O emprego de R na detección das características máis influentes na clasificación de pacientes infectados por COVID-19 en Galicia

VII Xornada de Usuarios de R en Galicia

Laura Davila Pena, Balbina Casas Méndez, Ignacio García Jurado

15 de outubro de 2020

Problemas de clasificación

Introdución

Problema de clasificación

Un **problema de clasificación** consiste en predicir o valor dunha variable resposta cualitativa para un ou máis individuos, facendo uso dos valores que xa coñecemos de certas variables categóricas (ou atributos) de tales individuos.

Problemas de clasificación

Introdución

Problema de clasificación

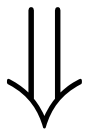
Un **problema de clasificación** consiste en predicir o valor dunha variable resposta cualitativa para un ou máis individuos, facendo uso dos valores que xa coñecemos de certas variables categóricas (ou atributos) de tales individuos.

Predicións $\longrightarrow$ Coñecemento obtido a través dunha mostra de individuos con valores coñecidos dos atributos e resposta.

Problema de clasificación

Un **problema de clasificación** consiste en predicir o valor dunha variable resposta cualitativa para un ou máis individuos, facendo uso dos valores que xa coñecemos de certas variables categóricas (ou atributos) de tales individuos.

Predicións $\longrightarrow$ Coñecemento obtido a través dunha mostra de individuos con valores coñecidos dos atributos e resposta.



Machine learning

Problemas de clasificación

Clasificadores

- Moitos clasificadores, ademais de clasificar, permiten avaliar a importancia que os diversos atributos tiveron na clasificación dun individuo concreto.
- En Strumbelj & Kononenko (2010) introdúcese un procedemento xeral para avaliar dita importancia.
- Este procedemento baséase no valor de Shapley para xogos cooperativos.

Strumbelj, E. & Kononenko, I. (2010) An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11, 1–18.

Importancia de atributos na clasificación

Strumbelj & Kononenko (2010)

1. Consideramos un individuo $\boldsymbol{x} = (x_1, \ldots, x_n)$ onde x_i é o valor do atributo i.
2. Dado un subconxunto de atributos, S, calcúlase a diferencia entre a predición cando só coñecemos eses valores do individuo, cuxos atributos pertencen a dito subconxunto, e a predición cando non se coñece ningún atributo.

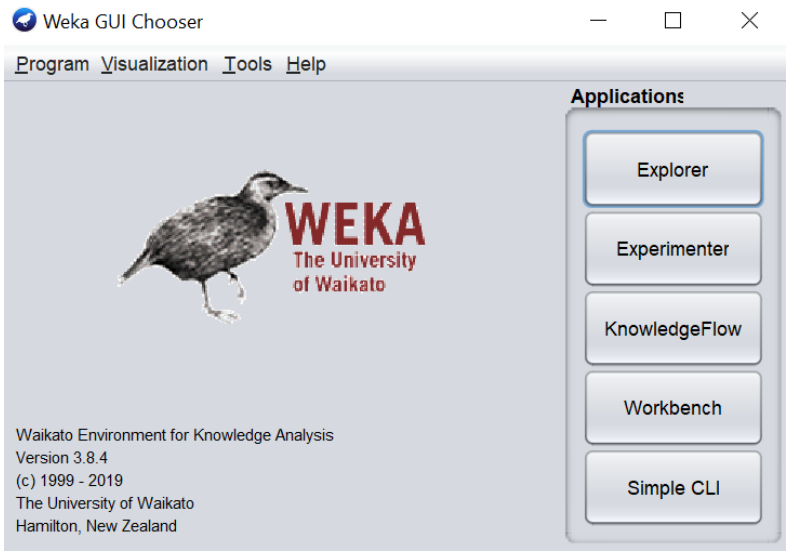
$$\Delta(S) = \frac{1}{|\mathcal{A}_{N\setminus S}|} \sum_{y \in \mathcal{A}_{N\setminus S}} f_c(\tau(x, y, S)) - \frac{1}{|\mathcal{A}_N|} \sum_{y \in \mathcal{A}_N} f_c(y)$$

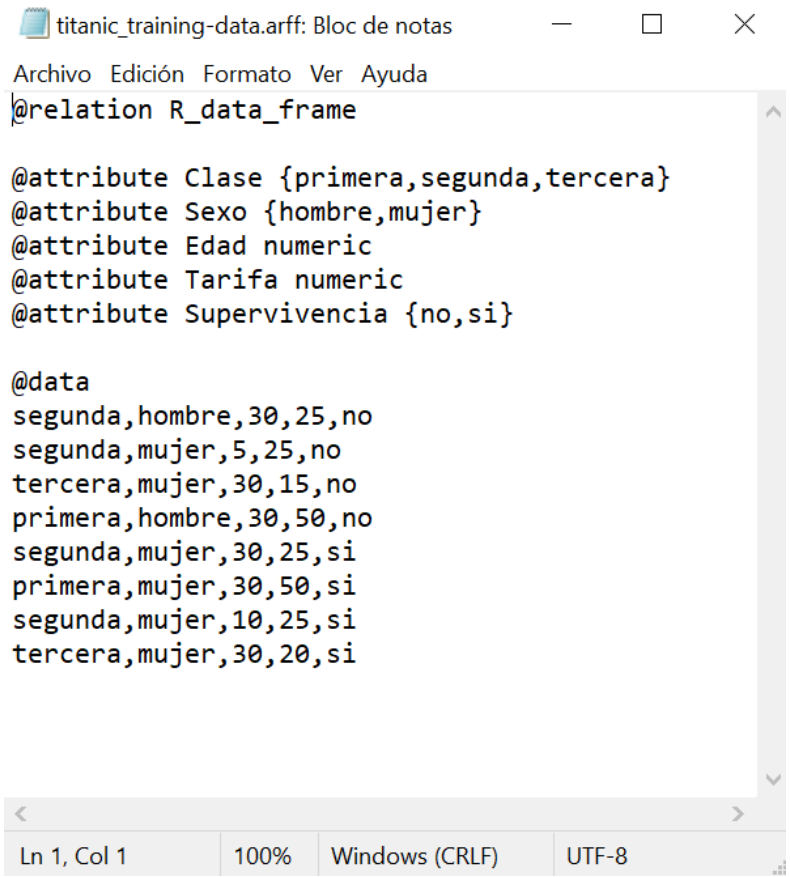
$$\tau(x, y, S) = (z_1, \ldots, z_n) \text{ con } z_i = \begin{cases} z_i = x_i & se \quad i \in S \\ \\ z_i = y_i & se \quad i \notin S \end{cases}$$

3. Calculamos o valor de Shapley do xogo anterior: cada coordenada representa a influencia dese atributo na clasificación.

Exemplo sinxelo: con Weka

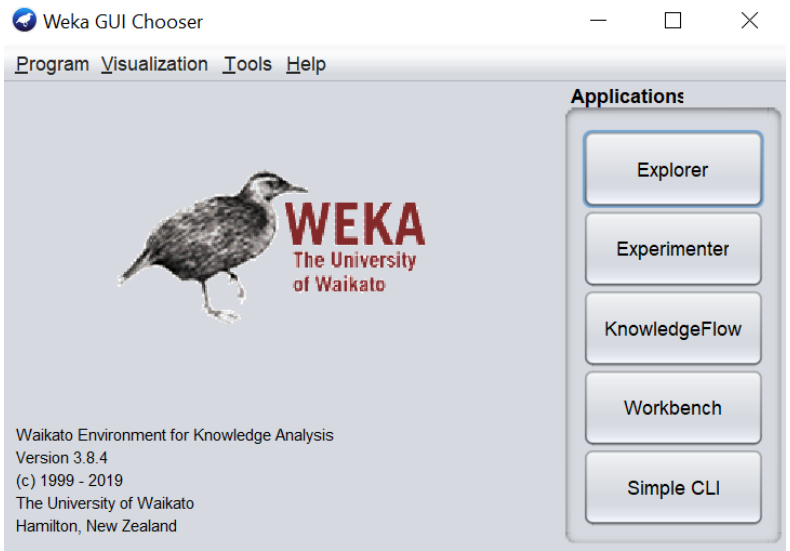
Titanic

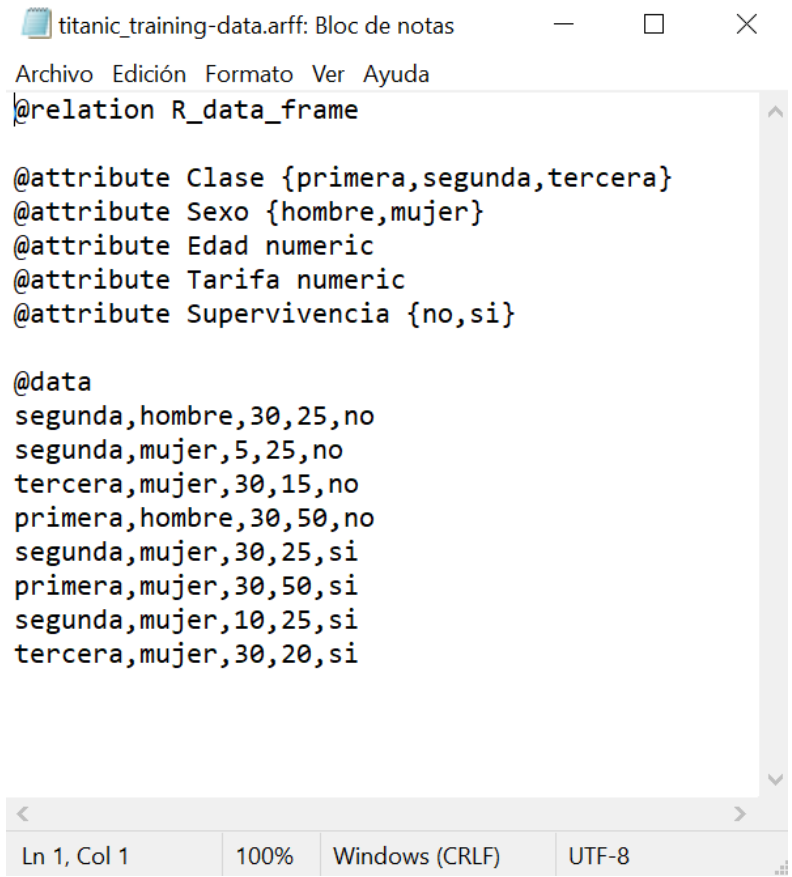




Exemplo sinxelo: con Weka

Titanic

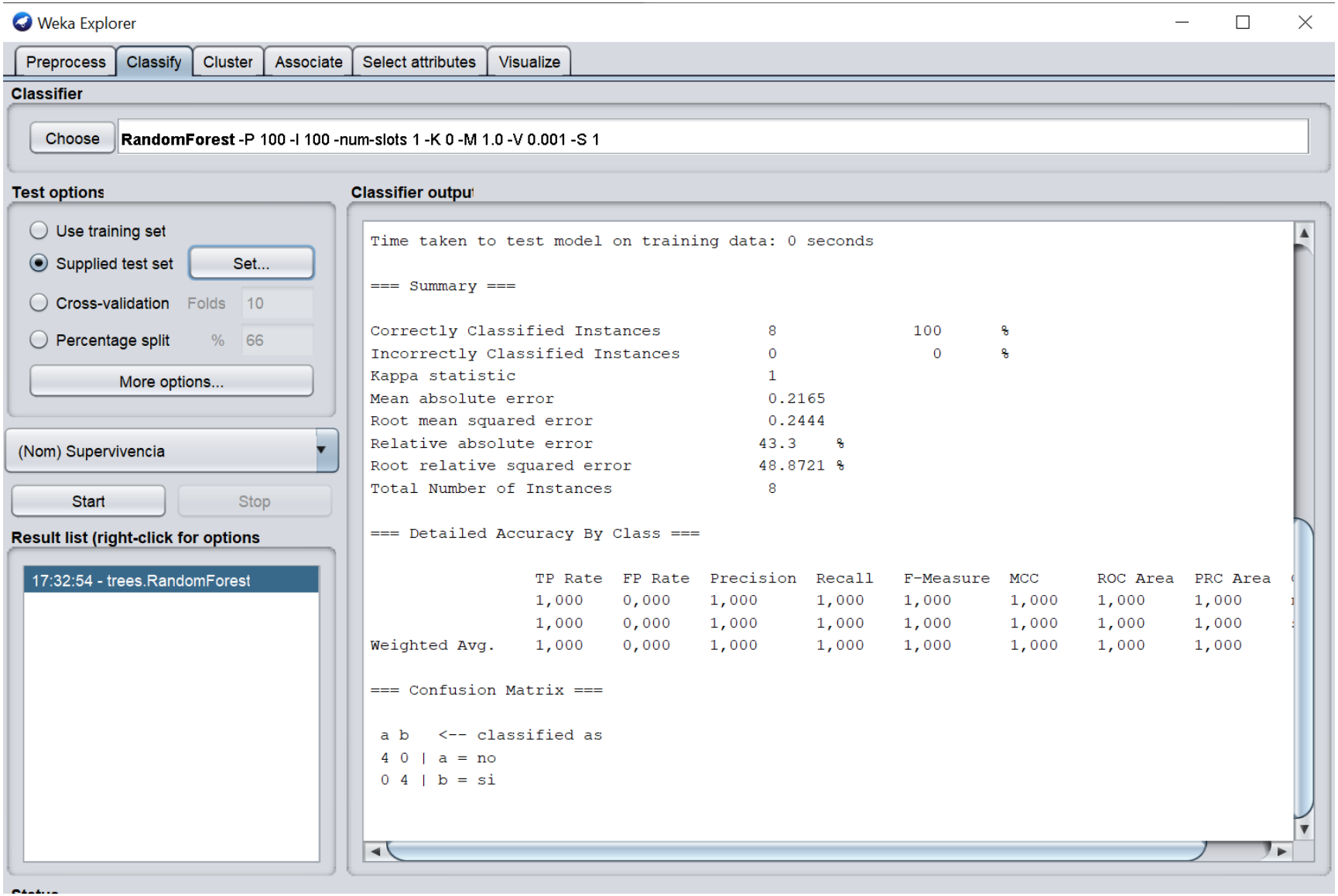




$$x = (\text{primera}, \text{mujer}, 30, 50)$$

Exemplo sinxelo: con Weka

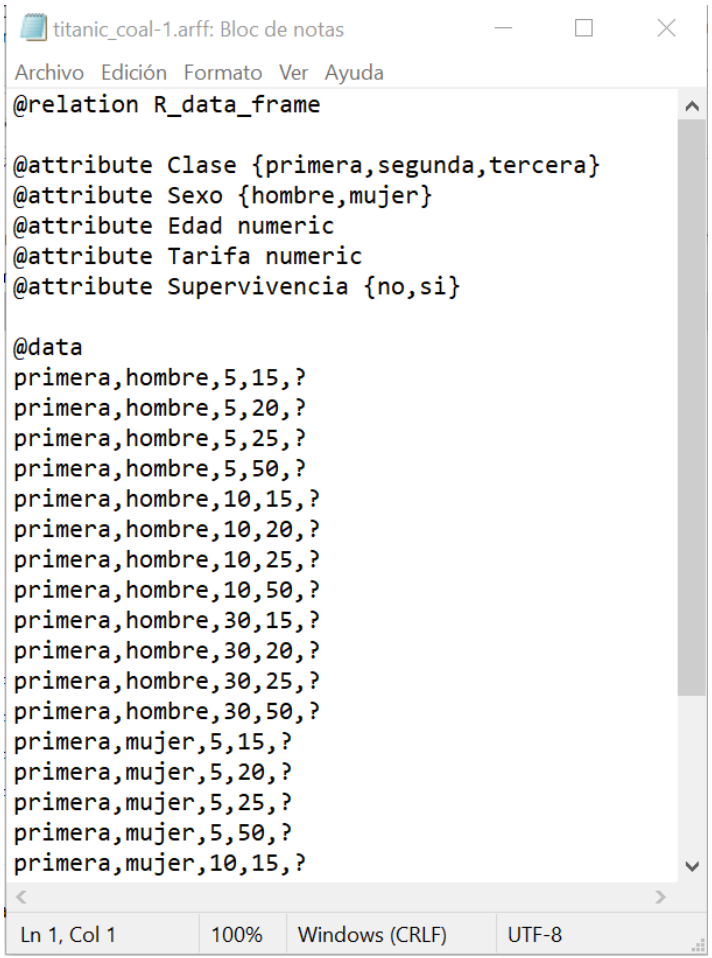
Titanic

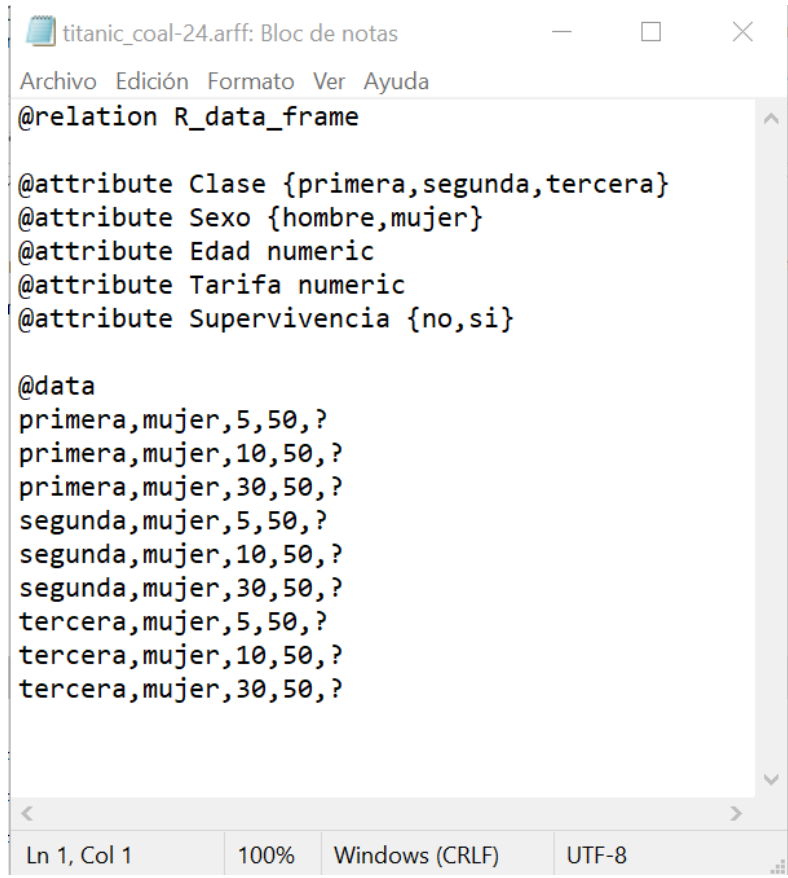


Exemplo sinxelo: con Weka

Titanic

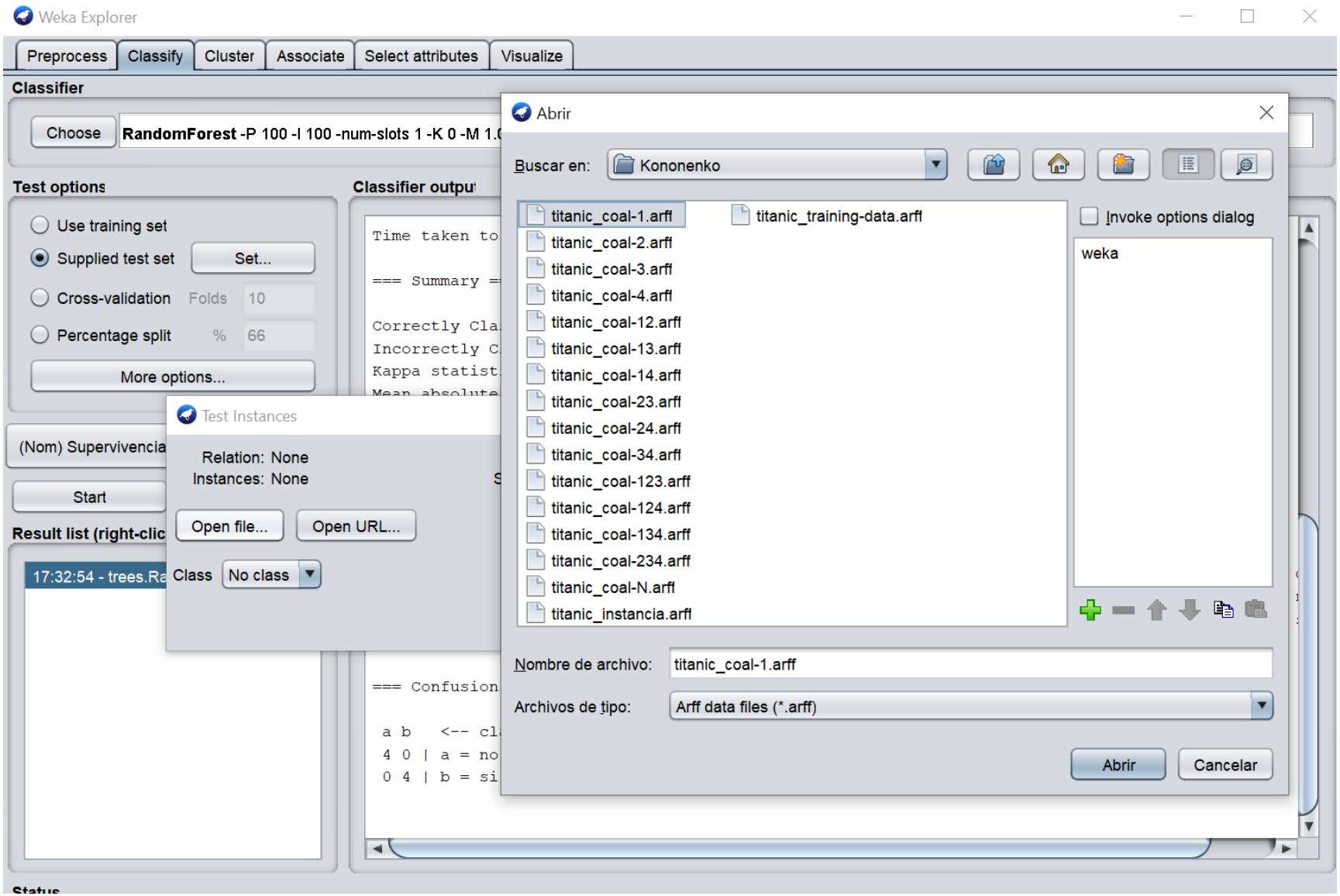
$$S \in \{\{1\},\{2\},\{3\},\{4\},\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\},$$
$$\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\},\{1,2,3,4\}\}$$





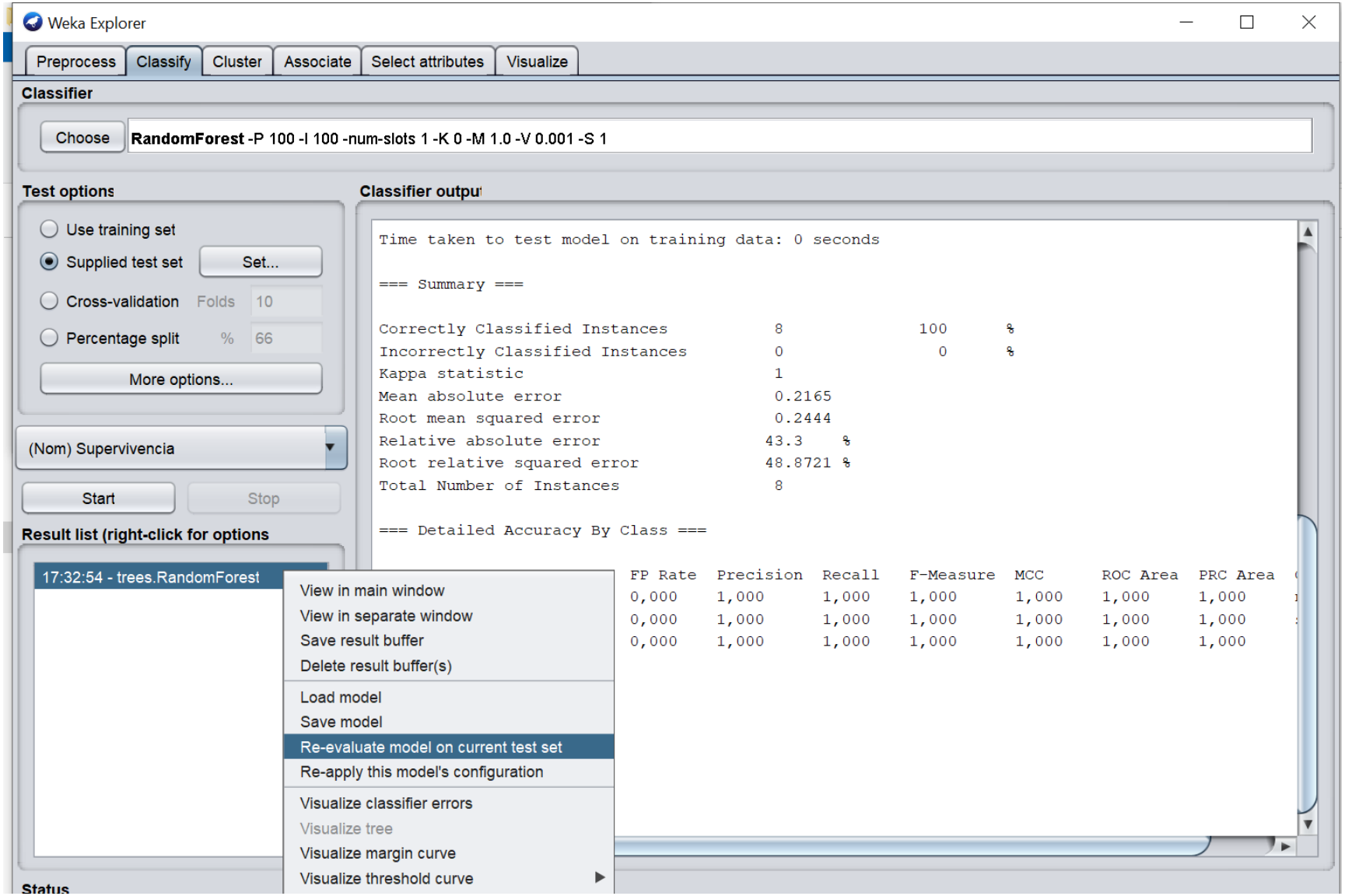
Exemplo sinxelo: con Weka

Titanic



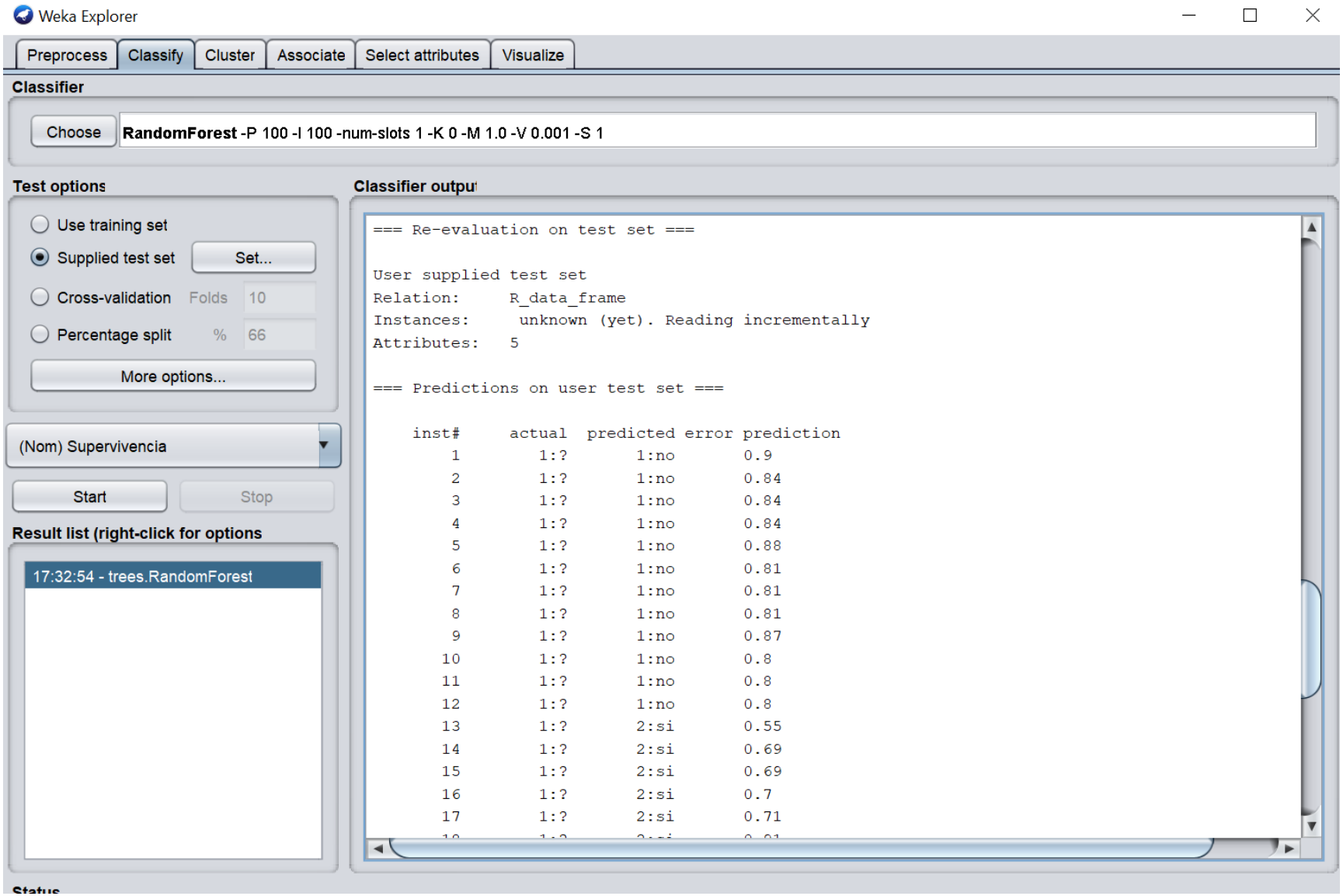
Exemplo sinxelo: con Weka

Titanic



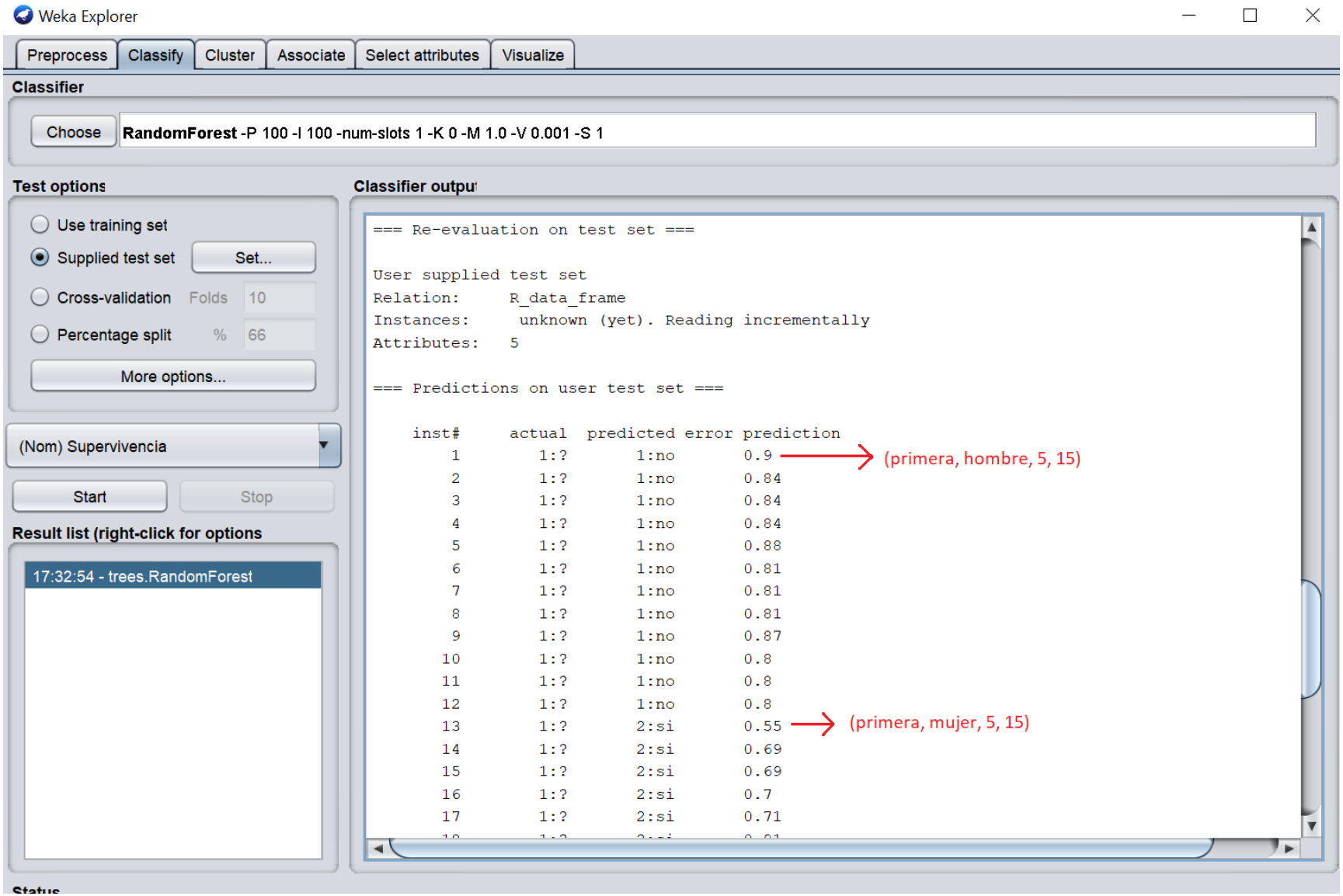
Exemplo sinxelo: con Weka

Titanic



Exemplo sinxelo: con Weka

Titanic



Exemplo sinxelo: con R

Titanic

$$\Delta(S) = \frac{1}{|\mathcal{A}_{N\setminus S}|} \sum_{y \in \mathcal{A}_{N\setminus S}} f_c(\tau(x, y, S)) - \frac{1}{|\mathcal{A}_N|} \sum_{y \in \mathcal{A}_N} f_c(y)$$

```
> library(Rweka)
> RF <- make_Weka_classifier("weka/classifiers/trees/RandomForest")
> modelo_rf <- RF(muestra_weka$Supervivencia ~ ., data = muestra_weka)
```

```
> library(ggm)
> S <- powerset(1:dim(X)[2], nonempty=T, sort=T)
```

```
> predict(modelo_rf, newdata=<X_S[[i]]>, type = c("class"))
> v[[i]] <- sum(pred[[i]][,5])==classlabel)/dim(pred[[i]])[1] - factor_fixo
```

```
> library(GameTheoryAllocation)
> Shapley <- Shapley_value(unlist(v), game = "profit")
```

COVID-19

Base de datos

- Temos un conxunto de 10454 pacientes de Galicia infectados con COVID-19 dende o 6 de marzo de 2020 ata o 7 de maio de 2020.
- O obxectivo é estudar a infuencia de varias características/atributos dos pacientes en tres variables resposta binarias de especial interese:
  - Necesidade de hospitalización.
  - Necesidade de ingreso en UCI.
  - Falecemento.
- Os atributos considerados son:
  - Idade: 0 (0-49 anos); 1 (50-64 anos); 2 (65-79 anos); 3 (80 anos en adiante).
  - Sexo: 0 (muller); 1 (home).
  - Patoloxías cardíacas: 0, 1, 2.
  - Patoloxías respiratorias: 0, 1, 2.
  - Patoloxías metabólicas: 0, 1, 2.
  - Patoloxías urinarias: 0, 1.

Consideramos o seguinte xogo:

$$v_x(S) = \frac{1}{|\mathcal{A}_{N\setminus S}|} \sum_{y \in \mathcal{A}_{N\setminus S}} f_c^p(\tau(x, y, S))$$

Consideramos o seguinte xogo:

$$v_x(S) = \frac{1}{|\mathcal{A}_{N\setminus S}|} \sum_{y \in \mathcal{A}_{N\setminus S}} f_c^p(\tau(x, y, S))$$

```
> predict(modelo_rf, newdata=<X_S[[i]]>, type = c("probability"))
```

Consideramos o seguinte xogo:

$$v_x(S) = \frac{1}{|\mathcal{A}_{N\setminus S}|} \sum_{y \in \mathcal{A}_{N\setminus S}} f_c^p(\tau(x, y, S))$$

```
> predict(modelo_rf, newdata=<X_S[[i]]>, type = c("probability"))
```

1. Para cada un dos atributos, j, e o seu valor, a_j, fixamos a submostra $\mathcal{M}_{a_j}$ cos individuos que teñen esas características.
2. Calculamos o xogo v_x para cada individuo x de $\mathcal{M}_{a_j}$.
3. Calculamos o valor de Shapley do xogo v_x, $\phi(v_x)$.
4. Promediamos os valores de Shapley, obtendo a nosa medida de influencia $I_j^{\Phi} = \frac{1}{|\mathcal{M}_{a_j}|} \sum_{x^i \in \mathcal{M}_{a_j}} \phi(v_{x^i})$.

$$T^{\Phi} := \sum_{k \in N} I_k^{\Phi}.$$

A cantidade T^{Φ} pertence a $[0, 1]$ e pódese interpretar como unha estimación da probabilidade de que a correspondente resposta dun individuo con atributo j igual a a_j sexa positiva.

Observación

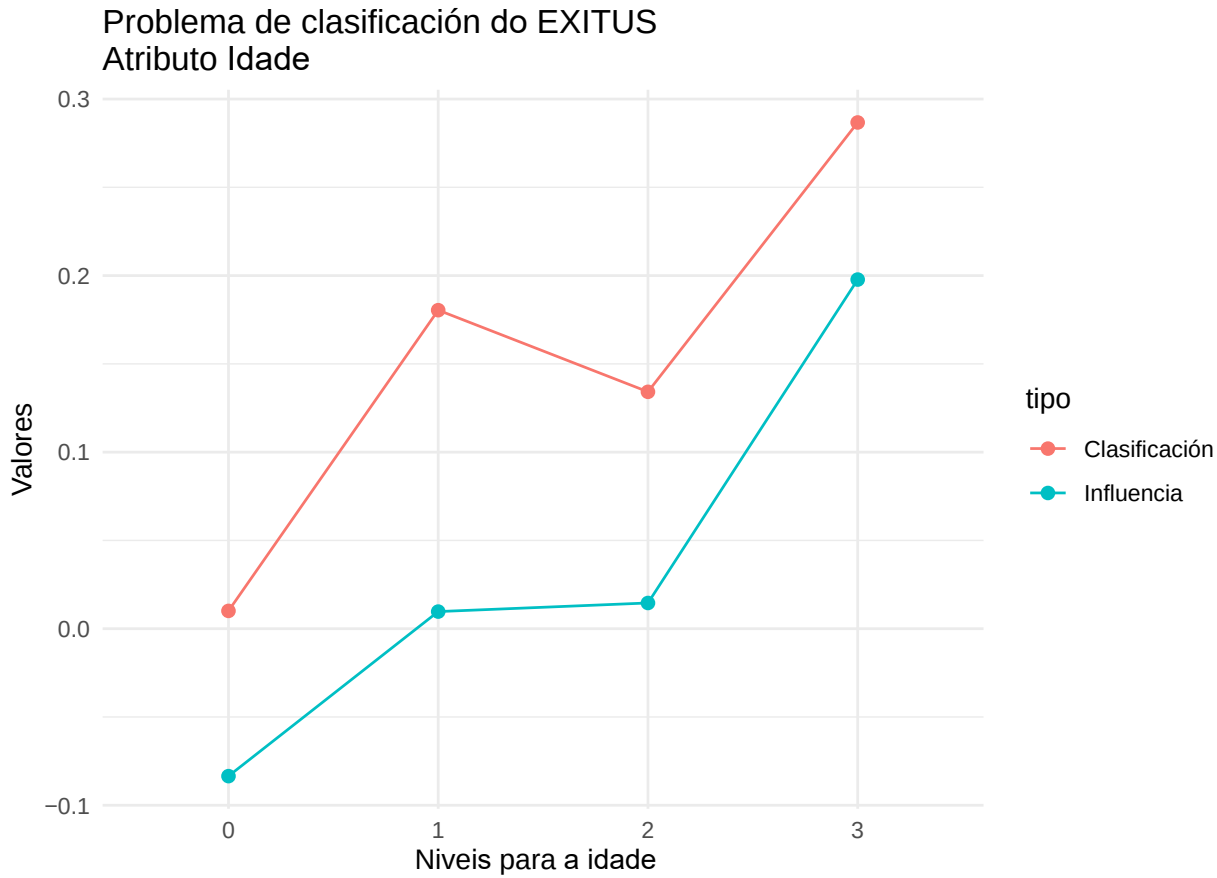
Nótese que I_j^{Φ} é a parte correspondente ao atributo j cando repartimos a cantidade T^{Φ} entre todos os atributos.

Deste xeito, a evolución dos números $\{I_j^{\Phi}\}$ e $\{T^{\Phi}\}$ é moi ilustrativa da influencia que os distintos valores de j teñen na resposta.

Por exemplo, se para un determinado valor observamos que ambos valores son próximos, é a vez que T^{Φ} é cercano a 1, podemos concluir que os individuos co atributo j igual a a_j teñen unha alta probabilidade de ser clasificados como positivos, e que iso débese principalmente ao atributo j.

COVID-19

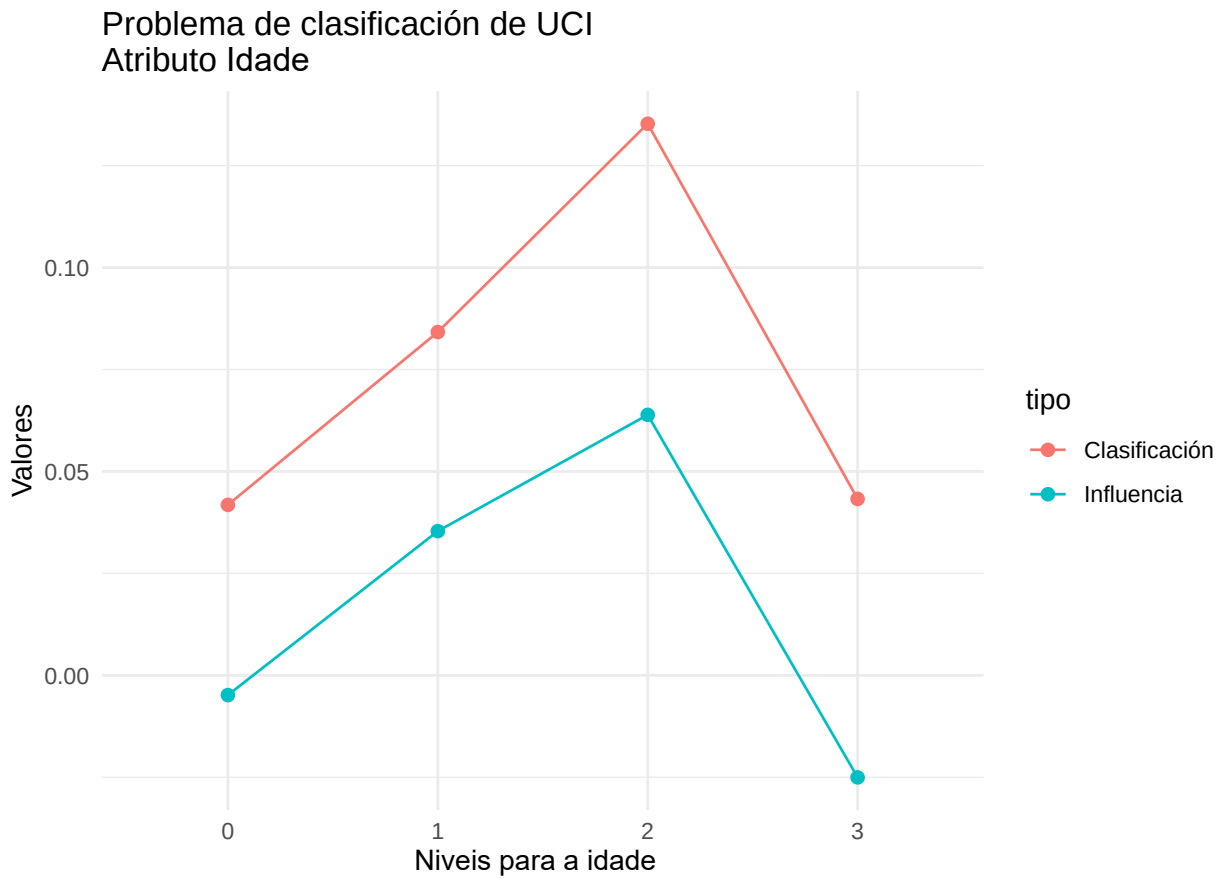
Idade - falecemento




```
> library(ggplot2)
```

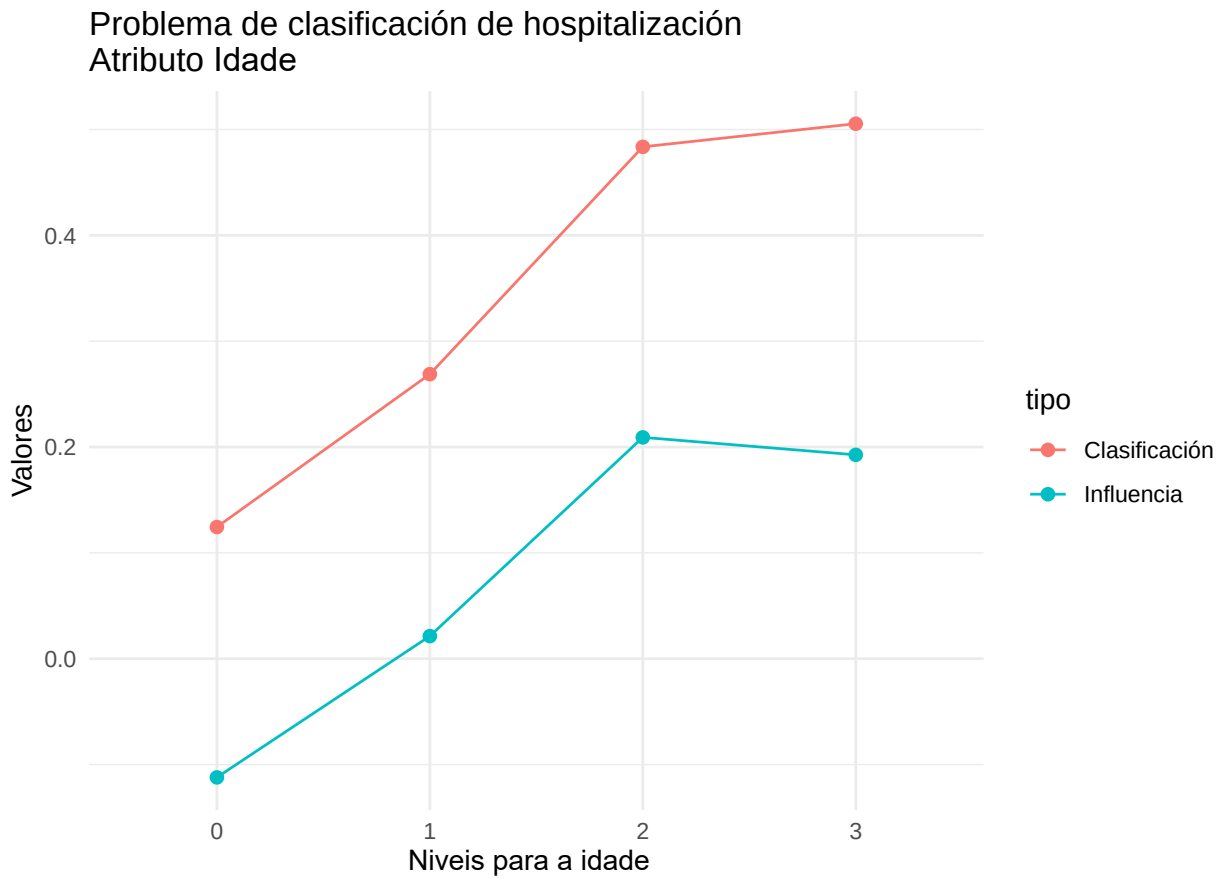
COVID-19

Idade - necesidade de ingreso en UCI



COVID-19

Idade - necesidade de hospitalización



O emprego de R na detección das características máis influentes na clasificación de pacientes infectados por COVID-19 en Galicia

VII Xornada de Usuarios de R en Galicia

Laura Davila Pena, Balbina Casas Méndez, Ignacio García Jurado

15 de outubro de 2020