Inteligencia en Sistemas Electrónicos

GUIÓN PRÁCTICA 1.1 Introduccion a Weka y sistemas de aprendizaje

Alumno 1 - Jezael Pérez Herrera_____

Alumno 2 - Elena Cantero Molina_____

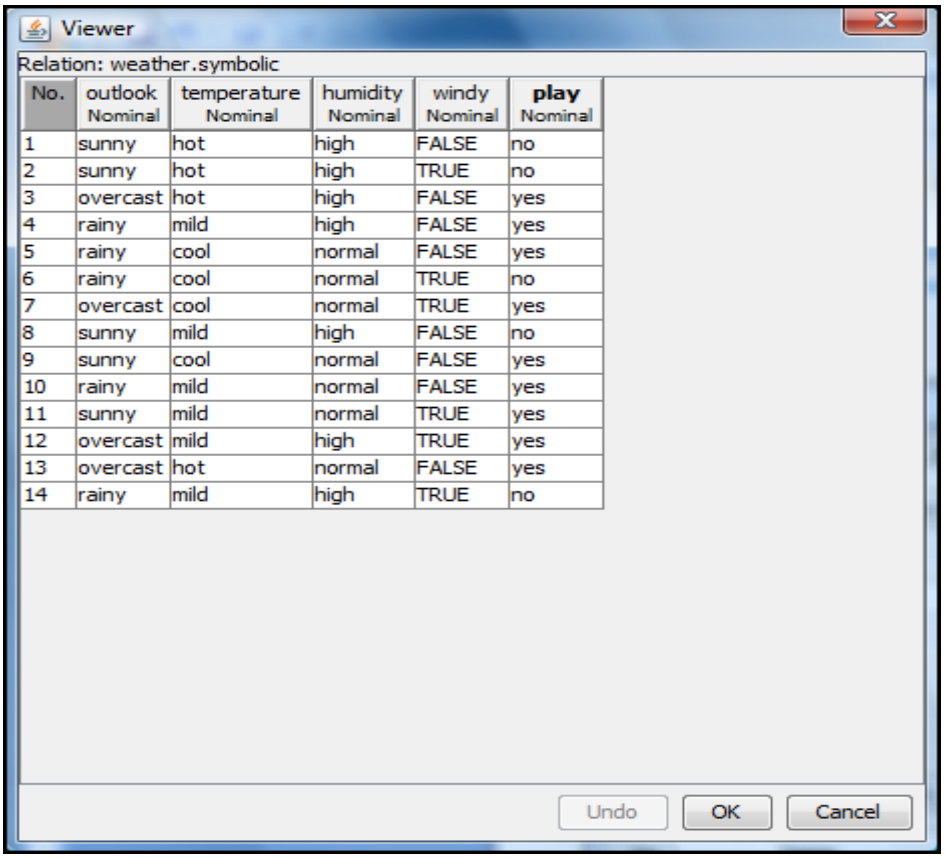
Grupo INSE15

Carga de datos

1. Abra el explorer y en la pestaña *Preprocess* teclee *Open file weather.nominal.arff*. Busque el directorio en c: Archivos de Programa/Weka/Data.
    a. Explore los atributos de cada registro de datos
        i. ¿Cuál son los valores posibles que el atributo temperatura puede tener?
        Hot, mild y cool
        ii. ¿y el de humidity?
        High y normal
    b. Abra un nuevo fichero de datos, *iris.arff.*
        i. ¿Cuántos ejemplos tiene esta base de datos? 150
        ii. ¿Cuántos atributos? Recuerde que la clase NO es un atributo. 5
        iii. ¿Cuál es el rango de posibles valores de *petallenght* ·? 1-6.9
    c. Abra un nuevo fichero de datos, *breast-cancer.arff.*
        i. ¿Cuántos ejemplos tiene esta base de datos? 286
        ii. ¿Cuántos atributos? 10

Editor de datos

2. Se puede ver y editar un conjunto completo de datos en Weka. Cargue el fichero weather.nominal.arff. pulse la pestaña Edit. Se abre la ventana Viewer en la que aparecen todos los ejemplos

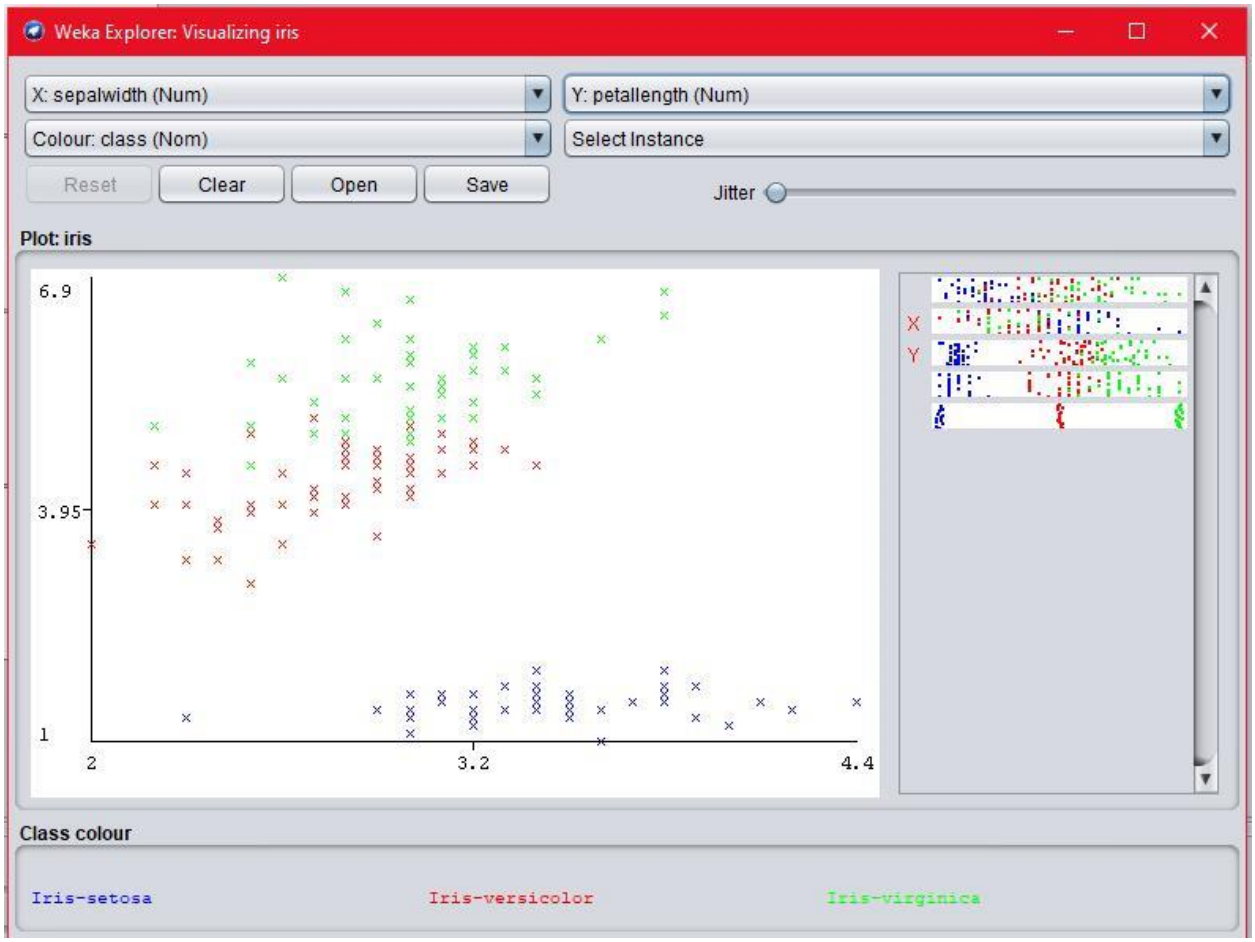


No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	play Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

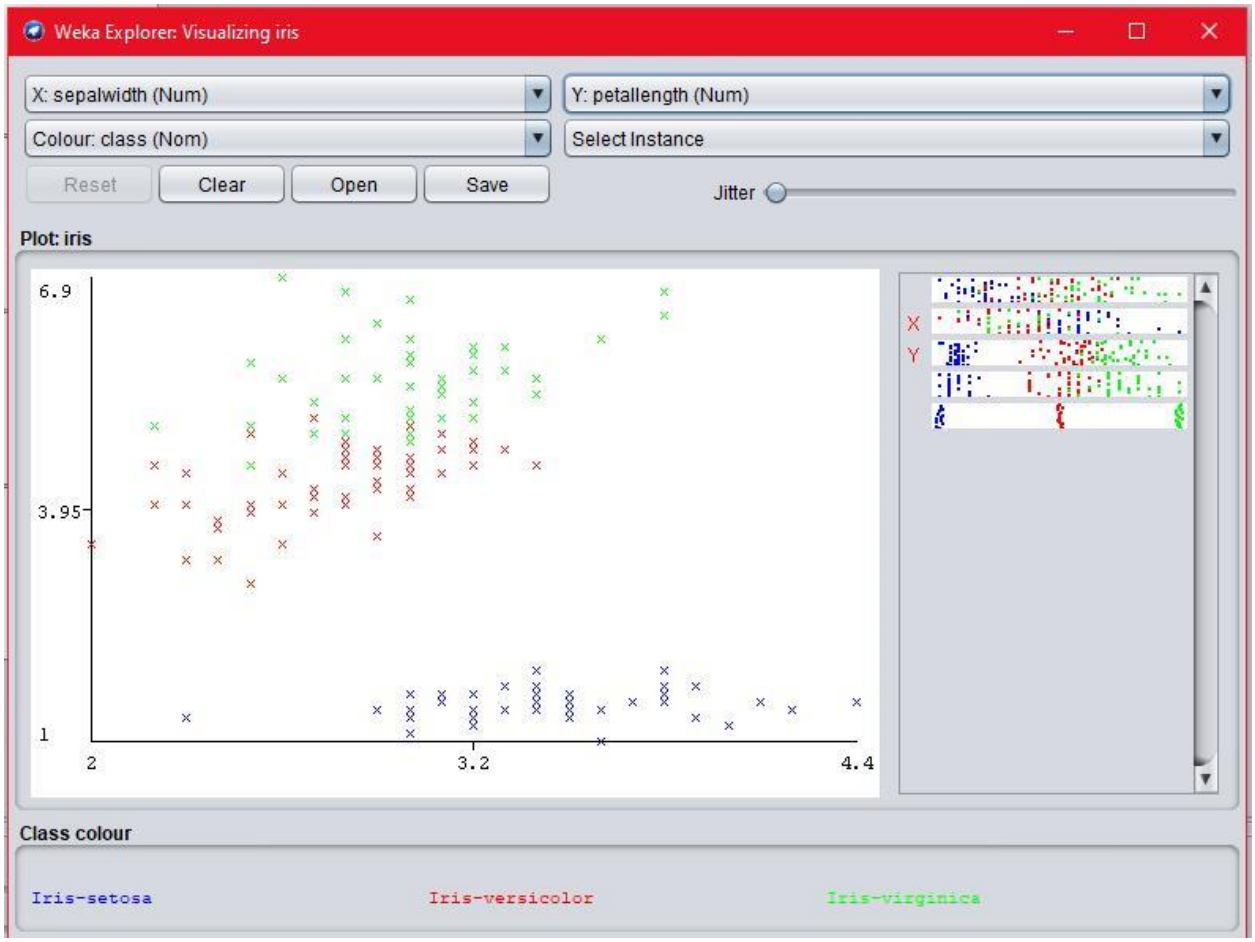
a. ¿Qué está contenido en la primera columna? Valores nominales del atributo Outlook.
b. ¿Cuál es el valor de la clase en el ejemplo 8?No
c. ¿Cuántos atributos nominales hay y cuántos numéricos? Cuatro nominales y ninguno numérico
d. ¿Cuántos ejemplos hay en esta serie de datos? 14

Visualización

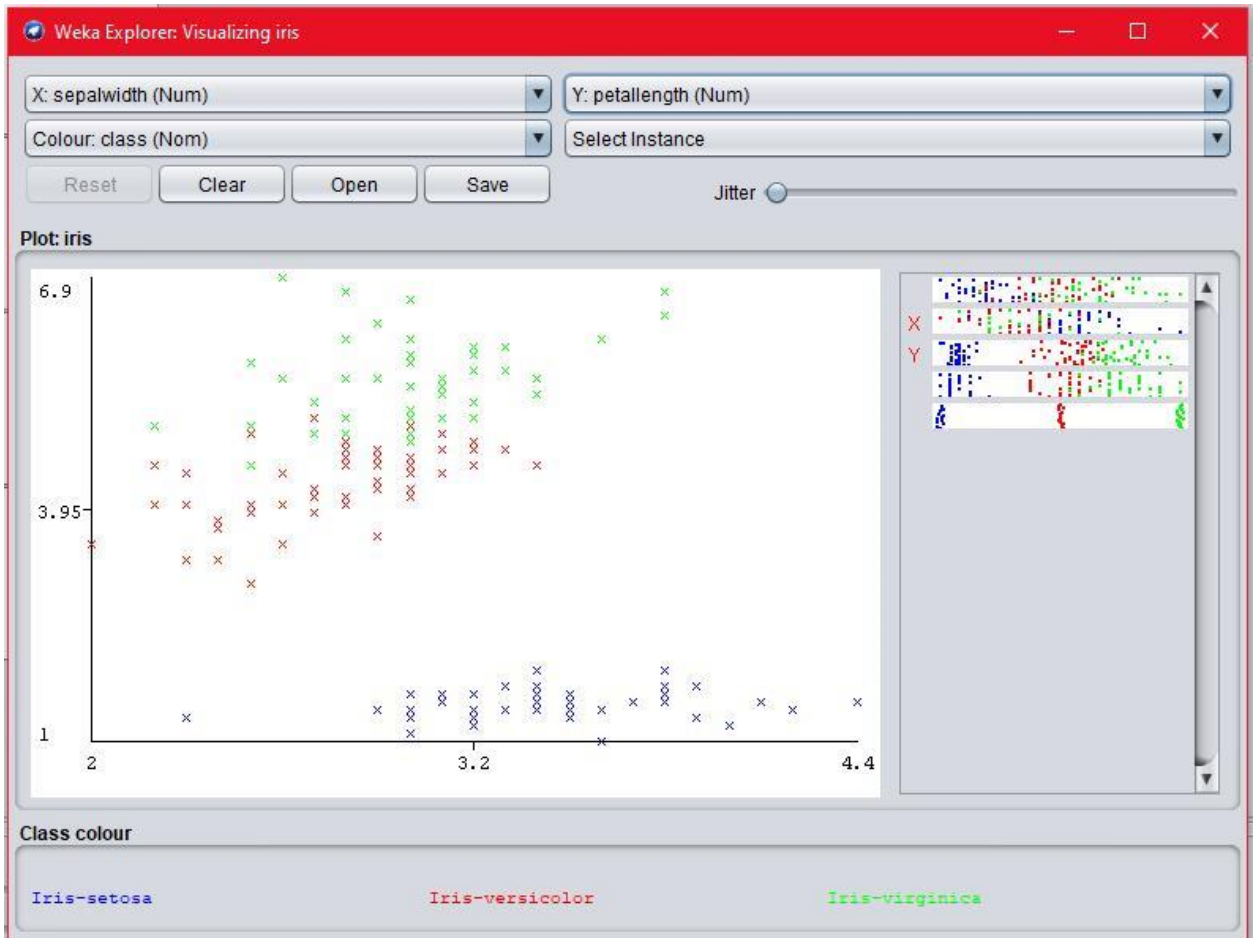
3. Cargue los datos *iris.arff*. Apriete la pestaña *Visualize*. Apriete en el primer recuadro de la segunda fila. Los ejemplos se muestran con pequeñas cruces con distinto color para cada clase. El eje x es *sepallength* y el eje y *petalwidth*. Si pulsamos el ratón en una de las cruces nos aparece el dato ligado a ese ejemplo. Si hacemos click en una de las barras que hay a la derecha, se cambia el eje X a ese atributo. Si clicamos en la ventana de arriba donde dice eje Y, podemos elegir el contenido del eje Y. Cambie el eje X a sepalwidth y el eje Y a petallenght. Incluya abajo una captura de pantalla de esa ventana. Use la opción recortar de Windows para capturar la ventana o la Opción ALT-ImprPant para copiar la ventana en el portapapeles.



a. La barra de *Jitter* sirve para cambiar aleatoriamente un poco los puntos mostrados para ver si en el mismo punto hay más de uno, puede haber más de uno dado que por la definición de la pantalla puede no caber todos. Experimente un poco con este potenciómetro. La pestaña *Select Instance* sirve para seleccionar un conjunto de datos eliminar los demás. Estos datos pueden guardarse para un estudio posterior. Seleccione la opción *Rectangle*, recuadre una parte de los datos. Haga una captura de pantalla en inclúyala.

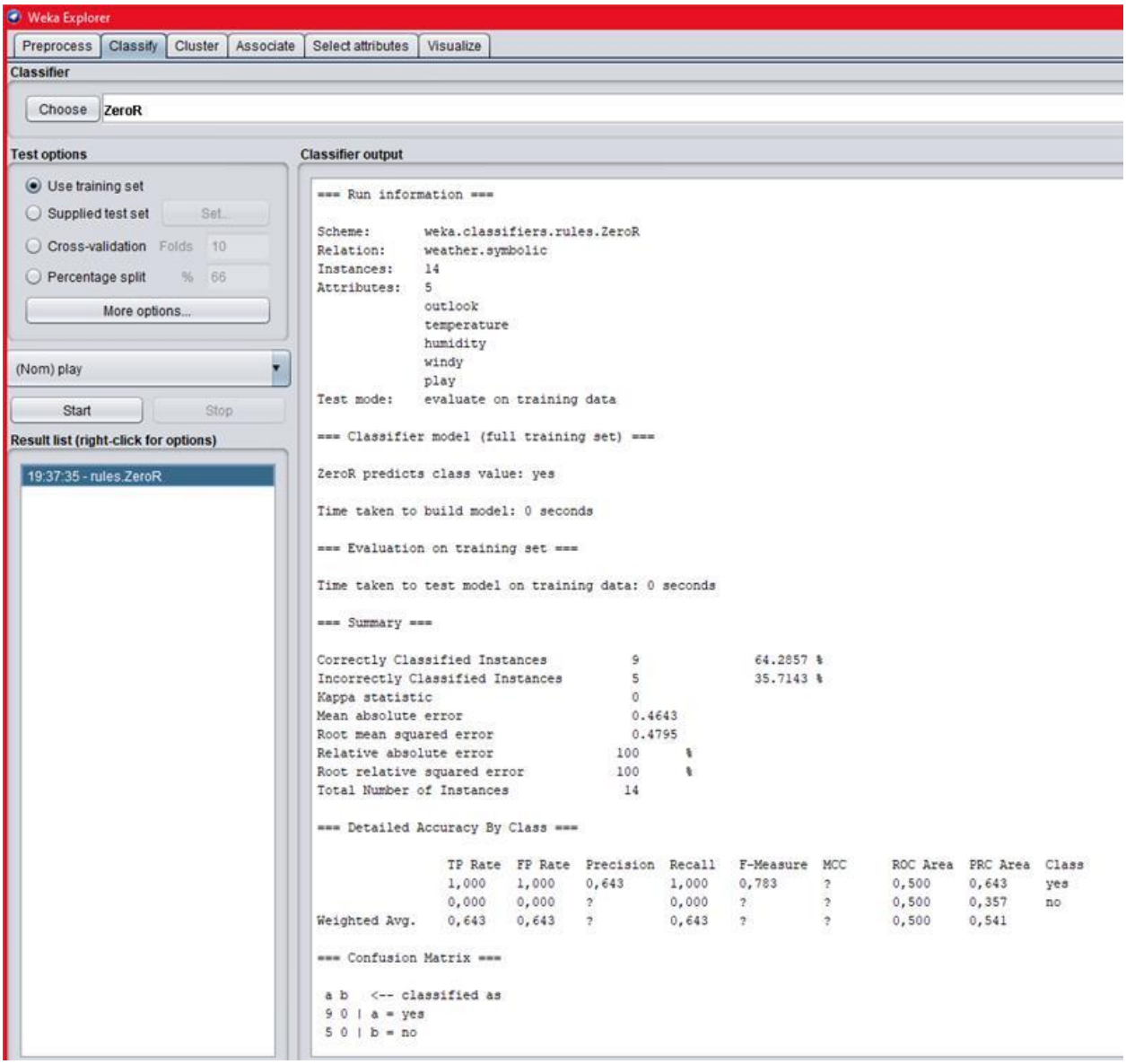


b. Haga *Submit* y verá que solo esos datos son los que se analizan ahora. Realice una nueva captura de pantalla e inclúyala. Vea que en la pestaña Save, puede guardar esta selección de datos para su análisis posterior



Clasificación

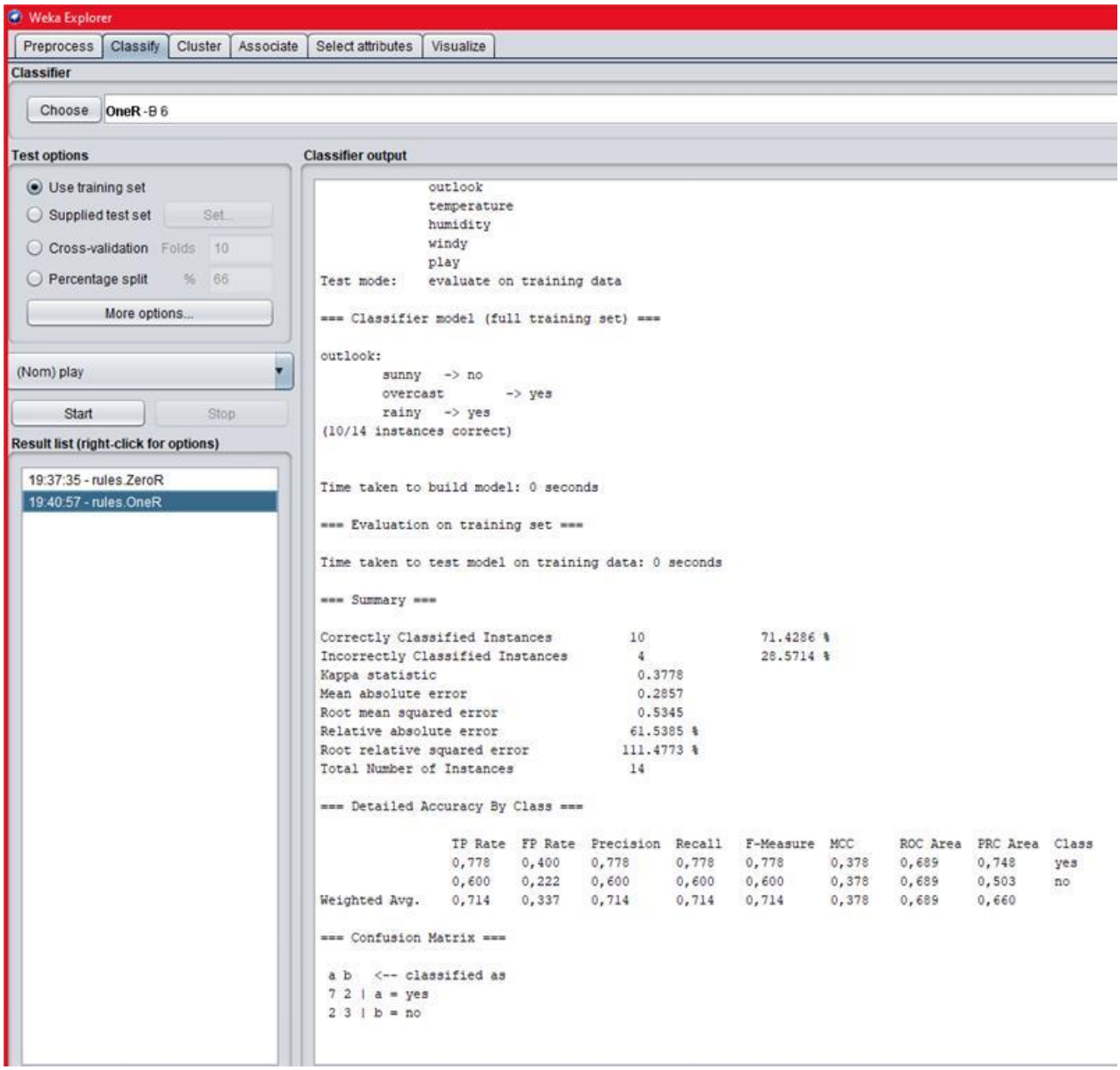
4. Aplicaremos ahora un clasificador a los datos *weather.nominal.arff.* Carguemos esos datos con la pestaña *Openfile*. Después pasaremos a la pestaña *Classify*. Use el clasificador 0R y en las opciones de test “Use training set”. Incluya una captura de Pantalla



a. ¿Cuál es la tasa de aciertos? 64.2857%

b. Fíjese ahora en la Matriz de confusión. ¿Cuántos se clasifican como a? ¿Cuántos como b? Diga dónde se producen los errores.
Se clasifican como a los 14 ejemplos y como b ninguno. Los errores residen en la clase no ya que 5 de esos 14 deberían de ser clasificados como b.

5. Con estos mismos datos, utilice el clasificador 1-R para el test “Use training set”. Incluya una captura de pantalla



a. ¿Cuál es la tasa de aciertos ahora? 71.4286%
b. Fíjese ahora en la Matriz de confusión. ¿Cuántos se clasifican como a? ¿Cuántos como b? Diga dónde se producen los errores. Explique según lo visto en clase por qué la tasa es más elevada.
   Se clasifican como a 9 y como b 5. Porque se tiene en cuenta tanto las clases como el atributo con menor número de errores en cambio en la regla zero no se aplican reglas a los atributos observados.

6. Cargue ahora los datos de weather numeric.
   a. Visualice todos los datos con la pestaña “edit” y haga una captura de pantalla

Viewer

Relation: weather

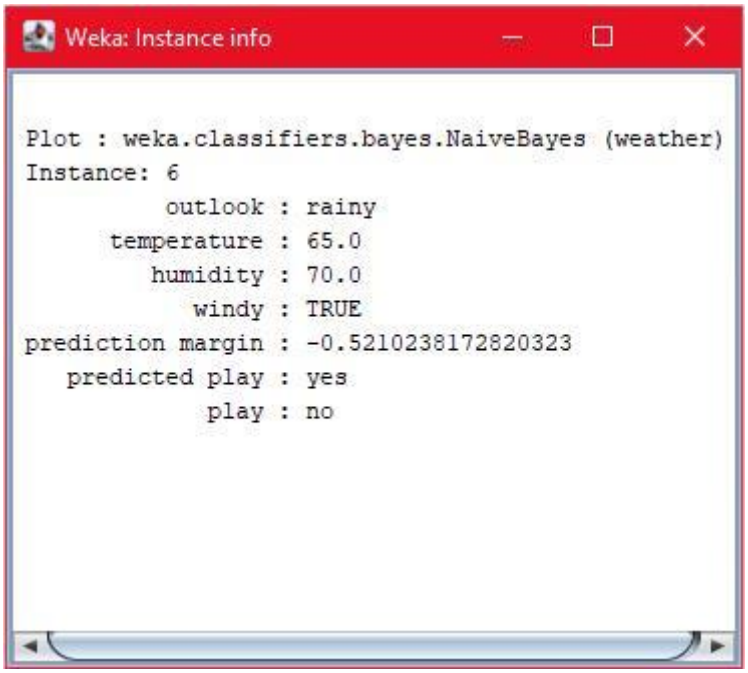
No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: **play** Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

b. Ponga los resultados del clasificador para 0R,1R y Naive Bayes
   - *0R – Tasa de acierto -> 64,2857%, Tasa de error -> 35.7143%*
   - *1R – Tasa de acierto -> 71,4286%, Tasa de error -> 28,5714%*
   - *Bayes – Tasa de acierto -> 92,8571%, Tasa de error -> 7.1429%*

c. Explique cuál es la justificación de que un método sea mejor que otro.
   En 0R no se aplican reglas a los atributos observados , en 1R se tiene en cuenta tanto las clases como el atributo con menor número de errores, y en Naive Bayes tiene en cuenta todos los atributos por lo que podemos tener a nuestra disposición mayor información.

d. Enumere los errores para el caso de Naïve Bayes.
   Observando la matriz de confusión vemos cómo el único error existente es un caso clasificado como a cuando debería de haber sido clasificado como b.

e. Pulse el ratón derecho sobre la ventana de resultados y visualice los errores. Capture una pantalla con el o los ejemplos concretos en los que se produce el error (haga click en el o los errores)



7. Fijémonos ahora en el clasificador Naive Bayes y vamos a usar la opción de “Test options” “percentage split”. El percentage Split coge un porcentaje de datos para aprender el modelo y otro porcentaje complementario para hacer el test.
    a. Haga pruebas con percentage Split de 66% y enumere el resultado usando en las opciones “Ramdom seed for XVal” de 1 y de 2.
    En ambos casos es del 80% de acierto.
    b. Haga pruebas con percentage Split de 80% y “Ramdom seed for XVal” de 2 y ponga el resultado
    Tasa de acierto -> 66,667%
    c. Razone qué es lo que está pasando
    El percentage Split produce una reducción de instancias comprobadas, por ejemplo en el caso con el percentage Split de 66% solo se tuvo en cuenta 5 instancias y en el percentage Split de 80% sólo 3 instancias.

P.S. Guarde el archivo como pdf y súbalo a la Entrega en Moodle