Minería de Datos – Trabajo Práctico.
Tema: Diabetes.

Carlos Crosetti (cvjh@chevron.com)
Febrero 9, 2017.

Contents

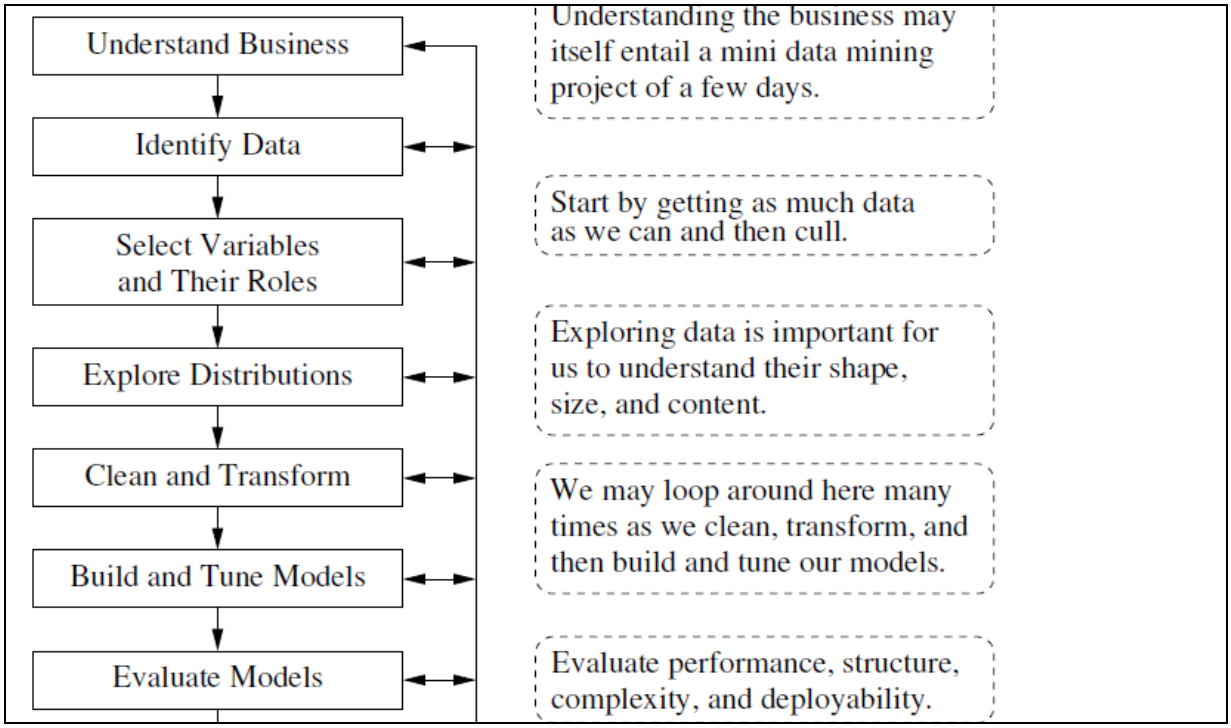
Objetivo.

Se busca corroborar resultados de exámenes de sangre de una muestra de pacientes con riesgo de diabetes empleando tres modelos predictivos y elegir el de mejor desempeño.

Como objetivo secundario, se propone el la adopción de Rattle para sacar provecho de la integración de R con funciones visuales y los diversos paquetes que Rattle carga bajo demanda.

Metodología.

Para lograr el objetivo de este TP se emplearon los conocimientos adquiridos en el curso de "Data Mining" (UTN, Prof. Lic. Ignacio Urteaga). El método se sintetiza en el siguiente esquema de procedimientos, segun Graham Williams [Williams-1]. Los dos últimos pasos (Deploy Model and Monitor Performance) no están contemplados en este TP.



Software utilizado.

Se han empleado Microsoft Excel 2010 y Rattle, un front-end de código abierto para el lenguaje de programación "R", version 3.3.2 de 64-bits.

Hardware

Se utiizó una laptop Lenovo T430 (Intel I5) con 8gb de RAM.

Identificando los datos.

El dataset contiene una muestra de pacientes de sexo femenino proveniente de una etnia llamada “Pima”, localizeda en el estado de Arizona, EEUU. Esta comunidad potencialmente desarrolla la enfermedad de la diabetes. La comunidad Pima es muy estudiada y el dataset se suele encontrar a menudo en competencias de ciencias de los datos, como por ejemplo “kaggle.com”. Fué originalmente donada al repositorio de la UCI [UCI-1]. El dataset presenta ocho (8) variables y posee en paricular una de respuesta que califica a cada observación estadística con una variable caegórica cuyo valor indica si el paciente presenta (o no) la enfermedad, según parámetros de la Organización Mundial de la Salud. Si bien el dataset invita a desarrollar un modelo de clasificación, en este TP se tomará como supuesto principal y didáctico que los pacientes han pasado por un test de sangre y todos los resultados (si presentan diabetes o no) han sido tomados por buenos.

A los efectos ilustrativos la muestra posee la siguiente estructura tabular:

	A	B	C	D	E	F	G	H	I
1	pregnant	plasma	diastolic	triceps	insulin	bmi	pedigree	age	class
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0

Limpieza y transformación del dataset.

En el primer lugar, tomando contacto con el dataset se ha hallado una tabla de 9 columnas con claros encabezados identificatorios de las variables y con 768 miembros-filas. Por la mera inspección visual no se detectaron outliers. El dataset no presenta ninguna dificultad al ser cargado tanto en Excel como en R. En segundo lugar, se ha realizado una validación de los atributos para ser considerados como potenciales predictores a través de una constatación de evidencias de diabetes reconocidas por la organización National Institutes of Health "www.nih.org". La descripción individual de tales atributos se detalla en el Anexo B.

De rigor, se realizó un primer chequeo estadístico, que permitió descubrir datos faltantes en varias de las variables.

```
> frame <- read.csv("C:/Users/cvjh/Desktop/DM/pimatxt.csv")
> summary(frame)
    pregnant         plasma         diastolic         triceps
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
    insulin           bmi           pedigree          age
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00
 Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
 Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
     class
 Min.   :0.000
 1st Qu.:0.000
 Median :0.000
 Mean   :0.349
 3rd Qu.:1.000
 Max.   :1.000
```

Al respecto, se han realizado imputaciones con Rattle que no han servido para remediar el dataset ya que no resultaron de utilidad a expensas de un gran esfuerzo de trabajo manual. Por este motivo, se tomó la decisión de suprimir los miembros del conjunto con datos faltantes, evitando no crear alteraciones en atributos donde el cero es válido en del rango de la variable. Por ejemplo, donde la variable "pregnant" tenía un valor cero, se conservó la observación, pero cuando "bmi" (índice de masa corporal) [BMI-1] mostraba un cero, se removió la fila completa. El dataset limpio quedó reducido a 392 elementos.

Establecer variables y sus roles.

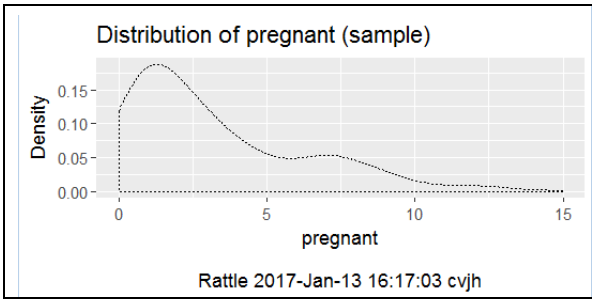
De todas las variables reconocidss en el dataset, la denominada "class" fué determinada como variable de respuesta (todos los valores eran 0 ó 1) y a las restantes se les asignaron al rol de entrada o predictores.

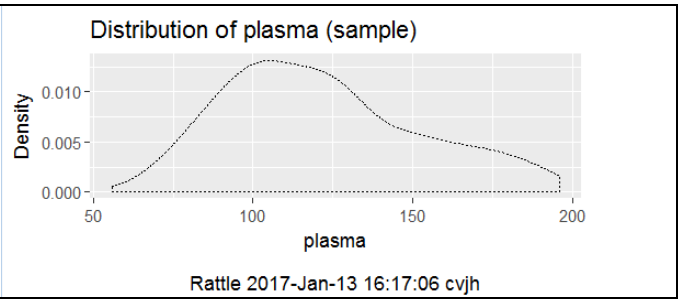
La carga del dataset "limpio" resultó trivial y correctamente tipificado por Rattle. Se particionó el dataset en tres conjuntos, a saber: 70% de las filas para entrenamiento, 15% para validción y 15% para prueba.

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	pregnant	Numeric	◉	○	○	○	○	○	Unique: 17
2	plasma	Numeric	◉	○	○	○	○	○	Unique: 117
3	diastolic	Numeric	◉	○	○	○	○	○	Unique: 37
4	triceps	Numeric	◉	○	○	○	○	○	Unique: 48
5	insulin	Numeric	◉	○	○	○	○	○	Unique: 184
6	bmi	Numeric	◉	○	○	○	○	○	Unique: 194
7	pedigree	Numeric	◉	○	○	○	○	○	Unique: 331
8	age	Numeric	◉	○	○	○	○	○	Unique: 43
9	class	Numeric	○	◉	○	○	○	○	Unique: 2

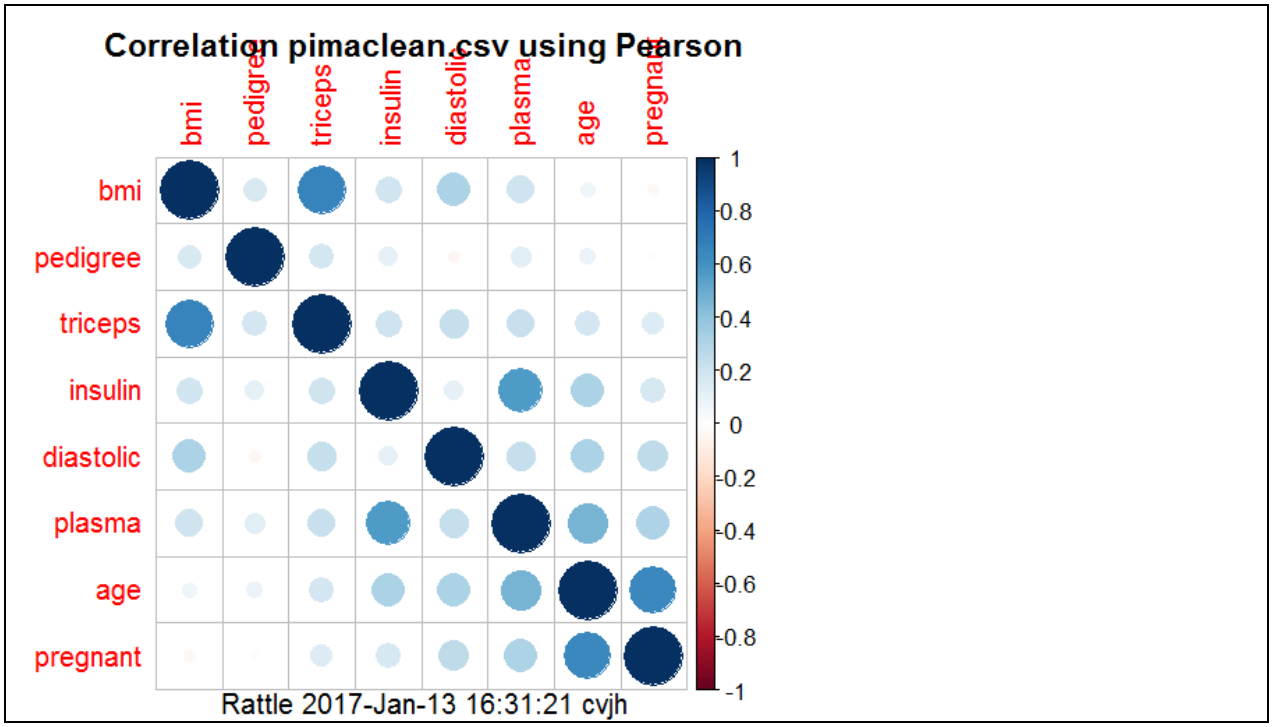
Exploración.

Se realizó una revisión de los rangos de cada variable a través de la función de Exploración de Rattle empleaando una aproximación intuitiva, visualizando la densidad de cada una de ellas y estableciendo gráficos de dispersión para empezar a tomar una idea general de las correlaciones (aquí solo veremos dos de ellas):

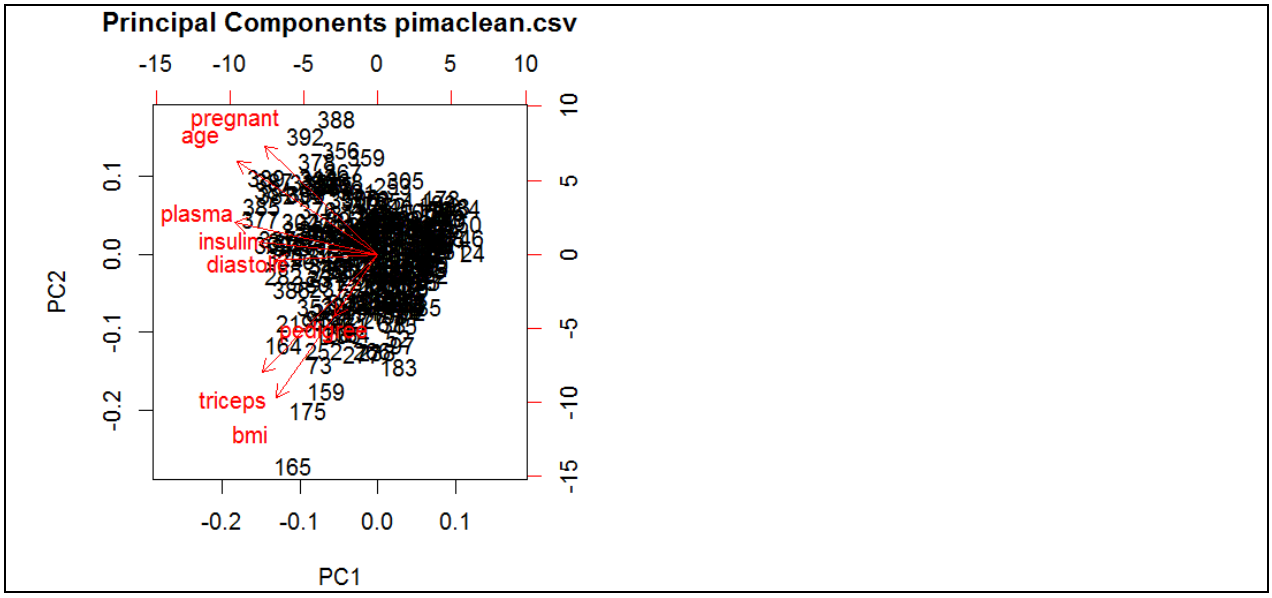




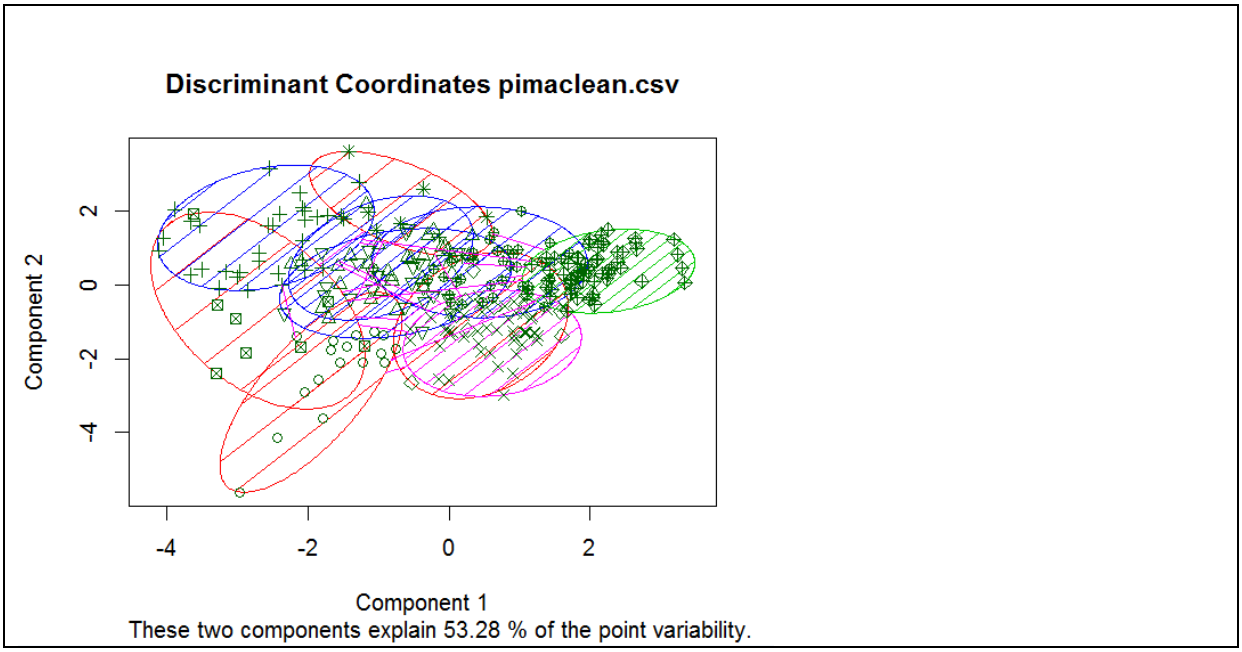
Seguidamente se obtuvo un mapa de correlación para entender el dataset a nivel de dependencias entre variables y se verificó **debil correlación** entre ellas, excepto en "pregnant", "plasma", "bmi", "pedigree" y "age".



A través del análisis de componentes principales [PCA-1] se puede reducir la dimensionalidad e identificar que variables incluir en el modelo. A pesar que unas pocas variables explican el modelo, no se van a retirar del mismo ninguna de las variables de entrada.



Solo a título informativo, se ejecutó el algoritmo de clustering "k-means" y como resultante se observa que dos componentes exhiben el 53% de la variabilidad.



Construir modelos predictivos.

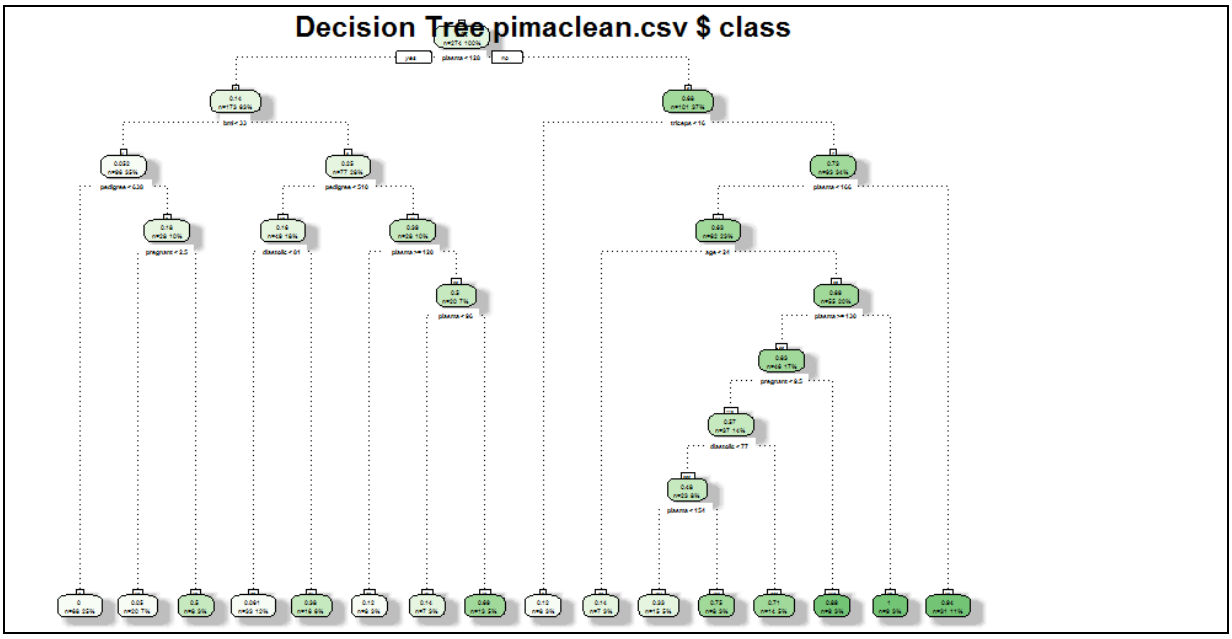
A partir de lo aprendido precedentemente se construyeron tres modelos (árbol de decisión, red neuronal y regresión logística) tomando como muestra el conjunto de entrenamiento. En el caso de la red neuronal se realizaron ajustes a la cantidad de neuronas en la capa intermedia.

Los modelos se construyeron con Rattle usando idéntico conjunto de predictores y la misma variable de respuesta ("class", valores 1 ó 0).

A continuación veremos un extracto del modelo de árbol de decisión:

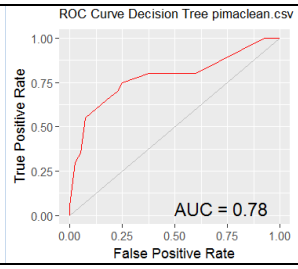
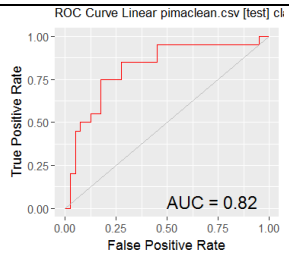
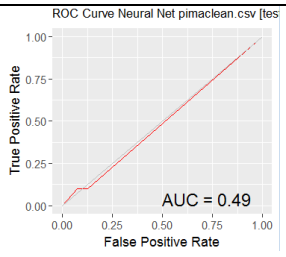
```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 274

node), split, n, deviance, yval
      * denotes terminal node

  1) root 274 61.4343100 0.33941610
    2) plasma< 127.5 173 20.6705200 0.13872830
      4) bmi< 33.25 96  4.7395830 0.05208333
        8) pedigree< 638.5 68  0.0000000 0.00000000 *
        9) pedigree>=638.5 28  4.1071430 0.17857140
         18) pregnant< 2.5 20  0.9500000 0.05000000 *
         19) pregnant>=2.5 8  2.0000000 0.50000000 *
      5) bmi>=33.25 77 14.3116900 0.24675320
```

```
       10) pedigree< 509.5 49  6.6938780 0.16326530
         20) diastolic< 81 33  1.8787880 0.06060606 *
         21) diastolic>=81 16  3.7500000 0.37500000 *
       11) pedigree>=509.5 28  6.6785710 0.39285710
         22) plasma>=119.5 8  0.8750000 0.12500000 *
         23) plasma< 119.5 20  5.0000000 0.50000000
           46) plasma< 95.5 7  0.8571429 0.14285710 *
           47) plasma>=95.5 13  2.7692310 0.69230770 *
    3) plasma>=127.5 101 21.8613900 0.68316830
      6) triceps< 16 8  0.8750000 0.12500000 *
      7) triceps>=16 93 18.2795700 0.73118280
       14) plasma< 165.5 62 14.4677400 0.62903230
         28) age< 24.5 7  0.8571429 0.14285710 *
         29) age>=24.5 55 11.7454500 0.69090910
           58) plasma>=130.5 46 10.7173900 0.63043480
            116) pregnant< 9.5 37  9.0810810 0.56756760
              232) diastolic< 77 23  5.7391300 0.47826090
                464) plasma< 154.5 15  3.3333330 0.33333330 *
                465) plasma>=154.5 8  1.5000000 0.75000000 *
              233) diastolic>=77 14  2.8571430 0.71428570 *
            117) pregnant>=9.5 9  0.8888889 0.88888890 *
           59) plasma< 130.5 9  0.0000000 1.00000000 *
       15) plasma>=165.5 31  1.8709680 0.93548390 *

Regression tree:
rpart(formula = class ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)], method = "anova", parms = list(split = "information"),
    control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] age       bmi       diastolic pedigree  plasma    pregnant  triceps

Root node error: 61.434/274 = 0.22421

n= 274

         CP nsplit rel error  xerror     xstd
1  0.307685      0   1.00000 1.00714 0.041314
2  0.044060      1   0.69232 0.70394 0.062591
3  0.031592      2   0.64825 0.70352 0.068134
4  0.030360      3   0.61666 0.70732 0.068071
5  0.026357      4   0.58630 0.74128 0.069585
6  0.017016      5   0.55995 0.88858 0.080756
7  0.016734      9   0.49188 0.90466 0.082185
8  0.014565     10   0.47515 0.90050 0.081873
9  0.012166     12   0.44602 0.88920 0.081482
10 0.011318     13   0.43385 0.90848 0.082483
11 0.010000     15   0.41121 0.90296 0.081811

Time taken: 0.02 secs

Rattle timestamp: 2017-01-13 18:16:25 cvjh
================================================================================
```



Evaluación de modelos.

Para la construcción de los tres modelos se usó el suconjunto de entrenamiento. Para ajustes, el subconjunto de validación y para prueba el mismo que lleva su nombre. A los efectos de comparar los resultados entre modelos se emplearon la matriz de error y el área bajo la curva (AUC) del gráfico ROC.:

Conjunto	*Arbol de Decisión (rpart)*	*Regresión Logística (glm)*	*Red Neuronal (nnet)*	*Ganador*
Error	23%	25%	37%	rpart
AUC	0.78	0.82	0.49	glm
				

El extracto de Rattle de la matrix de error de encuentra en el anexo D.

El modelo de árbol de decisión resultó ser el ganador por exhibir la tasa de error menor.

Conclusiones.

Se logró constatar que los resultados de los exámenes de sangre obtenidos de la muestra de los pacientes con riesgo de diabetes se pueden predecir con un modelo de árbol de decision, exhibiendo la menor tasa de error (23%).

El empleo de Rattle permitió ejecutar diversos modelos, sintonizarlos, graficar y compararlos con agilidad y sin necesidad de ejecutar comandos R desde la consola.

Referencias.

[Williams-1] "Data mining with Rattle and R: The art of excavating data for knowledge discovery", Graham Williams (Springer, 2011). e-ISBN 978-1-4419-9890-3.

[UCI-1] Univ. of California, Machine Learning Repository http://archive.ics.uci.edu/ml/index.html

[PCA-1] "Applied Predictive Modeling", Kuhn, Max, Johnson, Kjel (Springer, 2013). e-ISBN 978-1-4614-6849-3.

[BMI-1] Body Mass Index calculator @ NIH - https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmi-m_sp.htm

Anexo A. Detalles del dataset.

Source:

Original Owners:

National Institute of Diabetes and Digestive and Kidney Diseases

Donor of database:

Vincent Sigillito (vgs '@' aplcen.apl.jhu.edu)
Research Center, RMI Group Leader
Applied Physics Laboratory
The Johns Hopkins University
Johns Hopkins Road
Laurel, MD 20707
(301) 953-6231

Data Set Information:

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron-like devices. It is a unique algorithm; see the paper for details.

Attribute Information:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

** **UPDATE:** Until 02/28/2011 this web page indicated that there were no missing values in the dataset. As pointed out by a repository user, this cannot be true: there are zeros in places where they are biologically impossible, such as the blood pressure attribute. It seems very likely that zero values encode missing data. However, since the dataset donors made no such statement we encourage you to use your best judgement and state your assumptions.

Relevant Papers:

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care} (pp. 261--265). IEEE Computer Society Press.

Anexo B. Lista de predictores.

Atributo	Descriptón
pregnant	Numero de veces que la paciente estuvo embarazada, rango 0 a N – el cero es válido.
plasma	Concentración de glucosa en sangre, rango aprox. 60-500, debe ser mayor a cero.
diastolic	Presión aterial, debe ser mayor a cero.
triceps	Espesor del pliegue cutáneo en el músculo triceps, debe ser mayor a cero.
insulin	Nivel de insulina en sangre, debe ser mayor a cero y puede ser cercano a cero.
bmi	Body Mass Index, índice de masa corporal, debe ser mayor a cero.
pedigree	Es una función que expresa la relación hereditaria y genética del paciente con la enfermedad.
age	Edad, debe ser mayor a cero.
class	Resultado del test de glucosa en sangre, 1 presenta diabetes, 0 no presenta diabetes.

Anexo C. Archivos.

Archivo	Descripción
Pimatxt.csv	Dataset tal cual fue bajado del reposirorio.
Pimaclean.csv	Dataset limpio de elementos con datos faltantes, base del experimento.

Anexo D. Extracto de la matriz de error.

```
Error matrix for the Decision Tree model on pimaclean.csv [test] (counts):

      Predicted
Actual  0  1
     0 33  7
     1  7 13

Error matrix for the Decision Tree model on pimaclean.csv [test] (proportions):

      Predicted
Actual    0    1 Error
     0 0.55 0.12  0.18
     1 0.12 0.22  0.35

Overall error: 23%, Averaged class error: 26%

Rattle timestamp: 2017-02-09 13:51:07 cvjh
=========================================================================
Error matrix for the Linear model on pimaclean.csv [test] (counts):

      Predicted
Actual  0  1
     0 34  6
     1  9 11

Error matrix for the Linear model on pimaclean.csv [test] (proportions):

      Predicted
Actual    0    1 Error
     0 0.57 0.10  0.15
     1 0.15 0.18  0.45

Overall error: 25%, Averaged class error: 30%

Rattle timestamp: 2017-02-09 13:51:07 cvjh
=========================================================================
Error matrix for the Neural Net model on pimaclean.csv [test] (counts):

      Predicted
Actual  0  1
     0 36  4
     1 18  2

Error matrix for the Neural Net model on pimaclean.csv [test] (proportions):

      Predicted
Actual   0    1 Error
     0 0.6 0.07   0.1
     1 0.3 0.03   0.9

Overall error: 37%, Averaged class error: 50%
```

<<Fin del Documento>>