INTRODUCCIÓN A LA MINERÍA DE DATOS

MEMORIA DE PRÁCTICAS
PARA LA ASIGNATURA INTRODUCCIÓN A LA MINERÍA DE DATOS
DEL GRADO DE INGENIERÍA INFORMÁTICA
EN LA UNIVERSIDAD DE CÓRDOBA

Ángel Sevilla Molina
Septiembre 2018

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Preprocesamiento y exploración de datos

1.1. Introducción

En este capítulo se mostrará brevemente los efectos que tienen algunos filtros de procesamiento de datos que dispone Weka sobre algoritmos de aprendizaje. Para facilitar su comprensión se utiliza como ejemplo el uso de un clasificador tras aplicar cada uno de los filtros sobre la base de datos *Pima Indians Diabetes* [1], cuyos valores originales se muestran en la figura 1.1.

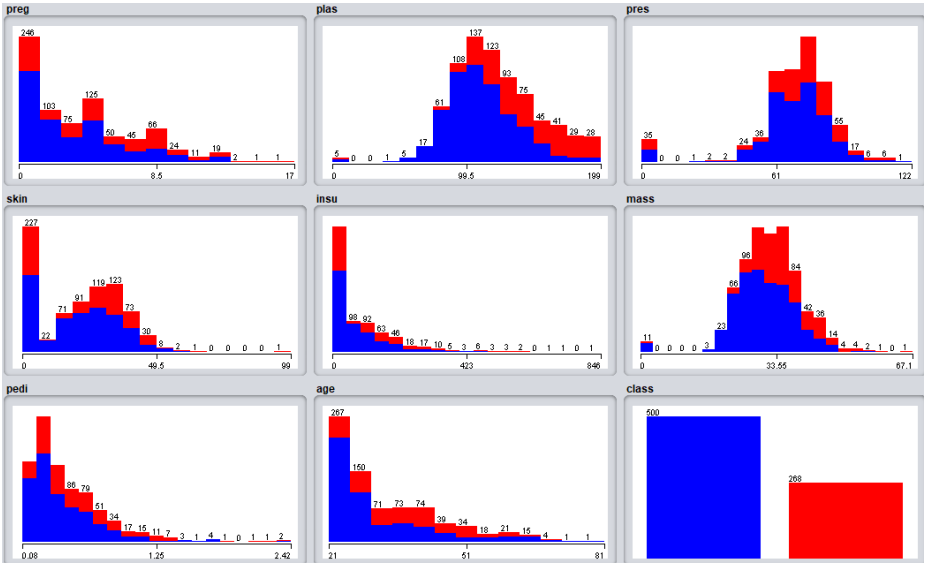

Figura 1.1: Visualización de la base de datos *Pima Indians Diabetes*

En orden de hacer la comparación se muestra el funcionamiento del algoritmo $C4{,}5$ (nombrado como $J48$ en Weka) para la generación de un árbol de decisión sobre la base de datos con los valores de validación cruzada por defecto (10 folds).

El árbol de decisión obtenido se muestra en la figura 1.2.

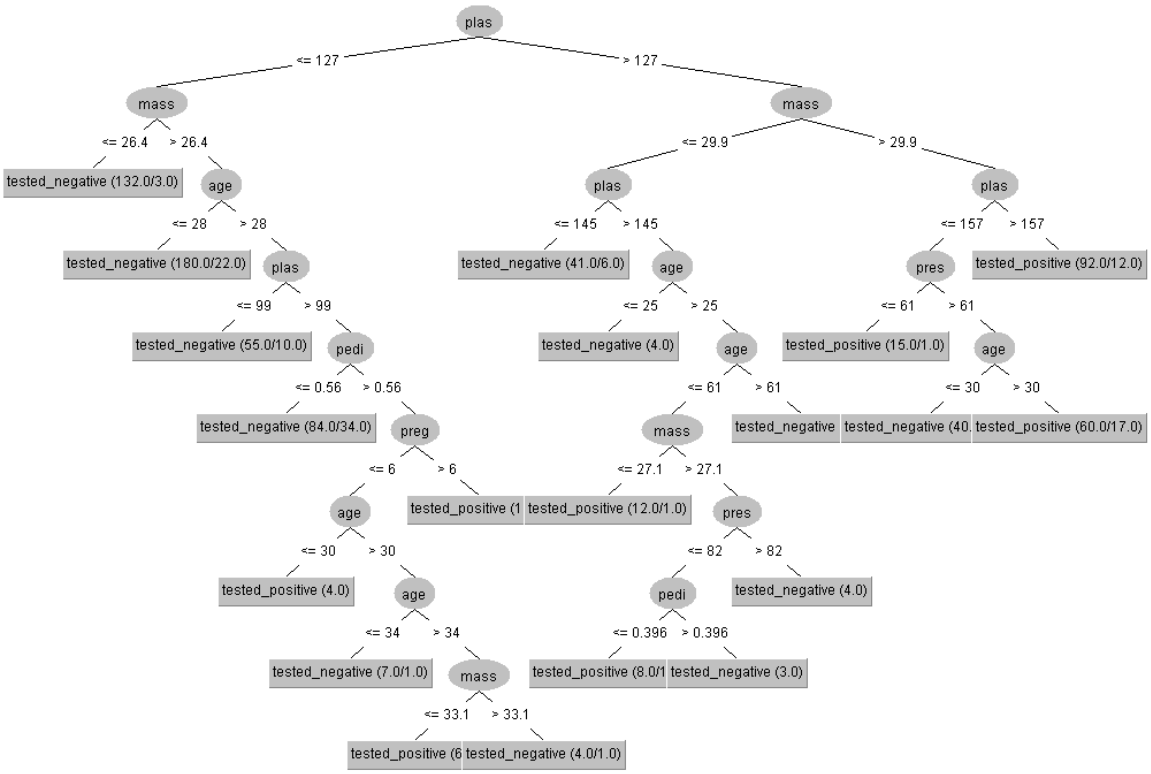

Figura 1.2: Árbol de decisión J48 sobre *Pima Indians Diabetes*

El tamaño del árbol es de 39 y está formado por 20 hojas. Hay dos variables que no aparecen en el árbol, que son *skin* e *insu*.

Los resultados de la evaluación se muestran en las tablas 1.1 y 1.2.

Correctly Classified Instances	567	73.8281 %
Incorrectly Classified Instances	201	26.1719 %
Kappa statistic	0.4164	
Mean absolute error	0.3158	
Root mean squared error	0.4463	
Relative absolute error	69.4841 %	
Root relative squared error	93.6293 %	
Total Number of Instances	768	

Tabla 1.1: Resumen de *J48* sobre *Pima Indians Diabetes*

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.814	0.403	0.790	0.814	0.802	0.417	0.751	0.811	tested_negative
0.597	0.186	0.632	0.597	0.614	0.417	0.751	0.572	tested_positive
0.738	0.327	0.735	0.738	0.736	0.417	0.751	0.727	

Tabla 1.2: Precisión detallada de *J48* sobre *Pima Indians Diabetes*

1.2. Ejercicio 1

1.2.1. filters/unsupervised/attributes/RandomProjection

El filtro *RandomProjection* reduce la dimensionalidad de los datos proyectándolos en un subespacio de menor dimensionalidad utilizando una matriz aleatoria, preservando las propiedades originales.

En la figura 1.3 se muestran los resultados tras realizar una proyección a un subespacio de cinco dimensiones con una distribución gaussiana.

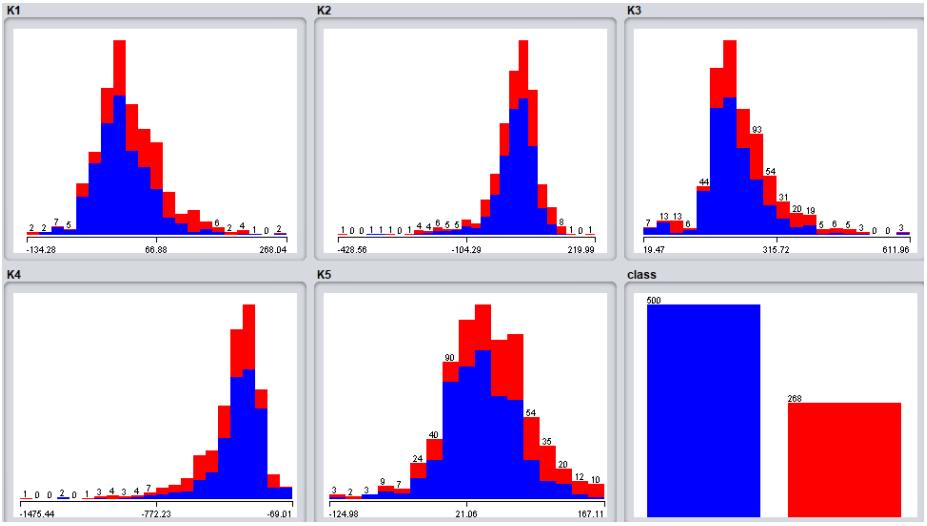

Figura 1.3: Visualización de la base de datos *Pima Indians Diabetes* tras aplicar el filtro *RandomProjection*

El árbol de decisión obtenido se muestra en la figura 1.4.

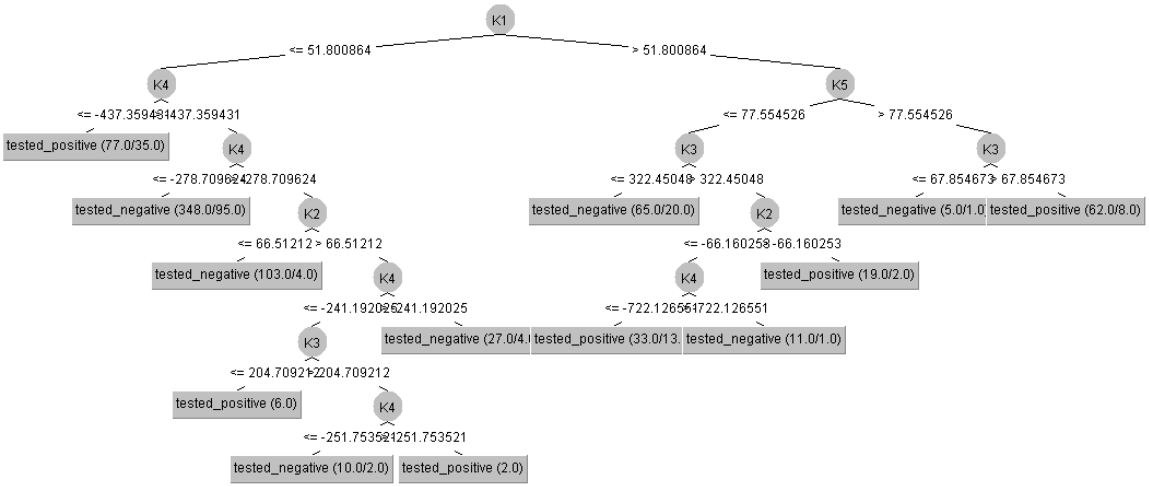

Figura 1.4: Árbol de decisión J48 sobre *Pima Indians Diabetes* con el filtro *RandomProjection*

Al disponer de menos atributos el tamaño final del árbol se reduce. Su tamaño pasa a ser de 25 y está formado por 13 hojas. Otro aspecto a tener en cuenta es que, al estar

Los resultados de la evaluación se muestran en las tablas 1.3 y 1.4.

Correctly Classified Instances	524	68.2292 %
Incorrectly Classified Instances	244	31.7708 %
Kappa statistic	0.2515	
Mean absolute error	0.3892	
Root mean squared error	0.4614	
Relative absolute error	85.6218 %	
Root relative squared error	96.7991 %	
Total Number of Instances	768	

Tabla 1.3: Resumen de *J48* sobre *Pima Indians Diabetes* con el filtro *RandomProjection*

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.832	0.597	0.722	0.832	0.773	0.259	0.649	0.740	tested_negative
0.403	0.168	0.563	0.403	0.470	0.259	0.649	0.513	tested_positive
0.682	0.447	0.666	0.682	0.667	0.259	0.649	0.661	

Tabla 1.4: Precisión detallada de *J48* sobre *Pima Indians Diabetes* con el filtro *RandomProjection*

1.2.2. filters/unsupervised/attributes/RemoveUseless

El filtro *RemoveUseless* elimina aquellos atributos que muestren poca o mucha variación entre sus valores.

Aplicar dicho filtro sobre *Pima Indians Diabetes*, con el límite de porcentaje de variación máxima permitida fijado a 0.99, no provoca ningún cambio como se puede comprobar en la figura 1.5.

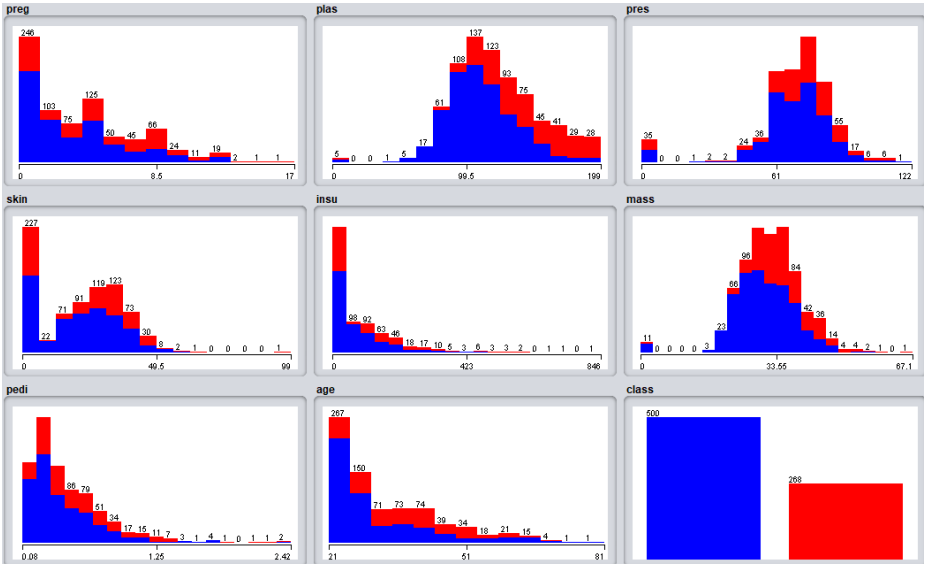

Figura 1.5: Visualización de la base de datos *Pima Indians Diabetes* tras aplicar el filtro *RemoveUseless*

Por este motivo, la ejecución de *J48* tras aplicar *RemoveUseless* es semejante a la de no aplicarlo, es decir, el árbol de decisión obtenido es el mismo que se muestra en la figura 1.2 y los resultados de la evaluación se muestran en las tablas 1.1 y 1.2.

1.3. Ejercicio 2

Bibliografía

[1] Ryan A. Rossi and Nesreen K. Ahmed. The network data repository with interactive graph analytics and visualization. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.