TITÀNIC

APA 2016 – 2017 Q1, FIB UPC

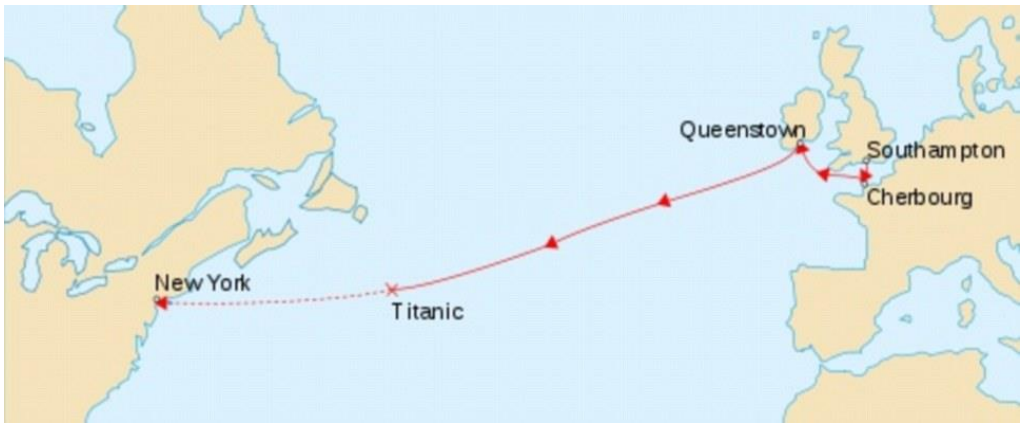
23 DE GENER DE 2017

RICARD MEYERHOFER PARRA – ALEJANDRO MARTINEZ ZAMORANO

Índex

Introducció

El nostre problema tracta sobre l'històric i desafortunat accident del Titànic (succeït l'any 1912) on van perdre la vida 1514 persones de les 2223 que van embarcar.



El nostre objectiu serà mitjançant l'anàlisi de les dades dels passatgers, crear un model que sigui capaç de classificar aquest segons els seus atributs i ens pugui dir si sobreviuria o no al accident.

Les dades que emprarem són el llistat de passatgers del Titànic on per cada passatger tenim el següent conjunt de dades:

Nom	**Valors**
survival	Survival (0 = No; 1 = Yes)
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Treballs anteriors

S'ha parlat molt sobre el arxiconegut accident del Titànic, però no s'ha parlat gaire sobre el que consisteix aquest treball, que es la classificació i índex de supervivència dels passatgers.

Un article bastant interessant és el de la universitat de Florida[3] que tracta bàsicament sobre el mateix que aquest treball, però ho analitza d'una manera més matemàtica i formal.
L'article planteja tres hipòtesis, que són:

- La primera hipòtesis tracta sobre que les dones de classe alta, tenen més probabilitats de sobreviure
- La segona és que la probabilitat de supervivència dels homes decreix a mesura de que decreix la classe
- I per últim, que la probabilitat de salvar-se és inversament proporcional a la edat de la persona.

La conclusió és, com un s'espera, que les dones de classe alta tenen una alta probabilitat de salvar-se, sobretot si són joves.
En canvi els homes, excepte si són molt joves, ho tenen bastant més complicat, sobretot quan baixa la seva classe en el vaixell.

Hi ha altres treballs i altres articles amb més o menys rigor, que també tracten sobre aquest tema, però bàsicament tots arriben a aquesta conclusió.

Preprocesament de dades

Al analitzar les dades originals, hem observat diversos problemes que calen tractar

- hi ha missings en les edats dels passatgers
- hi ha missings en el preu pagat pel bitllet (fare)
- hi ha missings en el port on van embarcar els passatgers.

Per tractar els missings de dades de les edats i dels fare, hem aplicat una regressió lineal per a cadascuna per predir els valors que falten

En canvi, per tractar el port on van embarcar els passatgers, hem obtingut la proporció que tenen els tres ports i hem vist que amb diferencia, el port mes comú és el de Southampton, per tant hem assignat la S als ports que tenien missing values.

Hem fet feature extraction i hem creat dos noves variables:

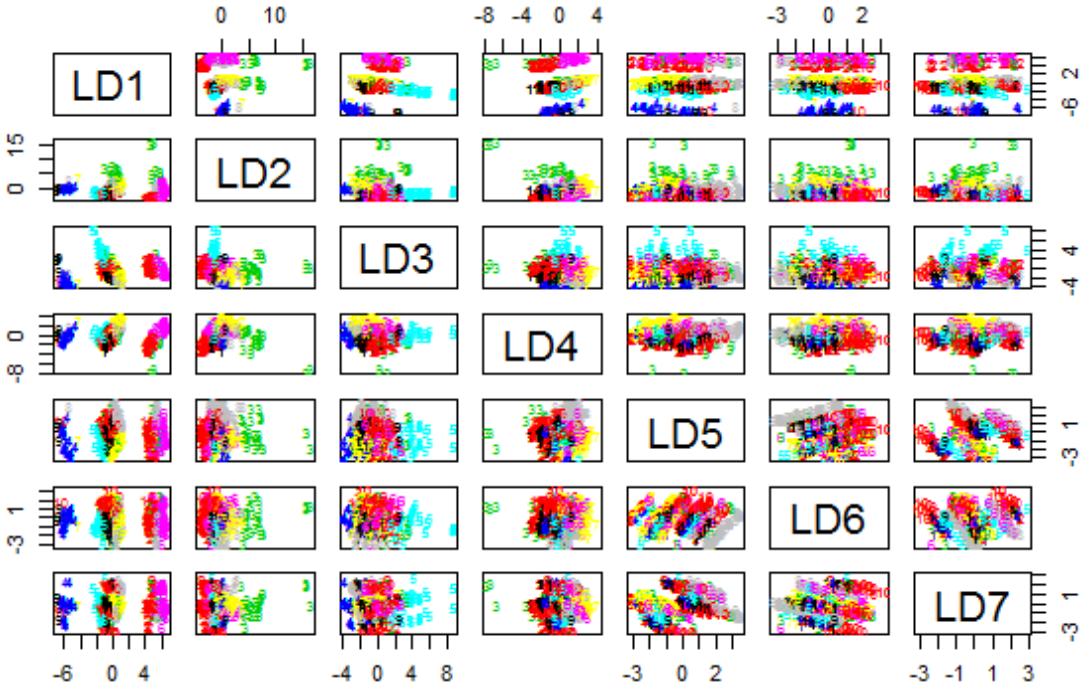
- **AgeGroup**: hem aplicat K-means a les edats per agrupar-les. Hem experimentat fins arribar a la conclusió que amb 7 clusters de edats obtenim el millor rendiment.
- **FamilySize**: hem agrupat les variables Parch i SibSp que entre les dos, ens deien el nombre total de familiars del passatger que anaven a bord.
- **Title**: Hem parsejat els noms dels passatgers per agafar el seu títol. El títol és, per exemple, mr, ms, master... doncs hem fet 4 categories.

Hem cregut convenient, almenys per ara abans d'una experimentació més profunda amb mètodes lineals i no lineals, no tenir en compte el nom dels passatgers i tampoc tenir en compte la cabina on s'allotjaven els passatgers, aquesta deguda a la gran quantitat de missings que tenim i que resulta impossible predir-los o obtenir-los d'alguna manera.
Tampoc ens hem quedat amb el PassengerId ja que simplement és un numero del 1 fins al total de passatgers que enumera cada passatger per saber la posició que ocupa en la taula de dades, cosa que no ens aporta cap tipus d'informació.

Clustering

Hem fet diversos experiments amb les dades i com agrupar-les.
La manera que hem trobat mes profitosa es fer Kmeans, assignar els clusters a la nostre matriu de dades i després fer la visualització amb el anàlisis discriminant lineal (lda).

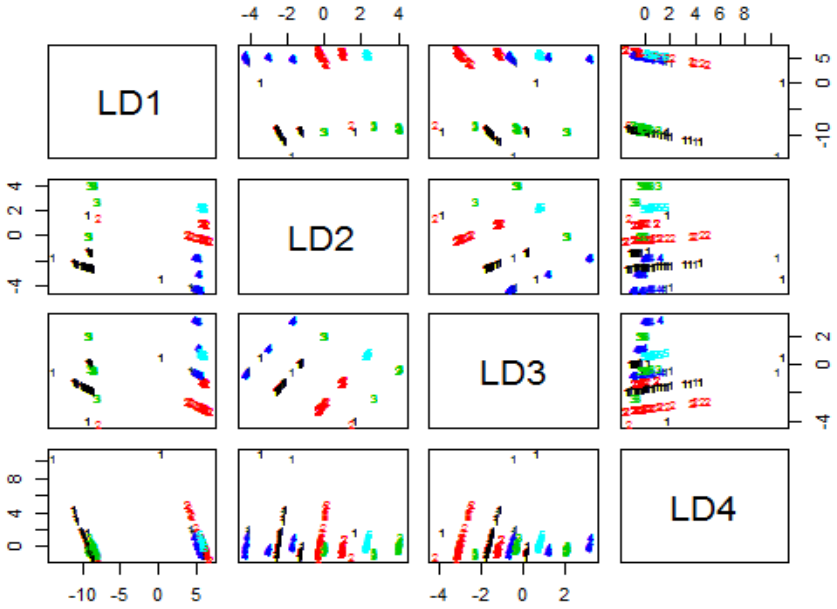
Primerament hem provat amb totes les dades i 10 clusters, ja que hem aplicat el índex de Calinski que ens mesura com de bona es la clusterització.



Hem utilitzat el mateix raonament per veure la importància de les variables predictores. Per a veure això, hem aplicat un subconjunt d'aquestes per a veure com interactuen entre si i si els clusters es veuen ben definits.

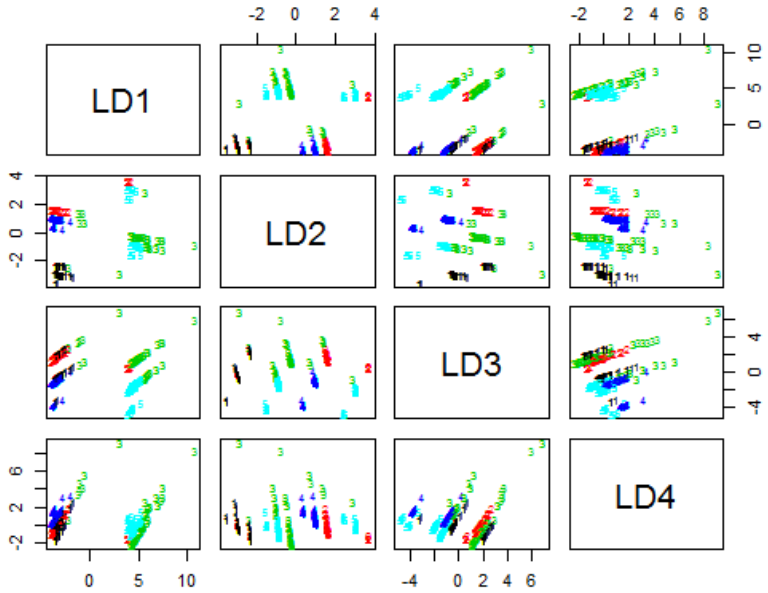
Aquí podem veure les variables:

Survived, Sex, Pclass i Fare



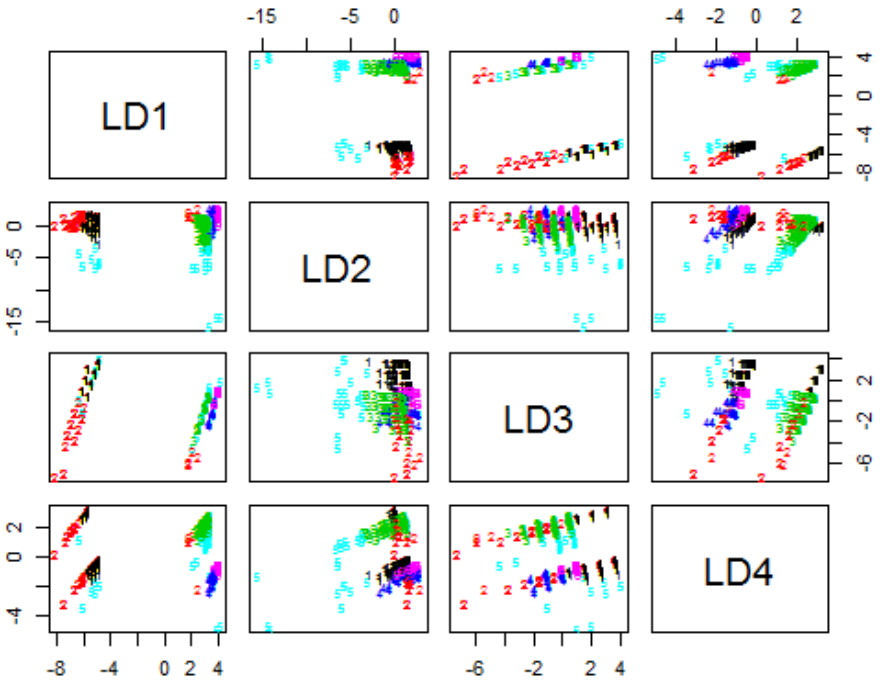
Podem observar 5 clusters ben definits, això ens indica que són variables bastant predictives ja que les podem observar nítidament.

Survived Sex Embarked i Fare



Igualment com a l'anterior, podem veure-les ben definides i separades pels clusters sense que es barregin.

Survived Sex FamilySize Fare



Aquí veiem les dades més barrejades, això ens pot indicar que familySize no és una variable predictora tant bona com les vistes anteriorment.

Metodologia emprada

Per avaluar el rendiment dels mètodes utilitzats, hem seguit aquests tres passos:

1. dividir el conjunt de dades en un conjunt de training (⅔ del total) i un conjunt de test (⅓ restant)
2. realitzar una 10x10 Cross-Validation del conjunt d'entrenament, per a veure el comportament dels models i identificar possibles casos d'overfitting / underfitting.
3. Optimitzar el model tant com ha sigut possible mitjançant l'ajust dels paràmetres dels mètodes. Això ho hem fet amb les xarxes neuronals.

Predicció amb models Lineals o Quadràtics

En aquesta secció, hem posat a prova els anàlisis QDA, Naive Bayes i Regressió Logística, que son els que ens han semblat que serien descriptius en els mètodes Lineals.

Quadratic Discriminant Analysis (QDA):

És de la família del Naive Bayes (NB), amb la diferencia que QDA estima cada matriu de covariància en comptes de fer les suposicions de Linear Discriminant Analysis (LDA) o NB.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 248  23
         1  58  89

               Accuracy : 0.8062
                 95% CI : (0.765, 0.843)
```

Ens dona un error de predicció d'un 19.32%

Naive Bayes:

Aquest mètode es basa en la aplicació del teorema de Bayes i la "ingènua" assumpció d'independència entre un parell de característiques. LDA i QDA pertanyen a aquesta família.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 248  23
         1  76  71

               Accuracy : 0.7632
                 95% CI : (0.7194, 0.8031)
```

Ens dona un error de predicció d'un 23.68%

Regressió Logística:

És un tipus d'anàlisi de regressió per predir el resultat d'una variable categòrica en funció de les variables independents o predictores.

```
     0   1
0 234  39
1  37 108
```

En aquest cas, el error és d'un 18,19%

Predicció amb models no Lineals

Xarxes Neuronals:

Es basa en la cerca de regressors adaptatius a un problema. La implementació que hem utilitzat es el Perceptró multi capa. Per optimitzar aquest mètode, hem buscat el "decay" òptim dels pesos per regularitzar la xarxa.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 250  21
         1  54  93

               Accuracy : 0.8206
                 95% CI : (0.7804, 0.8562)
```

Amb les xarxes neuronals, l'error que obtenim es d'un 17.94%

SVM amb RBF:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 214  57
         1  42 105

               Accuracy : 0.7632
                 95% CI : (0.7194, 0.8031)
```

En el cas de les maquines de vector suport amb kernel rbf, tenim que ens dona un error del 23.68%

Random Forest:

Meta estimador que s'ajusta a una sèrie de classificadors que son arbres de decisió en diverses submostres del conjunt de dades i utilitza el promig per millorar la precisió predictiva.

En primer lloc, per tal de comprovar el comportament correcte de RF en les nostres dades (mirant si overfitting o underfitting les dades) que vam decidir fer un 10x10 Cross-Validation. Un cop comprovat que el model es comporta correctament, entrenem amb totes les dades d'entrenament disponibles i tracten de predir les dades de prova. Aquests són els resultats obtinguts a partir d'aquest procés:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 241  30
         1  42 105

               Accuracy : 0.8278
                 95% CI : (0.7881, 0.8627)
```

Veiem un error d'un 17.22%, que com hem pogut veure, és el mes baix que hem arribat a trobar en tota la nostra experimentació.

	0	1	MeanDecreaseAccuracy	MeanDecreaseGini
Pclass	114.00033	87.595078	146.78135	24.43212
Sex	134.01971	88.420176	153.73869	51.87902
AgeGroup	-10.41171	102.247007	78.68530	19.55531
Fare	120.74865	103.972054	178.53272	62.52111
Embarked	42.40665	5.414004	40.33758	9.44159
Title	143.09073	111.904584	177.41448	70.34244
FamilySize	86.69883	44.315057	107.18947	22.33874

Hem extret també la importància de les variables. Podem veure que les mes importants son el sexe, el preu pagat pel bitllet (Fare) i el titol, la qual es una variable creada per nosaltres.

Conclusions i treball futur

Observant els diferents resultats, podem concloure que els millors mètodes que hem trobat son el random forest amb un 17.22% d'error, les xarxes neuronals amb un 17.94% i la regressió logística amb un 18.19%. És molt complicat decidir-se per una ja que totes tenen avantatges i inconvenients.

En el cas de la regressió logística, es mes simple i ràpida de calcular que les altres dues, però les xarxes neuronals i els boscos aleatoritzats, son més robustes i en general donen millor resultat sempre i quan ens puguem permetre el seu càlcul.

Com era de suposar, el extreure el coneixement de les dades per crear noves variables o agrupar d'altres és clau per una bona predicció. Sense les variables de title, age group i family size, els rendiments son molt mediocres. Amb el random forest, per exemple, tenia més d'un 35% d'error, per tant, es en el preprocesament i en la feature extraction on hem concentrat la majoria del nostre esforç i al final hem vist que ha sigut clau.

Com a comentaris personals, ens ha agradat molt poder sintetitzar tot el que hem après al curs en una practica on es un cas real, en aquest cas el titànic, encara que hi havia molts altres datasets que ens cridaven molt l'atenció.

Les principals dificultats que hem tingut, com ja hem dit, ha sigut en la part de la primera entrega, sobretot amb extraure informació de les dades per a tenir millor prediccions. Ens hem hagut de documentar bastant sobre el tema i llegir molt sobre l'accident per saber ben be quina informació era la mes important per a extreure-la.

Com a dificultats, bàsicament errors amb el R pròpiament, sobretot amb els mètodes de xarxes neuronals i random forest, també algunes dificultats amb la cross validation, però creiem que ens hem en sortit prou bé.

Com a treball futur, doncs nosaltres pensem que no es pot fer gaire més, no diem que sigui perfecte ni molt menys, però se'ns acut gaire cosa. Es podria intentar millorar la feature extraction i intentar donar-li un altre enfoc per a veure si es fan millors prediccions, o en comptes del 10x10 cros validation, fer un leave one out, que te un cost computacional molt més gran, però pot donar valors mes acurats.

Bibliografia

1. *Titanic Data Set, From Kaggle machine learning repository*,

   https://www.kaggle.com/c/titanic
2. *Titanic – Machine Learning From Disaster,*

   http://murphy.wot.eecs.northwestern.edu/~xto633/xiaodong/fullreport.pdf
3. *THE TITANIC SHIPWRECK: WHO WAS MOST LIKELY TO SURVIVE? A STATISTICAL ANALYSIS,*

   http://myweb.fsu.edu/tzuehlke/5427/files/titanic.pdf
4. *Titanic Data Analysis - Did Passengers Get Their Money's Worth?*

   https://www.tableau.com/blog/titanic-data-analysis-from-tableau-customer