Tipologia i cicle de vida de les dades

XAVIER JORDÀ MURRIA

ANNA SERENA LATRE

Màster Data Science UOC

MAIG 2019

Índex

PRESENTACIÓ

'TITANIC: *Machine Learning from Disaster*. '

Disponible a: https://www.kaggle.com/c/titanic

En aquesta pràctica procedim a una descàrrega de bases de dades en relació a 'TITANIC: Machine Learning from Disaster' en el repositori kaggle.com en una proposta analítica a concurs orientada a la recerca del model predictiu que maximitzi la capacitat predictiva.

Perquè és important i a quina pregunta es pretén respondre?

La pregunta a la qual es pretén respondre en aquest exercici d'integració, neteja, validació i anàlisis de les dades és l'adequació del dataset per a la seva disposició a algorismes de classificació, regressió i predicció per tal que sigui el més eficient possible en favor d'un rendiment òptim i que posarem a prova en les tècniques d'aprenentatge supervisat train & test en una avaluació dels resultats. En aquest cas tenim en consideració que el conjunt de dades en el fitxer train ve etiquetat en la variable classe en la classificació de supervivents i ofegats en l'esdeveniment tràgic de l'enfonsament del Titànic el qual disposarem a tasques de neteja, integració i validació, procedirem a una analítica descriptiva i finalment amb el qual crearem els models predictius els resultats dels quals posarem a prova.

BASE DE DADES

Les bases de dades descarregades en el repositori 'TITANIC: Machine Learning from Disaster' del repositori kaggle.com són :

- Train.csv 60 KB
- Test.csv 28 KB
- gendersubmission 4 B

CARACTERÍSTIQUES DELS DATASETS

Train.csv 60KB

Dataset dimension: 891 rows x 12 columns

Test.csv 28 KB

Dataset dimension: 418 rows x 11 columns

gendersubmission.csv 4 B

Dataset dimension: 418 rows x 2 columns

Variables en les columnes:

- PassengerID : numer assignat al passatger. Tipus: numèric. [0-891]
- Survived : supervivent =1, no supervivent = 0. Tipus: numèric. [1,0]
- Pclass : Tipus de viatge, estatus socio-econòmic. Tipus: numèric. [1,2,3]

- Name : Nom del passatger. Tipus char.
- Sex : sexe dels passatgers; Masculí 'male' i Femení 'female'. Tipus: char.
- Age : edat dels passatgers. Tipus : float. [0-80]
- Sibsp : Nº de relacions de cònjugues i germans abord. Tipus : numèric. [0,1,2,3,4,5,8]
- Parch : Nº de progenitors o descendència abord. Tipus : numèric. [0,1,2,3,4,5,6]
- Ticket : Número del bitllet. Tipus char.
- Fare : Tarifa del bitllet. Tipus: float. [0-513]
- Cabin : cabina assignada al passatger. Tipus : char. [A, B,C,D,E,F] codi alfanumèric.
- Embarked : port d'embarcació. Tipus : char. ['C','Q','S']

En el dataset test la variable Survived està omesa.

INTEGRACIÓ DE DADES

Proposta d'integració de dades que s'aplica als fitxers train.csv i test.csv:

1. Fusió dels fitxers test.csv i gendersubmision.csv.
2. Supressió de les variables nom del passatger, número del bitllet i del passengerID.
3. Conversió de les variables categòriques a factorial.

NETEJA DE DADES

Tasques de neteja de dades que s'aplica als fitxers train.csv i test.csv:

1. Duplicitat en les dades.
2. Consistència de les variables: gestió dels valors NA.

   Imputació de valors usant la mitja amb la informació de totes les variables numèriques, categòriques i semi-continues.

3. Validesa de les variables: valoració i tractament de la inconsistència de les dades.

VALIDESA DE LES DADES

1. Resum estadístic per a les variables continues.
   - Tendències centrals de dispersió.
2. Valors atípics: valoració d'outliers en variables quantitatives.
   - Boxplots.
   - Shapiro-Wilk test.
   - Gràfics per a la normalitat.
3. Valoració de la validesa de les categories.
4. Identificació de les categories en la factorització.

Podeu trobar les tasques d'integració, neteja i validesa de les dades amb els comentaris annexos en els documents:

Fitxer_Train.html 917 KB

Fitxer_test.html 918 KB

Fitxer_Train.rmd 12 KB

Fitxer_Test.rmd 13 KB

Els datasets que resulten de les tasques d'integració, neteja i validació aptes per a la fase analítica són:

Train_T.csv 41 KB

Test_T.csv 19 KB

ANÀLISI DESCRIPTIVA:

En aquesta fase d'anàlisi descriptiva desestimem les proves inferencials donat que les variables numèriques més rellevants Age i Fare no entren en un comportament d'una distribució normal, i optem per els arbres de decisió que són models algorísmics d'agrupament amb caràcter regressiu en aquest cas, i que disposen d'un fort potencial visual per a la descripció de tendències en les variables en una fase inicial d'aproximació.

La metodologia de treball és la de presentar una arbre de decisió amb totes les dades completes en un gràfic que resulta en un arbre complex inicialitzat en la variable classe Survived en funció del succés èxit o fracàs. La solució ens permet procedir amb arbres de menys profunditat per ordre d'importància segons els nivells presentats en l'arbre inicial. D'aquesta manera ens aproximem al cas amb resolucions gràfiques a mode d'arbres que també les hem acompanyat d'altres solucions gràfiques en les variables categòriques per tal d'il.lustrar les principals estructures en les dades.

Podeu trobar les tasques d'analítica descriptiva en els fitxers:

Pràctiques_analítiques.html 1044 KB

Pràctiques_analítiques.rmd 13 KB

MODELS DE REGRESSIÓ, CLASSIFICACIÓ I PREDICCIÓ:

El models algorísmics proposats per a tasques de predicció i les proves Train&Test són: model de regressió logística i el model SVM Suported Vectorial Machine.

<u>Model de regressió logística:</u>

- Entrenament del model de regressió logística.
- Valoració del model: Accuracy.
- Proves de predicció Train&Test per al model:
  - Matriu de confusió.
  - Analítica ROC.

La diversitat en la tipologia de dades en les variables ens aproxima a una complexitat que resolem en el model de regressió logística, és un model que per la seva simplicitat inicialment ens permet una coneixement més profund de la tipologia i comportament de les variables en una analítica de més fàcil comprensió.

Model SVM: Machine Learning.

- Entrenament del model Suported Vectorial Machine.
- Valoració del model: Accuracy.
- Proves de predicció Train&Test per al model:
  Matriu de confusió.
  Analítica ROC.

L'algorisme SVM junt amb les Xarxes Neuronals són els més utilitzats, i optem per el SVM per les seves prestacions de rapidesa en l'execució. L'encert i la capacitat optima d'aquest model en els resultats de la prova Train&Test fa que donem per acabada la recerca d'un model que resolgui satisfactòriament amb les tasques de classificació i predicció.

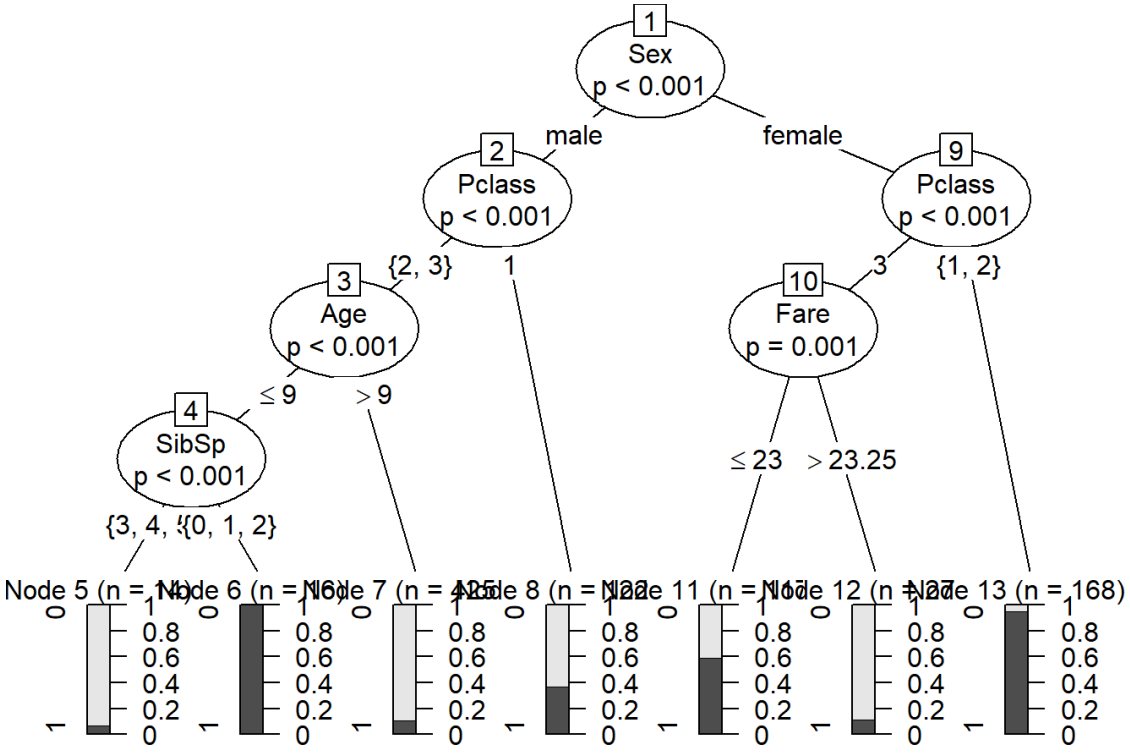
Podeu trobar les tasques de modelització en els fitxers:

Pràctiques_analítiques.html	1044 KB
Pràctiques_analítiques.rmd	13 KB

PRESENTACIÓ DELS RESULTATS EN TAULES I GRÀFIQUES

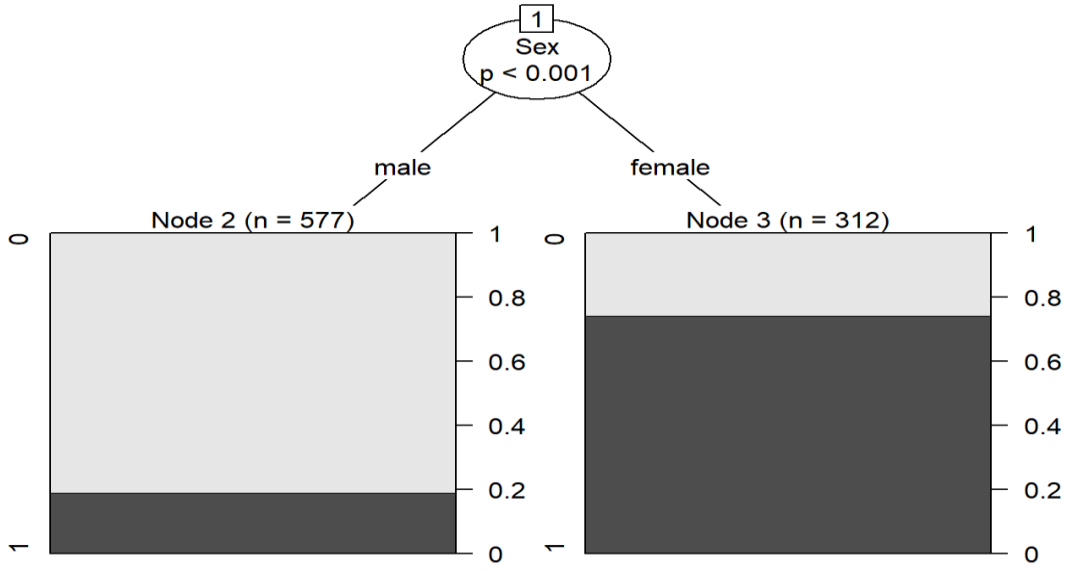
Analítica descriptiva en el conjunt Train_T.csv: arbres de decisió.

Arbre de decisió en la totalitat del conjunt de dades Train_T partint de la variable Survived com a punt de partida:



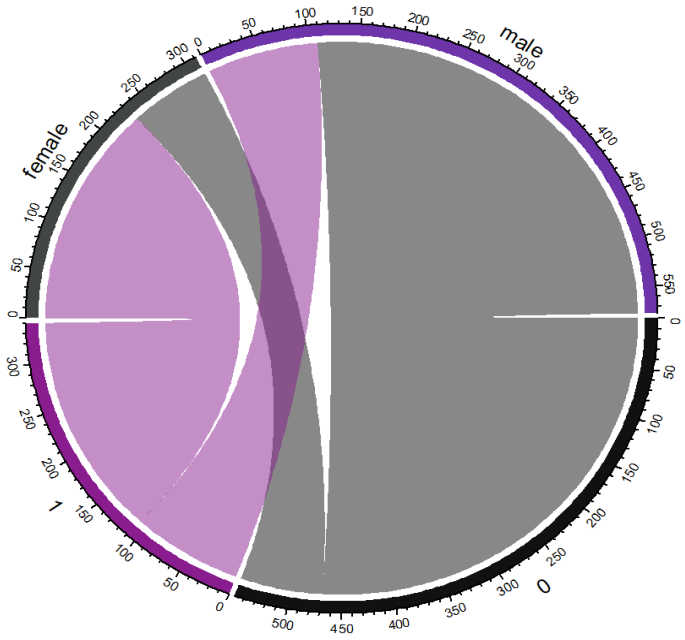
Com hem comentat anteriorment ens interessa la seqüenciació de les variables en els nivells per ordre d'importància en relació a la variable Survived i presentades en ordre descendent.

En el ***primer nivell*** trobem la variable Sex i per tant procedim amb l'arbre de classificació en les variables Survived i Sex:

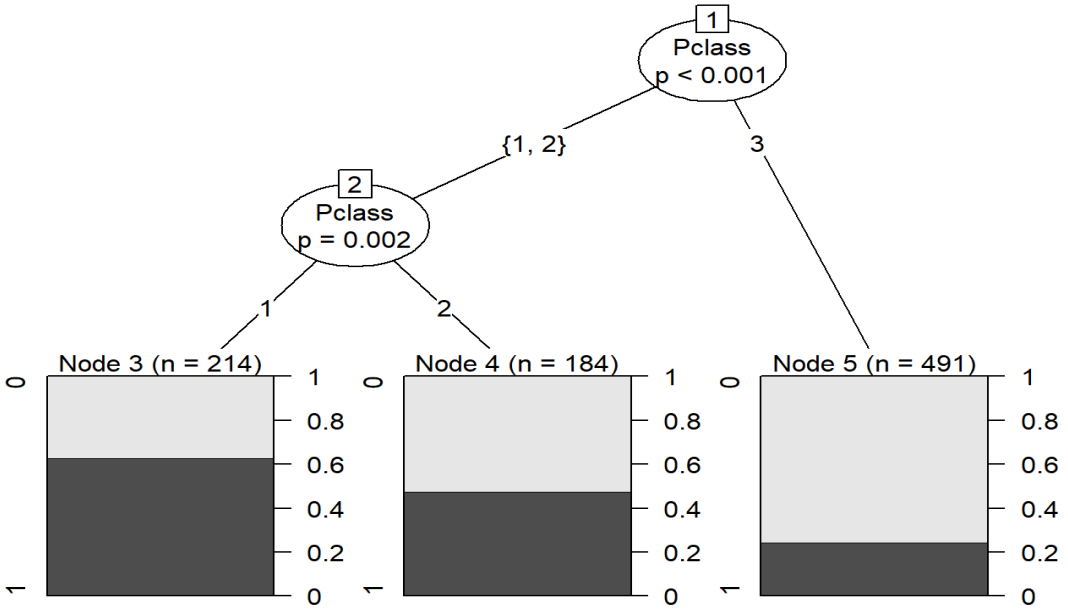


El sector femení disposa d'un 78% aproximadament en el succés èxit de supervivència, mentre que en el sector masculí aquest succés és d'un 20% aprox.

Gràfic circular en les variables Survived i Sex:

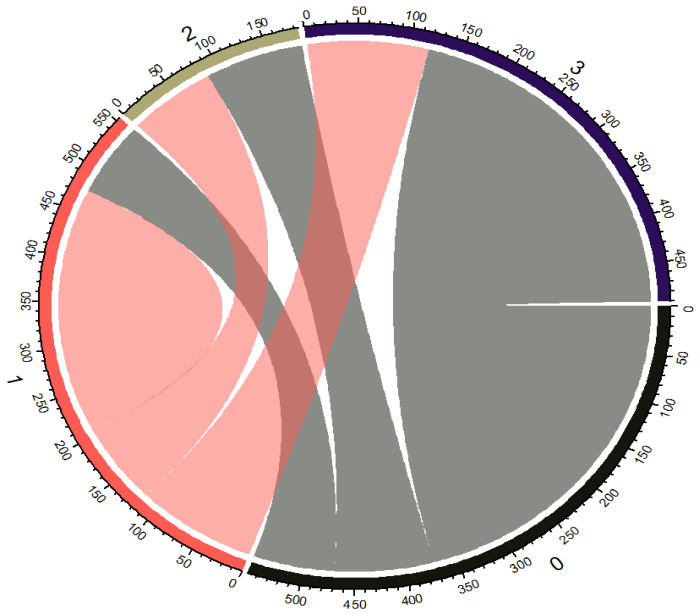


En un ***segon nivell*** trobem la variable Pclass i per tant procedim amb l'arbre de decisió en les variables Survived i Pclass:

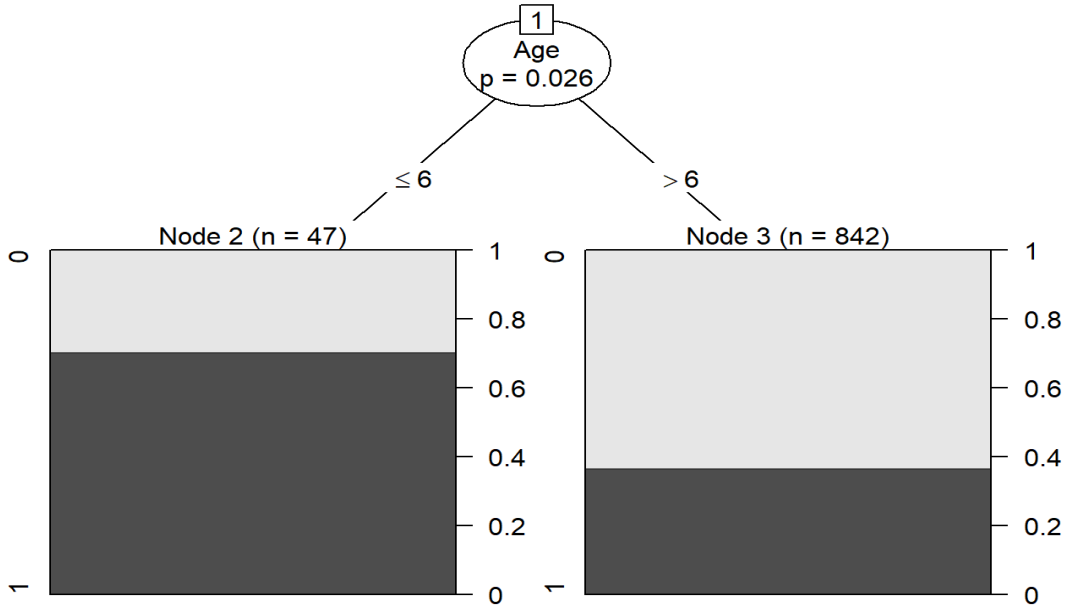


Observem com l'èxit es concentra en el segement de primera classe amb poc més d'un 60% de supervivència, mentre que en en la segona classe aconsegueix prop d'un 50% de supervivència però és un segment minoritari de 184 individus, i finalment el segment tercera classe disposa només d'un 23% d'èxit en la supervivència i és el segment majoritari en 491 individus.

Gràfic circular en les variables Survived i Pclass:

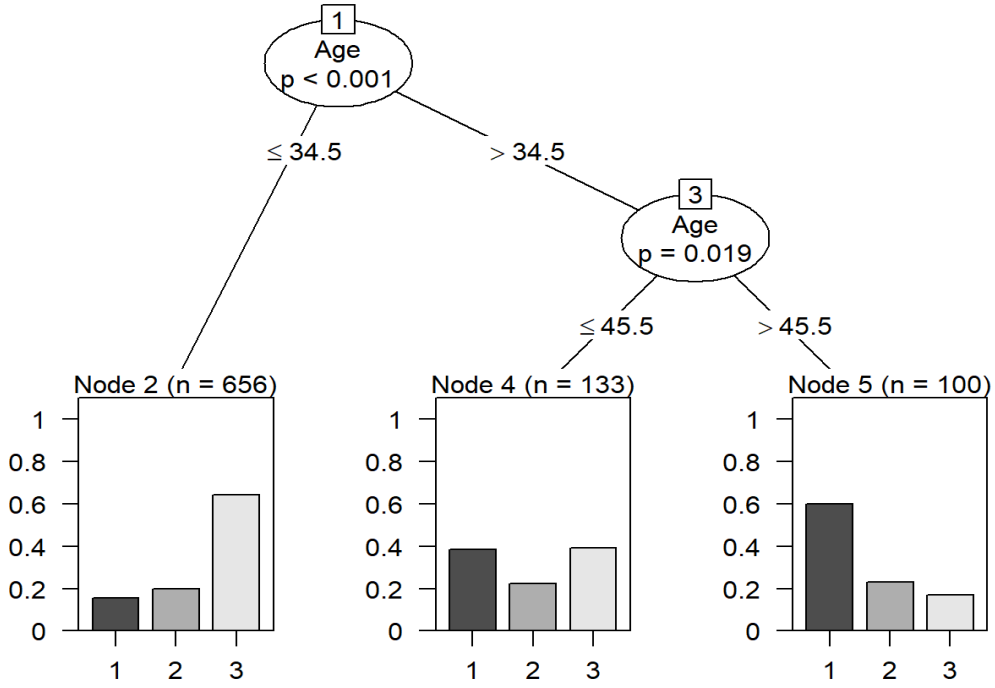


En un ***tercer nivell*** trobem la varaible Age i per tant procedim amb la representació gràfica de l'arbre decisió amb les variables Survived i Age:



En el gràfic observem com els infants de 6 o menys anys obtenen un èxit de supervivència en un 70%, mentre que a partir de sis anys les probabilitats de supervivència són de menys d'un 40%.

Arbre de decisió en les variables Pclass i Age:



Hem volgut també incloure l'arbre de decisió per a les variables Pclass i Age, en el qual constatem que els passatgers de tercera categoria es concentren en el primer segment de persones de menys de 35 anys, mentres que en l'interval de 35 a 46 anys existeix una diversitat en les tres classes en un repartiment de la densitat de manera molt equitativa entre els tres grups de les classes, i finalment en el segment de persones de més de 46 existeix una forta concentració en la classe primera.

Models de regressió, classificació i predicció: resultats en la regressió logística.

En el model de regressió logística presentem com a variable independent Survived i les variables explicatives són: Pclass,, Sex, Age, SibSp, Fare i Embarked.

Hem hagut de suprimir la variable Parch perquè no era vàlida per a les proves Train&Test.

Resultat del model de regressió logística:

```
glm(formula = Survived ~ ., family = binomial(link = "logit"),
    data = Train_T)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7993  -0.6002  -0.4125   0.6083   2.4862
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.925681   0.477894   8.215  < 2e-16 ***
Pclass2      -0.983963   0.298553  -3.296 0.000982 ***
Pclass3      -2.126366   0.295880  -7.187 6.64e-13 ***
Sexmale      -2.645883   0.197588 -13.391  < 2e-16 ***
Age          -0.040708   0.008081  -5.038 4.72e-07 ***
SibSp1        0.140612   0.212147   0.663 0.507456
SibSp2       -0.184671   0.521856  -0.354 0.723434
SibSp3       -2.009086   0.712034  -2.822 0.004778 **
SibSp4       -1.589729   0.736990  -2.157 0.031001 *
SibSp5      -15.921629 960.694512  -0.017 0.986777
SibSp8      -15.796329 758.696121  -0.021 0.983389
Fare          0.001942   0.002309   0.841 0.400443
EmbarkedQ     0.048253   0.381772   0.126 0.899422
EmbarkedS    -0.381258   0.242040  -1.575 0.115213
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.8  on 888  degrees of freedom
Residual deviance:  773.7  on 875  degrees of freedom
```

AIC:

```
AIC: 801.7

Number of Fisher Scoring iterations: 15
```

Matriu de confussió:

```
titanic_predict   0   1
              0 262  30
              1   4 122
```

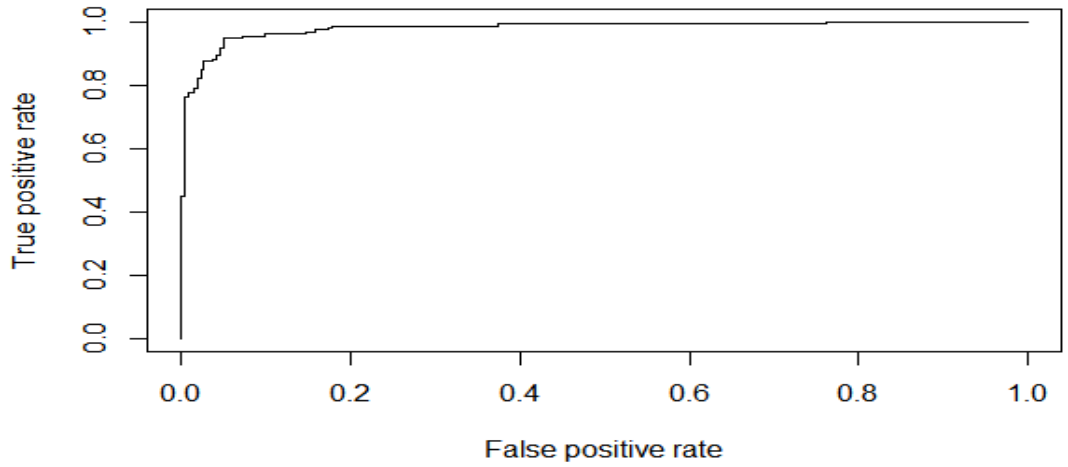
Precisió del model de regressió logística:

```
[1] "Accuracy 0.91866028708134"
```

La precisió del model de regressió logística és molt bona amb una Accuracy del 0'92 i això vol dir que la capacitat predictiva del model registra un 92% d'encert aprox. en la classificació.

En la matriu de confusió observem com el model classifica errròniament 4 ofegats que resulten ser supervivents i altrament classifica errròniament 30 supervivent que resulten ser ofegats.

Corba ROC del model de regressió logística:



En la proximitat de la corba ROC a l'angle 1.0 de l'esquerra en les ordenades True Positive rate significa que el model es troba en uns resultats òptims molt bons, i que en relació a una hipotètic angle de 45º en diagonal amb els eixos el model té una capacitat explicativa molt vàlida en una àrea extensa, i l'AUC per valor 0.97 amb bones expectatives d'encert en els TP.

Valor AUC:

```
[1] 0.975836
```

Models de regressió, classificació i predicció: resultats en el model SVM.

Resultat del model SVM:

```
Support Vector Machine object of class "ksvm"

SV type: eps-svr  (regression)
 parameter : epsilon = 0.1  cost C = 1

Linear (vanilla) kernel function.

Number of Support Vectors : 410

Objective Function Value : -353.5994
Training error : 0.825748
```

Matriu de confussió:

```
svm.predict   0   1
          0 266   0
          1   0 152
```

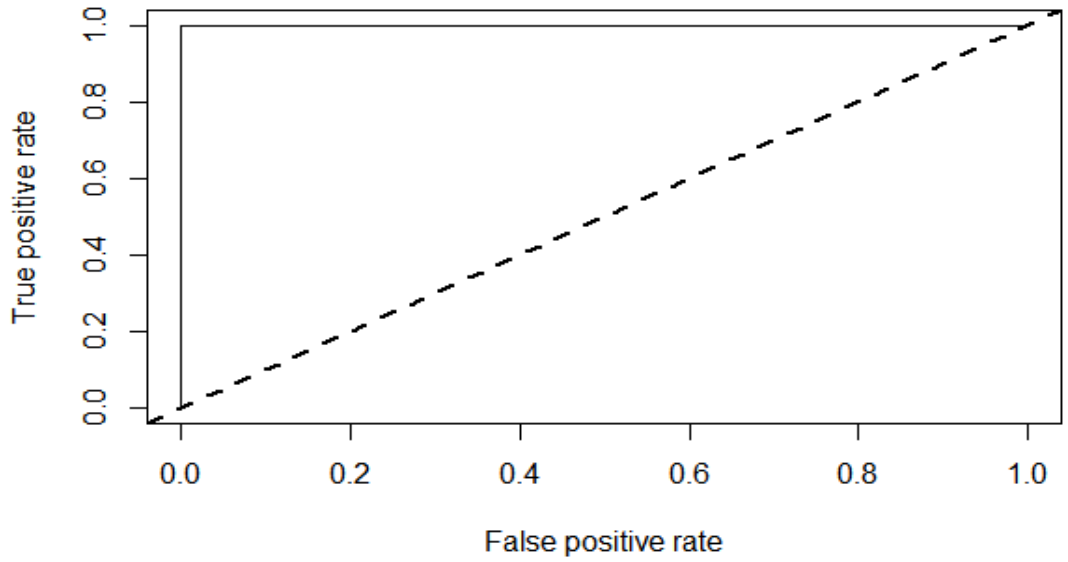
Accuracy:

```
[1] "Accuracy 1"
```

La matriu de confusió del model SVM mostra que no hi ha error en la classificació TP i TN, el model procedeix amb una classificació òptima sense component erràtica.

L'accuracy del model SVM és òptima ja que pren un valor d'1, i d'aquesta manera podem afirmar que el model té una capacitat de classificació i predicció de les dades en un 100% de fiabilitat segons les proves train&test dutes a terme.

En la corba ROC observem com el punt òptim de la corba es troba ubicada en l'angle esquerre proper a 1.0 de manera que s'optimitza les possibilitats classificadores i predictives de l'algorisme SVM en la mostra Train_T. Traçant una diagonal la corba ROC assumeix tota l'àrea dels vertaders positius de manera que el seu rendiment és òptim en aquest model, i així ho corrobora el valor AUC que pren un valor d'1 i significa que el model és propici a l'encert.



En la corba ROC observem com el model es troba en l'anlge 1 de l'esquerra optimmitzant la capacitat predictiva i miniitzant l'error a 0.

Valor AUC del model SVM:

```
Formal class 'performance' [package "ROCR"] with 6 slots
  ..@ x.name      : chr "None"
  ..@ y.name      : chr "Area under the ROC curve"
  ..@ alpha.name  : chr "none"
  ..@ x.values    : list()
  ..@ y.values    :List of 1
  .. ..$ : num 1
  ..@ alpha.values: list()
```

AUC:

```
## [1] 1
```

Podeu veure els resultats amb els comentaris annexos en els fitxers :

Pràctiques_analítiques.html 1044 KB

Pràctiques_analítiques.rmd 13 KB

RESOLUCIÓ DEL PROBLEMA

Conclusions.

Els arbres de decisió són models que no aprenen en el sentit que no optimitzen el rendiment en la creació de l'algorisme, però altrament resulten de gran utilitat en les tasques de classificació i regressió perquè despleguen un gran potencial per a la comprensió visual de la segmentació i ordenació de les dades, i en aquest cas ens hem servit d'aquest algorisme en una fase d'analítica descriptiva en una aproximació a les estructures subjacents que hi ha en les dades.

Segons aquesta fase d'analítica descriptiva, en un primer nivell situem la variable sexe com la més rellevant per a l'explicació del succés èxit o fracàs segons la supervivència o no dels individus, que en aquest cas es manifesta favorable en el segment femení amb un 60% de probabilitat de supervivència aproximadament tot i que és un segment inferior en nombre al segement masculí. En segon nivell s'ha situat la variable Pclass com a variable rellevant i que resulta determinant el l'èxit de supervivència sobretot en les persones de primera classe, en el segment de segona classe també és un factor substancial en la determinació de l'èxit o fracàs a parts iguals, mentre que en la tercera categoria té més pes el succés el fracàs. Finalment en un tercer nivell de rellevància s'ha situat la variable edat com a factor crític i l'arbre resol que per menors de 6 o menys anys la probabilitat de supervivència és d'un 70% aproximadament mentre que per a majors de 6 anys la probabilitat d'èxit se situa en un 20% escàs.

Tot fa pensar que les consignes de dones i nens primer de tot a bord dels bots salvavides varen tenir el seu fruit en la representació del segment femení i també per als infants de sis o menys anys. Altres consideracions al respecte es que en moltes ocasions va prevaldre la classe de les persones i així queda reflectit en el segment d'infants de sis o menys anys amb les probabilitat d'èxit d'un 70% aproximadament entre els quals només un infant en primera categoria va morir.

El model de regressió logística ofereix un bon rendiment encara que no òptim i es tracta d'una modelització que serveix de gran utilitat per a entendre i valorar la naturalesa de les variables que entren en joc i per a disposar-les per a altres algorismes de més precisió o complexitat.

El model que presenta millor rendiment és el SVM que junt amb les xarxes neuronals són els més utilitzats per el bon rendiment que ofereixen en quant a eficiència perquè són optimitzadors, si bé el SVM és un algorisme més simple però que funciona amb molta rapidesa en la realització i processament de còmputs en a màquina. En aquest cas el model SVM amb una accuracy d'un 100% té una plena capacitació per a tasques de classificació de les dades i també en tasques predicció. Es podria tractar d'un model sobreentrenat però en tractar amb la població universal de les dades és un aspecte que no ha de preocupar i queda resolt.

Quina és la finalitat del nostre estudi ?

L'estudi pretén una adequació de les dades per a les tasques analítiques que contempla les fases d'integració neteja i validació de les dades, que posteriorment han de servir a una fase d'analítica descriptiva en l'exploració de la naturalesa de les dades per tal d'aproximar-nos a la complexitat del cas i finalment ens habilitin en la construcció de models algorísmics amb capacitat de classificació i predicció.

La finalitat del nostre estudi es centra en la cerca d'un model òptim en quant a capacitat de representació de la variabilitat de les dades i que pugui funcionar amb eficiència en les tasques de classificació, regressió o bé predicció,... la intenció és la de procurar les tècniques i mecanismes que millor puguin funcionar per a la optimització del rendiment dels algorismes aplicats en els casos en qüestió.

Podem afrontar la resolució del problema ?

Concloem l'informe de l'estudi en una valoració positiva, afirmant que sí podem afrontar la resolució del problema, i davant d'una entrada aleatòria d'un vector podem predir si aquest input es tracta d'un supervivent o bé un nàufrag amb una precisió del 100%. Per altra banda acomplim amb la petició encomanda de dur a terme tasques d'integració, neteja i validació amb rigorositat, així com d'utilitzar almenys tres algorismes diferents en les tasques analítiques.

CODI

Codi R per a la integració, neteja i validació de la base de dades Train.csv disponible a:

https://github.com/AnnaSerenaLatre/TipologiaPac2/blob/master/Rmd_Files/Fitxer_Train.Rmd

https://github.com/XavierJordaMurria/TipologiaPac2/blob/master/Rmd_Files/Fitxer_Train.Rmd

Codi R per a la integració, neteja i validació de la base de dades Test.csv disponible a:

https://github.com/AnnaSerenaLatre/TipologiaPac2/blob/master/Rmd_Files/Fitxer_Test.Rmd

https://github.com/XavierJordaMurria/TipologiaPac2/blob/master/Rmd_Files/Fitxer_Test.Rmd

Codi R per a les pràctiques analítiques:

https://github.com/AnnaSerenaLatre/TipologiaPac2/blob/master/Rmd_Files/Pràctiques_analítiques.Rmd

https://github.com/XavierJordaMurria/TipologiaPac2/blob/master/Rmd_Files/Pràctiques_analítiques.Rmd

DOCUMENTS ANNEXOS

Trobareu els datasets descarregats en la plataforma kaggle.com en el directori input_data.

- Train_T.Rmd 12 KB
- Test_T.Rmd 13 KB
- Pràctiques_analítiques.Rmd 13 KB
- Train_T.html 918 KB
- Test_T.html 918 KB
- Pràctiques analítiques.html 1044 KB
- PAC2_TCVD.pdf 267 KB

DATASETS RESULTANTS

Trobareu els datasets resultants en el directori output_data.

- Test_T.csv 19 KB
- Train_T.csv 41 KB

LINKS REPOSITORI GITHUB

Els arxius, fitxers i documentació són disponibles en el repositori github.com:

https://github.com/XavierJordaMurria/TipologiaPac2/

https://github.com/AnnaSerenaLatre/TipologiaPac2/

REFERÈNCIES

Stackoverflow (2019) Solucions data mining i analítiques. Disseny del lloc web/logo©2019 Stack Exchange Inc: contribucions d'usuaris llicenciats en cc by-sa 3.0. Disponible a:

https://es.stackoverflow.com/

Statmethods.net (2019) Quick R – by DataCamp. Kaabacoff, R.I.; 'ANOVA' . Copyright © 2012 Robert I. Kabacoff, Ph.D. Disponible a:

https://www.statmethods.net/stats/anova.html

Rpubs.com (2019) Brought to you by RStudio. Li, E.C.; 'Titanic Survival Analysis Using Logisic Regression'. Easy web publishing. RStudio™ Support. Disponible a:

https://rstudio-pubs-static.s3.amazonaws.com/283447_fd922429e1f0415c89b93b6da6dc1ccc.html

Taula de contribucions al treball:

Contribucions	Signa
Recerca prèvia	X.J.M, A.S.L.
Redacció de resposta	X.J.M, A.S.L.
Desenvolupament del codi	X.J.M, A.S.L.