Práctica Clasificadores Combinados Grupo 11

1 Descripción del problema

El problema consiste en predecir si un paciente tiene muchas o pocas posibilidades de sufrir un infarto. El fichero heart.csv consta de múltiples diagnósticos de pacientes relacionados con enfermedades del corazón, contiene las siguientes variables:

- "*age*": edad.
- "*sex*": sexo (1 = hombre; 0 = mujer).
- "*cp*": tipo de dolor de pecho:
    - Valor 1: Angina típica.
    - Valor 2: Angina atípica.
    - Valor 3: Dolor no anginal.
    - Valor 4: Asintomático.
- "*trtbps*": presión sanguínea en reposo (en mm Hg).
- "*chol*": colesterol (en mg/dl).
- "*fbs*": nivel de azúcar en ayunas > 120 mg/dl (1 = Sí; 0 = No).
- "*restecg*": resultados del electrocardiograma en reposo:
    - Valor 0: Normal.
    - Valor 1: Anormalidad en la onda ST-T.
    - Valor 2: Muestra una hipertrofia ventricular izquierda por el criterio de Estes.
- "*thalach*": máximas pulsaciones alcanzadas.
- "*exng*": angina provocada por ejercicio (1 = Sí; 0 = No).
- "*oldpeak*": depresión de la onda ST inducida por el ejercicio en relación con el reposo.
- "*slp*": pendiente del segmento ST durante ejercicio máximo:
    - Valor 1: Ascendente.
    - Valor 2: Plano.
    - Valor 3: Descendente.
- "*caa*": número de vasos sanguíneos principales (0-3).
- "*thall*": talasemia:
    - Valor 1: Defecto fijo (sin flujo de sangre en alguna parte del corazón).
    - Valor 2: Flujo normal de sangre.
    - Valor 3: Defecto reversible (se observa un flujo de sangre pero no es normal).
- "*output*": posibilidad de tener un infarto (1 = Más posibilidades de sufrir infarto; 0 = Menos posibilidades de sufrir infarto).

2 Carga y análisis de los datos

1. En la pestaña <**Preprocess**> de Weka, abrir el fichero heart.csv. ¿Cuántos atributos e instancias contiene? ¿De qué tipo son los atributos?

   14 atributos y 303 instancias. Todos los atributos son numéricos.

2. ¿Cuántos valores toma el atributo “*cp*”?

   4

3. ¿De qué tipo es la clase? ¿Cuántos y qué valores toma?

   Númerica. 2 valores. 1 y 0.

4. Para realizar la clasificación con Weka, es necesario que la clase sea de tipo nominal. Para esto, seleccionar el filtro unsupervised > attribute > NumericToNominal y aplicarlo en la clase.

5. ¿De qué tipo es ahora la clase?

   Nominal

3 Bagging

1. En la pestaña <**Classify**>, elegir el clasificador meta > Bagging.
2. Pulsar en el nombre del filtro y en el botón <**More**> para ver la sintaxis del clasificador. ¿Qué significan los parámetros <**bagSizePercent**>, <**classifier**> y <**numIterations**>?

   bagSizePercent: Size of each bag, as a percentage of the training set size.
   classifier: The base classifier to be used.
   numIterations: The number of iterations to be performed.

3. Completar las siguientes tablas con las precisiones obtenidas:

Classifier: REPTree (78.8779%)

bagSizePercent numIterations	5	20	50	80	100
10	71.9472	77.8878	83.1683	80.198	81.1881
50	78.8779	81.8482	82.8383	82.5083	82.5083
100	78.8779	82.5083	83.1683	82.8383	81.8482
500	78.8779	82.1782	81.8482	82.1782	81.5182

Media: 80.9406

Classifier: J48 (78.5479%)

bagSizePercent numIterations	5	20	50	80	100
10	75.9076	81.5182	77.8878	81.1881	79.5380
50	82.5083	83.4983	80.5281	80.1980	80.5281
100	84.1584	82.8383	81.5182	80.1980	80.1980
500	85.4785	84.4884	82.8383	80.8581	80.1980

Media: 81.3036

Classifier: PART (78.2178%)

bagSizePercent numIterations	5	20	50	80	100
10	75.9076	82.5083	81.5182	80.1980	79.5380
50	82.1782	81.8482	81.8482	81.8482	82.1782
100	83.8284	85.1485	83.8283	82.5083	80.5281
500	85.8086	84.8185	82.8383	82.1782	81.5182

Media: 82.1287

Classifier: IBk (KNN=8 → 83.1683%)

bagSizePercent numIterations	5	20	50	80	100
10	74.9175	81.5182	80.5281	80.8581	81.1881
50	76.5677	79.8680	80.8581	82.1782	81.8482
100	77.8878	79.8680	81.5182	82.5083	81.8482
500	79.5380	79.8680	80.5281	80.8581	80.8581

Media: 80.2806

4. ¿Qué parámetros obtienen una mejor precisión?

Classifier: PART; bagSizePercent: 5; numIterations: 500

3.1 Random Forest

1. En la pestaña <**Classify**>, elegir el clasificador trees > RandomForest.

2. Pulsar en el nombre del filtro y en el botón <**More**> para ver la sintaxis del clasificador. ¿Qué significa el parámetro <**numFeatures**>?

numFeatures: Sets the number of randomly chosen attributes. If 0, int(log_2(#predictors) + 1) is used.

3. Completar las siguientes tablas con las precisiones obtenidas:

numFeatures: 0 (int(log2(13) + 1) = 4)

bagSizePercent numIterations	5	20	50	80	100
200	82.8383	82.1782	82.1782	81.8482	81.8482
500	81.8483	82.5083	82.8383	82.1782	82.8383
1000	82.1782	82.1782	82.1782	81.8482	82.1782
2000	82.1782	82.8383	81.8482	81.8482	81.5182

numFeatures: 3 (sqrt(13))

bagSizePercent numIterations	5	20	50	80	100
200	82.5083	82.8383	81.8482	82.8383	82.5083
500	83.4983	82.8383	82.8383	82.8383	82.5083
1000	82.1782	82.8383	82.8383	83.1683	82.1782
2000	82.5083	83.1683	82.8383	82.8383	82.1782

numFeatures: 13

bagSizePercent numIterations	5	20	50	80	100
200	83.1683	82.1782	81.5182	80.1980	80.1980
500	82.8383	81.8482	81.5182	80.5281	80.1980
1000	82.1782	81.5182	81.5182	80.5281	80.1980
2000	82.5083	81.5182	81.5182	80.5281	80.5281

4. ¿Qué parámetros obtienen una mejor precisión?

numFeatures: 3; bagSizePercent: 5; numIterations: 500

4 Boosting

1. En la pestaña Classifier, elegir el clasificador meta > AdaBoostM1, ¿cuál es la precisión usando REPTree, J48, OneR, ZeroR y NaiveBayes?

REPTree	J48	OneR	ZeroR	NaiveBayes

5 Stacking

1. En la pestaña <**Classify**>, elegir el clasificador meta > Stacking, y ajustar los parámetros, añadiendo en clasifiers J48 y NaiveBayes, y como meta clasificador PART. ¿Cuál es la precisión?

2. Cambiar el metaclasificador a OneR, ZeroR y NaiveBayes y comparar sus precisiones. ¿Cuál es la más alta?

OneR	ZeroR	NaiveBayes