Titanic Survival Prediction using Classical and Machine Learning Models

Oriol Jiménez Asensi[a,1], **Fèlix Sáiz von Fraunberg**[b,2], **Eduardo Pérez Motato**[c,3] **and Roger Guitart Casals**[d,4]

[a] *1641014*
[b] *1620854*
[c] *1709992*
[d] *1711342*

Abstract—In this project, we analyze the Titanic dataset to predict passenger survival using classical and machine learning models. After data cleaning and feature engineering, we evaluate Logistic Regression, SVM, and Random Forest classifiers through cross-validation. The Random Forest achieved the best accuracy (0.83), showing robust performance with balanced precision and recall.

Keywords—*Super vector machine, logistic regression, random forest, titanic*

Contents

1. Exploratory data analysis

La base de dades consta de 12 columnes, incloent la variable target. Aquesta variable, anomenada Survived, és binària i conté 549 mostres falses i 342 mostres verdaderes, indicant que està relativament equilibrada.

Table 1. Tipus de les variables independents

Nom	Tipus
PassengerId	enter
Pclass	categorica
Name	cadena de caracters
Sex	binaria
Age	punt flotant
SibSp	enter
Parch	enter
Fare	punt flotant
Cabin	cadena de caracters
Embarked	categorica

Pel que fa als NaNs, la mostra d'entrenament presenta 177 valors nuls a la columna Age, 687 a la columna Cabin i 2 a la columna Embarked.

Fent una matriu de correlació al quadrat de les variables numèriques observem que Survived està relacionada amb Pclass, Sex, Fare. Age està relacionada amb Pclass i Pclass està molt relacionada amb Flare.

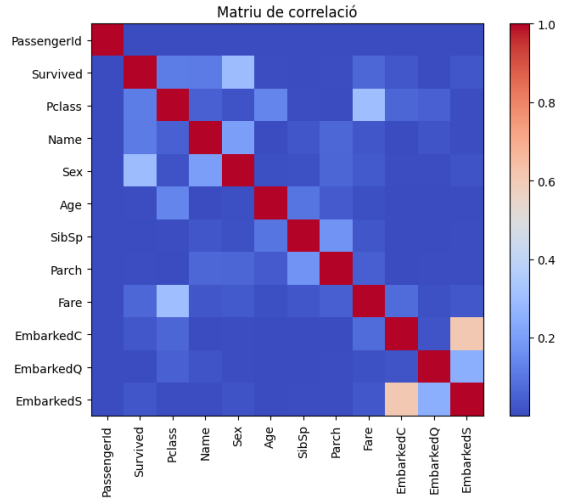

Figure 1. Matriu de correlació al quadrat de les variables numèriques

Quant a les etiquetes de les variables categòriques (com amb la variable target), no sembla que les etiquetes estiguin prou poc representades per causar "problemes de categories rares".

Table 2. Etiquetes d'Embarked

Etiqueta	Nombre d'instàncies
S	644
C	168
Q	77

Table 3. Etiquetes de Pclass

Etiqueta	Nombre d'instàncies
1	216
2	184
3	491

2. Preprocessing

Les dades no estan normalitzades. Malgrat que les instàncies numèriques no semblen ser excessivament grans, considerem que normalitzar les dades sempre és una cosa positiva. En aquest cas, utilitzarem una estandarització Z. PassengerId i Name semblen inútils a l'hora de realitzar la predicció. Malgrat això, sovint la longitud del nom és considerada una mostra de prestigi com es veu a l'aristocràcia, per això també hem decidit provar d'afegir la mida del nom com a variable. Curiosament, a la matriu de correlació al quadrat es pot veure una correlació amb la variable target. Per tractar les variables categoòriques, podem aprofitar la monotonietat de la variable Pclass per interpretar-ho com una variable numèrica i repartir Embarked en tres variables binàries. Aquestes evidentment apareixen correlacionades a la matriu de covariàncies. En quant als

Nans, els hem substituït per -1 a les variables numèriques (Age) i a Embarked simplement quedarà 0 en totes les categories. Considerem que no té sentit filtrar-los tenint en compte que al test set també n'hi ha i els necessitarem per fer la predicció. No hem realitzat pca perquè teniem molt poques variables i per tant era irrellevant.

3. Metric Selection

Hem entrenat un model logístic a partir de les dades preprocessades i sense tocar hiper-paràmetres. Hem creat la matriu de confusió del model i hem provat l'accuracy, la f1_score i l'average_precission_score. Despres hem realitzat la corba roc i la precission-recall curve. A partir d'aquí hem determinat que la variable target no està gaire descompensada, utilitzarem l'accuracy_score.

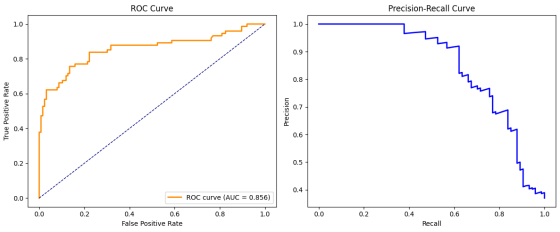

Figure 2. Corba ROC i Precission-Recall Curve

Table 4. Matriu de confusió (test)

		Prediccions	
		0	**1**
Reals	**0**	105	21
	1	17	57

4. Model Selection amb validació creuada

Hem entrenat els següents models per a triar quin encerta més, aquestes han estat SVM i LogisticRegrsion ja que són els que s'han fet a classe i com a 3r hem triat el RandomForests, ja que el coneixem de l'assignatura de OOP. Els accuracy scores dels models amb hiperparàmetres per defecte són:

Table 5. Accuracy scores dels models

Model	Accuracity-score
SVM	0.813722
RandomForest	0.804733
LogisticRegrsion	0.775551

Malgrat que el temps d'entrenament per aquesta base de dades és negligible, si l'escalèssim el que donaria millors resultats seria el logistic (0.011358s) seguit de el SVM(0.093215s) i RF (0.588249s). En quant a la cerca d'hiperparàmetres ens hem plantejat utilitzar grid search (GridSearchCV) però era molt costós i ens preocupava que, degut la petita mida de la mostra, es produís overfiting implícit. Per tant hem acabat utilitzant random search (RandomizedSearchCV) amb 45 iteracions per model. El que ha donat millors resultats pels models ha sigut:

Table 6. Resultats de la cerca d'hiperparàmetres dels models

Model	Accuracy mitjà	Temps (s)
RandomForest	0.829364	9.052283
SVM	0.821543	2.532295
LogisticRegression	0.785619	1.517158

Millors hiperparàmetres utilitzats:

- **RandomForest**: `{'n_estimators': 100, 'min_samples_split': 5, ...}`
- **SVM**: `{'kernel': 'rbf', 'gamma': 0.1, 'C': 1.6681005...}`
- **LogisticRegression**: `{'solver': 'liblinear', 'penalty': 'l1', 'C': ...}`

5. Anàlisi Final

La mètrica principal (Accuracy = 0.89) mostra un bon rendiment global. En un context pràctic, el model podria servir per assignar una probabilitat de supervivència i prioritzar recursos o suports segons risc. També podria usar-se com a eina d'anàlisi per entendre quines variables influeixen més en la supervivència.

Es podria millorar per exemple, extreure el títol del nom o crear la mida de la família, imputacions més precises dels valors nuls i ajust d'hiperparàmetres amb més iteracions.

6. Acknowledgements

Tau LaTeXtemplate built by Guillermo Jimenez.