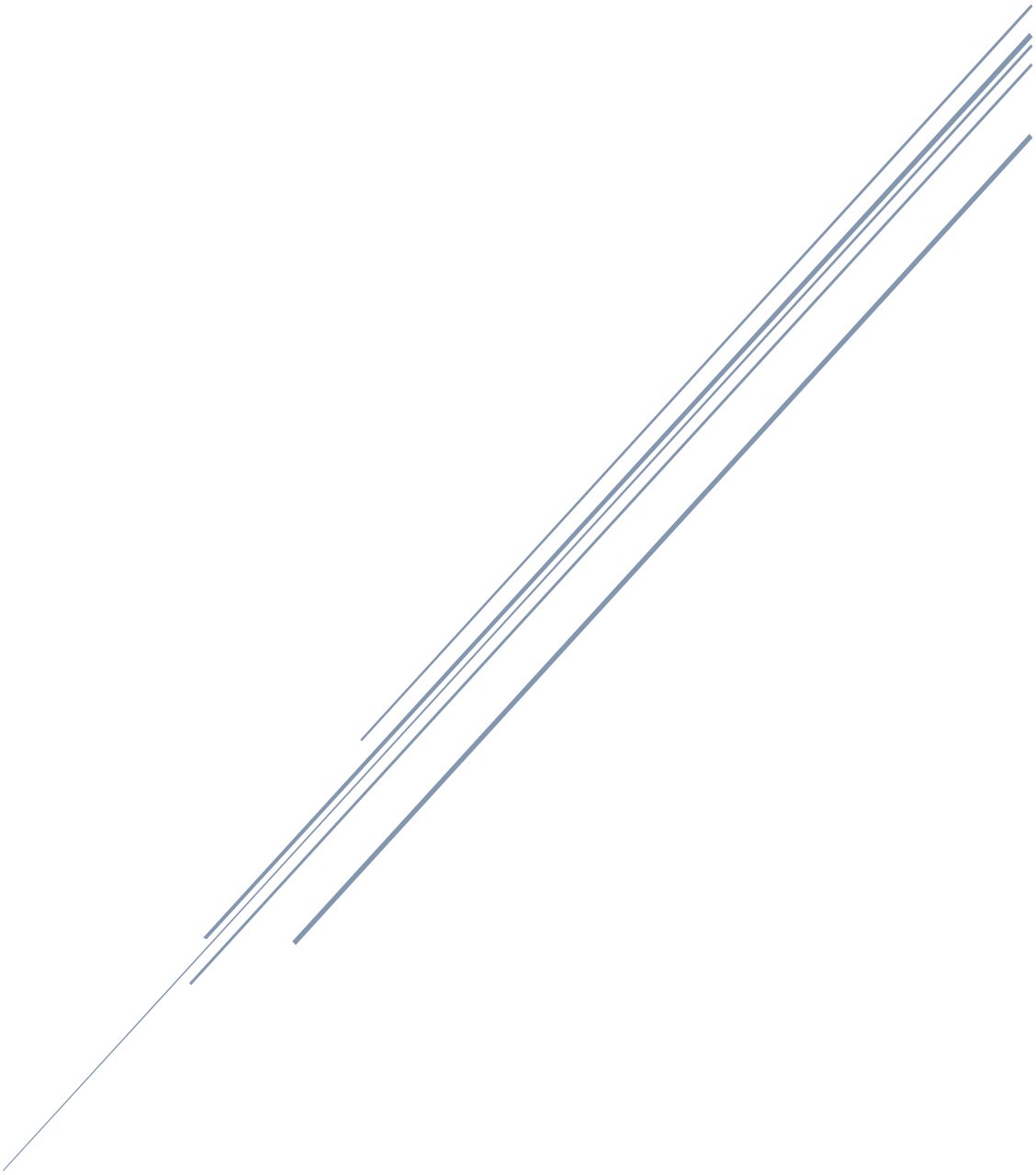
MEMORIA PRÁCTICA 5

TECNICAS DE APRENDIZAJE AUTOMÁTICO



CURSO 2022/23

JAVIER ABAD HERNÁNDEZ

DESCRIPCIÓN DE DATOS:

Contact-lenses: Conjunto de datos dedicado a un problema de clasificación relacionado con la predicción de si las personas necesitarán lentes de contacto blandas, lentes de contacto duras o no necesitarán lentes en función de su edad y los diferentes problemas de visión con los que cuente.

→ https://archive.ics.uci.edu/ml/datasets/Lenses
→ 24 instancias
→ 4 atributos (1+ clase)
→ 3 clases

Iris: conjunto de datos de clasificación que contiene información sobre tres especies de iris (setosa, virginica y versicolor) con 50 muestras cada una y algunas propiedades sobre cada flor.

→ https://archive.ics.uci.edu/ml/datasets/iris
→ 150 instancias (50 de cada clase)
→ 4 atributos (numéricos)
→ 3 clases

Soybean: El conjunto de datos Soybean contiene información sobre la presencia o ausencia de varias enfermedades en plantas de soja.

→ https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)
→ 683 instancias
→ 35 atributos (todos nominales)
→ 19 clases

Vote: Por otro lado, el conjunto de datos Vote contiene información sobre los registros de votación de los miembros del Congreso de EE. UU. sobre ciertas cuestiones

→ https://archive.ics.uci.edu/ml/datasets/congressional+voting+records
→ 435 instancias (50 de cada clase)
→ 16 atributos (todos nominales)
→ 1 clase (democrat o republican)

Thoracic_surgery: Conjunto de datos dedicado a un problema de clasificación relacionado con la esperanza de vida postoperatoria en pacientes con cáncer de pulmón.

→ https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data
→ 470 instancias
→ 17 atributos
→ 1 clase

Adult: El conjunto de datos Adult es un conjunto de datos de censo que contiene información sobre personas y su ingreso anual. El objetivo del conjunto de datos es predecir si el ingreso anual de una persona supera los $50K al año basándose en los datos del censo.

→ https://archive.ics.uci.edu/ml/datasets/Adult
→ 48842 instancias
→ 14 atributos (+1 clase)
→ 2 clases

EJERCICIO 1:

Conjunto de datos: contact-lenses.arff

	Algoritmos (Validación cruzada 10 particiones)				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa Error	0,16667	0,291667	0,291667	0,25	0,16667

Vamos a realizar la comparación respecto a los apuntes:

- OneR: Coincide totalmente.
- Prism: Coinciden totalmente.
- PART: No coinciden en principio porque no se han tomado los mismos atributos sino únicamente 3 en los apuntes, mientras que en Weka he utilizado todos.

EJERCICIO 2:

Conjunto de datos: iris.arff

	Algoritmos (Validación cruzada 10 particiones)				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa Error	0,04	0,08	-	0,046667	0,06

No se puede utilizar Prism porque son atributos de tipo real, como se puede ver en el link de donde he obtenido el dataset (https://archive.ics.uci.edu/ml/datasets/iris)

Conjunto de datos: soybean.arff

	Algoritmos (Validación cruzada 10 particiones)				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa Error	0,084919	0,600293	-	0,077599	0,080527

No se puede utilizar Prism porque presenta valores desconocidos, como se puede ver en el dataset (https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/soybean.arff)

Conjunto de datos: vote.arff

	Algoritmos (Validación cruzada 10 particiones)				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa Error	0,036782	0,043678	-	0,045977	0,052874

No se puede utilizar Prism porque son atributos de tipo categórico y además presenta valores desconocidos, como se puede ver en el link de donde he obtenido el dataset (https://archive.ics.uci.edu/ml/datasets/congressional+voting+records)

Conjunto de datos: thoracic_surgery.arff

	Algoritmos (Validación cruzada 10 particiones)				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa Error	0,155319	0,165957	-	0,153191	0,208511

No se puede utilizar Prism porque son atributos de tipo integer y real, como se puede ver en el link de donde he obtenido el dataset (https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data)

Conjunto de datos: adult.arff

	Algoritmos (Validación cruzada 10 particiones)				
	J48	**OneR**	**Prism**	**JRIP**	**PART**
Tasa Error	0,13768	0,190934	-	0,154602	0,146586

No se puede utilizar Prism porque son atributos de tipo integer y además presenta valores desconocidos, como se puede ver en el link de donde he obtenido el dataset (https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult)

Prism:

Para entender porque el algoritmo Prism no permite interacción, podemos ver en la página de WEKA que lo trata con atributos nominales y no puede tratar con valores desconocidos.
(https://weka.sourceforge.io/doc.stable/weka/classifiers/rules/Prism.html)

```
public class Prism
extends Classifier
implements TechnicalInformationHandler

Class for building and using a PRISM rule set for classification. Can only deal with nominal attributes. Can't deal with missing values. Doesn't do any pruning.
```

Además, Prism no hace ninguna poda.

CONCLUSIONES:

Tras varios experimentos y analizar los resultados de ambos ejercicios, se pueden observar las siguientes características: J48 obtiene mejores resultados en promedio. OneR es el que obtiene peores resultados en promedio en comparación con las otras opciones. PRISM solo se puede utilizar en el conjunto de datos contact-lenses debido a su simplicidad y porque no permite la entrada de atributos no nominales. En este conjunto de datos, PRISM presenta resultados similares a OneR. JRIP tiene tasas de error similares a su homónimo basado en árboles de decisión (J48), pero con tasas de error algo peores en promedio. PART obtiene resultados aceptables. La elección entre las diferentes estrategias de aprendizaje no es arbitraria y depende de muchos factores como la estructura del conjunto de datos, la cantidad de instancias de entrenamiento que se posean, las limitaciones computacionales o la tasa de error admisible en la clasificación de resultados.