DECSAI
Departamento de Ciencias de la Computación e I.A.
Universidad de Granada

Tratamiento Inteligente de Datos

Guión de Prácticas

Práctica 4

Clasificación

FICHEROS DE DATOS

Datos de iris.csv,bankloan.csv

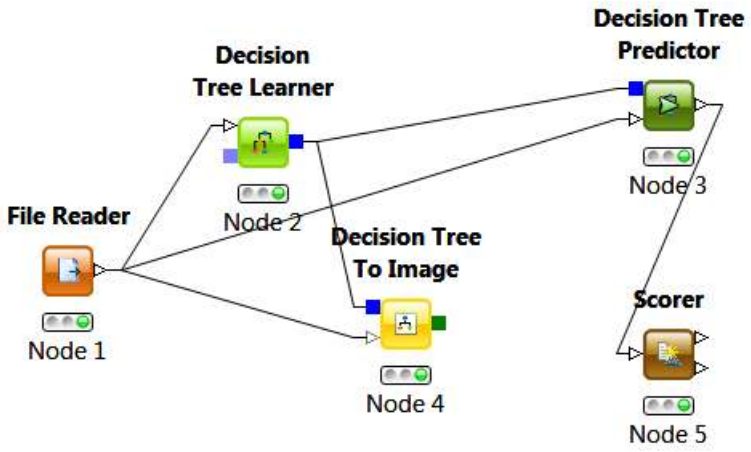
Clasificacion en Knime

La clasificación en Knime permite aplicar los algoritmos más comunes, permite clasificar mediante árboles de decisión, Naive Bayes y K-vecinos más cercanos entre otros. También incluye métodos para medir la calidad de la clasificación como medidas de calidad, validación cruzada y curvas ROC.

Los nodos de Weka ofrecen una variedad mucho más amplia de métodos , dando distintas variantes de cada uno de los métodos básicos. Como en el caso del clustering se recomienda utilizar los métodos básicos, probando algunos de ellos y utilizar variantes Weka para mejorar algún resultado que no sea del todo satisfactorio

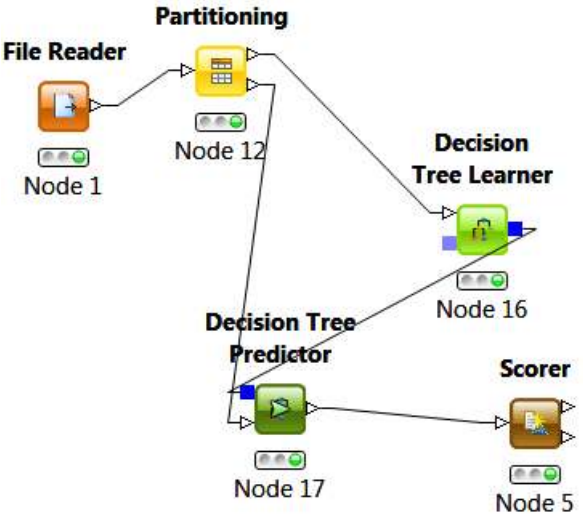
Clasificación mediante árboles de decisión.

Leer de iris dataset y realizar el siguiente flujo:

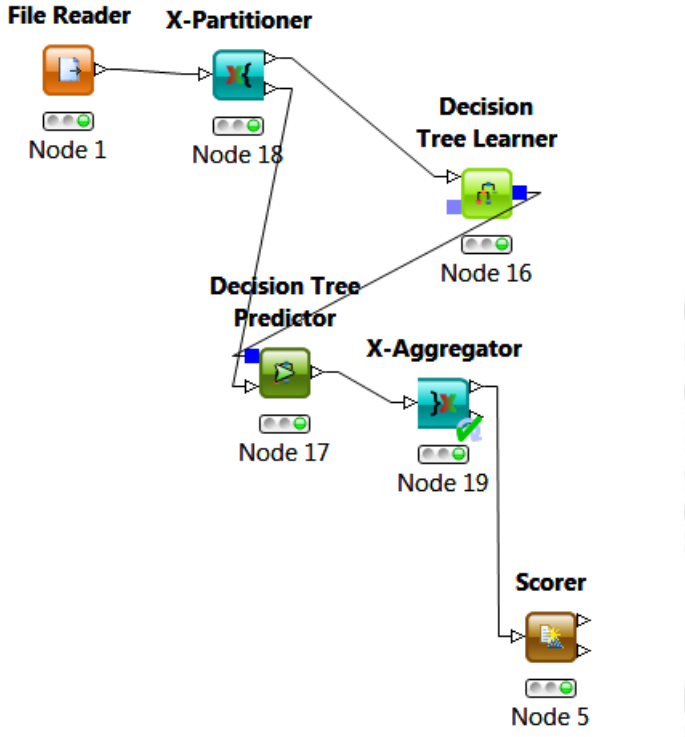


La salida del árbol de decisión se puede visualizar y también la verificación de la predicción directa que como vemos es muy buena.

Verificar que no estamos sobreaprendiendo se debe ahora trabajar con datos. Particionados, podemos simplemente dividir entre test y entrenamiento, o bien realizar una validación cruzada, utilizando los nodos adecuados. La primera opción seria:



La segunda:



Pregunta:

A la vista del árbol de decisión se puede encontrar una variable que prediga mejor que las otras?.

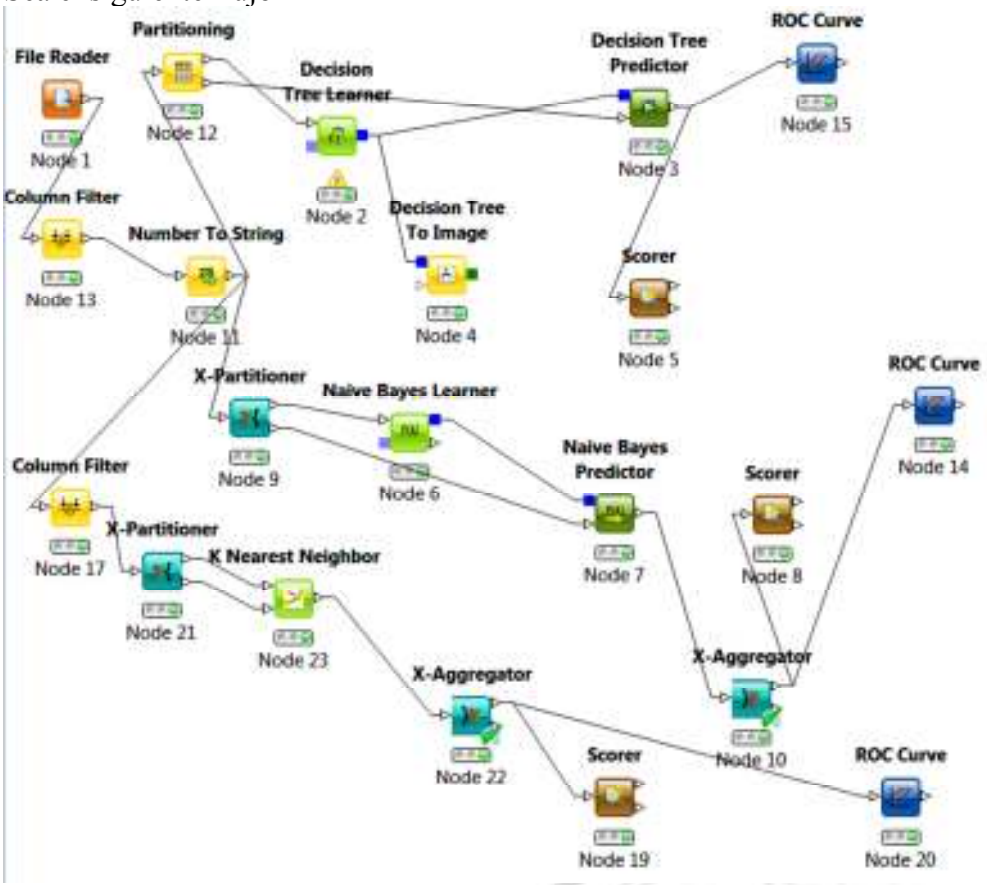
¿ Qué pasaría si utilizamos como variables los factores obtenidos en PCA?. Ver si mejora la clasificación, ver también si un factor es más determinante que otro.

Ejercicio.1

Probar el método Naive Bayes y el KNN para este ejemplo.

Ejemplo de predicción de variable binaria y uso de curva ROC

Sea el siguiente flujo



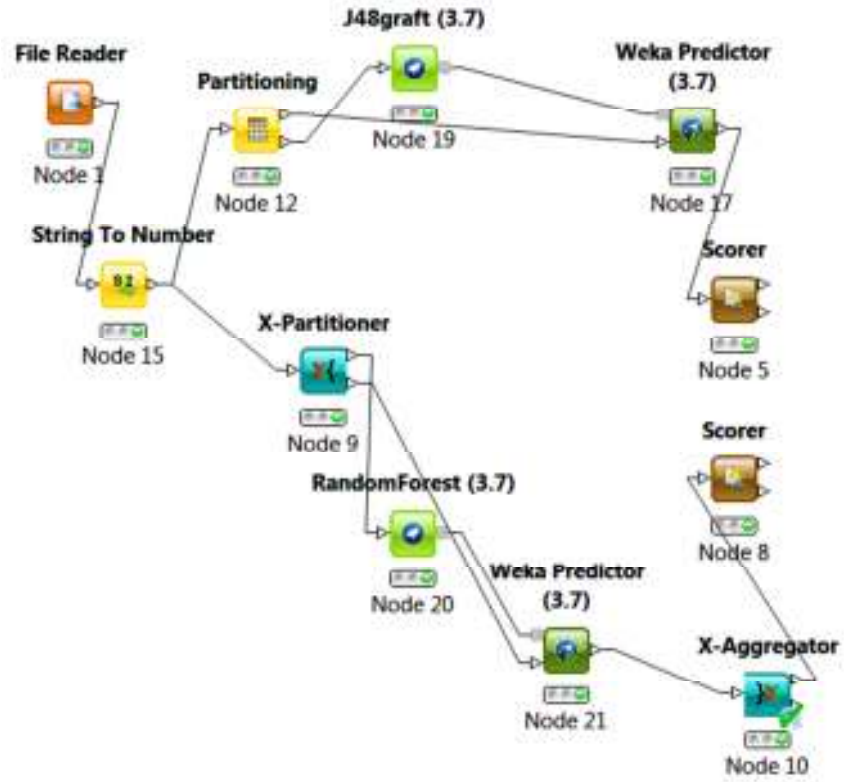
Trabajando con bankloan.csv realizar el proceso de clasificación para la variable binaria impago, por tres métodos disitntos. Comprobar si los resultados son similares, puesto que es una variable binaria, estudiar las curvas ROC. A la vista de las mismas ¿Cuáles son las variables que mejor predicen?. ¿Te parece razonable el resultado?

Ejercicio 3

Tratar de predecir la edad en bankloan.csv . Analizar resultados. ¿Podrías hacer una selección de variables que fuera adecuada para mejorar resultados?

Otros ejemplos de uso

Para probar con alguna de las herramientas de Weka para estos problemas. Hay que seleccionar previamente las variables y que hay que usar predictores para obtener los datos clasificados por el modelo.



La figura anterior nos muestra cómo se reproduce el ejemplo del iris utilizando dos nodos de Weka, uno de ellos utiliza cono clasificador el C4.5 de Quinlan el otro utiliza un random forest.

Clasificación con R-studio

Como es lógico la clasificación es problema que ha recibido mucha atención en R. No obstante no existe una librería única bien organizada y las posibles opciones se encuentran algo dispersas. Es de hacer notar que las aproximaciones de Weka son, en la parte de clasificación, mucho más completas que las que R ofrece. Pasamos a las más comunes.

- Para obtener árboles de decisión se pueden usar las siguientes librerías:
  - ***Tree*** obtiene un árbol de decisión/regresión binario, si la variable objetivo es nominal, el árbol obtenido es de decisión binario en caso contrario de regresión, basado en medidas de impureza, se puede encontrar un ejemplo de uso en el script ***Arboloes de decision 2.R*** en el apartado **Scripts de clasificación/Clasificacion-iris**
  - ***Party*** obtiene un árbol de decisión/regresión binario, si la variable objetivo es nominal, el árbol obtenido es de decisión binario en caso contrario de regresión,, basado en test de independencia estadísticos, se puede encontrar un ejemplo de uso en el script ***Arboloes de decision 1.R*** en el apartado **Scripts de clasificación/Clasificacion-iris**
  - ***Rpart*** obtiene un árbol de decisión/regresión binario de forma muy general, si la variable objetivo es nominal, el árbol obtenido es de decisión binario en caso contrario de regresión, y se puede seleccionar la medida de selección entre el índice de Gini y la ganancia de información, se puede encontrar un ejemplo de uso en el script ***Arboloes de decision rparty*** en el apartado **Scripts de clasificación/Clasificacion-iris**

- La clasificación KNN se encuentra en la librería ***Knn*** y se puede encontrar un ejemplo de uso en el script ***KNN en R.R*** en el apartado **Scripts de clasificación/Clasificacion-iris.**
- La clasificación Naive Bayes se encuentra en la librería ***e1071*** y se puede encontrar un ejemplo de uso en el script ***naiveBAyes.R*** en el apartado **Scripts de clasificación/Clasificacion-iris.**
- Por último, entre otros, la clasificación mediante Random Forest se encuentra en la librería ***RandomForest*** se puede encontrar un ejemplo de uso en el script ***Arboloes de decisión randomforest2.R*** en el apartado **Scripts de clasificación/Clasificacion-iris.**

Con respecto al cálculo de medidas de bondad, no se encuentran muy organizadas en las distintas librerías; pero la facilidad de programación de R permite sin grandes problemas programar la realización de dichos cálculos. En todos los sripts antes mencionados se han añadido sentencias calculan:

- Matriz de confusión para el conjunto de tests y el de entrenamiento en los casos en que se puede calcular este.
- El porcentaje de ítems bien y mal clasificados
- Precision, Recall y F-measure por clases
- F-measure total

El sript ***Comparacion-iris.R*** recoge las globales calculadas por los scripts anteriores y las muestra en conjunto para comparar.

Los scripts que se muestran en el apartado **Scripts de clasificación/Clasificacion-bankloan** muestran ejemplos adicionales de árboles de decisión con bankloan, en ellos se incluye predice el impago como variable de clase (binaria) y se dan dos formas de construir la curva ROC con dos librerías distintas ROC y pROC

Clasificación en SPSS

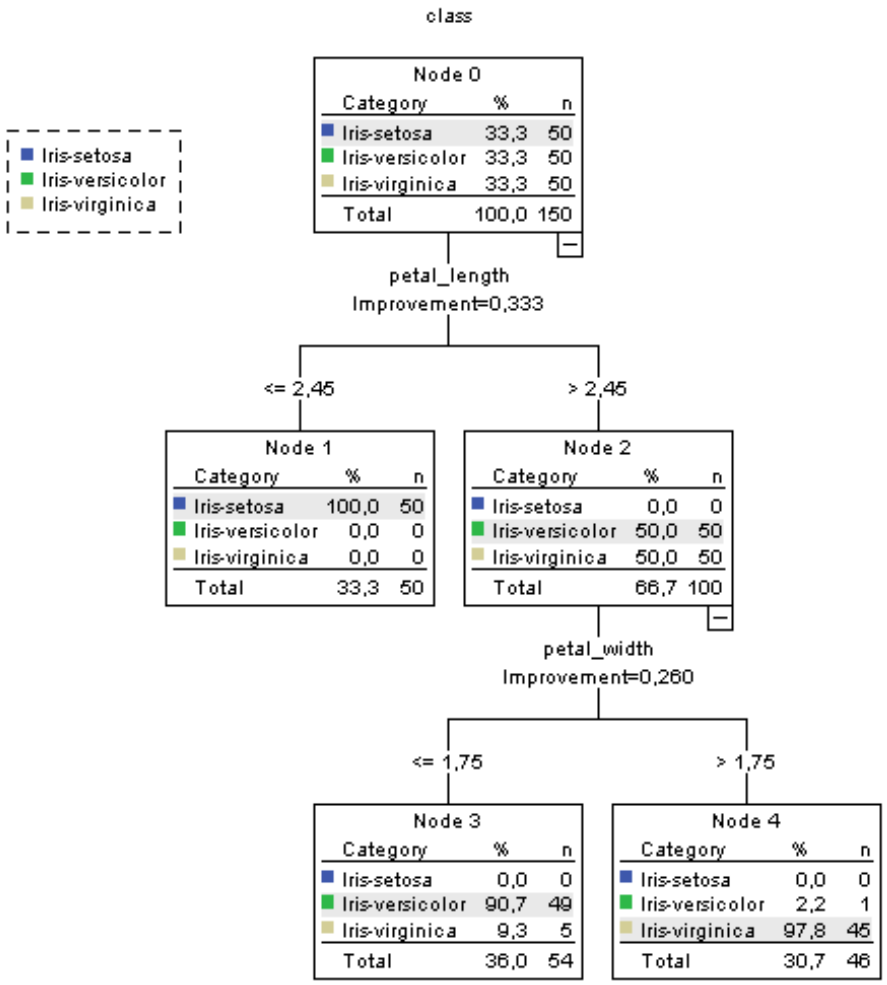
SPSS permite trabajar con árboles de decisión, especificando distintas opciones, realizar análisis discriminante y con el vecino más cercano.

Clasificación mediante árboles

En SPSS se puede elegir el método de crecimiento, habitualmente CRT que es una derivación de C4.5, y también seleccionar criterios como cuantos casos se toman en los nodos hijo y padre, también permite hacer validación cruzada.

Ejemplo

Haciendo una clasificación mediante arboles en SPSS *(Analyze>Classify>Tree)* del dataset iris, ajustar los valores mínimos por nodo padre e hijo en 50 25, y hacer validación cruzada se obtiene el siguiente árbol



Con la siguiente tabla de clasificación:

Classification

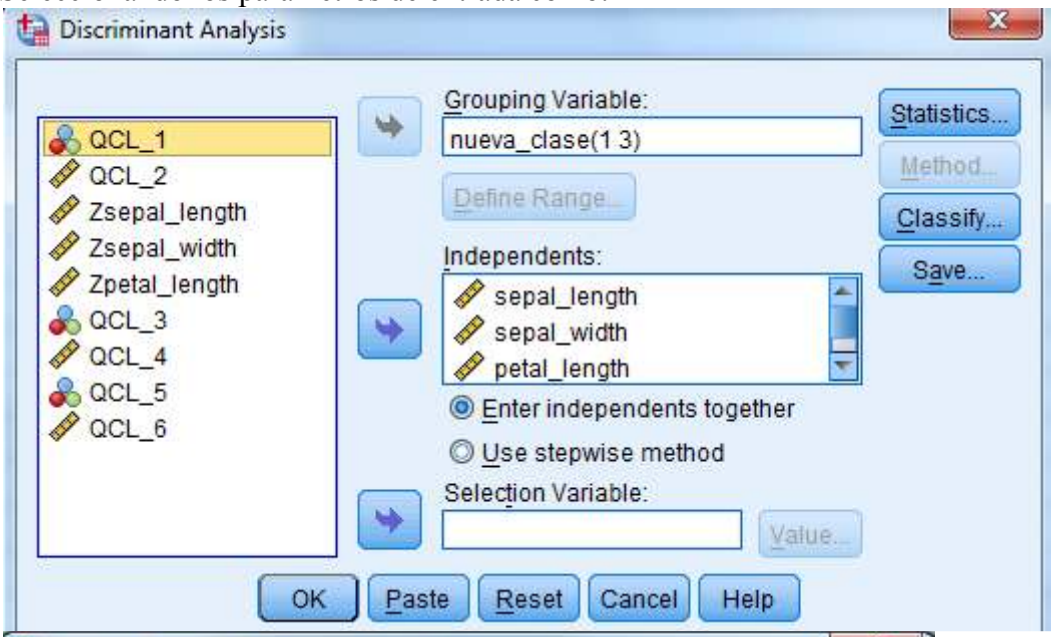
Observed	Predicted			
	Iris-setosa	Iris-versicolor	Iris-virginica	Percent Correct
Iris-setosa	50	0	0	100,0%
Iris-versicolor	0	49	1	98,0%
Iris-virginica	0	5	45	90,0%
Overall Percentage	33,3%	36,0%	30,7%	96,0%

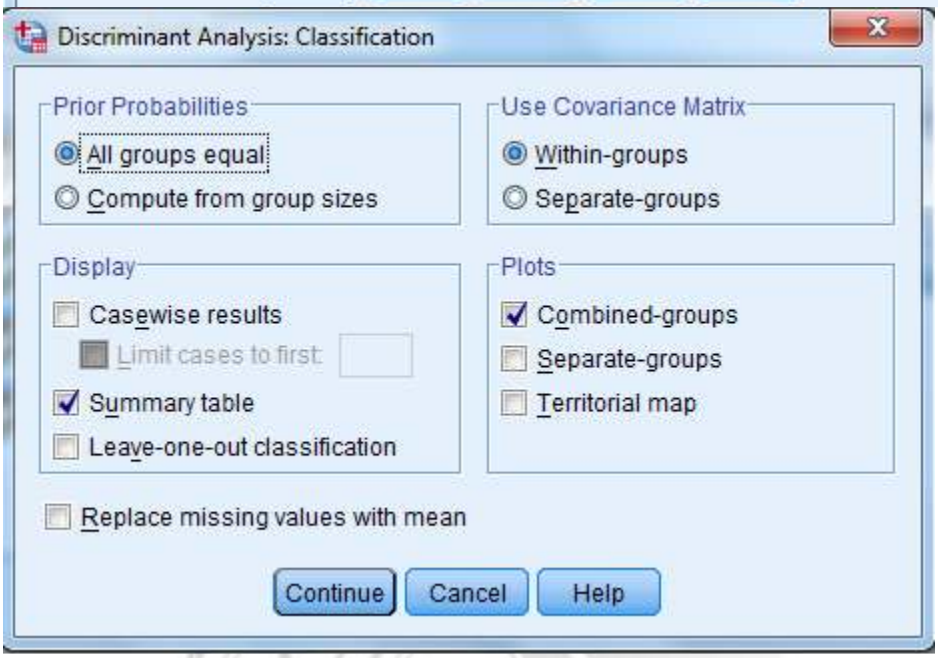
Growing Method: CRT

Dependent Variable: class

Tambien se pueden aplicar KNN y Analisis discriminante. Nos centramos en este último porque es el que no se contempla en Knime.

Para utilizar Analisis Discrminante, la variable clase ha de ser numérica, para ello utilizamos *transform>automatic recode* generando una nueva variable, por ejemplo, clase_nueva. Se puede observar en los datos que aparece con los valores 1,2 3.
Esta variable es la que nos sirve para entrar en *Analyze>Classify>Discriminant* como variable de agrupamiento.
Seleccionando los parámetros de entrada como:





Obtenemos la siguiente salida:

Discriminant

Notes

Output Created	03-NOV-2014 17:49:21

Comments		
Input	Data	C:\Users\Amparo\Dropbox\Docencia\Data-Mining\Datos\iris-kmedias-spss.csv
	Active Dataset	DataSet1
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	150
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing in the analysis phase.
	Cases Used	In the analysis phase, cases with no user- or system-missing values for any predictor variable are used. Cases with user-, system-missing, or out-of-range values for the grouping variable are always excluded.
Syntax		DISCRIMINANT /GROUPS=nueva_clase(1 3) /VARIABLES=sepal_length sepal_width petal_length petal_width /ANALYSIS ALL /PRIORS EQUAL /STATISTICS=MEAN STDDEV UNIVF TABLE /PLOT=COMBINED /CLASSIFY=NONMISSING POOLED.
Resources	Processor Time	00:00:00,27
	Elapsed Time	00:00:00,27

[DataSet1]

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		150	100,0
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Total	0	,0
Total		150	100,0

Group Statistics

nueva_clase		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
Iris-setosa	sepal_length	5,006	,3525	50	50,000
	sepal_width	3,428	,3791	50	50,000
	petal_length	1,462	,1737	50	50,000
	petal_width	,246	,1054	50	50,000
Iris-versicolor	sepal_length	5,936	,5162	50	50,000
	sepal_width	2,770	,3138	50	50,000
	petal_length	4,260	,4699	50	50,000
	petal_width	1,326	,1978	50	50,000
Iris-virginica	sepal_length	6,588	,6359	50	50,000
	sepal_width	2,974	,3225	50	50,000
	petal_length	5,552	,5519	50	50,000
	petal_width	2,026	,2747	50	50,000
Total	sepal_length	5,843	,8281	150	150,000
	sepal_width	3,057	,4359	150	150,000
	petal_length	3,758	1,7653	150	150,000
	petal_width	1,199	,7622	150	150,000

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
sepal_length	,381	119,265	2	147	,000

sepal_width	,599	49,160	2	147	,000
petal_length	,059	1180,161	2	147	,000
petal_width	,071	960,007	2	147	,000

Analysis 1

Summary of Canonical Discriminant Functions

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	32,192[a]	99,1	99,1	,985
2	,285[a]	,9	100,0	,471

a. First 2 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,023	546,115	8	,000
2	,778	36,530	3	,000

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
sepal_length	-,427	,012
sepal_width	-,521	,735
petal_length	,947	-,401
petal_width	,575	,581

Structure Matrix

	Function	
	1	2
petal_length	,706*	,168
sepal_width	-,119	,864*
petal_width	,633	,737*
sepal_length	,223	,311*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

Functions at Group Centroids

nueva_clase	Function	
	1	2
Iris-setosa	-7,608	,215
Iris-versicolor	1,825	-,728
Iris-virginica	5,783	,513

Unstandardized canonical discriminant functions evaluated at group means

Classification Statistics

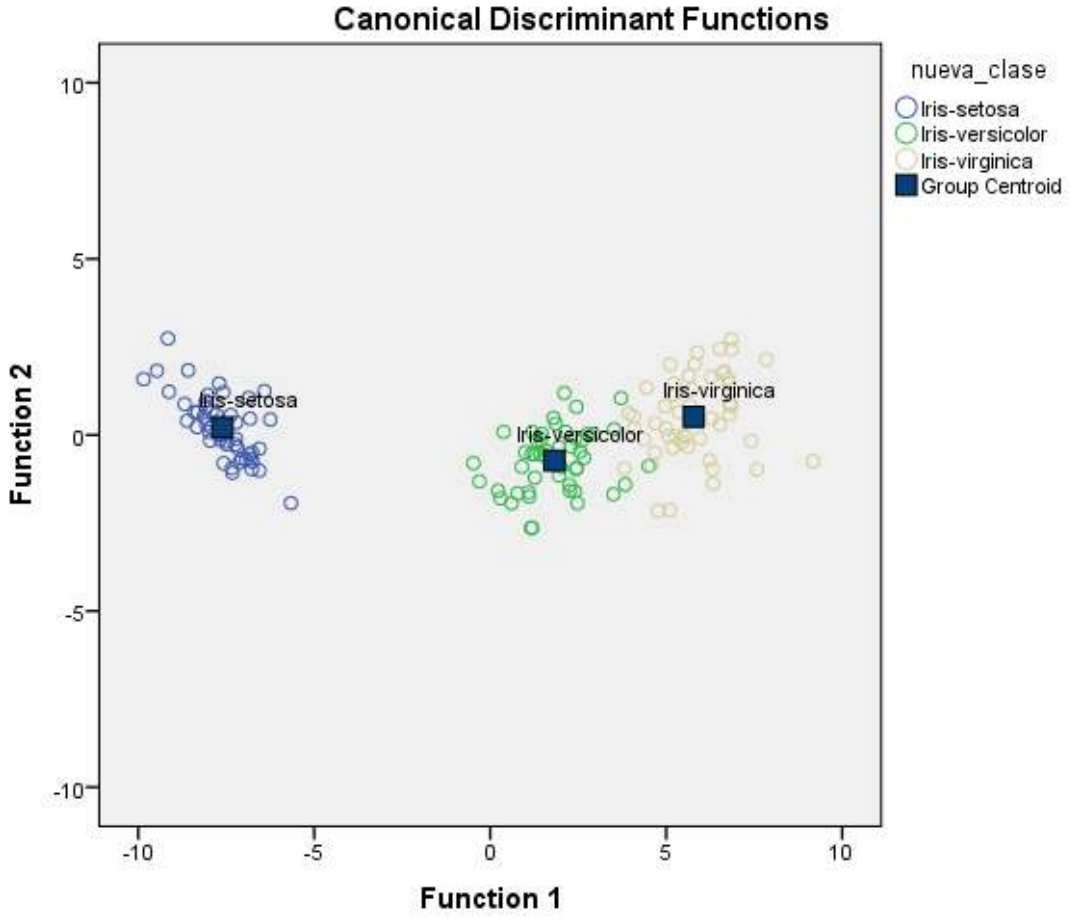
Classification Processing Summary

Processed		150
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0

Used in Output	150

Prior Probabilities for Groups

nueva_clase	Prior	Cases Used in Analysis	
		Unweighted	Weighted
Iris-setosa	,333	50	50,000
Iris-versicolor	,333	50	50,000
Iris-virginica	,333	50	50,000
Total	1,000	150	150,000



Classification Results[a]

		nueva_clase	Predicted Group Membership			Total
			Iris-setosa	Iris-versicolor	Iris-virginica	
Original	Count	Iris-setosa	50	0	0	50

	Iris-versicolor	0	48	2	50
	Iris-virginica	0	1	49	50
	Iris-setosa	100,0	,0	,0	100,0
%	Iris-versicolor	,0	96,0	4,0	100,0
	Iris-virginica	,0	2,0	98,0	100,0

a. 98,0% of original grouped cases correctly classified.

Que nos permite analizar la clasificación con gran detalle. Recordemos que el analisis discriminante general funciones lineales que parten el espacio para separar las clases. En este caso obtenemos dos funciones para separar, cuyos coeficientes están en la tabla **Function Coefficients,** la representación de los puntos según estas funciones lineales es el gráfico que aparece y muestra cómo divide las clases.

Ejercicio 4:

Realizar un análisis discriminante para predecir la variable impago en función de ingresos y distintos tipos de deuda en bankloan_csv.

Bibliografía

- C.Perez Técnicas de Análisis Inteligente de datos. Aplicaciones con SPSS (Pearson 2004)
- F.Berzal, J.C. Cubero Guiones de prácticas de S.I. de gestión (DECSAI)
- G.Bakor Knime Essentials Packt Publishing 2013
- IBM SPSS Documentos de ayuda (2014)
- J.P. Verma Data Analysis in Management with SPSS Software Springer (2013)
- R. Silipo KNIME Beginner's Luck Knime Press (2014)