecuado para nuestro propósito, indicando las ventajas que pueda tener pero también sus posibles debilidades.

Caso de que se destaquen dos o más clasificadores en la selección final es interesante considerar qué criterios usar y por qué razón en el caso en el que los resultados a los que llevaran fueran contradictorios. (Probablemente esto conlleve poner el foco en algún indicador de calidad concreto de especial consideración por alguna razón : esta es la clave )

No podrá superarse la práctica si:

El informe no emite una recomendación JUSTIFICADA de uso de un clasificador.

Si no se comentan los resultados más relevantes de la comparaciones.

Si resulta evidente que no se han realizado todas las tareas requeridas.

Su estructura

* no deja claro el proceso y método de trabajo y la razón de ello: Preproceso, comparación de métodos , uso de diferentes herramientas ( Explorer/ experimenter) y diferente software ( Weka / R)
* no se consideran las particularidades que pueden tener las distintas opciones del análisis : Los distintos métodos o técnicas ( No todos pueden usar datos de la misma naturaleza). Las distintas herramientas: Explorer/ Experimenter/ R no hacen las misma cosas. Tampoco los resultados se analizan y muestran igual.