Tipologia i cicle de vida de les dades

XAVIER JORDÀ MURRIA

ANNA SERENA LATRE

Màster Data Science UOC

MAIG 2019

Índex

	Pàg.
PRESENTACIÓ	**1**
BASE DE DADES	**1**
CARACTERÍSTIQUES DELS DATASETS	**1**
INTEGRACIÓ DE DADES	**2**
NETEJA DE DADES	**2**
VALIDESA DE LES DADES	**2**
ANÀLISI DESCRIPTIVA I INFERENCIAL	**2**
MODELS DE REGRESSIÓ, CLASSIFICACIÓ I PREDICCIÓ	**3**
PRESENTACIÓ DELS RESULTATS EN TAULES I GRÀFIQUES	**3**
RESOLUCIÓ DEL PROBLEMA	**10**
CODI	**11**
DOCUMENTS ANNEXOS	**12**
DATASETS TRACTATS	**12**
LINKS REPOSITORIS GITHUB	**12**
REFERÈNCIES	**12**

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. Validesa de les variables: gestió dels valors NA.

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

3. Valoració i tractament de la inconsistència de les dades.
4. Valors atípics: valoració d'outliers en variables quantitatives.
   - Boxplots.
   - Taula d'estimació de tendències centrals de dispersió.

VALIDESA DE LES DADES

1. Valoració de la validesa dels outliers en els intervals.
2. Valoració de la validesa de les categories.
3. Identificació de les categories en la factorització.

ANÀLISI DESCRIPTIVA I INFERENCIAL:

Arbres de classificació: Els arbres de classificació són models algorísmics de classficicació que tenen un fort potencial visual per a la descripció de les tendències en les variables contínues, aquest cas Age i Fare en relació als casos d'èxit, és a dir procedim a una valoració de les tendències en l'edat i les tarifes de tiquet dels supervivents.

ANOVA: En aquesta analítica volem constatar que les ordres que es van donar en el moment de l'enfonsament del vaixell en última instància, que eren que pugen a bord dels bots d'emergència les dones i els nens en primer lloc, si es tracta d'una ordre real o bé tan sols va ser una consigna.

- ANOVA : Sex vs. Age.
- ANOVA: Sex vs. Pclass.

MODELS DE REGRESSIÓ, CLASSIFICACIÓ I PREDICCIÓ:

Machine Learning. El models proposats per a tasques de predicció i les proves Train&Test són:

Model de regressió logística:

- Entrenament del model de regressió logística.
- Valoració del model.
- Proves de predicció Train&Test per al model.

Model SVM: Machine Learning.

- Entrenament del model Suported Vectorial Machine.
- Valoració del model.
- Proves de predicció Train&Test per al model.

Una de les tècniques d'aplicació en aquest cas és la visualització del conjunt de dades train basat en els mètodes probabilístics en arbres de probabilitats o arbres de regressió.

PRESENTACIÓ DELS RESULTATS EN TAULES I GRÀFIQUES

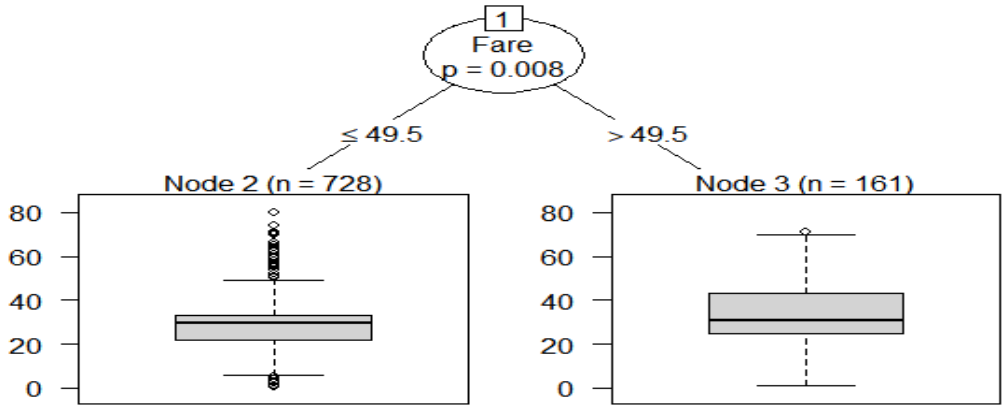
Analítica descriptiva i inferencial: resultats en els arbres de classificació.

Resultats de l'arbre de classificació en les variables Age i Fare.

```
         Conditional inference tree with 2 terminal nodes

Response:  Age
Input:  Fare
Number of observations:  889

1) Fare <= 49.5; criterion = 0.992, statistic = 6.971
  2)*  weights = 728
1) Fare > 49.5
  3)*  weights = 161
```

En la mostra la majoria d'individus a bord del Titanic tenen un bitllet igual o inferior a 49.5, són un total de 728 individus en la mostra Train.csv i suposen el 81'889%, mentre que una minoria a bord un total de 161 individus disposen d'una tarifa superior a 49.5 i suposen un 18'11% del total de la mostra.

Arbre de classificació en les variables Age i Fare:



En el gràfic observem com la mostra de la dreta que paga una tarifa superior a 49.5 és una minoria un 18'11% i representa una població més envellida, mentre que la mostra de l'esquerra que paga una tarifa igual o inferior a 49.5 és una població més jove i suposa la representació d'una majoria en un 88'89% llevat d'algunes excepcions representades en els outliers.

Resultats de l'arbre de classificació en les variables Survived i Age.

```
Conditional inference tree with 2 terminal nodes

Response:  Survived
Input:  Age
Number of observations:  889

1) Age <= 6; criterion = 0.974, statistic = 4.952
  2)*  weights = 47
1) Age > 6
  3)*  weights = 842
```

En la classificació ctree per a les variables Survived i Age es posa de manifest que en la mostra Train_T de 889 individus els menors amb sis o menys anys són un total de 47.

Resultats de l'arbre de classificació en les variables Survived i Fare.

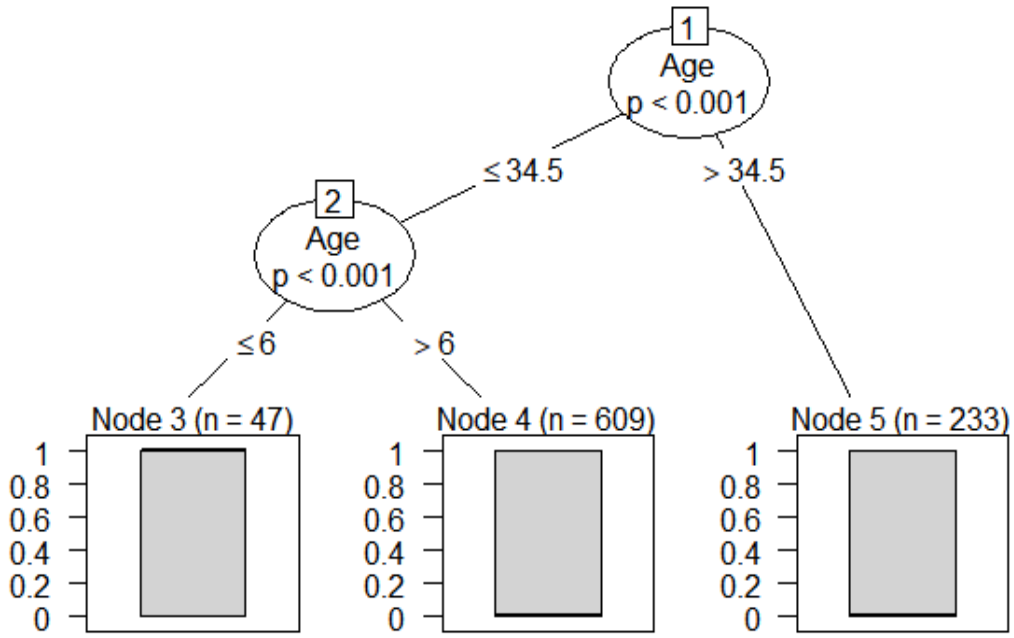
```
         Conditional inference tree with 3 terminal nodes

Response:  Survived
Input:  Fare
Number of observations:  889

1) Fare <= 10.4625; criterion = 1, statistic = 57.874
  2)*  weights = 339
1) Fare > 10.4625
  3) Fare <= 73.5; criterion = 1, statistic = 18.982
    4)*  weights = 455
  3) Fare > 73.5
    5)*  weights = 95
```

En la prova ctree per a les variables Survived i Fare es posa de manifest que en la mostra Train_T de 889 individus que les persones que paguen més de 73.5 per un bitllet són una minoria un 10'68% sumant un total de 95 individus, mentre que la majoria de passatgers es concentra en una tarifa d'entre 10.46 i 73.5 en un 51'181% sumant un total de 455 i en la tarifa de menys de 10.46 en un 38'13% amb un total de 339 individus.

Resultats de l'arbre de classificació en les variables Survived, Fare i Age.

```
        Conditional inference tree with 3 terminal nodes

Responses: Survived, Fare
Input:  Age
Number of observations:  889

1) Age <= 34.5; criterion = 1, statistic = 15.963
  2) Age <= 6; criterion = 0.999, statistic = 14.589
    3)*  weights = 47
  2) Age > 6
    4)*  weights = 609
1) Age > 34.5
  5)*  weights = 233
```

Arbre de classificació en les variables Survived, Fare i Age:

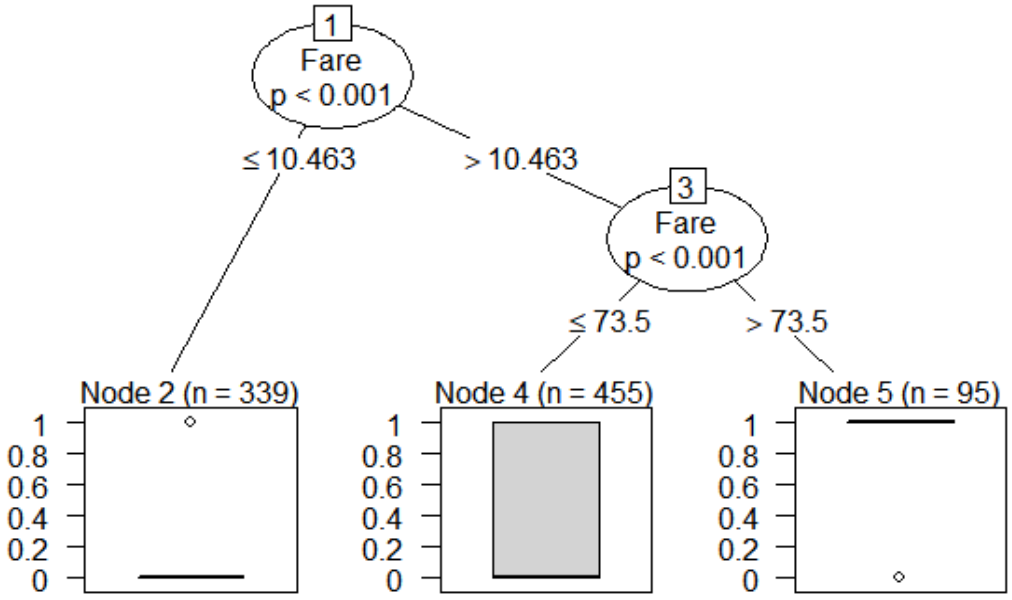


Les persones de sis o menys anys són 47 individus i suposen un 5'28% dels passatgers, mentre que les persones entre 6 i 34 anys suposen un 68'5% dels passatgers i finalment les persones que tenen més de 34 anys són un total de 233 i suposen el 26'21% dels passatgers.

Resultats en l'arbre de classificació de les variables Survived, Age i Fare.

```
Conditional inference tree with 3 terminal nodes

Responses: Survived, Age
Input:  Fare
Number of observations:  889

1) Fare <= 10.4625; criterion = 1, statistic = 68.226
  2)*  weights = 339
1) Fare > 10.4625
  3) Fare <= 73.5; criterion = 1, statistic = 23.771
    4)*  weights = 455
  3) Fare > 73.5
    5)*  weights = 95
```

Arbre de classificació per a les variables Survived, Age i Fare.



En l'arbre de classificació per a les variables Survived, Age i Fare observem com per un bitllet inferior a 10.46 les probabilitats de supervivència són poques properes a 0 llevat d'una excepció (veure outlier), mentre que les persones que paguen una tarifa mitja entre 10.46 i 73.5 tenen una probabilitat més diversificada, i finalment les persones que suposen una minoria i que paguen una tarifa superior a 73.5 tenen una alta probabilitat de supervivència.

Analítica descriptiva i inferencial: model ANOVA.

Resultats del model ANOVA en les variables Survived, Sex, Pclass i Age.

```
Call:
   aov(formula = Survived ~ Sex * Pclass * Age, data = Train_T)

Terms:
                       Sex    Pclass       Age Sex:Pclass   Sex:Age
Sum of Squares   61.58609  15.14444   3.34922    2.55158   0.49493
Deg. of Freedom         1         1         1          1         1
                Pclass:Age Sex:Pclass:Age Residuals
Sum of Squares     0.00631        0.06636 126.76732
Deg. of Freedom          1              1       881

Residual standard error: 0.3793287
Estimated effects may be unbalanced
```

En el model ANOVA la variables sexe és més rellevant en quant a la classificació Survived, seguida de la variable PClass i finalment la variable Age.

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

Resultats del model de regressió logística:

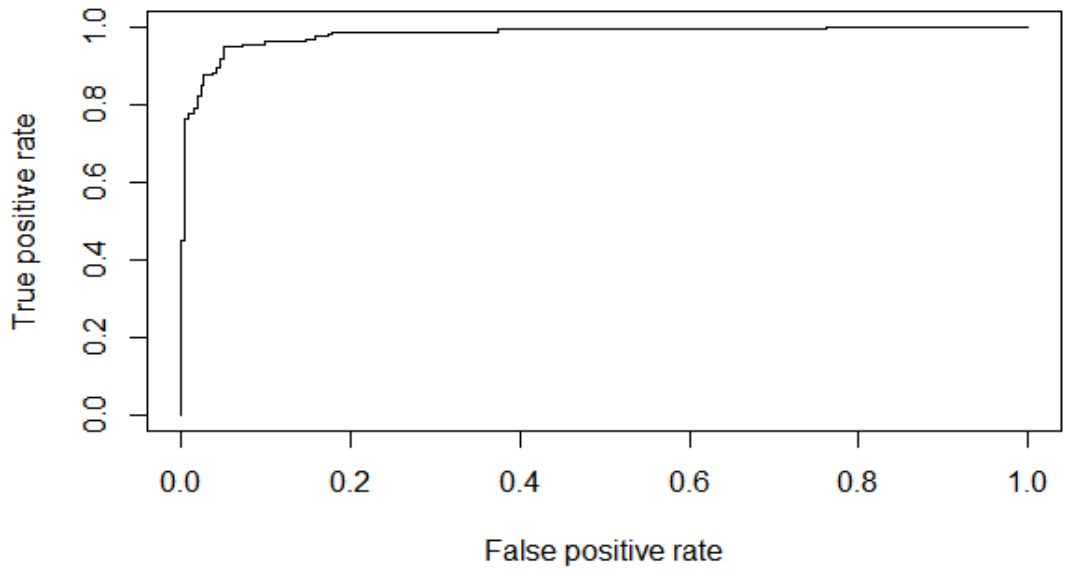
```
Call:
glm(formula = Survived ~ ., family = binomial(link = "logit"),
    data = Train_T)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6446  -0.5907  -0.4230   0.6220   2.4431

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.285188   0.564778   9.358  < 2e-16 ***
Pclass      -1.100058   0.143529  -7.664 1.80e-14 ***
Sexmale     -2.718695   0.200783 -13.540  < 2e-16 ***
Age         -0.039901   0.007854  -5.080 3.77e-07 ***
SibSp       -0.325777   0.109384  -2.978   0.0029 **
Parch       -0.092602   0.118708  -0.780   0.4353
Fare         0.001918   0.002376   0.807   0.4194
EmbarkedQ   -0.034076   0.381936  -0.089   0.9289
EmbarkedS   -0.418817   0.236794  -1.769   0.0769 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.82  on 888  degrees of freedom
Residual deviance:  784.19  on 880  degrees of freedom
AIC: 802.19

Number of Fisher Scoring iterations: 5
```

Precisió del model de regessió logística:

```
[1] "Accuracy 0.913669064748201"
```

La precisió del model de regressió logística és molt bona amb una Accuracy del 0'91 i això vol dir que la capacitat predictiva del model registra un 91'36% d'encert en la classificació.

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 1 amb bones expectatives d'encert en els TP.

Valor AUC:

```
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()


[1] 1
```

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

Resultats 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 : 414

Objective Function Value : -353.5994
Training error : 0.82558
```

Els algorismes SVM i Xarxes neuronals són els més utilitzats, per la seva efectivitat essent SVM un algorisme més ràpid en l'execució per la seva simplicitat en els càlculs.

Resultats en el model predictiu SVM:

```
        [,1]
1 0.04865698
2 0.95087662
3 0.04853553
4 0.04860953
5 0.95094333
6 0.04870747
```

Matriu de confusió del model SVM:

```
svm.predict   0   1
          0 265   0
          1   0 152
```

La matriu de confusió del model SVM mostra que no hi ha error en la classificació TP i TN.

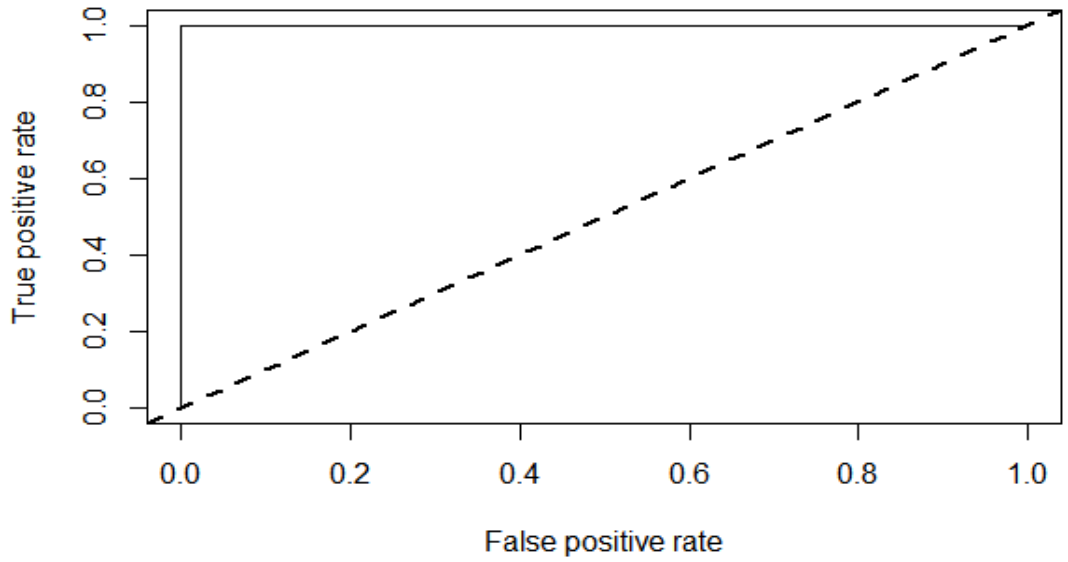
Precisió del model SVM:

```
"Accuracy 1"
```

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.

Corba ROC per a la classificació de vertaders i falsos positius:

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.



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()

[1] 1
```

RESOLUCIÓ DEL PROBLEMA

Conclusions.

Els arbres de classificació 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 per a una comprensió visual de la segmentació i ordenació de les dades, en aquest cas per a la valoració de l'èxit en les variables continues Age i Fare.

En la fase analítica hem procedit a unes proves ANOVA per tal de valorar el succés èxit en la variable Sex, Age i Pclass per als supervivents i la rellevància de les tres variables.

Segons aquesta fase d'analítica descriptiva i inferencial, sembla ser que la variable sexe és la més determinant per a les probabilitats d'èxit en el factor femení i més probabilitats de fracàs per al sector masculí, altrament en segon lloc és la variable Pclass la que és mostra favorable a l'èxit en un segment minoritari especialment per a la gent que ha pagat molt per a un bitllet. Finalment constatem que en la variable edat el segment d'infants de menys de sis anys es veu discriminat així com el segment d'adults, 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 però no per als infants entre els quals va significar primer la classe social.

El model de regressió logística ofereix un bon rendiment encara que no òptim i es tracta d'una modelització que pot servir 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 com pot ser les xarxes neuronals.

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 aquest model amb una accuracy d'un 100% té una plena capacitació per a la classificació de les dades.

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 passen a una fase analítica i inferencial d'exploració de la naturalesa de les dades per tal d'aproximar-nos a 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, en una modelització regressiva o en algorismes més complexos com pot ser SVM o com a funció logística activadora en un model de xarxa neuronal, ... intentarem buscar les tècniques i mecanismes que millor puguin funcionar per a la optimització del rendiment de l'algorisme 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 fiabilitat del 100%.

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/Fitxer_Train.Rmd

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

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

Codi R per a les pràctiques analítiques:

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

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

DOCUMENTS ANNEXOS

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

- Train_T.Rmd 8 K
- Test_T.Rmd 8 K
- Pràctiques_analítiques.Rmd 6 KB
- Train_T.html 861 KB
- Test_T.html 857 KB
- Pràctiques analítiques 945 KB
- PAC2_TCVD.pdf 267 KB

DATASETS TRACTATS

Trobareu els datasets resultants en el directori output_data.

- Test_T.csv 17 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

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.