THE INTERSTELLAR DREAM

Giuliano Cucina

Machine Learning

THE PROCESS

UNDERSTANDING DATA

Merging csv
Summary Statistics

DATA PROCESSING

Fill NA
Data Cleaning
Feature Engineering

BUILD MODELS

KNN
Random Forest
Logistic Regression

MODEL EVALUATION

F1
Recall
Precision
Accuracy

DEMOSTRATION

Streamlit

INTERESTELLARS DATASETS

- billing_info.csv
- passenger.csv
- travel_data.csv
- test.csv

- train.cvs

SUMMARY STATISTICS:

VIP
True 199

False 8494

CRYOSLEEP
True 3037

False 56564

TRANSPORTED
True 4378

False 4315

USER
8693

AGE
0 to 79 years old

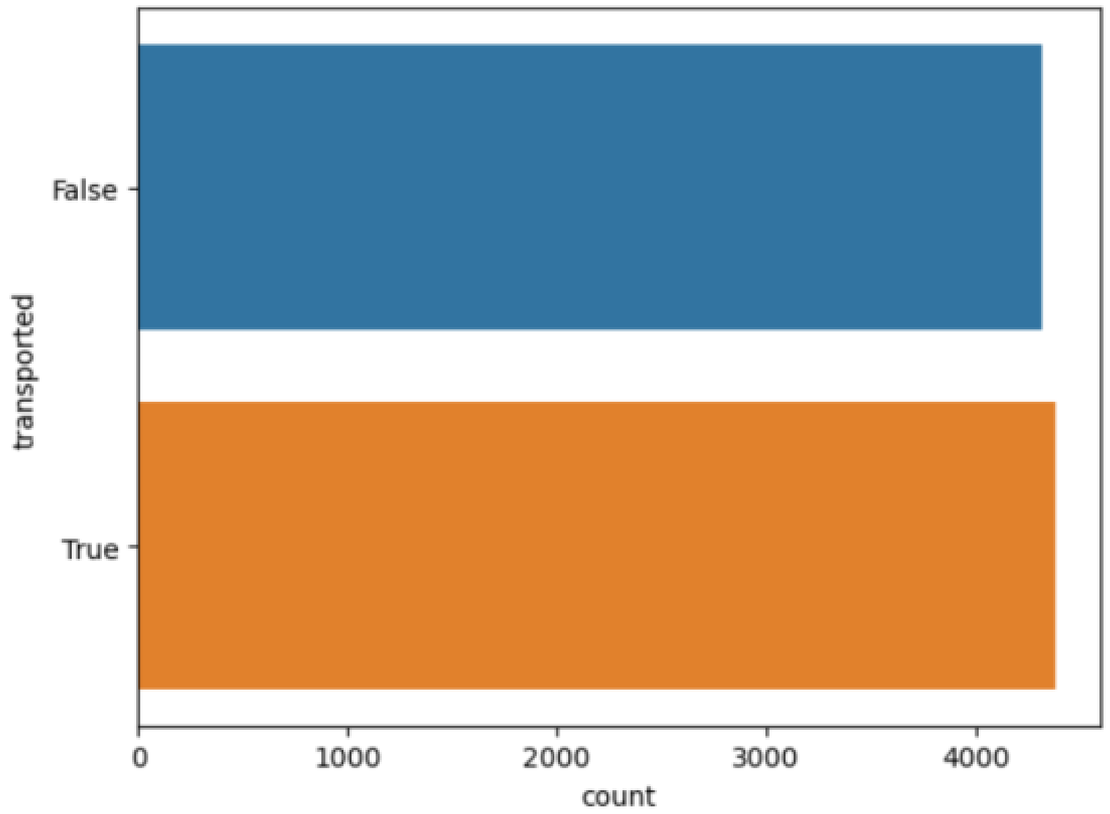
DESTINATION
55 Cancri e

PSO J318.5-22

TRAPPIST-1e

COLUMNS

'age'
'vip'
'spa'
'cabin'
'name'
'vrdeck'
'cryosleep'
'foodcourt'
'destination'
'homeplanet'
'transported'
'passengerid'
'roomservice'
'shoppingmall'



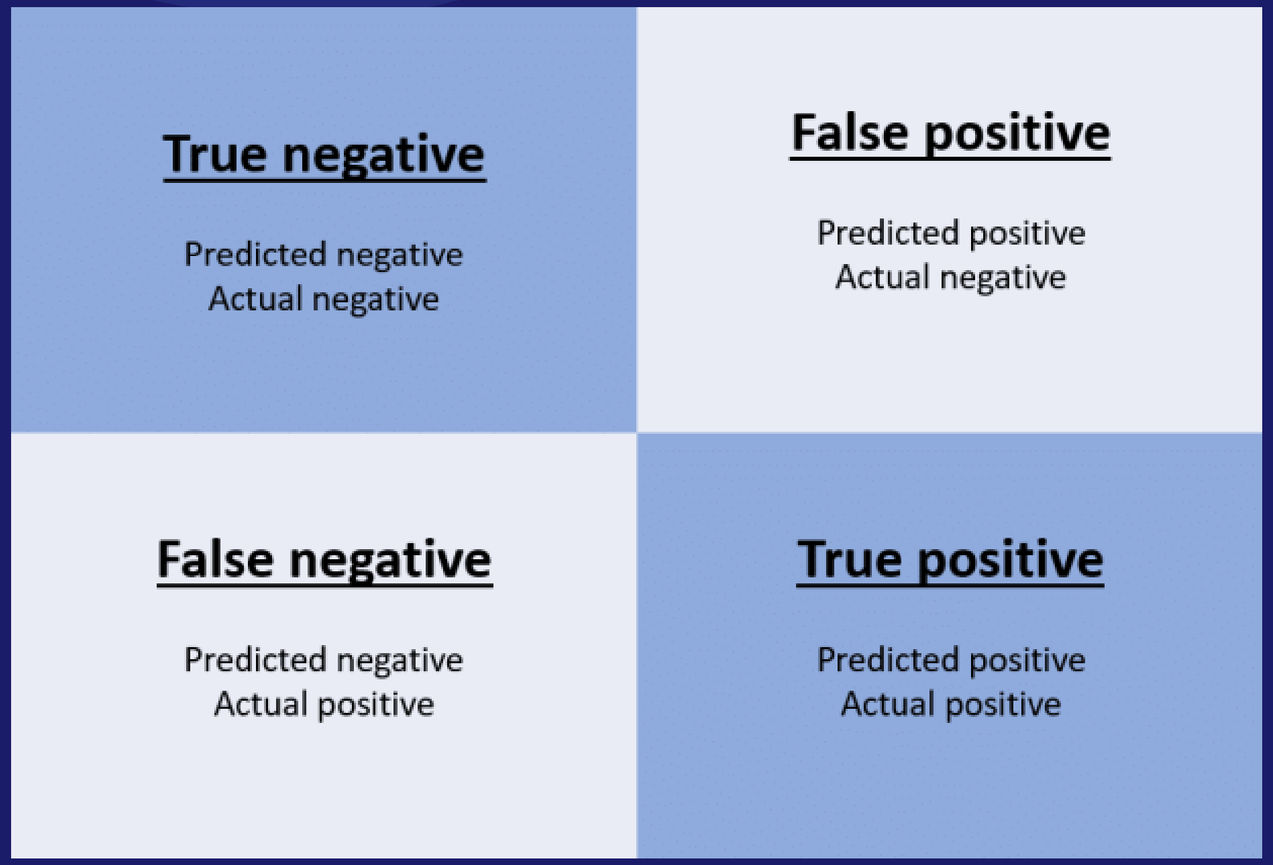
COUNT MISSING VALUES

column_name	percent_missing
cryosleep	2.496261
shoppingmall	2.392730
vip	2.335212
homeplanet	2.312205
name	2.300702
cabin	2.289198
vrdeck	2.162660
foodcourt	2.105142
spa	2.105142
destination	2.093639
roomservice	2.082135
age	2.059128
passengerid	0.000000
transported	0.000000

column_name	percent_missing
foodcourt	2.478373
spa	2.361468
cabin	2.338087
shoppingmall	2.291326
name	2.197802
cryosleep	2.174421
vip	2.174421
destination	2.151040
age	2.127660
homeplanet	2.034136
roomservice	1.917232
vrdeck	1.870470
passengerid	0.000000

PRECISION

Precision = TP / (TP + FP)



I don't care if I miss a 1, as long as no 0 (FP) slips in as if it were 1. That when I predict it as 1, it is 1. The focus must be on minimizing the FP.

RESULTS:

	F1	Recall	Accuracy	Precision
rf	0.782217	0.791572	0.777458	0.773081
lg	0.783380	0.805239	0.775158	0.762675
knn	0.770950	0.785877	0.764232	0.756579

THANK YOU!