**Tipologia i cicle de vida de les dades**

**XAVIER JORDÀ MURRIA**

**ANNA SERENA LATRE**

**Màster Data Science UOC**

**MAIG 2019**

**PRESENTACIÓ**

‘TITANIC: *Machine Learning from Disaster*. ‘

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

**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 que es pot posar 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 amb el qual crearem els models que posarem a prova en el fitxer test en una fase analítica.

**BASE DE DADES**

* 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 classe 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.

1. Valoració i tractament de la inconsistència de les dades.
2. 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 :**

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 abord 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.

Machine Learning. El models proposats per a les proves Train&Test són:

Model de regressió logística:

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

Model SVM: Machine Learning.

* Entrenament del model Suported Vectorial Machine.
* Valoració del 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**

Proves test: matrius de confusió.

**RESOLUCIÓ DEL PROBLEMA**

**Conclusions.**

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

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.

Els arbres de probabilitats 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.

**Quina és la finalitat del nostre estudi ?**

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 un model de regressió 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 ?**

La resolució del problema passa per unes tasques d’integració i neteja adequades per tal de poder procedir amb la resolució algorísmica que demanda de precisió.

**CODI**

**LINKS REPOSITORI GITHUB**

[**https://github.com/XavierJordaMurria/TipologiaPac2/**](https://github.com/XavierJordaMurria/TipologiaPac2/)

[**https://github.com/AnnaSerenaLatre/TipologiaPac2/**](https://github.com/AnnaSerenaLatre/TipologiaPac2/)

**DOCUMENTS ANNEXOS**

**DATASETS TRACTATS**

**WEBGRAFIA**

<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. |