

Immagine/Titolo

Christian Gambardella, Vincenzo Offertucci

March 12, 2022

[Link GitHub](#)

Index

- 1 Business Understanding, 3
 - 1.1 Introduzione al problema, 3
 - 1.2 Obiettivi di business, 3
 - 1.3 Descrizione dell'ambiente, 4
 - 1.4 Business success criteria, 4
 - 1.5 Tool da utilizzare, 5
- 2 Data Understanding, 5
 - 2.1 Scelta del dataset, 5
 - 2.2 Analisi del dataset, 5
 - 2.3 Data quality, 6
 - 2.4 Data exploration, 6
- 3 Data Preparation, 8
 - 3.1 Data cleaning, 8
 - 3.2 Define data and target variables, 8
 - 3.3 Data balancing, 8
 - 3.4 Feature engineering, 9
 - 3.4 Feature scaling, 9
- 4 Data Modeling, 9
- 5 Evaluation, 13

1 Business Understanding

1.1 Introduzione al problema

Uno dei fondamenti dell'astronomia è la classificazione degli astri celesti in stelle, galassie e quasar (anche detta classificazione stellare). In particolare i quasar sono stati argomento di dibattito all'interno della comunità scientifica per tutta la seconda metà del XX secolo: questi astri, che sembravano stelle, erano troppo luminosi per essere così lontani dal nostro pianeta.

1.2 Obiettivi di business

L'obiettivo del nostro progetto è realizzare un modello di machine learning che sia capace di classificare gli astri celesti, in particolare i quasar, sulla base di dati spettroscopici.

1.3 Descrizione dell'ambiente

PEAS	
Performance	La misura di performance del modello è la sua capacità di avvicinarsi il più possibile alla corretta classificazione dei tre astri celesti
Environment	<p>L'ambiente di riferimento del nostro modello è l'astronomia, inoltre è:</p> <ul style="list-style-type: none">• completamente osservabile in quanto l'agente in ogni momento ha accesso allo stato completo dell'ambiente;• episodico in quanto le azioni del modello in un dato istante non sono influenzate dalle precedenti;• statico in quanto l'ambiente rimane invariato mentre l'agente sta deliberando;• discreto in quanto l'agente può ricevere un numero ben definito di percezioni ed effettuare un numero ben definito di azioni.
Actuators	L'agente agisce sull'ambiente tramite lo stream di output del nostro computer fornendo così la tipologia di astro celeste che stiamo valutando
Sensors	L'agente percepirà l'ambiente tramite uno stream di input del nostro computer

1.4 Business success criteria

Per validare il nostro modello adotteremo i seguenti criteri: puntiamo innanzitutto ad avere un accuracy almeno del 90% in quanto i dati a nostra disposizione sono sufficientemente numerosi e molto precisi, si parla comunque di misurazioni effettuate con appositi strumenti. Vogliamo inoltre massimizzare i valori di precision e recall per quanto riguarda l'individuazione dei quasar, che sono l'astro più interessante del nostro problema, in particolare puntiamo a raggiungere l'80% in entrambi i casi.

1.5 Tool da utilizzare

I tool che utilizzeremo per realizzare il nostro modello sono i seguenti:

- **Python**
- **Anaconda**
- **ScikitLearn**
- **Pandas**
- **Kaggle**
- **JupyterLab**
- **Mathplot**
- **TeXStudio**
- **MikTeX**

2 Data Understanding

2.1 Scelta del dataset

Per la realizzazione del nostro progetto, dopo svariate ricerche in rete, abbiamo deciso di adottare [questo](#) dataset per la realizzazione del nostro modello di machine learning.

2.2 Analisi del dataset

Nel dataset in questione i dati sono stati collezionati nell'ultimo trentennio da parte della SDSS (Sloan Digital Sky Survey) che si è occupata di processare le foto degli astri celesti in dati, in particolare noi stiamo usando il data release 17 della SDSS-IV. Da notare che il dataset usato da noi non contiene tutte le colonne dell'originale bensì è stata fatta una selezione di 18 (a partire dai 153 iniziali). Nel dataset sono presenti 17 colonne:

- **obj_ID**: è un valore unico che identifica l'oggetto all'interno del catalogo di immagini processate da SDSS.
- **alpha**: ascensione retta, una misura analoga alla longitudine ma proiettata sulla sfera celeste anziché sulla superficie terrestre.
- **delta**: angolo di declinazione, rappresenta una delle coordinate equatoriali per determinare l'altezza di un astro della sfera celeste (analogo alla latitudine).
- **u**: filtro ultravioletto del sistema fotometrico.
- **g**: filtro verde del sistema fotometrico.
- **r**: filtro rosso del sistema fotometrico.
- **i**: filtro vicino all'infrarosso del sistema fotometrico.

- **z**: filtro infrarosso del sistema fotometrico.
- **run_ID**: è un valore unico che identifica la scansione utilizzata.
- **rerun_ID**: è un valore unico che identifica la modalità con cui l'immagine è stata processata.
- **cam_col**: è un valore che indica quale colonna della camera è stata utilizzata nella scansione.
- **field_ID**: è un valore unico che identifica ogni campo.
- **spec_obj_ID**: è un valore unico che identifica l'astro all'interno del catalogo di immagini processato da SDSS (nel dataset originale erano presenti più oggetti relativi allo stesso astro).
- **class**: è la nostra variabile target/dipendente, può assumere i valori "STAR", "GALAXY" o "QUASAR".
- **redshift**: è il valore assunto dal redshift dell'astro basato sull'incremento della lunghezza d'onda (lo spostamento di un astro è da noi percepito come una variazione dello spettro elettromagnetico tendente verso il rosso).
- **plate**: è un valore unico usato come identificatore all'interno dei sistemi SDSS.
- **MJD**: è una versione modificata della data giuliana, in particolare corrisponde a 2400000.5 dopo il giorno 0 del calendario giuliano.
- **fiber_ID**: è un valore unico che identifica la fibra ottica che ha puntato la luce all'interno del piano focale.

Il dataset è di 16.76 Mb e contiene 100000 righe.

2.3 Data quality

Nel dataset non sono presenti dati mancanti, inoltre i dati presentano tutti lo stesso formato (numerico) ma su scale eterogenee, sarà dunque fondamentale prestare attenzione a questo aspetto in fase di data preparation.

2.4 Data exploration

Non sono state individuate relazioni tra i dati del dataset.

	obj_ID	alpha	delta	u	\
count	1.000000e+05	100000.000000	100000.000000	100000.000000	
mean	1.237665e+18	177.629117	24.135305	21.980468	
std	8.438560e+12	96.502241	19.644665	31.769291	
min	1.237646e+18	0.005528	-18.785328	-9999.000000	
25%	1.237659e+18	127.518222	5.146771	20.352353	
50%	1.237663e+18	180.900700	23.645922	22.179135	
75%	1.237668e+18	233.895005	39.901550	23.687440	
max	1.237681e+18	359.999810	83.000519	32.781390	

	g	r	i	z	\
count	100000.000000	100000.000000	100000.000000	100000.000000	
mean	20.531387	19.645762	19.084854	18.668810	
std	31.750292	1.854760	1.757895	31.728152	
min	-9999.000000	9.822070	9.469903	-9999.000000	
25%	18.965230	18.135828	17.732285	17.460677	
50%	21.099835	20.125290	19.405145	19.004595	
75%	22.123767	21.044785	20.396495	19.921120	
max	31.602240	29.571860	32.141470	29.383740	

	run_ID	rerun_ID	cam_col	field_ID	spec_obj_ID	\
count	100000.000000	100000.0	100000.000000	100000.000000	1.000000e+05	
mean	4481.366060	301.0	3.511610	186.130520	5.783882e+18	
std	1964.764593	0.0	1.586912	149.011073	3.324016e+18	
min	109.000000	301.0	1.000000	11.000000	2.995191e+17	
25%	3187.000000	301.0	2.000000	82.000000	2.844138e+18	
50%	4188.000000	301.0	4.000000	146.000000	5.614883e+18	
75%	5326.000000	301.0	5.000000	241.000000	8.332144e+18	
max	8162.000000	301.0	6.000000	989.000000	1.412694e+19	

	redshift	plate	MJD	fiber_ID
count	100000.000000	100000.000000	100000.000000	100000.000000
mean	0.576661	5137.009660	55588.647500	449.312740
std	0.730707	2952.303351	1808.484233	272.498404
min	-0.009971	266.000000	51608.000000	1.000000
25%	0.054517	2526.000000	54234.000000	221.000000
50%	0.424173	4987.000000	55868.500000	433.000000
75%	0.704154	7400.250000	56777.000000	645.000000
max	7.011245	12547.000000	58932.000000	1000.000000

obj_ID	0
alpha	0
delta	0
u	1
g	1
r	0
i	0
z	1
run_ID	0
rerun_ID	0
cam_col	0
field_ID	0
spec_obj_ID	0
class	0
redshift	0
plate	0
MJD	0
fiber_ID	0
dtype: int64	

	obj_ID	alpha	delta	u	g	r	i	z	run_ID	rerun_ID	\
79543	False	False	False	True	True	False	False	True	False	False	

	cam_col	field_ID	spec_obj_ID	class	redshift	plate	MJD	fiber_ID
79543	False	False	False	False	False	False	False	False

['n/a', 'na', '--', 'nan', 'NaN', -9999]
--

Eliminiamo l'unica riga che presenta dati mancanti.

3 Data Preparation

3.1 Data cleaning

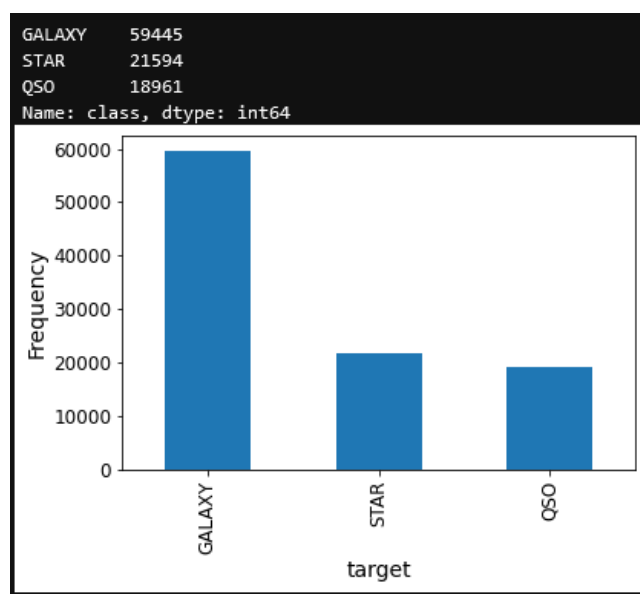
Dati i risultati ottenuti in fase di data exploration non troviamo necessario l'utilizzo di nessuna tecnica o algoritmo di data imputation.

3.2 Define data and target variables

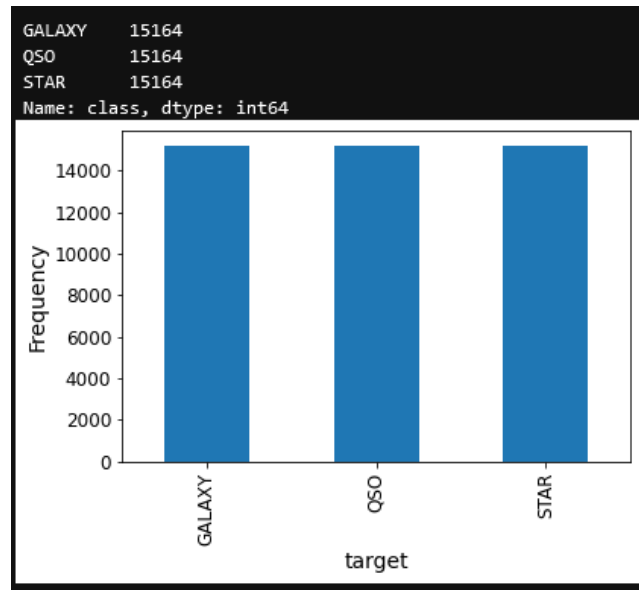
La variabile target è **class** ed è la quattordicesima colonna del dataset.

3.3 Data balancing

Il nostro dataset è sbilanciato, le righe riguardanti le galassie sono ben maggiori rispetto a stelle e quasar:



Data la soddisfacente quantità di istanze all'interno del nostro dataset riteniamo più opportuno applicare una tecnica di Under-sampling, in particolare valuteremo RandomUnderSampler e NearMiss. Pensiamo che il secondo ci permetterà di ottenere risultati più soddisfacenti ma decidiamo comunque di testarli entrambi per poi decidere in fase di valutazione quale utilizzare sul nostro modello (i risultati saranno inoltre valutati anche senza bilanciare il dataset). Prima di applicare una di queste tecniche dividiamo i dati in training set e test set, rispettivamente 80% e 20% del dataset originale, per poi ottenere il seguente risultato:



3.4 Feature scaling

I dati in nostro possesso, seppure nello stesso formato, riguardano range di valori tutti differenti tra loro. Per evitare che il nostro machine learner interpreti li interpreti male verranno tutti normalizzati in un range di valori $[0,1]$ con la tecnica Min-Max, in fase di validazione sceglieremo tra il modello con o senza normalizzazione.

3.5 Feature engineering

Analizzando il dominio del problema e i dati in nostro possesso, nonché le informazioni ricavate fin'ora, abbiamo deciso di rimuovere le colonne **obj_ID** e **rerun_ID** in quanto non utili alla classificazione (la prima è solo una chiave utilizzata nel dataset originale e la seconda si ripete identica in tutte le righe). Abbiamo inoltre deciso di utilizzare l'algoritmo SelectKBest per fare feature selection con $k=10$

4 Data Modeling

Per l'implementazione del nostro machine learner utilizzeremo un algoritmo basato su entropia: il **Decision Tree**. Testeremo su questo algoritmo le varie configurazioni discusse in questo documento per poi decidere tra queste il modello migliore.

Report dei risultati senza data balancing, normalizzazione e feature selection:

```

[[11640  165    99]
 [  806 3004    0]
 [    0   0 4286]]
precision    recall  f1-score   support

   GALAXY      0.94      0.98      0.96    11904
    QSO      0.95      0.79      0.86     3810
    STAR      0.98      1.00      0.99     4286

 accuracy              0.95    20000
  macro avg              0.95      0.92      0.94    20000
weighted avg              0.95      0.95      0.94    20000

Accuracy: 0.9465

```

Report dei risultati con normalizzazione e senza data balancing e feature selection:

```

[[2642 9260    2]
 [  27 3783    0]
 [4219   0  67]]
precision    recall  f1-score   support

   GALAXY      0.38      0.22      0.28    11904
    QSO      0.29      0.99      0.45     3810
    STAR      0.97      0.02      0.03     4286

 accuracy              0.32    20000
  macro avg              0.55      0.41      0.25    20000
weighted avg              0.49      0.32      0.26    20000

Accuracy: 0.3246

```

Report dei risultati con feature selection e senza data balancing e normalizzazione:

```

[[11640  165    99]
 [  806 3004    0]
 [    0   0 4286]]
precision    recall  f1-score   support

   GALAXY      0.94      0.98      0.96    11904
    QSO      0.95      0.79      0.86     3810
    STAR      0.98      1.00      0.99     4286

 accuracy              0.95    20000
  macro avg              0.95      0.92      0.94    20000
weighted avg              0.95      0.95      0.94    20000

Accuracy: 0.9465

```

Report dei risultati con feature selection e normalizzazione e senza data balancing:

[[2642 9260 2]					
[27 3783 0]					
[4219 0 67]]					
	precision	recall	f1-score	support	
GALAXY	0.38	0.22	0.28	11904	
QSO	0.29	0.99	0.45	3810	
STAR	0.97	0.02	0.03	4286	
accuracy			0.32	20000	
macro avg	0.55	0.41	0.25	20000	
weighted avg	0.49	0.32	0.26	20000	
Accuracy: 0.3246					

Report dei risultati con RandomUnderSampler e senza normalizzazione e feature selection:

[[11441 363 100]					
[677 3133 0]					
[0 0 4286]]					
	precision	recall	f1-score	support	
GALAXY	0.94	0.96	0.95	11904	
QSO	0.90	0.82	0.86	3810	
STAR	0.98	1.00	0.99	4286	
accuracy			0.94	20000	
macro avg	0.94	0.93	0.93	20000	
weighted avg	0.94	0.94	0.94	20000	
Accuracy: 0.943					

Report dei risultati con RandomUnderSampler e normalizzazione e senza feature selection:

[[2378 9524 2]					
[22 3788 0]					
[4015 0 271]]					
	precision	recall	f1-score	support	
GALAXY	0.37	0.20	0.26	11904	
QSO	0.28	0.99	0.44	3810	
STAR	0.99	0.06	0.12	4286	
accuracy			0.32	20000	
macro avg	0.55	0.42	0.27	20000	
weighted avg	0.49	0.32	0.26	20000	
Accuracy: 0.32185					

Report dei risultati con RandomUnderSampler e feature selection e senza normalizzazione:

[[11441	363	100]			
[677	3133	0]			
[0	0	4286]]			
	precision		recall	f1-score	support
GALAXY	0.94		0.96	0.95	11904
QSO	0.90		0.82	0.86	3810
STAR	0.98		1.00	0.99	4286
accuracy				0.94	20000
macro avg	0.94		0.93	0.93	20000
weighted avg	0.94		0.94	0.94	20000
Accuracy: 0.943					

Report dei risultati con RandomUnderSampler e normalizzazione e feature selection:

[[2378 9524 2]					
[22 3788 0]					
[4015 0 271]]					
	precision	recall	f1-score	support	
GALAXY	0.37	0.20	0.26	11904	
QSO	0.28	0.99	0.44	3810	
STAR	0.99	0.06	0.12	4286	
accuracy			0.32	20000	
macro avg	0.55	0.42	0.27	20000	
weighted avg	0.49	0.32	0.26	20000	
Accuracy: 0.32185					

Report dei risultati con NearMiss e senza normalizzazione e feature selection:

[[11440	364	100]			
[676	3134	0]			
[0	0	4286]]			
	precision		recall	f1-score	support
GALAXY	0.94		0.96	0.95	11904
QSO	0.90		0.82	0.86	3810
STAR	0.98		1.00	0.99	4286
accuracy				0.94	20000
macro avg	0.94		0.93	0.93	20000
weighted avg	0.94		0.94	0.94	20000
Accuracy: 0.943					

Report dei risultati con NearMiss e normalizzazione e senza feature selection:

```

[[2279 9531  94]
 [  22 3788   0]
 [  29   0 4257]]

```

	precision	recall	f1-score	support
GALAXY	0.98	0.19	0.32	11904
QSO	0.28	0.99	0.44	3810
STAR	0.98	0.99	0.99	4286
accuracy			0.52	20000
macro avg	0.75	0.73	0.58	20000
weighted avg	0.85	0.52	0.49	20000

Accuracy: 0.5162

Report dei risultati con NearMiss e feature selection e senza normalizzazione:

```

[[11440  364  100]
 [  676 3134   0]
 [    0   0 4286]]

```

	precision	recall	f1-score	support
GALAXY	0.94	0.96	0.95	11904
QSO	0.90	0.82	0.86	3810
STAR	0.98	1.00	0.99	4286
accuracy			0.94	20000
macro avg	0.94	0.93	0.93	20000
weighted avg	0.94	0.94	0.94	20000

Accuracy: 0.943

Report dei risultati con NearMiss e normalizzazione e feature selection:

```

[[2378 9524  2]
 [  22 3788   0]
 [4015   0 271]]

```

	precision	recall	f1-score	support
GALAXY	0.37	0.20	0.26	11904
QSO	0.28	0.99	0.44	3810
STAR	0.99	0.06	0.12	4286
accuracy			0.32	20000
macro avg	0.55	0.42	0.27	20000
weighted avg	0.49	0.32	0.26	20000

Accuracy: 0.32185

5 Evaluation

Osservando i risultati ottenuti notiamo che la feature selection è stata efficace nella misura in cui, una volta applicata, le performance del modello rimangono invariate, caratteristica desiderabile in quanto permette al nostro modello di lavorare efficacemente con meno percezioni a disposizione. Un altro risultato soddisfacente è rappresentato dall'applicazione del data balancing che non solo non peggiora le performance ma ci permette anche di superare, per quanto riguarda la recall dei quasar, la soglia decisa in fase di business success criteria: come ci aspettavamo il numero di galassie del nostro dataset è sufficiente anche dopo il bilanciamento ad assicurare un'adeguata classificazione di quest'ultime. Seppur non abbiamo ottenuto buoni risultati nelle configurazioni in cui normalizzavamo i dati possiamo ritenerci soddisfatti per aver ottenuto risultati in conformità con i nostri obiettivi in ben 4 configurazioni (underSampler e UnderSampler con feature selection, Near Miss e NearMiss con feature selection). Dal momento che la scelta dell'algoritmo di under-sampling non incide sui risultati ottenuti dal classificatore possiamo concludere scegliendo la configurazione NearMiss con feature selection per il nostro modello.