Università degli Studi di Milano - Bicocca
Scuola di Scienze
Dipartimento di Informatica, Sistemistica, Comunicazione
Informatica Magistrale

Exoplanet
Progetto Machine Learning

Autori:
829470 Federica Di Lauro
829827 Davide Cozzi
829835 Gabriele De Rosa

19 gennaio 2021

Repository link:
github.com/derogab/exoplanet

Indice

Capitolo 1

Introduzione

Durante la realizzazione di questo progetto ci siamo proposti di studiare alcuni dei modelli per il riconoscimento di esopianeti.

Il dataset da noi scelto per le analisi è messo a disposizione dalla NASA[1] e si tratta di un aggregatore di dati che sono stati raccolti dal telescopio **Kepler Space Telescope** per lo studio e la ricerca di esopianeti.
Nel dettaglio sono fornite una serie di informazioni relative a varie osservazioni selezionate come potenziali esopianeti che successivamente vengono quindi classificate come tali o meno; nel dettaglio è a loro applicata l'etichetta di *confermato*, *falso positivo* o, se ancora in fase di studio, *candidato* o *non classificato*.

Si è quindi iniziato con uno studio esplorativo del dataset per poi passare all'utilizzo e allo studio di alcuni modelli di machine learning.

[1] https://www.kaggle.com/nasa/kepler-exoplanet-search-results

Capitolo 2

Analisi del dataset

Come anticipato, il dataset fornisce informazioni in merito alle varie osservazioni ottenute dal **Kepler Space Telescope** dei cosiddetti **Kepler Object of Interest (*KOI*)**.

La NASA fornisce una descrizione di ogni colonna presente nel dataset e pubblicamente accessibile a questo link.

2.1 Analisi dei dati

La colonna *target* è rappresentata da **koi_disposition**, la quale fornisce l'indicazione in merito alla classificazione dell'oggetto studiato. Come è già stato anticipato, questa colonna può assumere i seguenti 4 valori:

- **CONFIRMED**, per indicare che quell'osservazione ha portato al riconoscimento effettivo dell'esopianeta.
- **FALSE POSITIVE**, per indicare che quell'osservazione è stata riconosciuta come un falso positivo, non indicando quindi un esopianeta.
- **CANDIDATE**, per indicare che la comunità scientifica non si è ancora espressa in merito alla natura dell'osservazione in quanto mancano ancora diversi test sulle osservazioni, nonostante siano stati fatti già i test che escludano l'osservazione dall'essere catalogata come *FALSE POSITIVE*.
- **NOT DISPOSITIONED**, per indicare che non sono stati ancora eseguiti nemmeno i test he escludano l'osservazione dall'essere catalogata come *FALSE POSITIVE*.

Proseguendo l'analisi del dataset individuiamo alcune colonne che identificano l'osservazione con dei valori assegnati a priori, come un numero incrementale ed un identificativo alfanumerico.

Si ha anche, per le osservazioni che hanno portato all'identificazione effettiva di un esopianeta, l'indicazione del nome assegnato al corpo celeste.

Le suddette informazioni sono generalmente riconducibili a due categorie di dati:

- **Identification Columns**
- **Exoplanet Archive Information**

Un ulteriore gruppo di dati è rappresentato dalla categoria **Project Disposition Columns**, dove troviamo ulteriori informazioni riguardanti l'osservazione.

Nel dettaglio abbiamo il valore **koi_pdisposition** che indica le classi che potrebbero essere assegnate in modo probabilistico a partire dai dati (a differenza di **koi_disposition** che analizza lo stato attuale dello studio delle osservazioni).

Si ha inoltre un'indicazione (**koi_score**) indicante lo score della confidenza del valore presente in **koi_disposition**.

Nella medesima categoria si trovano anche quattro flag booleani che vale la pena analizzare in modo approfondito:

1. **Not Transit-Like Flag (*koi_fpflag_nt*)** indicante che il KOI ha una curva di luce non è coerente con quella di un pianeta in transito (anche se questo potrebbe essere ricondotto ad errori nella strumentazione).
2. **Stellar Eclipse Flag (*koi_fpflag_ss*)** indicante che il KOI ha l'evento simile al transito potrebbe essere causato da un sistema stellare binario (anche un eventuale *gioviano caldo* potrebbe avere questo flag settato).
3. **Centroid Offset Flag (*koi_fpflag_co*)** indicante che il sorgente del segnale proviene da una stella vicina.
4. **Ephemeris Match Indicates Contamination Flag (*koi_fpflag_ec*)** indicante che il KOI condivide lo stesso periodo e l'epoca di un altro oggetto e che viene ritenuto essere il risultato di una "contaminazione" durante l'analisi.

Tutti i dati delle colonne seguenti rappresentano valori non booleani relativi alle effettive misurazioni (eventualmente arricchite con errori) di una certa osservazione.

Nel dettaglio si riconsocono varie categorie tra cui informazioni in merito al transito,a i dati della stella etc...

2.2 Selezione iniziale dei dati

Dopo una prima analisi del dataset si è scelto di modificarlo al fine di poter ottenere un dataset più pulito, con le sole indicazioni utili ai fini dello studio.
Innanzitutto si è deciso di concentrarsi unicamente sulle etichette **CONFIRMED** e **FALSE**

POSITIVE in quanto le altre due categorie non permetterebbero una corretta analisi delle predizioni dei modelli di machine learning, non avendo un'opinione "sicura" della comunità scientifica in merito ai KOI. Il problema viene quindi ridotto ad un caso binario.
Proseguendo si è ovviamente scelto di rimuovere le colonne relative a nomi ed identificatori (che in alcuni casi sono direttamente legati alla label **CONFIRMED**) dei KOI. Viene rimosso anche lo score di punteggio della confidenza.
Dal punto di vista dei quattro booleani trattati precedentemente viene scelto di non tenere nessuno dei quattro attributi. Questa scelta è dovuta al fatto che, nonostante, come descritto, lascino spazio per eventuali riconoscimenti sia positivi che negativi, sono prettamente legati, a seconda del flag, ad una certa label.
Il dataset forniva inoltre un attributo *koi_vet_stat* (con associata anche la data) che certifica, con il valore *Done*, che la comunità scientifica è giunta ad una conclusione ufficiale sul KOI. Viene scelto, per poter avere un confronto più sicuro tra gli esiti dei modelli e i dati ufficiali, di lavorare quindi solo con tali valori (trascurando quindi i record con tale attributo a valore *Active*). La colonna relativa alla data di validazione viene ovviamente rimossa.
Per lavorare con un dataset più pulito si è anche scelto di eliminare le colonne relative ad attributi in cui l'intero dataset presentava, per ogni riga, il solo valore *NULL*.
Vengono rimosse anche le poche colonne con variabili categoriche.

Una seconda osservazione è in merito al **downsample**. Il dataset presentava uno sbilanciamento nei confronti dei KOI **FALSE POSITIVE**. Si è quindi scelto, non avendo motivo di dare più peso a questi KOI, di procedere con il downsample.

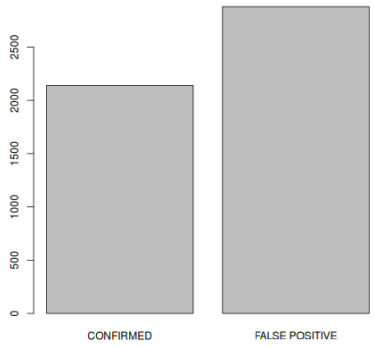

Figura 2.1: Distribuzione del target prima il downsampling

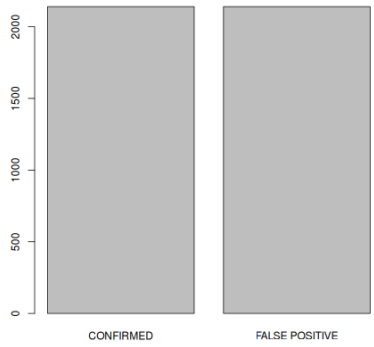

Figura 2.2: Distribuzione del target dopo il downsampling

2.3 Principal Component Analysis

Si è quindi proceduto con la PCA sul dataset intero. È stato scelto di tenere le feature per le quali, con la tecnica della PCA, si ottiene un autovalore maggiore o uguale a 1 in quanto è una delle metriche maggiormente utilizzate.

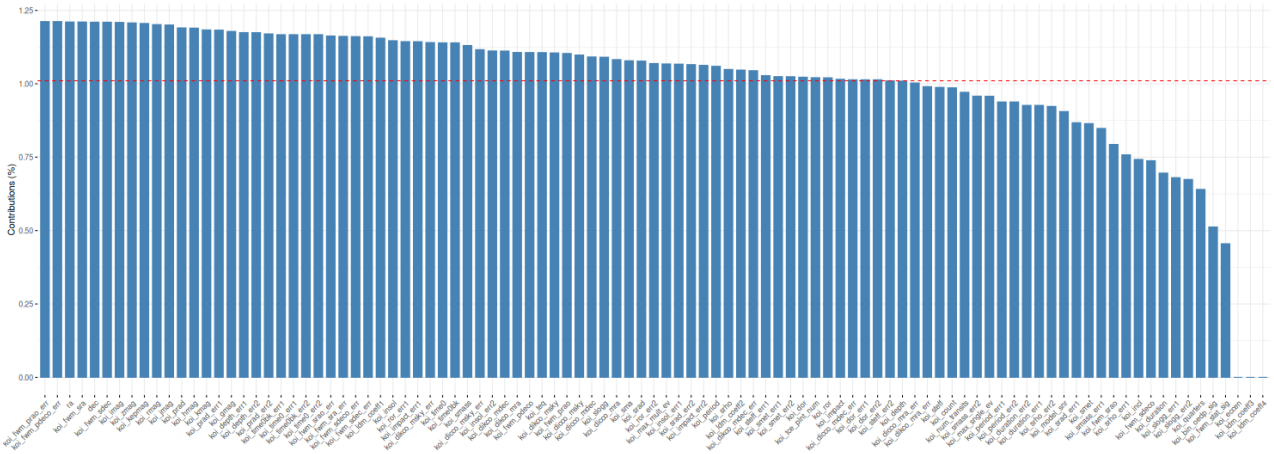

Figura 2.3: Barplot dei contributi delle variabili rispetto alle dimensioni della PCA. La linea rossa rappresenta il contributo medio atteso.

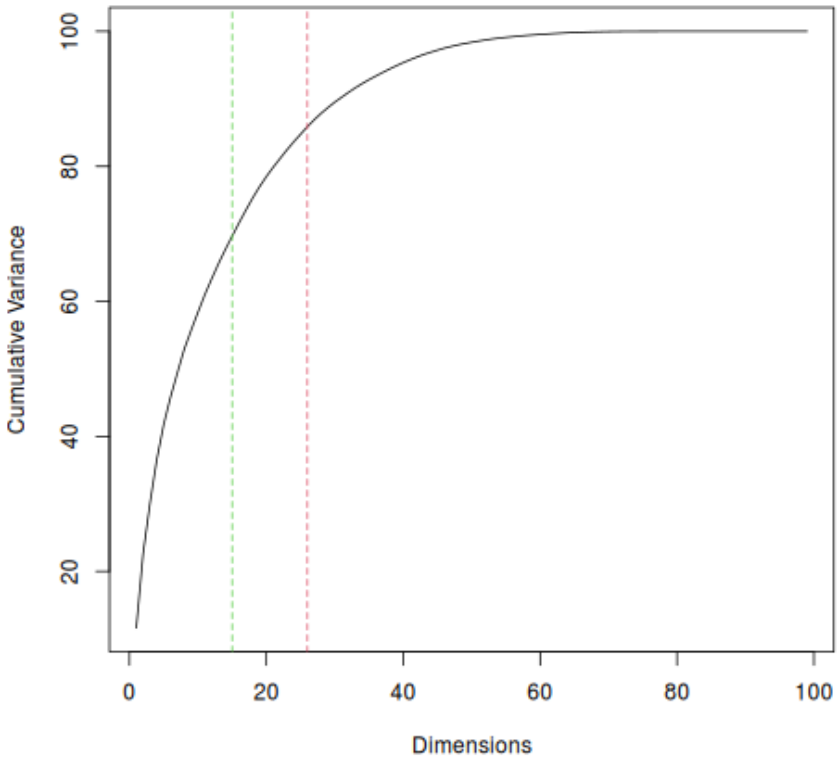

Figura 2.4: Plot della varianza cumulata delle componenti della PCA

Alla fine della PCA sono state quindi selezionate 26 feature più significative, a partire dalle iniziali 100, che verranno utilizzate per i modelli di machine learning. Queste numero di variabili comportano una varianza cumulata di $\sim 85\%$, come segnalato dalla linea verticale

rossa in figura 2.4 (mentre la linea verde segnala la varianza cumulata del $\sim 70\%$, un'altra metrica spesso usata al posto dell'autovalore maggiore di uno).