5/23/2022

Anonimo

**LABORATORIO APRENDIZAJE SUPERVISADO CON WEKA**

Inteligencia Artificial e Ingeniería del Conocimiento

**CONTENIDO**

[INTRODUCCIÓN 3](#_Toc104212252)

[OBJETIVO 3](#_Toc104212253)

[DESCRIPCIÓN DE LA ACTIVIDAD 3](#_Toc104212254)

[PREPARACIÓN DEL LABORATORIO 3](#_Toc104212255)

[PREGUNTA 1 4](#_Toc104212256)

[Pregunta 1.1 4](#_Toc104212257)

[Pregunta 1.2 5](#_Toc104212258)

[Pregunta 1.3 5](#_Toc104212259)

[Pregunta 1.4 5](#_Toc104212260)

[PREGUNTA 2 6](#_Toc104212261)

[Pregunta 2.1 6](#_Toc104212262)

[Pregunta 2.2 6](#_Toc104212263)

[Pregunta 2.3 8](#_Toc104212264)

[PREGUNTA 3 8](#_Toc104212265)

[SOFT: Lentes de contacto, suave. 9](#_Toc104212266)

[NONE: Lentes de contacto, ninguno 10](#_Toc104212267)

[PREGUNTA 4 10](#_Toc104212268)

[PREGUNTA 5 15](#_Toc104212269)

[PREGUNTA 6 15](#_Toc104212270)

[Pregunta 6.1 16](#_Toc104212271)

[Pregunta 6.2 16](#_Toc104212272)

[Pregunta 6.3 16](#_Toc104212273)

[Pregunta 6.4 17](#_Toc104212274)

[PREGUNTA 7 17](#_Toc104212275)

[Pregunta 7.1 17](#_Toc104212276)

[Pregunta 7.2 18](#_Toc104212277)

[Pregunta 7.3 18](#_Toc104212278)

[Pregunta 7.4 18](#_Toc104212279)

[PREGUNTA 8 19](#_Toc104212280)

[ANEXOS 20](#_Toc104212281)

[Anexo 1 20](#_Toc104212282)

# INTRODUCCIÓN

El presente documento contiene la presentación de la segunda actividad de la materia de Inteligencia Artificial e Ingeniería del Conocimiento (GII) - PER3288 2021-2022 de la Universidad Internacional de La Rioja (UNIR), impartida por Elena Verdú Pérez.

# OBJETIVO

**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.

|  |
| --- |
|  |

1. 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 que representan dichos valores.

|  |  |  |
| --- | --- | --- |
| **CLASES** | **VALORES** | **SIGNIFICADO** |
| AGE | young | Personas menores a 35 años. |
| pre-presbyopic | Personas de 35 a 45 años. |
| presbyopic | Personas mayores a 45 años. |
| SPECTACLE-PRESCRIP | myope | Gafas de corrección óptica para miopía. |
| hypermetrope | Gafas de corrección óptica para hipermetropía |
| ASTIGMATISM | yes | Tiene Astigmatismo. |
| no | No tiene Astigmatismo. |
| TEAR-PROD-RATE | reduced | Producción de lágrima es reducida. |
| normal | Producción de lágrima es normal. |
| CONTACT-LENSES | soft | Lentes de contacto, suave. |
| hard | Lentes de contacto, duro. |
| none | Lentes de contacto, ninguno |

## Pregunta 1.2

Número de instancias en total.

|  |  |
| --- | --- |
| **CLASES** | **NUMERO DE INSTANCIAS TOTAL** |
| AGE | 38 |
| SPECTACLE-PRESCRIP | 38 |
| ASTIGMATISM | 38 |
| TEAR-PROD-RATE | 38 |
| CONTACT-LENSES | 38 |

## Pregunta 1.3

Número de instancias pertenecientes a cada clase.

|  |  |  |
| --- | --- | --- |
| **CLASES** | **VALORES** | **NUMERO DE INSTANCIAS** |
| AGE | young | 15 |
| pre-presbyopic | 11 |
| presbyopic | 12 |
| SPECTACLE-PRESCRIP | myope | 18 |
| hypermetrope | 20 |
| ASTIGMATISM | yes | 19 |
| no | 19 |
| TEAR-PROD-RATE | reduced | 18 |
| normal | 20 |
| CONTACT-LENSES | soft | 7 |
| hard | 7 |
| none | 24 |

## Pregunta 1.4

Número de atributos de entrada y valores que pueden tomar.

|  |  |  |
| --- | --- | --- |
| **CLASES** | **VALORES** | **NUMERO DE ENTRADA** |
| AGE | young | 3 |
| pre-presbyopic |
| presbyopic |
| SPECTACLE-PRESCRIP | myope | 2 |
| hypermetrope |
| ASTIGMATISM | yes | 2 |
| no |
| TEAR-PROD-RATE | reduced | 2 |
| normal |
| CONTACT-LENSES | soft | 3 |
| hard |
| none |

1. 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?

ASTIGMATISM

## 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).

|  |  |
| --- | --- |
|  |  |

**AGE**

**SPECTACLE-PRESCRIP**

1.28

**ASTIGMATISM**

0.95

**TEAR-PROD-RATE**

Seleccionamos el valor menor a todos, el mismo sera la raiz del arbol.

## Pregunta 2.3

¿Qué atributo o atributos se comprueban en el siguiente nivel del árbol?

TEAR-PROD-RATE

# 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:

* 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

Diagram

Description automatically generated

# SOFT: Lentes de contacto, suave.

* 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.

Diagram

Description automatically generated

# NONE: Lentes de contacto, ninguno

# PREGUNTA 4

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

Para responder esta pregunta, arme un Excel, el mismo que adjunto en este literal. En él se encuentra todos los datos generados al cargar el fichero ***contact-lenses-2021-22.arff***, en el e puede filtrar según el camino que tome el árbol en los dos diagramas de la pregunta anterior.

|  |
| --- |
|  |

|  |
| --- |
| 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 |
| Diagram  Description automatically generated |
| **Todos los Datos** |
| Table  Description automatically generated |
| **Filtramos por ASTIGMATISM *-> no*** |
| Table  Description automatically generated |
| **Filtramos por *TEAR-PROD-RATE -> normal*** |
| Table  Description automatically generated |
| **Filtramos por *TEAR-PROD-RATE -> normal*** |
| Graphical user interface, table  Description automatically generated |
| Como se puede visualizar se tiene 4 Intancias con el valor de SOFT para el tipo de lentes de contacto a entregar. |

|  |
| --- |
| 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. |
| Diagram  Description automatically generated |
| Todos los Datos |
| Table  Description automatically generated |
| **Filtramos por ASTIGMATISM *-> yes*** |
| Table  Description automatically generated |
| **Filtramos por *TEAR-PROD-RATE -> reduced*** |
| Table  Description automatically generated |
| **Filtramos por *AGE -> presbyopic*** |
|  |
| **Filtramos por *SPECTACLE-PRESCRIP -> myope*** |
|  |
| Como se puede visualizar se tiene 1 Intancia con el valor de NONE para el tipo de lentes de contacto a entregar. |

# 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.

El número de instancias correctas es la suma de los valores en la diagonal de la matriz, mientras que la suma de los demás números de la matriz sin incluir la diagonal son las instancias incorrectas.

|  |  |  |
| --- | --- | --- |
| **CLASE** | **MATRIZ** | **INSTANCIAS CORRECTAS/INCORRECTAS** |
| AGE | Text  Description automatically generated | Instancias Correctas: 4 + 0 + 3 = 7  Instancias Incorrectas: 9 + 2 + 9 + 2 + 5 + 4 = 31 |
| SPECTACLE-PRESCRIP | Text  Description automatically generated | Instancias Correctas: 5 + 5 = 10  Instancias Incorrectas: 13 + 15 = 28 |
| ASTIGMATISM | Text  Description automatically generated with low confidence | Instancias Correctas: 13 + 9 = 22  Instancias Incorrectas: 6 + 9 = 15 |
| TEAR-PROD-RATE | Text  Description automatically generated | Instancias Correctas: 13 + 14 = 27  Instancias Incorrectas: 5 + 6 = 11 |
| CONTACT-LENSES | A picture containing text  Description automatically generated | Instancias Correctas: 6 + 5 + 21 = 32  Instancias Incorrectas: 1 + 2 + 1 + 2 = 6 |

# PREGUNTA 6

Fíjate en la métrica TP Rate obtenida para cada una de las clases.

|  |  |  |
| --- | --- | --- |
| **CLASE** | **MATRIZ** | **TP RATE** |
| AGE | Text  Description automatically generated | A picture containing table  Description automatically generated |
| SPECTACLE-PRESCRIP | Text  Description automatically generated | Table  Description automatically generated |
| ASTIGMATISM | Text  Description automatically generated with low confidence | Table  Description automatically generated |
| TEAR-PROD-RATE | Text  Description automatically generated | Table  Description automatically generated |
| CONTACT-LENSES | A picture containing text  Description automatically generated | Table  Description automatically generated |

## Pregunta 6.1

Define en una frase (no mediante una fórmula) la métrica TP Rate:

TP Rate o Tasa de Éxito es la suma de los valores de la diagonal dividido por el total de las instancias.

## 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:

|  |  |
| --- | --- |
| A picture containing text  Description automatically generated |  |

## 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:

En la clase CONTACT-LENSES, ya que es el resultado al cual estamos apuntando llegar.

# PREGUNTA 7

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

|  |  |  |
| --- | --- | --- |
| **CLASE** | **MATRIZ** | **TP RATE** |
| AGE | Text  Description automatically generated | Table  Description automatically generated |
| SPECTACLE-PRESCRIP | Text  Description automatically generated | Text, table  Description automatically generated with medium confidence |
| ASTIGMATISM | Text  Description automatically generated with low confidence | Table  Description automatically generated with medium confidence |
| TEAR-PROD-RATE | Text  Description automatically generated | Table  Description automatically generated |
| CONTACT-LENSES | A picture containing text  Description automatically generated | Table  Description automatically generated |

## Pregunta 7.1

Define en una frase (no mediante una fórmula) la métrica FP Rate.

La tasa de de FP Rate o la tasa de falsos positivos es una instancia incorrectamente clasificada como afirmativa o positiva cuando es negativa

## 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.

|  |  |
| --- | --- |
| A picture containing text  Description automatically generated |  |

## 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.

No entiedo esta parte!

# 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).

Se presentan el número de clases de pruebas tal como se muestra en la imagen

Graphical user interface, text, application, chat or text message

Description automatically generated

Al seleccionar cualquiera de estas clases, se presentara una análisis de salida de la clase seleccionada, en la misma se encuentra la especificación del orden del árbol con un resumen como el número de instancias correctas y el número de instancias incorrectas con los porcentajes de estadística y error de este análisis del total de casos.

También se presenta la exactitud detallada de la clase junto con la matriz de confusión, tal como se muestra en la imagen adjunta.

A picture containing table

Description automatically generated

# ANEXOS

## 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****.*

Graphical user interface, text

Description automatically generated

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

Graphical user interface, text, application, email

Description automatically generated

Reiniciando **Weka** ya están disponibles algoritmos tales como ID3.