NASA EXOPLANET ARCHIVE:
Analisi delle Temperature degli Esopianeti

Introduzione:

Contesto:

Gli esopianeti, o pianeti extrasolari, sono corpi celesti che orbitano attorno a stelle diverse dal nostro Sole, offrendo uno sguardo su un universo estremamente ampio.

Alcuni esopianeti sono molto simili ai pianeti del nostro sistema solare, ma possono variare enormemente per dimensioni, composizione e distanza dalla loro stella madre.

Fino ad oggi, sono stati confermati più di 5.500 esopianeti in oltre 4.000 sistemi stellari. Di questi, in parte sono pianeti simili alla Terra, potenzialmente abitabili, ma molti sono completamente diversi da qualsiasi cosa si trovi nel nostro Sistema Solare.

Il dataset che abbiamo preso in esame contiene una lista di potenziali esopianeti identificati dalla missione spaziale TESS (Transiting Exoplanet Survey Satellite), lanciata dallac NASA nel 2018, con l'obiettivo principale di scoprire nuovi esopianeti.

Il satellite TESS utilizza un telescopio spaziale per monitorare e osservare il transito di pianeti che passano davanti alla loro stella madre. Quando il pianeta transita davanti alla stella (rispetto alla nostra linea di vista), blocca una piccola frazione della luce stellare. Questa riduzione viene registrata come un calo nella curva di luce della stella. L'analisi di questo fenomeno permette di stimare alcuni parametri come il raggio del pianeta, il periodo orbitale e la temperatura di equilibrio.

Dataset oggetto di studio:

Il dataset preso in osservazione, disponibile sul NASA exoplanet archive, contiene un elenco di TESS Objects of Interest (TOI), identificati dal progetto TESS. Esso è composto principalmente da candidati planetari (esopianeti), ma include anche pianeti in transito già noti e falsi positivi. Per ogni esopianeta riporta due tipologie di variabili: quelle relative ai singoli pianeti e quelle relative alle loro stelle madri.

Nome delle Variabili	Definizione
Pl_tranmid	Tempo medio di transito
Pl_orber	Periodo orbitale del pianeta (day)
Pl_trandep	Diminuzione del flusso stellare causato dal transito del pianeta
Pl_trandurh	Durata totale del transito
Pl_rade	Raggio del pianeta
Pl_insol	Insolazione - Energia solare ricevuta dal pianeta
Pl_eqt	Temperatura di equilibrio del pianeta (Kelvin)
St_tmag	Luminosità della stella ospitante
St_dist	Distanza del sistema planetario dalla nostra Terra
St_teff	Temperatura effettiva della stella ospite (Kelvin)
St_logg	Accelerazione gravitazionale (espressa in logaritmo)
St_rad	Raggio della stella
St_pmdec	Variazione angolare della declinazione nel tempo
St_pmra	Variazione angolare dell'ascensione retta nel tempo
dec	Declinazione del sistema planetario (gradi)
ra	Ascensione retta del sistema planetario (gradi)
Tfopwg_disp	PC, FA, FP, CP, KP, APC

Vediamo nel dettaglio la variabile Tfopwg_disp:
FA (False Alarm) si intende quando un segnale rilevato (ad esempio, un calo di luminosità della stella o una variazione nella velocità radiale) sembra indicare un pianeta, ma non soddisfa i criteri statistici o fisici per essere considerato un vero segnale planetario.
FP (False Positive) indica quando un segnale viene inizialmente attribuito a un pianeta, ma in realtà è causato da un altro fenomeno astrofisico.
KP (Known Planet) indica un oggetto che presenta segnali coerenti con quelli di un pianeta, ma che non è stato ancora confermato come tale attraverso ulteriori osservazioni o analisi.
CP (Confirmed Planet) è un oggetto che è stato definitivamente identificato come un esopianeta attraverso un'analisi approfondita.
PC (Planet Candidate) è un oggetto con alta probabilità di essere un pianeta, ma che non è ancora stato confermato e richiede osservazioni per escludere falsi positivi.
APC (Ambiguous Planetary Candidate) indica che l'oggetto ha caratteristiche che potrebbero suggerire sia una natura planetaria che un'origine non planetaria. I dati non sono sufficienti per confermare né escludere che si tratti di un pianeta.

Domande e Obiettivi:

L'obiettivo della nostra analisi è quello di studiare la temperatura di equilibrio dei pianeti (pl_eqt), essa infatti rappresenta una variabile molto significativa per la classificazione degli esopianeti. Per attuare questo tipo di studio, scegliamo di restringere la nostra analisi solamente ai "Pianeti Candidati". In particolare, implementeremo due modelli per analizzare la temperatura di equilibrio: il primo di tipo Model-based classification e il secondo di tipo Mixture of Expert Model.

Data cleaning e Data Transformation:

Data cleaning:

Procediamo eliminando tutti i valori mancanti (NA) e i valori nulli presenti nel dataset. Non usiamo alcun metodo di imputazione poiché essi rappresentano una piccola parte del dataset.

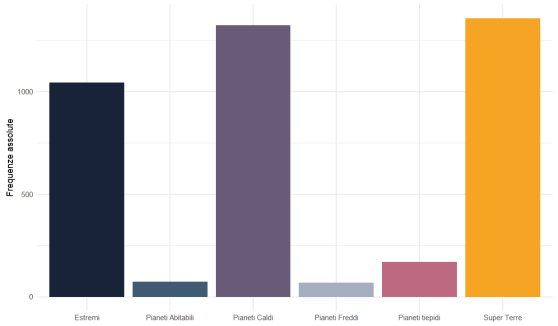
Data Transformation:

La temperatura di equilibrio di un esopianeta (espressa in Kelvin) dipende, a sua volta, dalla temperatura della stella ospite, dalla distanza del pianeta dalla stella e dalle caratteristiche atmosferiche del pianeta, come l'insolazione ovvero l'energia solare ricevuta dal pianeta.
A seguito valutiamo di creare delle classi in base alla temperatura di equilibrio di ogni pianeta. Questa classificazione è spesso utilizzata per raggruppare i vari pianeti.

La classificazione basata sulla temperatura misurata in Kelvin è la seguente:

Categoria	Intervallo di Temperatura
Pianeti Freddi	$T \leq 273$ K
Pianeti Abitabili	$273 \text{ K} < T \leq 373$ K
Pianeti Tiepidi	$373 \text{ K} < T \leq 500$ K
Pianeti Caldi	$500 \text{ K} < T \leq 1000$ K
Super Terre	$1000 \text{ K} < T \leq 1500$ K
Pianeti Estremamente Caldi	$T > 1500$ K

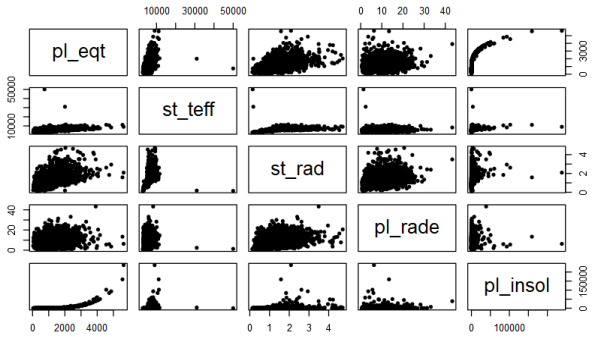
Il seguente istogramma mostra la distribuzione di osservazioni per ogni classe precedentemente creata.



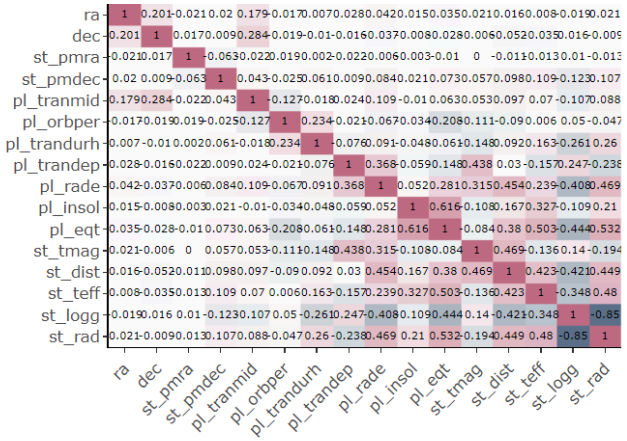
Notiamo la forte disparità nella numerosità delle classi. Esse non sono equamente rappresentate.

Analisi Esplorative:

Scatterplot per visualizzare le variabili:

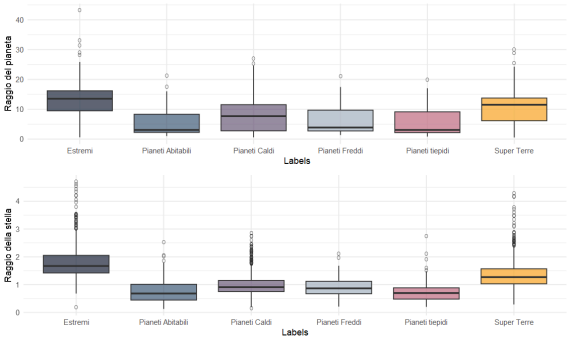


Prendiamo in considerazione le variabili maggiormente correlate con la temperatura di equilibrio. Come si evince dalla seguente matrice di correlazione, tali variabili sono: il raggio del pianeta (pl_rade), il raggio della stella madre (st_rad), la temperatura effettiva della stella (st_teff) e l'insolazione del pianeta (pl_insol).



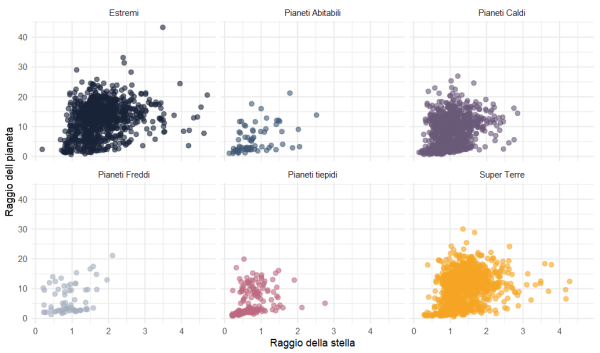
Come varia il Raggio del pianeta e della stella in base alla classe di Temperatura:

Nel grafico sotto riportato possiamo notare come le variabili prese in esame variano molto in base al gruppo di appartenenza. Il raggio del pianeta, per esempio, è in media più grande nelle classi caratterizzate da temperatura maggiore (“Estremamente caldi”, “Caldi” e “Super Terre”). Analogamente possiamo notare la stessa variazione per il raggio della stella madre.



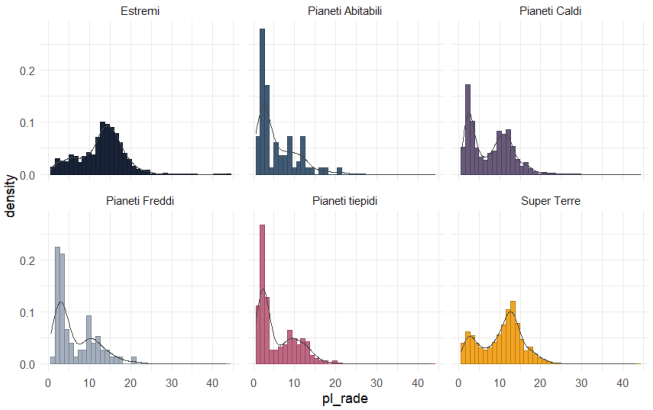
Il Raggio del pianeta è proporzionale al raggio della stella:

Prendiamo in esame nel seguente Scatterplot la relazione tra la dimensione del raggio del pianeta e quello della stella madre diviso per le diverse classi dei pianeti. Possiamo notare che con l’aumentare del raggio del sole aumenta anche quello dell’esopianeta.



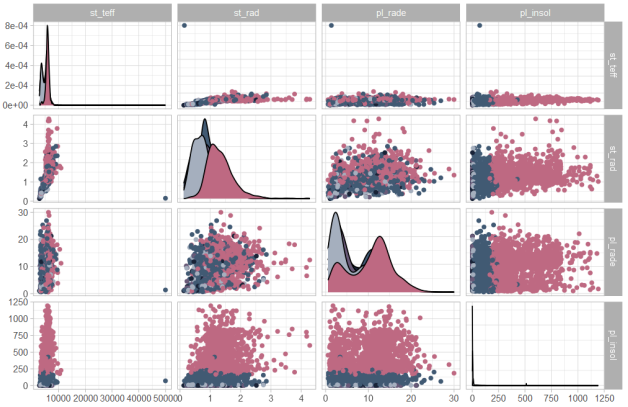
Analisi di densità della grandezza dei pianeti

Osserviamo che con un aumento della temperatura del pianeta, dato che sono presenti più osservazioni, la stima delle densità diventa sempre più precisa come si può notare nel grafico. Inoltre, notiamo vari picchi nelle stime di densità, evidenti maggiormente nelle classi con temperatura più bassa.



Visualizzazione delle classi per ogni coppia di variabile:

Dal seguente scatterplot possiamo notare una netta separazione tra le classi create. Questo risultato ci porta a studiare il fenomeno utilizzando un modello di classification (Model-based classification).



Model-Based Classification:

Per la seguente analisi non prendiamo in considerazione i pianeti nella classe “Pianeti Estremamente Caldi”, poiché sono caratterizzati da temperature fuori dal normale.

Cerchiamo il miglior modello attraverso la minimizzazione della media del MER (Missclassification Error Rate) stimata tramite Cross Validation (CV) e il BIC per classificare i pianeti nelle 5 classi (pianeti freddi, pianeta abitabili, pianeti tiepidi, pianeti caldi e super terre):

Ripetendo per 50 volte la stima del modello troviamo che quello migliore è il “Gaussian_pk_Lk_D_Ak_D”, ovvero il modello VVE (modello Ellissoidale, con volume e forma variabili e orientamento uniforme) secondo la scomposizione VSO (Volume Shape Orientation).

Infatti è possibile decomporre la matrice di varianze e covarianze sfruttando il Teorema Spettrale: $\Sigma = \lambda jDAjD$

Risultati:

Analizziamo il modello osservando le etichette predette e le confrontiamo con quelle vere del modello. Calcoliamo l'adjusted rand index (ARI) e otteniamo un risultato pari a 0.8045, il che evidenzia una buona vicinanza della partizione stimata rispetto a quella reale. La bontà di tale classificazione viene anche sottolineata nella Confusion Matrix, la quale riporta sulle righe le classi predette dal modello e sulle colonne le vere classi:

Confusion Matrix:

PIANETI	Abitabili	Caldi	Freddi	Tiepidi	S. Terre
Abitabili	23	0	3	3	0
Caldi	0	531	0	6	29
Freddi	1	0	24	0	0
Tiepidi	2	11	0	76	0
S. Terre	0	34	1	1	585

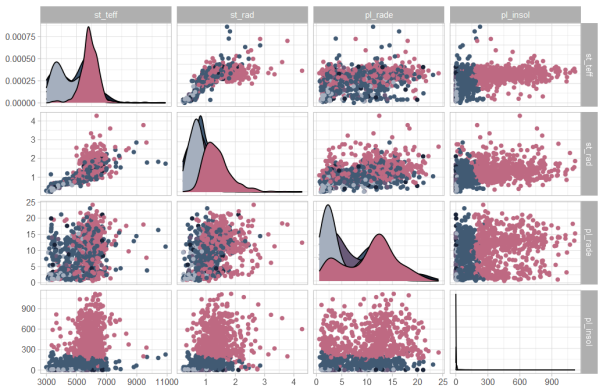
Otteniamo un accuracy pari a 93,16%, un risultato ottimo. Il modello riesce a classificare molto bene tutte le classi anche per quanto riguarda la "sensitivity" e "specificity"; non sembrano esserci problemi di Imbalanced Classes.

	Abitabili	Caldi	Freddi	Tiepidi	S. Terre
Sensitivity	0.808	0.906	0.857	0.849	0.958
Specificity	0.995	0.954	0.999	0.986	0.939

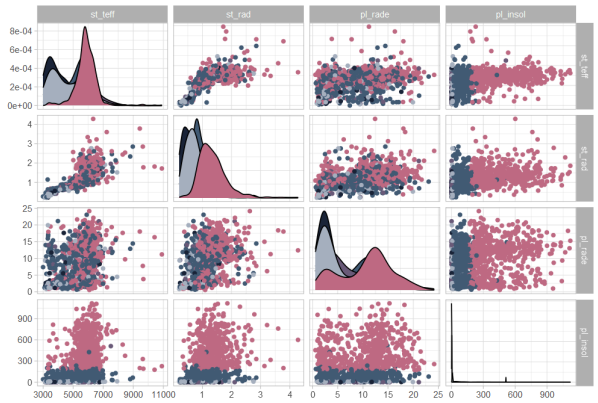
Visualizzazione dei risultati:

Visualizziamo i risultati della classification, confrontando le classi predette (della temperatura) sul test-set con le classi realmente osservate sempre sugli stessi dati. Confrontiamo sia gli scatterplot tra le variabili considerate sia le stime non parametriche della funzione di densità.

Classi predette per i dati di Test:



Classi reali per i dati di Test:



Notiamo un'ottima classificazione come avevamo precedentemente osservato dall'Accuracy e dall'Adjusted Rand Index.

Può la classe di Temperatura essere spiegata e predetta dalle variabili considerate?

L'analisi condotta ci ha permesso di classificare i pianeti in base alla loro temperatura di equilibrio con ottimi risultati. Il raggio del pianeta, il raggio della stella, la Temperatura effettiva della stella e l'insolazione del pianeta si sono rivelate variabili utili e sensate per spiegare e predire la classe di temperatura.

Mixture of Expert Models:

Prendiamo in esame la relazione tra la temperatura di equilibrio del pianeta e le covariate maggiormente correlate (secondo il criterio precedentemente utilizzato) tramite un modello di regressione di misture finite. Cerchiamo il numero di gruppi ottimale secondo il criterio del BIC (Bayesian Information Criterion) e dell'ICL (Integrated Complete Likelihood) e otteniamo come miglior risultato un numero di cluster pari 2.
Il modello giunge a convergenza in 214 iterazioni e presenta un BIC pari a 53258.67 e un ICL di 53276.84.

Parametri Stimati:

Otteniamo quindi due rette di regressione, una per ogni componente, con i seguenti parametri:

Componente 1:

	Parametri	Std. Error	Z value	Pr(>|z|)
Intercetta	5.0507e+02	1.1526e+01	43.8205	< 2.2e-16
st_rad	2.5684e+01	7.4389	3.4527	0.000555
pl_rade	3.7872e+00	5.5135e-01	6.8690	6.464e-12
st_teff	1.3814e-02	2.4127e-03	5.7253	1.033e-08
pl_insol	1.2909e+00	1.4310e-02	90.2097	< 2.2e-16

La prima componente presenta tutte le variabili significative al 1%.

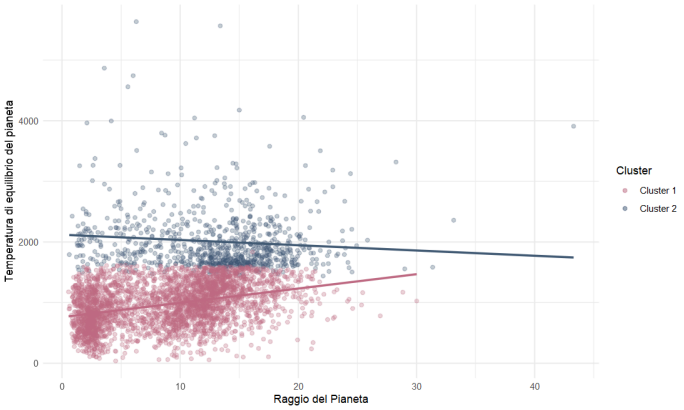
Componente 2:

	Parametri	Std. Error	Z value	Pr(>|z|)
Intercetta	1.3062e+03	5.4466e+01	23.9813	< 2.2e-16
st_rad	8.8753e+01	1.6698e+01	5.3154	1.064e-07
pl_rade	-5.3909	1.8082	-2.9813	0.00287
st_teff	5.7186e-02	7.8108e-03	7.3214	2.454e-13
pl_insol	3.0586e-02	8.5848e-04	35.6285	< 2.2e-16

Anche la seconda componente presenta tutte variabili significative per il modello di regressione.

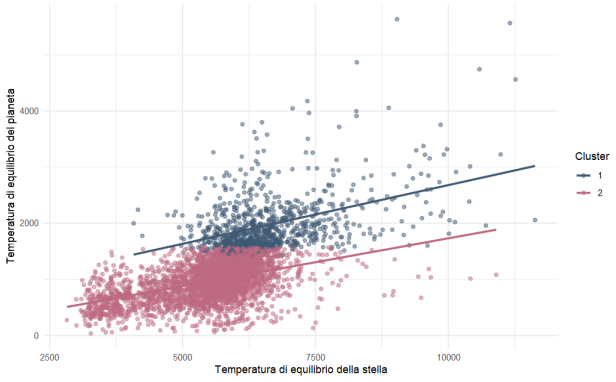
Possiamo notare delle particolarità nella stima dei parametri nelle due componenti: nello specifico le stime dei parametri del raggio del pianeta sono entrambe significative, ma con il segno opposto. Nelle due componenti troviamo quindi differenze (di segno o di grandezza) nell'effetto delle covariate sulla temperatura di equilibrio.

Come cambia l'effetto del raggio del pianeta sulla temperatura di equilibrio in base ai cluster?



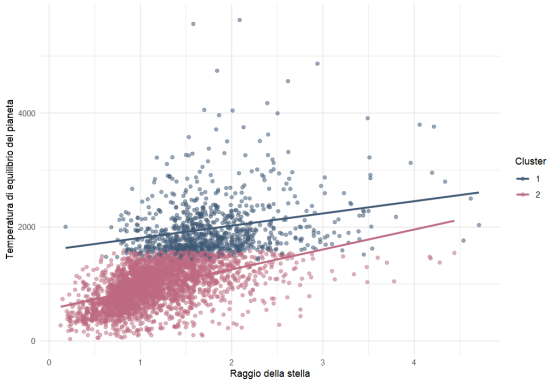
Come detto in precedenza l'effetto del raggio del pianeta sulla temperatura di equilibrio cambia molto in base al cluster, in questo caso hanno addirittura un effetto opposto.
Nel cluster 1 a seguito dell'incremento unitario del raggio del pianeta, a parità delle altre covariate, la temperatura di equilibrio aumenta in media di 3,78 K.
Invece nel cluster 2 a seguito dell'incremento unitario del raggio del pianeta, a parità delle altre covariate, la temperatura di equilibrio diminuisce in media di 5,39 K.

Come cambia l'effetto della temperatura effettiva della stella sulla temperatura di equilibrio in base ai cluster?



In questo caso l'effetto delle componenti è molto simile, infatti le due rette di regressione sembrano quasi parallele.

Come cambia l'effetto del raggio della stella sulla temperatura di equilibrio in base ai cluster?



Notiamo delle differenze nell'effetto del raggio della stella, non di segno ma di grandezza. Infatti per quanto riguarda il cluster 1, a seguito di un incremento unitario del raggio della stella e a parità delle altre covariate la temperatura di equilibrio aumenta in media di 88,75 K. Invece per quanto riguarda il cluster 2, a seguito di un incremento unitario del raggio della stella e a parità delle altre covariate la temperatura di equilibrio aumenta in media di 25,6K.

Si possono ottenere dai cluster individuati nel modello di regressione a misture finite informazioni utili per la temperatura di equilibrio?

Il modello di regressione a misture finite, applicato alla temperatura di equilibrio dei pianeti, ha permesso di identificare due sottogruppi distinti con caratteristiche specifiche, evidenziando diverse relazioni tra la variabile di risposta e le covariate considerate. Questa metodologia si è rivelata particolarmente efficace per catturare eterogeneità latenti nel dataset, fornendo una rappresentazione più accurata rispetto ai modelli di regressione tradizionali.

Sitografia:

- https://iopscience.iop.org/article/10.3847/538-3881/aba2cb
- https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=TOI
- https://exoplanetarchive.ipac.caltech.edu/docs/API_TOI_columns.html