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

Tratamiento Inteligente de Datos

Guión de Prácticas

Regresión

Amparo Vila

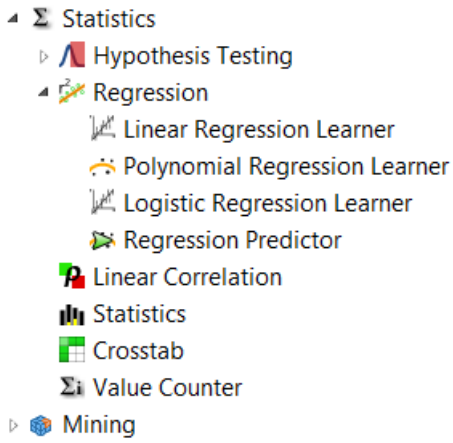
FICHEROS DE DATOS

```
Iris.csv
Bankloan.csv
```

Analisis de regresión utilizando Knime

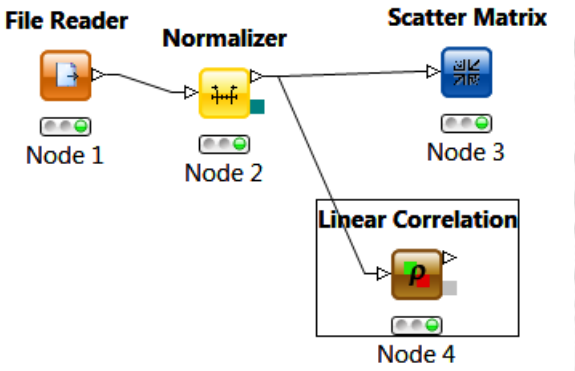
Regresión multivariante. Datos de Iris-Data set.

El análisis de regresión en Knime se encuentra en el módulo de estadística.

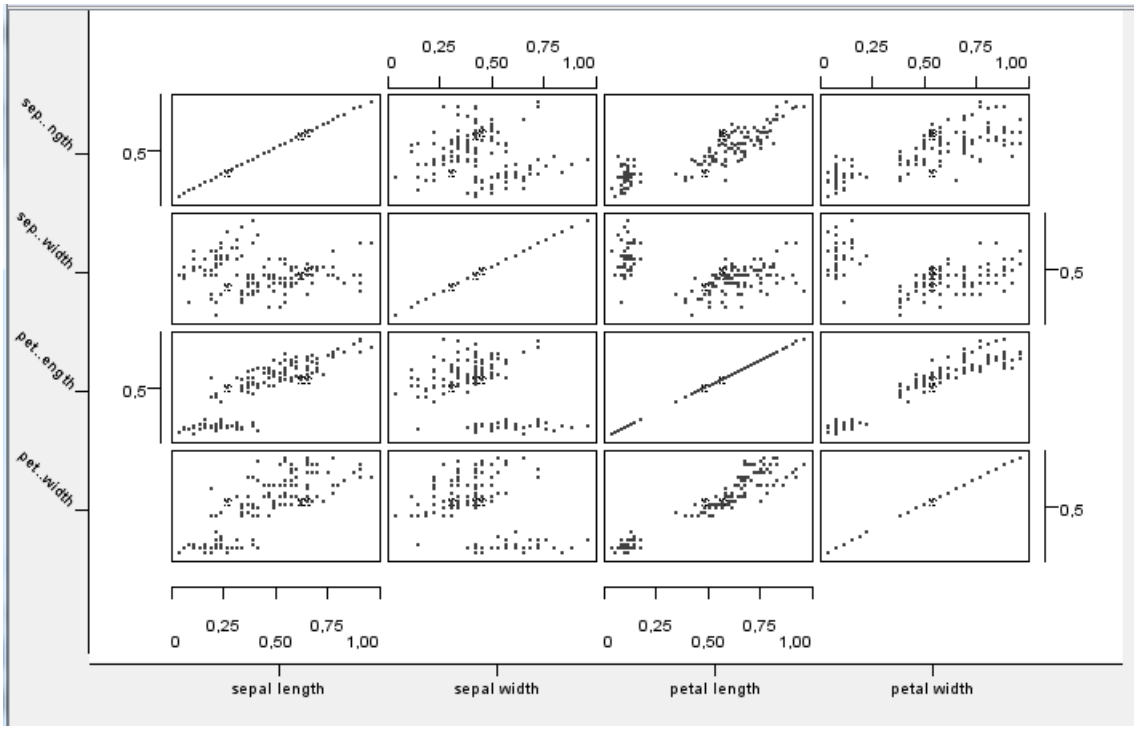


y como puede verse incluye los elementos más habitules de este tipo de análisis.

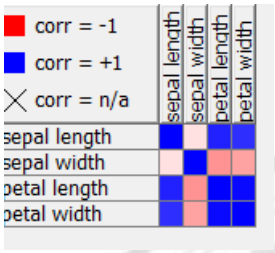
El análisis de regresión con los datos del Iris, comienza con un análisis exploratorio utilizando representaciones gráficas y estudio de correlaciones, obviamente normalizando antes para que las variables sean equiparables :



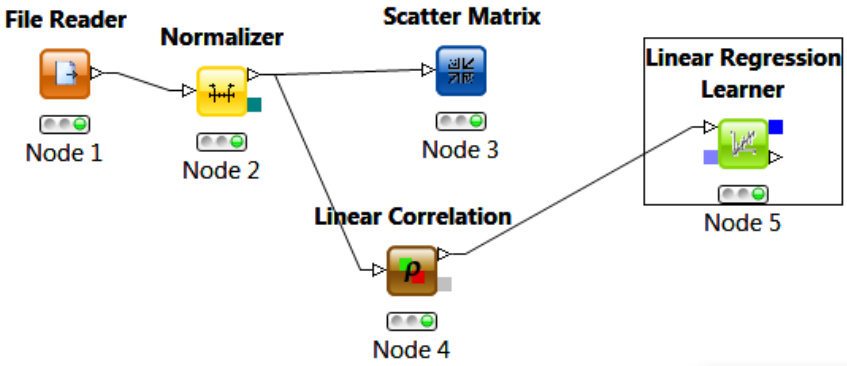
Que da como salida:



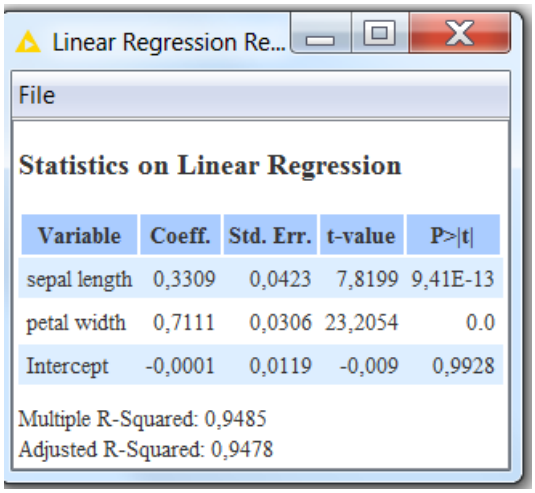
y



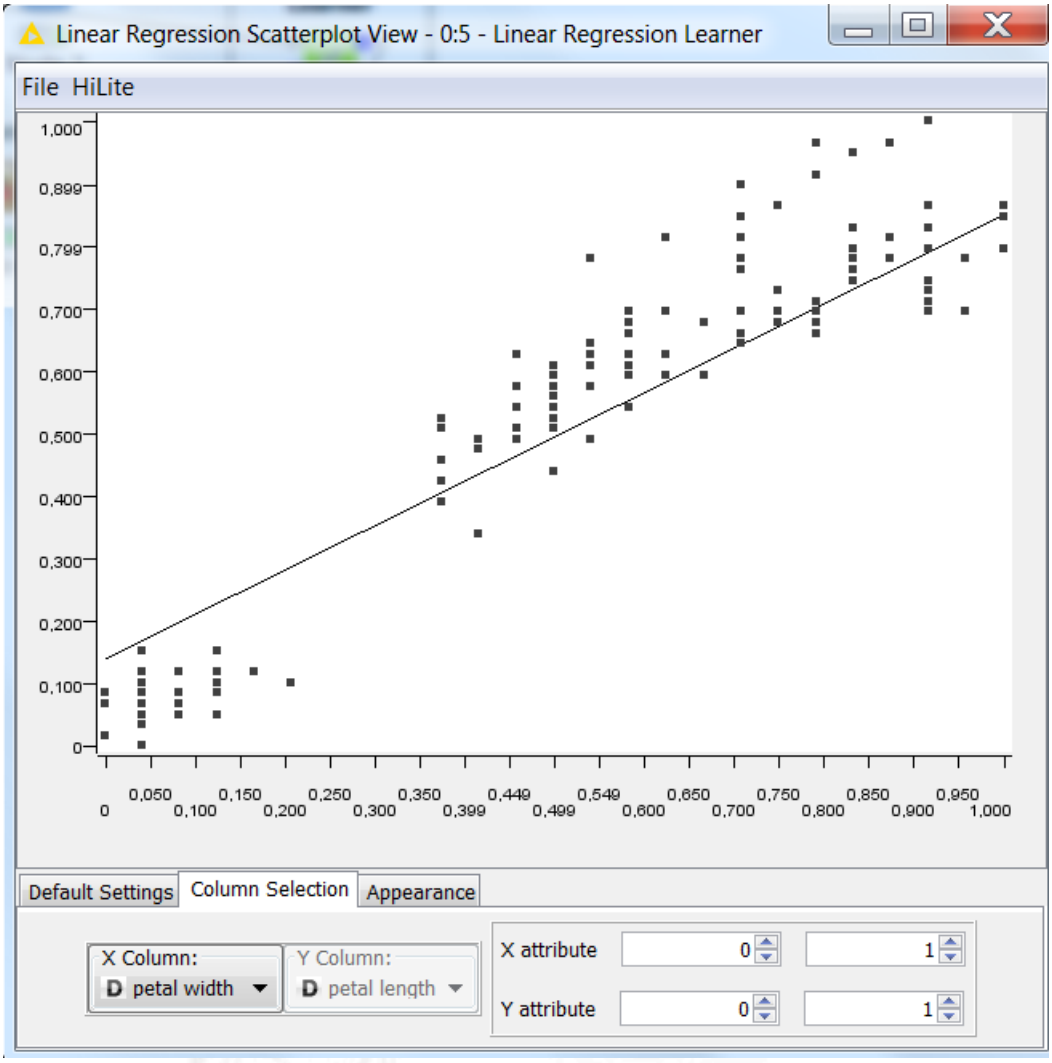
La matriz de gráficos sugiere que la dependencia lineal más clara está en la variable "petal length" como variable dependiente y en "sepal length" y "petal width" como variables indpendientes (las nubes de puntos se parecen más a una recta en estos casos). Además la correlación confirma que "sepal width " no tiene mucha relación con las otras. Con esta información hacemos el análisis de regresión, tomando "petal length" como variable dependiente y "petal width " y "sepal lenght" como independientes.



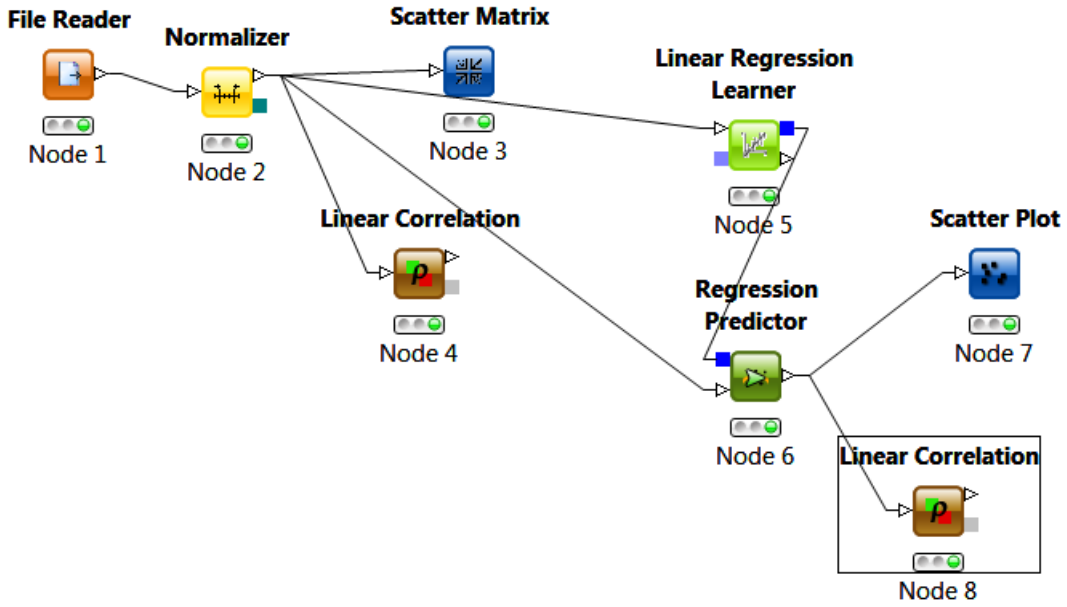
Lo que nos dá las siguientes salidas:



Donde se puede ver que la R^2 ajustada es muy buena y el peso de las variables en la predicción. También pueden verse las rectas que se ajustan, según una u otra variable. Por ejemplo:



Para terminar el análisis trabajo con el predictor y comparamos los valores de ambos



Viendo que los datos predichos y los iniciales se distribuyen a largo de la recta y=x y que la correlación entre ambos es de 0.97.

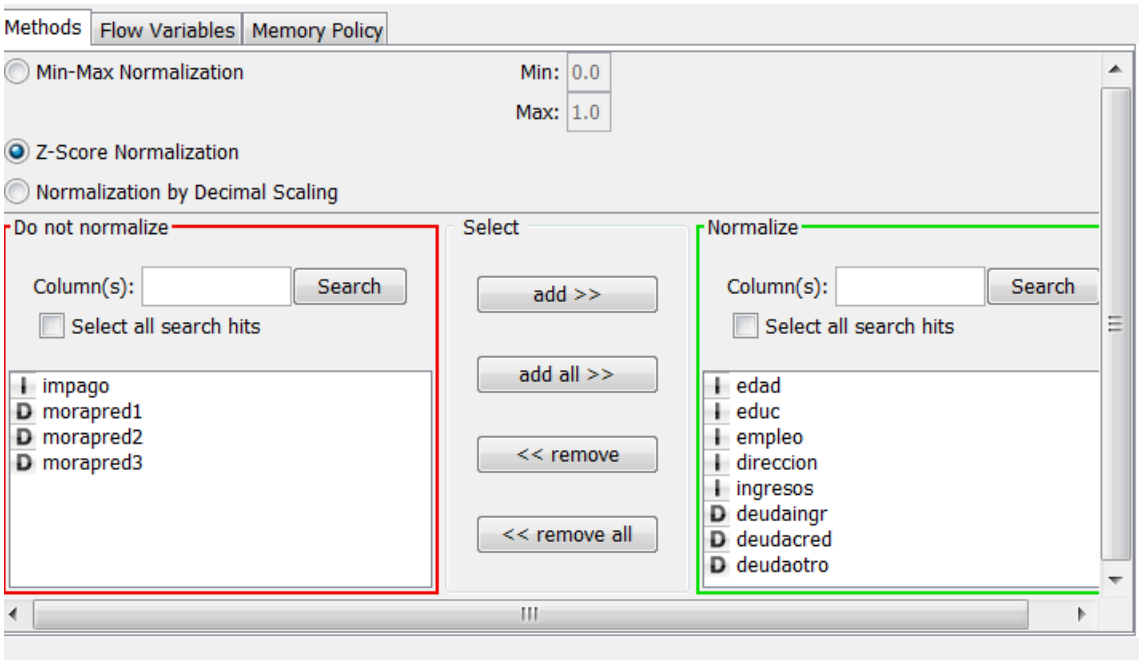
Ejercicio propuesto:

Trabajar con los datos de bankloan.csv y ver si los datos personales tales como: edad, años domicilio, años en el empleo etc. permiten predecir los tipos de deuda ó una variable que sea la media de los tres tipos de deuda. Ver si datos como edad y años en el empleo permiten predecir los ingresos. Justificar las conclusiones.

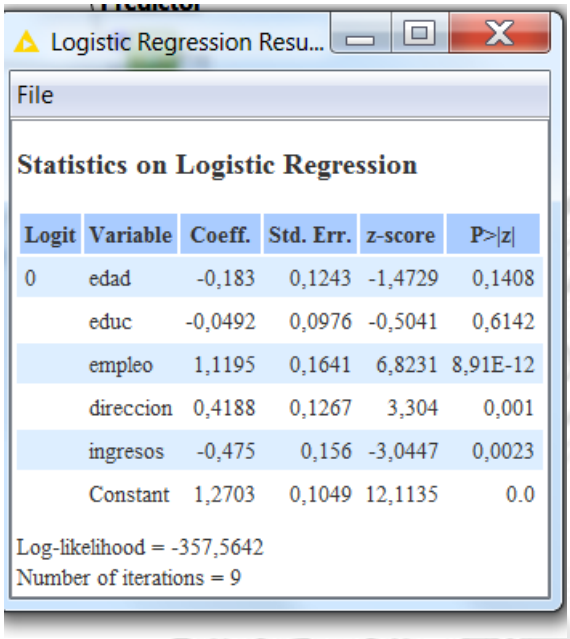
Regresion logística. Datos de bankloan-data set.

Como se conoce de la teoría, la regresión logística sirve para predecir la probabilidad de una variable bivaluada, por ello la utilizaremos para predecir la probabilidad de impago del data set bankloan. Para ello:

- Cambiaremos a numéricas las variables que no se hayan leído bien. Puede ocurrir que alguna variable decimal se haya leído como string
- Cambiaremos a string la variable impago que es la que se predice
- Normalizaremos las variables que pueden ser independientes en este caso

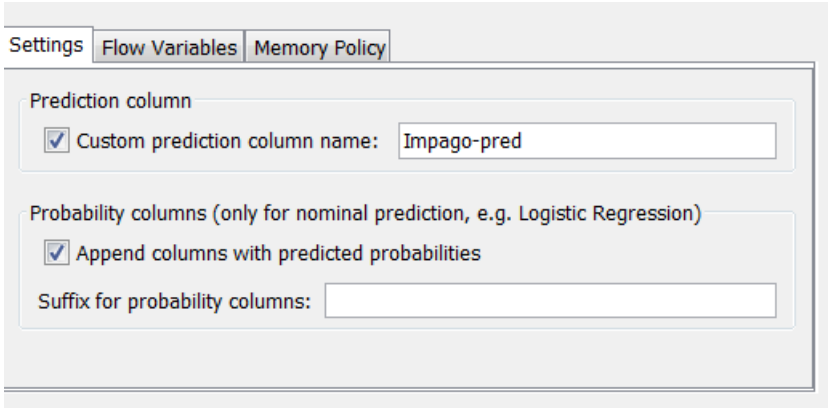


- Filtramos los valores perdidos de la variable impago.
- Hacemos la regresión logística obteniendo como resultado:

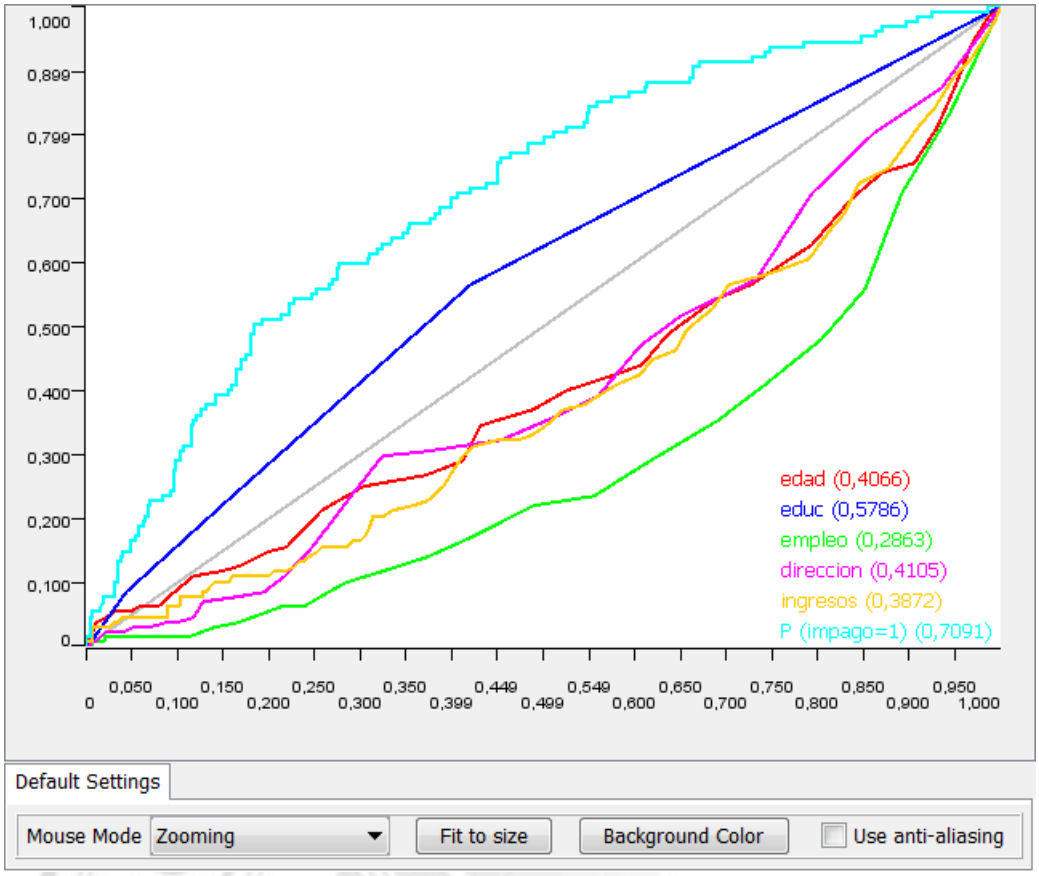


A la vista de los valores de z-score vemos que las variables que más influyen son empleo, dirección e ingresos, esta última negativamente, lo cual es lógico ya que estamos prediciendo la probabilidad de impago.

- Para ver cómo predice el modelo añadimos un nodo de predicción que en este caso hay que configurar un poco más

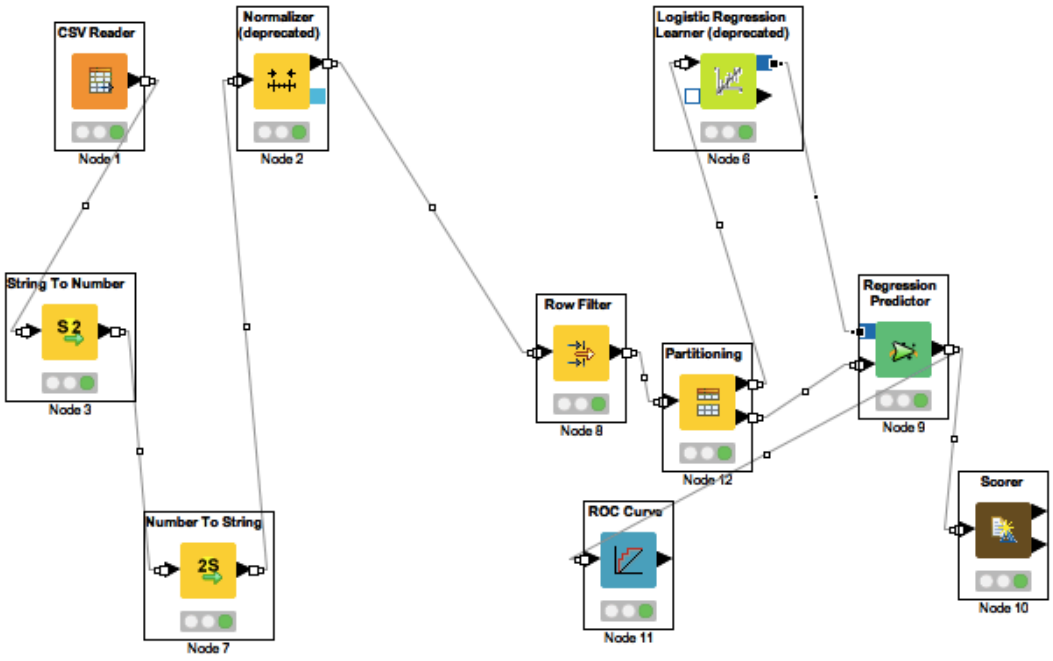


- Se obtiene un impago predicho que puede compararse con el real mediante:
  - El nodo scorer que nos dá la matriz de confusión y donde se puede ver que el error de predicción /clasificación es del 26% aproximadamente
  - Las curvas ROC que nos dan como salida:



Donde puede verse que ninguna de las variables predice muy bien.

El flujo completo del ejemplo es:



Ejercicio propuesto:

Realizar el mismo análisis utilizando como variables independientes los tres tipos de deuda. Realizar el mismo análisis utilizando las probabilidades que da el data-set. Como sería muy simple si se tuviera el flujo Knime del ejercicio anterior, no se incluye en esta práctica dicho flujo sino las indicaciones para hacerlo

Analisis de regresión utilizando R-studio.

En R la regresión se hace utilizando las funciones ***lm*** y ***glm*** que se encuentran dentro del paquete ***stat*** como material se incluen los scripts:

- ***Ejemplo-simple de regresión.R*** donde se rehace el análisis de regresión referente a los datos de Iris
- ***Logistica.R*** donde se predice mediante regresión logística el impago en los datos de bankloan. Dado que realmente es una clasificación en este scripts se calculan las medidas de bondad de la clasificación y las curvas ROC para la predicción. Un ejercicio es el comparar los resultados de este enfoque con los

que proporcionan los scripts de predicción de impago basados en árboles, Knn y Naive Bayes. Como resultado se tiene lo siguiente.

Datos de bondad:

Medida/metod.	***Arbol***	***Knn***	***Naive***	***Logisti***
%-Bien-clasi.	78.89	73.94	75.64	70.79
F-med-tot	.65	.62	.63	.58
Area-ROC	.69	.66	.73	.72

Con estos datos podemos asumir que el mejor clasificador es el Naïve Bayes aunque no es mucho mejor que otros. Ninguno de ellos se puede considerar un buen clasificador

Analisis de regresión utilizando SPSS

Regresión multivariante. Datos de Iris-Data set.

Como era de esperar el SPSS ofrece muchas opciones para hacer regresión, incluyendo en modelo general de regresión multivariante, y el modelo clásico de regresión.

Se puede reproducir el primer ejemplo del apartado anterior utilizando *analizar>regresión>lineales* utilizando como variable dependiente Petal Length y variables independientes Sepal Length y Petal Width. Revisamos las opciones pidiendo todos los gráficos y que guarde los valores predichos no tipificados. Puede comprobarse la salida que es muy parecida a la que proporciona Knime.

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Sepal_length, Petal_width[b]	.	Introducir

a. Variable dependiente: Petal_length

b. Todas las variables solicitadas introducidas.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,974[a]	,949	,948	,4032

a. Variables predictoras: (Constante), Sepal_length, Petal_width

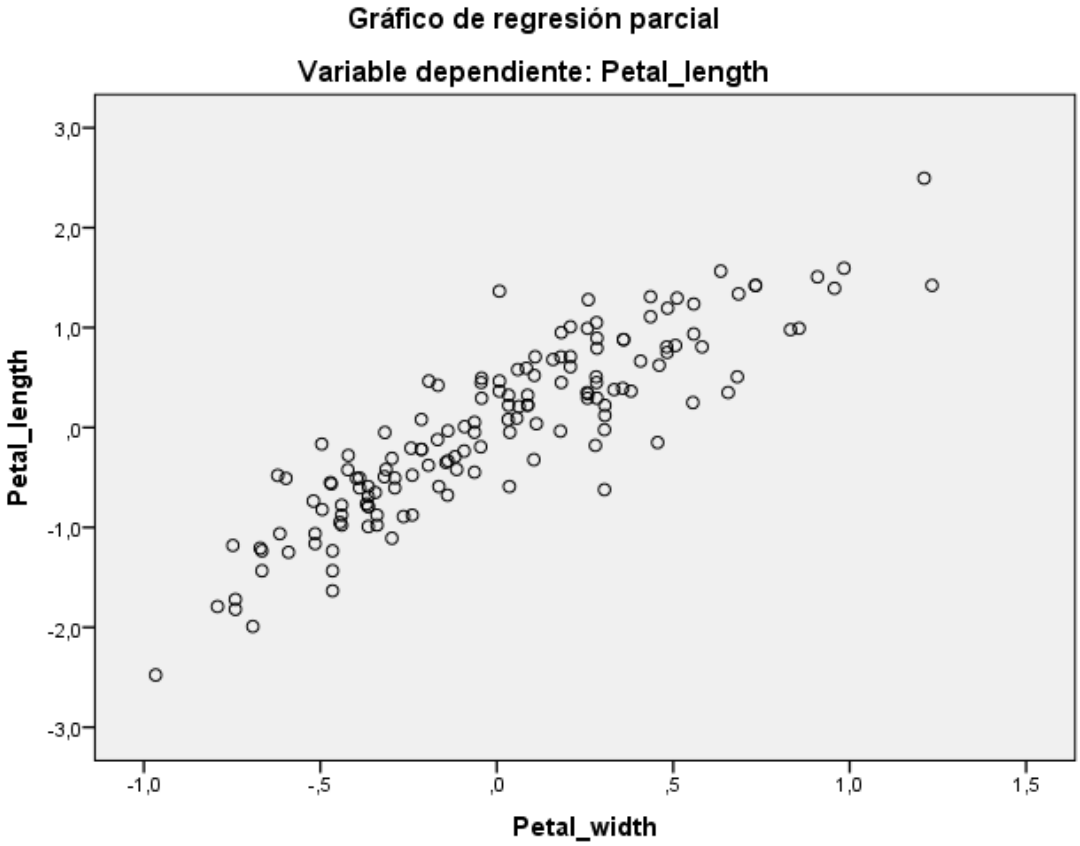
b. Variable dependiente: Petal_length

Coeficientes[a]

Modelo	Coeficientes no estandarizados	Coeficientes tipificados	t	Sig.

		B	Error típ.	Beta		
1	(Constante)	-1,507	,337		-4,473	,000
	Petal_width	1,748	,075	,755	23,205	,000
	Sepal_length	,542	,069	,254	7,820	,000

a. Variable dependiente: Petal_length



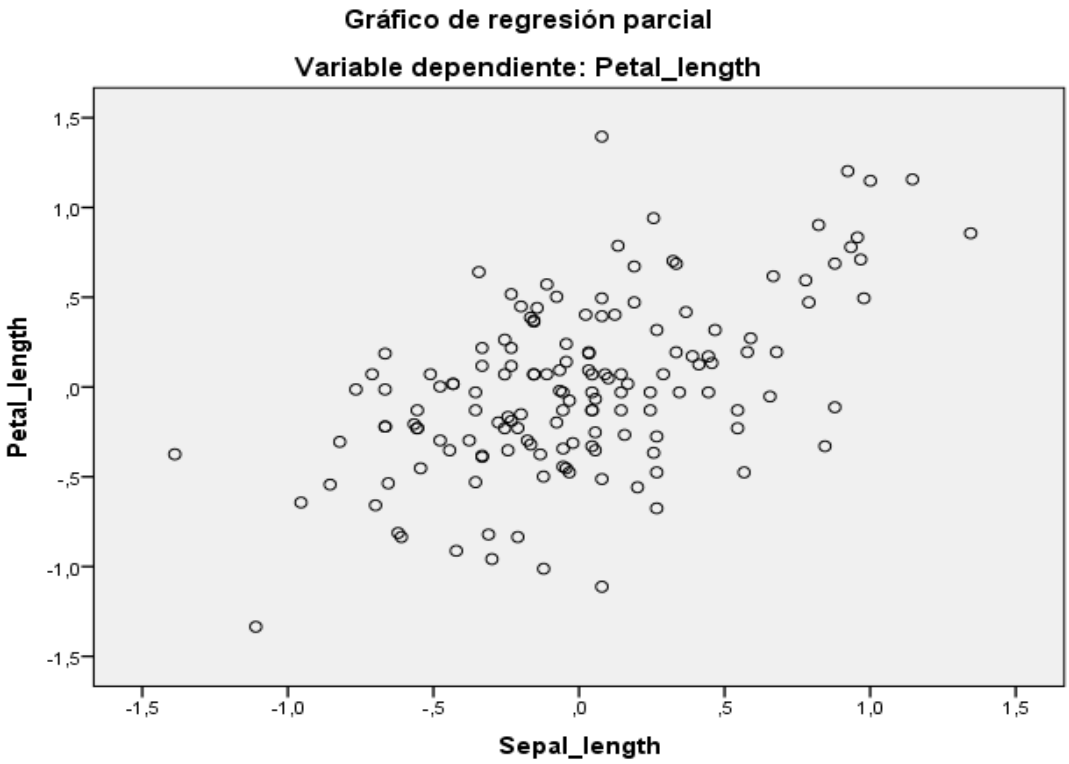
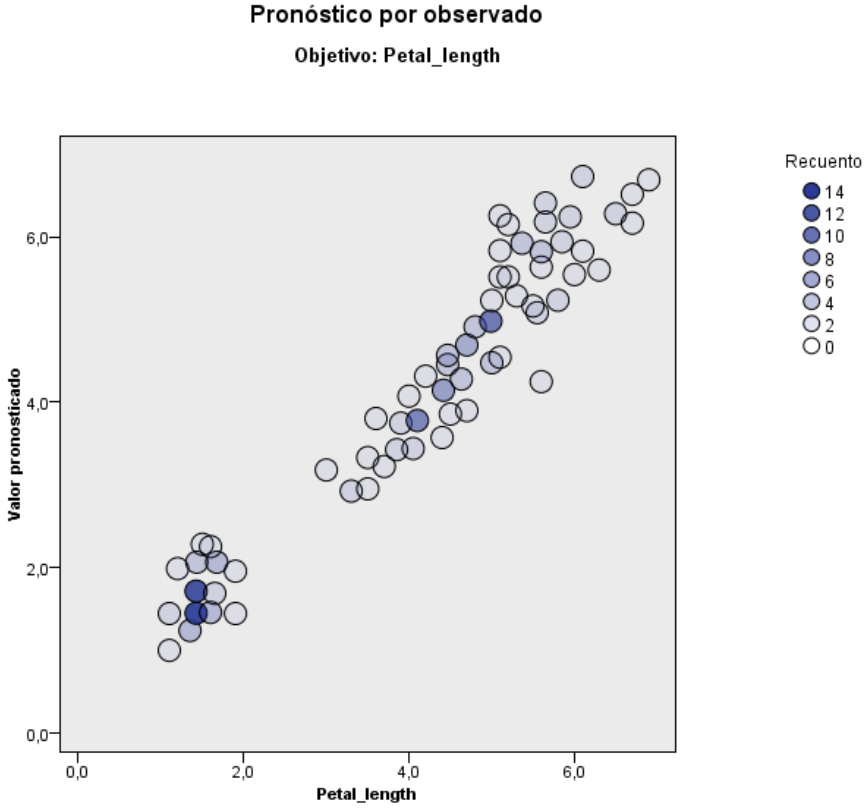
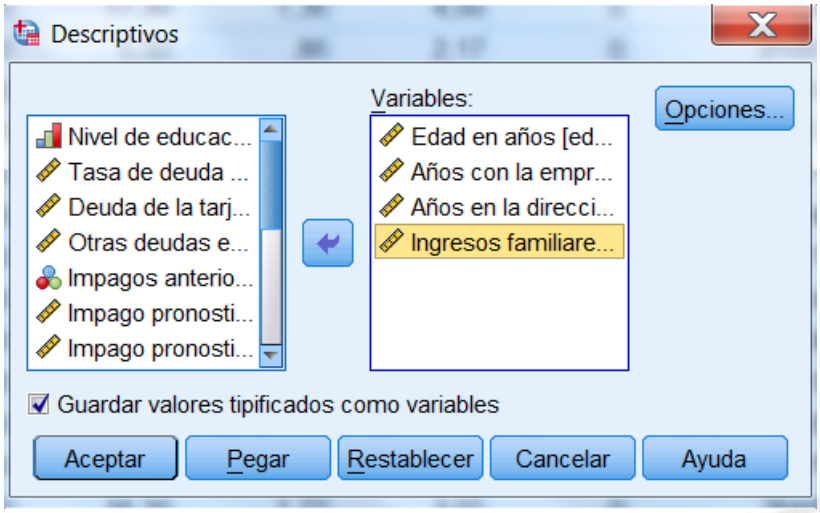


Gráfico de regresión parcial
Variable dependiente: Petal_length
Petal_length
1,5
1,0
,5
,0
-,5
-1,0
-1,5
-1,5
-1,0
-,5
,0
,5
1,0
1,5
Sepal_length


Pronóstico por observado
Objetivo: Petal_length
Valor pronosticado
6,0
4,0
2,0
0,0
0,0
2,0
4,0
6,0
Petal_length
Recuento
14
12
10
8
6
4
2
0

Regresion logística. Datos de bankloan-data set.

Para hacer la regresión logística hacemos *analizar>regresión>logística binaria* utilizaremos bankloan.sav como data-set y vamos a predecir impago a partir de edad, años en el empleo, años en la misma dirección e ingresos. Puesto que tenemos variables enteras, con distintos rangos, normalizaremos antes.



Un resumen de los resultados obtenidos aplicando regresión logística es el siguiente:

Resumen del procesamiento de los casos

Casos no ponderados[a]		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	700	82,4
	Casos perdidos	150	17,6
	Total	850	100,0
Casos no seleccionados		0	,0
Total		850	100,0

a. Si está activada la ponderación, consulte la tabla de clasificación para ver el número total de casos.

Codificación de la variable dependiente

Valor original	Valor interno
No	0
Sí	1

Tabla de clasificación[a]

	Observado	Pronosticado	
		Impagos anteriores	Porcentaje correcto

			No	Sí	
Paso 1	Impagos anteriores	No	497	20	96,1
		Sí	163	20	10,9
	Porcentaje global				73,9

a. El valor de corte es ,500

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Zedad	,182	,124	2,144	1	,143	1,199
	Zempleo	-1,147	,156	54,382	1	,000	,318
	Zdireccion	-,417	,127	10,825	1	,001	,659
	Zingresos	,509	,142	12,791	1	,000	1,664
	Constante	-1,269	,105	146,666	1	,000	,281

a. Variable(s) introducida(s) en el paso 1: Zedad, Zempleo, Zdireccion, Zingresos.

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	715,381[a]	,119	,175

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Como puede verse el modelo funciona mejor para predecir los pagos que los impagos, hay muchos impagos que no se predicen y las variables que más influyen son la edad y los ingresos.

Ejercicio propuesto:

Realizar el mismo análisis utilizando como variables independientes los tres tipos de deuda