Actividad: Laboratorio Aprendizaje Supervisado con Weka

**Objetivos**

**Aplicar algoritmos de aprendizaje supervisado** sobre un conjunto de datos para generar un modelo clasificador; **interpretar el modelo** de clasificación generado; **interpretar los resultados de validación** a partir de diferentes métricas habitualmente utilizadas, y **evaluar** la actividad realizada por compañeros, en base a unos criterios establecidos, desarrollando la competencia crítica y reflexionando así sobre la propia ejecución de la actividad.

**Descripción de la actividad**

La actividad consiste en aplicar el algoritmo de aprendizaje supervisado ID3 sobre un conjunto de datos para generar modelos clasificadores:

Imagina que trabajas en una empresa dedicada al análisis de datos y un centro médico te contrata para ayudarle a prescribir lentes de contacto. Te proporciona para ello un conjunto de datos que constituye un histórico de pacientes, en el que se puede observar diferentes parámetros de los pacientes como, por ejemplo, la franja de edad o la prescripción de gafas. Además, dicho histórico informa del tipo de lentes de contacto que resultaron adecuadas para el paciente. Debes proporcionar al centro médico un modelo predictor que le permita predecir las lentes adecuadas para un futuro paciente en base a sus parámetros.

Por tanto, mediante el uso de Weka, generarás un modelo clasificador, específicamente un árbol de decisión, interpretarás el modelo generado y lo evaluarás.

**Preparación del laboratorio**

Previo a acudir al laboratorio debes repasar el tema 6. Para realizar el proceso de aprendizaje automático se utilizará la herramienta Weka. El *software* está disponible en la sección de descargas de la página oficial de Weka:

Accede al recurso a través del aula virtual o desde la siguiente dirección web:

<http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

A continuación, se describen los pasos para realizar el laboratorio:

1. Descárgate e instala Weka en tu ordenador si no lo tienes instalado.

2. Ejecuta el Explorer de Weka y abre el fichero contact-lenses-2021-22.arff que te ha proporcionado el profesor.

3. Haciendo uso del interfaz de la pestaña *preprocess* de Weka, haz una exploración de los atributos y sus valores. Puedes además obtener información sobre el dataset abriendo el fichero *.*arff con un editor de texto estándar, puesto que contiene comentarios sobre los datos.

**Pregunta 1**: describe el conjunto de datos incluyendo:

* Pregunta 1.1: posibles valores de la clase, indicando qué representan dichos valores.
* Pregunta 1.2: número de instancias en total.
* Pregunta 1.3: número de instancias pertenecientes a cada clase.
* Pregunta 1.4: número de atributos de entrada y valores que pueden tomar.

4. A continuación ejecuta el algoritmo ID3 con las opciones por defecto de Weka para la validación (desde la pestaña *Classify*). *Nota: Sigue las instrucciones del anexo 1 de este documento si no aparece el algoritmo ID3 en la lista de clasificadores a la que se puede acceder a través de la pestaña “Classify” del Explorer (carpeta “trees”).*

**Pregunta 2:** incluye en el informe la salida obtenida con Weka. Es muy importante porque tus respuestas se verificarán contrastando con la salida que has obtenido.

* Pregunta 2.1. ¿Qué atributo aparece en el nodo raíz del árbol?
* Pregunta 2.2. Demuestra por qué ID3 ha seleccionado este atributo para ser la raíz del árbol (nota: esta es una pregunta teórico-práctica que implica hacer cálculos manualmente).
* Pregunta 2.3. ¿Qué atributo o atributos se comprueban en el siguiente nivel del árbol?

**Pregunta 3:** explica a tu cliente cómo puede utilizar el modelo generado por el algoritmo ID3 para predecir las lentes de contacto adecuadas. Una forma sencilla de explicarlo es a través de ejemplos. Responde, por tanto, a las siguientes cuestiones razonando la repuesta para que el cliente entienda cómo utilizar el modelo:

**3.1.** Un paciente presenta los siguientes datos. ¿Cuál es el tipo de lentes de contacto que se le debe prescribir?

Age = young

spectacle-prescrip = myope

astigmatism = no

tear-prod-rate = normal

**3.2.** A un paciente que presenta los siguientes datos. ¿Cuál es el tipo de lentes de contacto que se le debe prescribir?

Age = presbyopic

spectacle-prescrip = myope

astigmatism = yes

tear-prod-rate = reduced.

**Pregunta 4:** de acuerdo con la validación realizada, ¿qué porcentaje de instancias son clasificadas correctamente? Explica de dónde has obtenido el dato.

**Pregunta 5:** de acuerdo con la matriz de confusión obtenida durante la etapa de validación del modelo, para cada clase, ¿cuántas de las instancias son clasificadas correctamente como de esa clase? ¿Cuántas son incorrectamente clasificadas? Razona la respuesta.

**Pregunta 6:** fíjate en la métrica TP Rate obtenida para cada una de las clases.

* Pregunta 6.1: define en una frase (no mediante una fórmula) la métrica TP Rate. Pregunta 6.2: razona sobre el rendimiento del modelo en base a dicha métrica, de manera que consideres el comportamiento del modelo respecto de las diferentes clases.
* Pregunta 6.3: demuestra cómo se obtiene el valor de dicho parámetro para la clase none a partir de la matriz de confusión.
* Pregunta 6.4: fíjate en el significado de cada clase y explica en qué clase o clases consideras que es más importante tener un valor alto de TP Rate.

**Pregunta 7.** Fíjate en la métrica FP Rate obtenida para cada una de las clases.

* Pregunta 7.1: define en una frase (no mediante una fórmula) la métrica FP Rate. Pregunta 7.2: razona sobre el rendimiento del modelo en base a dicha métrica, de manera que consideres el comportamiento del modelo respecto de las diferentes clases.
* Pregunta 7.3: demuestra cómo se obtiene el valor de dicho parámetro para la clase none a partir de la matriz de confusión.
* Pregunta 7.4: fíjate en el significado de cada clase y explica en qué clase o clases consideras que es más importante tener un valor bajo de FP Rate.

**Pregunta 8.** Concluye explicando cómo se ha realizado el proceso de validación (con las opciones por defecto que puedes observar en la casilla “Test options” de la pestaña Classifier).

Debes entregar un informe que incluya la salida obtenida al ejecutar el algoritmo ID3, así como las respuestas a las cuestiones planteadas. **No incluyas tu nombre en el informe** para asegurar el anonimato durante el proceso de evaluación.

**Una vez realizada la entrega debes evaluar las actividades realizadas por tus compañeros** que te sean asignadas **utilizando la rúbrica específica** descrita a continuación en este documento (el profesor te facilitará un fichero Excel que deberás completar y entregar) y explicando al compañero el motivo si no se otorga la máxima calificación. Previo a la fase de evaluación el profesor corregirá la actividad en clase.

**Extensión máxima:** el informe, sin contar con las salidas proporcionadas por Weka, tendrá una extensión máxima de 5 páginas (Georgia 11 e interlineado 1,5).

**Rúbrica**

|  |  |  |  |
| --- | --- | --- | --- |
| Laboratorio Aprendizaje Supervisado con Weka | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Calificación de la actividad por el profesor | Calificación obtenida por la evaluación de la actividad por parte del profesor siguiendo la rúbrica específica. | 3,5 | 35% |
| Calificación de la actividad por los compañeros (\*) | Media de las calificaciones obtenidas de las evaluaciones de la actividad por los compañeros siguiendo la rúbrica específica. Es importante que las explicaciones sean sencillas y claras; pensad que vuestros compañeros van a valorar cómo de bien comprenden vuestras explicaciones. | 3,5 | 35% |
| Calificación por las evaluaciones de las actividades de otros compañeros (\*) | Una vez entregada la actividad se deberá evaluar las actividades asignadas de otros compañeros y se recibirá una calificación por estas evaluaciones, para las cuales se debe utilizar la rúbrica específica, así como ser justos y objetivos. | 3 | 30% |
|  |  | **10** | **100 %** |

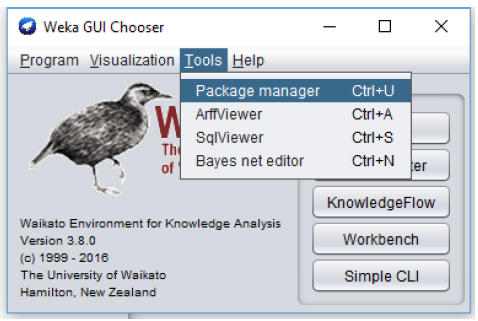
**Rúbrica específica**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Aspectos que evaluar** | **Niveles de desempeño** | | | | | **Calificación** |
| **5** | **4** | **3** | **2** | **1** |
| Pregunta 1. Descripción del conjunto de datos (0,75 puntos máximo) | Las respuestas a las 4 preguntas son correctas y completas. (0,75 puntos) | Las respuestas a 3 preguntas son correctas y completas. (0,50 puntos) | Las respuestas a 2 preguntas son correctas y completas. (0,30 puntos) | La respuesta a 1 pregunta es correcta y completa. (0,15 puntos) | Ninguna respuesta está correcta y completa. (0 puntos) |  |
| Pregunta 2. Incluye en el informe la salida obtenida con Weka (0,75 puntos máximo) (si no se incluye, se evalúan las respuestas contrastando con la salida proporcionada por el profesor) | Sí incluye en el informe la salida, es completa y es legible. (0,75 puntos) | Sí incluye en el informe la salida y es completa, pero no se puede leer claramente (por ejemplo, se ha realizado una captura de una imagen, y el texto, aunque se consigue leer, no es fácilmente legible). (0,30 puntos) | | No incluye en el informe la salida o sí la incluye, pero no es completa o no es legible. (0 puntos) | |  |
| Preguntas 2.1 y 2.3. Interpretación correcta del modelo (1 punto máximo) | Las respuestas a las preguntas 2.1 y 2.3 son correctas. (1 punto) | La respuesta a una de las preguntas 2.1 o 2.3 es correcta. (0,3 puntos) | | Ninguna de las respuestas a las preguntas 2.1 y 2.3 es correcta. (0 puntos) | |  |
| Pregunta 2.2. Demostración de selección de atributo raíz (1,5 puntos máximo) | Se demuestra correctamente, completamente y con claridad por qué ID3 ha seleccionado el atributo raíz. (1,5 puntos) | Se demuestra completamente y con claridad por qué ID3 ha seleccionado el atributo raíz, pero existe un error en los cálculos, siendo las expresiones correctas. (1 punto) | Se demuestra completamente y con claridad por qué ID3 ha seleccionado el atributo raíz, pero hay más de 1 error en los cálculos, siendo las expresiones correctas. (0,6 puntos) | Los cálculos numéricos son correctos, las expresiones matemáticas incluidas son correctas, pero el desarrollo no es completo, por lo que no se demuestra completamente y con claridad por qué ID3 ha seleccionado el atributo raíz (0,3 puntos). | Las expresiones son incorrectas y cualquier otro caso no contemplado en los elementos previos (0 puntos). |  |
| La claridad y completitud implica que las expresiones matemáticas sean claras y estén completamente desarrolladas | | | | |
| Pregunta 3. Explicación de cómo usar el modelo (1 punto máximo) | Las respuestas a las preguntas 3.1 y 3.2 son correctas. También, se incluye una explicación clara (comprensible por cualquier usuario no técnico), con los ejemplos de las preguntas 3.1 y 3.2, de cómo se puede utilizar el modelo para que el cliente pueda utilizarlo (1 punto). | | Las respuestas a las preguntas 3.1 y 3.2 son correctas, pero no se explica claramente cómo obtener dichas respuestas. La explicación no es suficiente para un usuario que desconozca los modelos de árboles de decisión o su representación en Weka. | La respuesta a la pregunta 1 o a la pregunta 2 es incorrecta. (0 puntos) | |  |
| Pregunta 4. % de instancias clasificadas correctamente (0,5 puntos máximo) | La repuesta es correcta y se explica de dónde se ha obtenido el dato. (0,5 puntos) | | La respuesta es correcta, pero no se explica de dónde se ha obtenido el dato. (0,2 puntos) | La respuesta es incorrecta. (0 puntos) | |  |
| Pregunta 5. Interpretación de matriz de confusión (0,5 puntos máximo) | Indica para cada clase correctamente el número de instancias correcta e incorrectamente clasificadas. El razonamiento es claro y correcto. (0,5 puntos) | Indica para cada clase correctamente el número de instancias correcta e incorrectamente clasificadas. El razonamiento no es correcto o no tiene la suficiente claridad para confirmar que es correcto (0,2 puntos). | | No es correcto el número de instancias correcta o incorrectamente clasificadas en una o más clases. (0 puntos) | |  |
|  |
| Preguntas 6.1 y 6.3. Definición TP Rate (0,5 puntos máximo) | La definición de TP Rate en una frase es correcta y la demostración es correcta y está completamente desarrollada. La definición debe darse en una frase, no mediante una fórmula. (0,5 puntos) | | La definición de TP Rate en una frase es correcta pero la demostración es incorrecta o a la inversa. Se considera incorrecta la demostración que no tiene la claridad suficiente para verificarla. La definición debe darse en una frase, no mediante una fórmula. (0,25 puntos). | Ni la definición es correcta ni la demostración es correcta. (0 puntos) | |  |  |
| Pregunta 6.2. Rendimiento según TP Rate (0,75 puntos máximo) | Razonamiento correcto y claro que considera el comportamiento del modelo respecto a esa métrica para todas las clases. (0,75 puntos) | | Razonamiento correcto y claro que considera el comportamiento del modelo respecto a esa métrica, pero no para todas las clases (0,4 puntos) | Razonamiento incorrecto o que no es lo suficientemente claro para verificar su corrección. (0 puntos) | |  |  |
| Pregunta 6.4. Relevancia del TP Rate para cada clase (0,5 puntos máximo) | Razonamiento correcto y claro que tiene en cuenta todas las clases. (0,5 puntos) | | Razonamiento correcto y claro, pero que no contempla todas las clases.(0,2 puntos) | Razonamiento incorrecto o que no es lo suficientemente claro para verificar su corrección. (0 puntos). | |  |  |
| Preguntas 7.1 y 7.3. Definición FP Rate (0,5 puntos máximo) | La definición de FP Rate en una frase es correcta y la demostración es correcta y está completamente desarrollada. La definición debe darse en una frase, no mediante una fórmula. (0,5 puntos) | | La definición de FP Rate en una frase es correcta, pero la demostración es incorrecta o a la inversa. Se considera incorrecta la demostración que no tiene la claridad suficiente para verificarla. La definición debe darse en una frase, no mediante una fórmula. (0,25 puntos). | Ni la definición es correcta ni la demostración es correcta. (0 puntos) | |  |  |
| Pregunta 7.2 Rendimiento según FP Rate (0,75 puntos máximo) | Razonamiento correcto y claro que considera el comportamiento del modelo respecto a esa métrica para todas las clases. (0,75 puntos) | | Razonamiento correcto y claro, que considera el comportamiento del modelo respecto a esa métrica, pero no para todas las clases (0,4 puntos) | Razonamiento incorrecto o que no es lo suficientemente claro para verificar la corrección. (0 puntos) | |  |  |
| Pregunta 7.4. Relevancia del FP Rate para cada clase (0,5 puntos máximo) | Razonamiento correcto y claro que tiene en cuenta todas las clases. (0,5 puntos) | | Razonamiento correcto y claro pero que no contempla todas las clases. (0,25 puntos). | Razonamiento incorrecto o que no es lo suficientemente claro para verificar la corrección. (0 puntos) | |  |  |
| Pregunta 8. Proceso de validación (0,5 puntos máximo) | La explicación es correcta y clara. (0,5 puntos) | | La explicación no es correcta o no es lo suficientemente clara para verificar su corrección. (0 puntos) | | |  |  |
|  |  |  |  |  | **TOTAL:** | 0 |  |
|  |  |  |  |  |  |

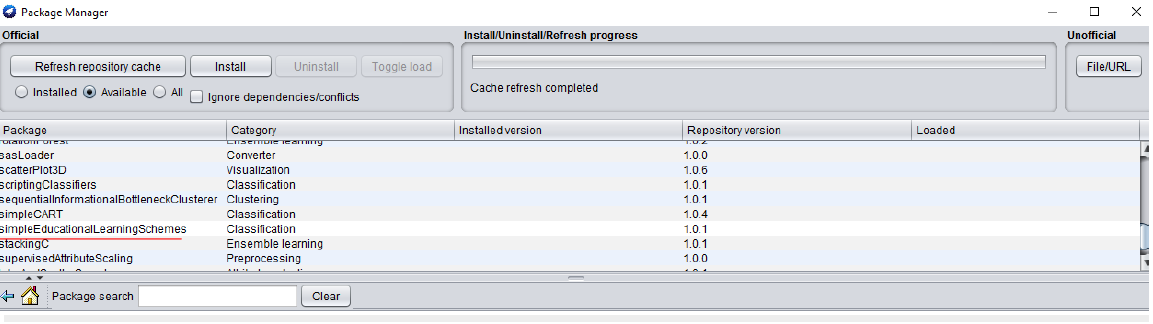
Anexo 1

En la última versión de Weka (Weka 3.8) por defecto no vienen incorporados algoritmos tales como Prism o ID3. Se pueden incorporar fácilmente tal y como se explica a continuación.

Primeramente, debes abrir el Package Manager disponible a través del menú “Tools->Package Manager”.



A continuación, sólo hay que buscar el paquete simpleEducationalLearningSchemes en la lista que aparece e instalarlo.



Reiniciando Weka ya están disponibles algoritmos tales como ID3.