Universitat Oberta de Catalunya (UOC)

Màster Universitari en Ciència de Dades (*Data Science*)

TREBALL FINAL DE MÀSTER

Àrea: 2

# Extracció atmosfèrica dels exoplanetes de la missió Ariel utilitzant mètodes de conjunt basats en arbres de decisió

---

Autora: Elisabet Ejarque Gonzalez

Tutora: Laura Ruiz Dern

Co-tutors: Octavi Fors, Andrea Butturini

---

Barcelona, 24 de juny de 2023

# FITXA DEL TREBALL FINAL

| | |
|---|---|
| Títol del treball: | Extracció atmosfèrica dels exoplanetes de la missió Ariel utilitzant mètodes de conjunt basats en arbres de decisió |
| Nom de l'autora: | Elisabet Ejarque Gonzalez |
| Nom de la tutora: | Laura Ruiz Dern |
| Nom dels co-tutors: | Octavi Fors, Andrea Butturini |
| Nom del/de la PRA: | Jordi Casas Roma |
| Data de lliurament: | 06/2023 |
| Titulació: | Màster Universitari en Ciència de Dades |
| Àrea del Treball Final: | Àrea 2 |
| Idioma del treball: | Català |
| Paraules clau | Aprenentatge en conjunts, Random Forest, Gradient boosting, Extracció atmosfèrica, missió Ariel |

# Resum

En les darreres dues dècades s'ha descobert la presència de fins a 5272 exoplanetes, els quals ja han començat a redefinir la nostra comprensió sobre la formació i evolució dels sistemes planetaris. Actualment, s'estan centrant esforços en passar de la detecció a la caracterització d'aquests exoplanetes, és a dir, entendre de què estan fets. Gràcies a properes missions com el recent James Webb Space Telescope, i la missió Ariel de l'Agència Espacial Europea (prevista per llançament el 2029), s'espera obtenir una quantitat de dades espectrals sense precedents, les quals permetran caracteritzar la composició i propietats físiques de les atmosferes d'aquests mons llunyans.

Malgrat tot, els mètodes actuals per interpretar els espectres atmosfèrics són computacionalment molt costosos i poden representar un coll d'ampolla a l'hora de processar tot el volum de dades que es preveuen generar en els propers anys. Davant d'aquest escenari, l'àmbit de l'aprenentatge automàtic s'està plantejant com una alternativa potencialment més flexible i amb menys requeriments computacionals. Recentment s'ha posat a disposició de la comunitat científica l'Atmospheric Big Challenge Database (ABC Database) (Changeat i Yip 2023), un conjunt de dades que simulen el volum i complexitat de dades que es mesuraran en la missió Ariel. Aprofitant l'oportunitat que representa aquest conjunt de dades, en aquest treball s'-ha explorat l'ús de models de conjunt (Random Forest, Gradient Boosting) per tal d'extreure informació de temperatura i composició atmosfèrica a partir de dades espectrals.

Els resultats obtinguts han demostrat que aquesta família de tècniques, aplicades a l'ABC Database, tenen una capacitat de precció més elevada que el mètodes tradicionals d'inferència Bayesiana Nested Sampling. A més, tots els models desenvolupats han presentat uns temps d'entrenament de màxim 1.5 minuts, posant de relleu la seva eficiència computacional. En conseqüència, els mètodes de conjunt basats en arbres de decisió es posicionen com una alternativa prometedora als mètodes actuals per fer front al gran volum de dades que s'adquirirà en les properes missions dedicades a la caracterització atmosfèrica d'exoplanetes.

*In the last two decades up to 5272 exoplanets have been discovered, which have already begun to redefine our understanding of the formation and evolution of planetary systems. Recently,*

*scientists have turned their attention from detection to the characterization of exoplanet atmospheres. Thanks to upcoming missions such as the recently launched James Webb Space Telescope, and the European Space Agency's Ariel Mission (due for launch in 2029), an unprecedented amount of atmospheric transmission spectra is to be obtained. These data will enable the characterization of the chemical composition and physical properties of these distant worlds.*

*However, current state-of-the-art methods for interpreting atmospheric spectral data are computationally expensive and may pose a bottleneck when processing the expected volume of data to be generated in the coming years. In this context, the field of machine learning is emerging as a promising alternative due to its high flexibility and rapid inference time. Recently, the Atmospheric Big Challenge Database (ABC Database) has been released to the community. This data set simulates the quantity and quality of data that will be measured in the Ariel mission. Seizing the opportunity presented by this dataset, this study explores the use of ensemble models (Random Forest, Gradient Boosting) to retrieve temperature and chemical composition information from spectral data.*

*The obtained results have demonstrated that this family of techniques, applied to the ABC Database, exhibit higher predictive capabilities compared to traditional Bayesian method Nested Sampling. Furthermore, all developed models have shown training times of up to 1.5 minutes, highlighting their computational efficiency. Consequently, ensemble methods based on decision trees emerge as a promising alternative to current methods in handling the large volume of data expected to be acquired in future missions dedicated to the atmospheric characterization of exoplanets.*

**Paraules clau / Keywords**: Ensemble models, Random Forest, Gradient Boosting, Atmospheric retrieval, Exoplanets, Ariel mission.

# Índex

# Índex de figures

# Índex de taules

# Capítol 1

# Introducció

## 1.1 Justificació de la proposta de treball

La descoberta del planeta 51 Peg b, anunciada per Mayor i Queloz (1995), va representar l'inici de les ciències exoplanetàries. Es tractava de la primera evidència científica de l'existència d'altres mons orbitant estrelles semblants a la nostra més enllà del sistema solar. Però la sorpresa no va quedar només en demostrar la seva existència, sinó que les característiques d'aquest nou planeta eren totalment inesperades. 51 Peg b tenia una massa similar a la de Júpiter, però en canvi, la seva òrbita era tan propera a la seva estrella, que completava una revolució sencera cada 4 dies. Aquesta troballa va ser tant sorprenent com inesperada, ja que segons les teories de formació planetària del moment, un planeta semblant a Júpiter s'hauria hagut de formar a distàncies molt més llunyanes de la seva estrella. Era una incògnita com sobrevivia aquest planeta tan a prop de la seva estrella, i com havia arribat fins allà.

La descoberta d'aquest planeta va ser possible gràcies a avenços tecnològics que van permetre mesurar petits moviments orbitals de la seva estrella induïts per la presència del propi planeta orbitant al seu voltant (velocitat radial). Des d'aleshores, aquest mètode ha permès descobrir fins a 1027 planetes (NASA Exoplanet Archive 2023). Però el gran pas endavant es va produir l'any 1999 amb la detecció del primer exoplaneta mitjançant la tècnica del trànsit (Charbonneau et al. 2000). Aquesta tècnica consisteix en mesurar la disminució de la brillantor d'una estrella per efecte del pas d'un exoplaneta, el qual bloqueja part de la llum emesa per l'estrella. Aquest mètode va suposar una acceleració de la detecció d'exoplanetes, especialment des del llançament de missions com CoRot (Convection, Rotation and planetary Transits) (Pätzold et al. 2012), Kepler (Borucki et al. 2010) i TESS (Transiting Exoplanet Survey Satellite) (Ricker et al. 2014), amb les quals s'han fet fins al 75% de les deteccions.

En l'actualitat hi ha fins a 5272 exoplanetes confirmats (NASA Exoplanet Archive 2023), dels quals se n'han pogut caracteritzar paràmetres macroscòpics com el període orbital o la dis-

tància a la seva estrella. Els resultats obtinguts fins ara, mostren no només que els exoplanetes són omnipresents, sinó que presenten un rang de diversitat molt més ampli del que existeix al nostre sistema solar. Tot i que hi ha un biaix metodològic pel fet que els planetes més grans i amb períodes més curts són els més fàcils d'observar, s'han detectat una gran quantitat de planetes amb característiques que no es troben al nostre sistema solar, com ara gegants gasosos en òrbita extremament propera a la seva estrella (*hot jupiters*) (Fortney, Dawson i Komacek 2021), o planetes amb una massa intermèdia entre la de la Terra i la de Neptú (Haghighipour 2013).

Per tant, la demografia dels exoplanetes descoberts fins ara suggereix que el nostre sistema solar no tindria una arquitectura típica dins del context còsmic (Gaudi, Christiansen i Meyer 2021). Per tal d'entendre millor els processos de formació i evolució de sistemes planetaris, necessitem comprendre millor les característiques d'aquests planetes, és a dir, de què estan fets.

### 1.1.1 De la detecció a la caracterització

El mètode del trànsit, consistent en analitzar la llum d'una estrella afectada pel pas d'un exoplaneta, representa una oportunitat per caracteritzar les propietats físiques i químiques d'una atmosfera. Quan la llum d'una estrella intercepta un planeta, una fracció dels seus rajos passaran a través de la seva atmosfera, i seran selectivament absorbits pels components químics que contingui (Figura 1.1). A més, característiques físiques com el perfil tèrmic i la presència de núvols alteraran com es transmet la llum a través del medi atmosfèric. Mesurant el trànsit d'un exoplaneta en funció de diferents longituds d'ona, estem capturant l'efecte conjunt de totes aquestes interaccions, obtenint un resum integrat de les propietats físiques i químiques d'aquesta atmosfera (Figura 1.2). Aquesta tècnica s'anomena espectroscòpia de transmissió.

Cada regió de l'espectre captura la presència de diferents tipus d'espècies químiques, així com de diferents regions de l'atmosfera (Madhusudhan 2019). Per exemple, la regió ultraviolada absorbeix principalment espècies atòmiques, presents a les regions més externes de l'atmosfera; mentre que en l'infrarroig es poden detectar espècies moleculars com l'$H_2O$, CO, $CO_2$ i el $CH_4$, presents en capes més baixes de l'atmosfera. Altres molècules com metalls pesants (TiO, VO, TiH) s'absorbeixen a la zona visible de l'espectre.

Addicionalment, propietats com la temperatura i la pressió afectaran la forma de l'espectre d'una forma més global, desplaçant l'escala i magnitud de les característiques espectrals. A majors temperatures, l'atmosfera es dilata i per tant la senyal atmosfèrica en la seva globalitat serà major. En canvi, a major pressió, l'atmosfera es contraurà i per tant la senyal espectral serà menor (Changeat i Yip 2023).

Així doncs, l'estudi de les atmosferes exoplanetàries consisteix en interpretar un espectre de transmissió, i extreure'n la informació sobre els processos subjacents que l'han generat. És

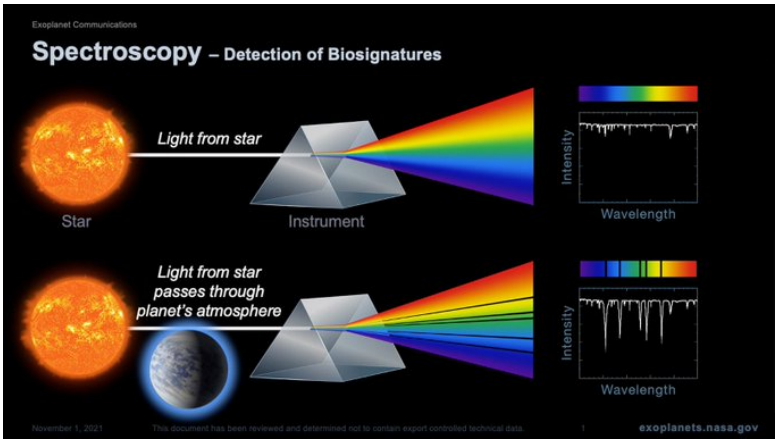


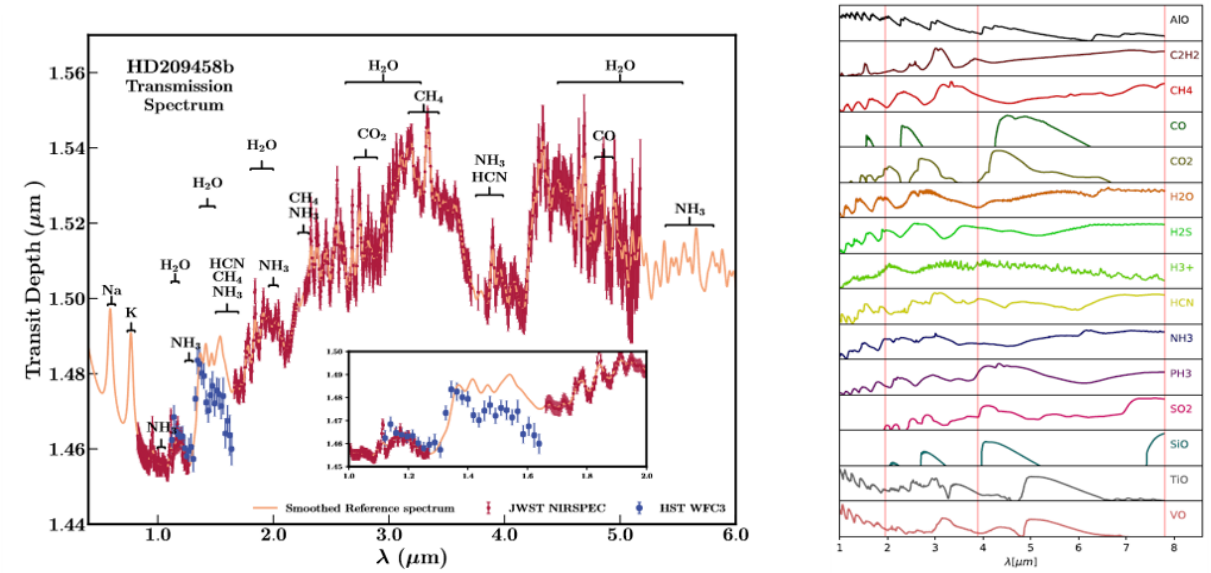

Figura 1.1: Mesura d'un espectre de transmissió. Quan la llum d'una estrella passa a través d'una atmosfera planetària, algunes bandes de l'espectre electromagnètic són selectivament absorbides, en funció de les molècules presents en aquella atmosfera. Font: exoplanets.nasa.gov

el que s'anomena problema invers (Potthast 2006). L'observador té accés a l'efecte (espectre de transmissió) produït per un procés (definit per paràmetres atmosfèrics). L'objectiu del problema invers és recuperar les causes amagades que han produït l'efecte observat, sovint amb la dificultat que l'efecte només s'observa parcialment a causa de pèrdua d'informació (soroll, rang espectral mesurat limitat, etc). Específicament en el cas de l'estudi d'espectres atmosfèrics, aquest problema s'anomena extracció atmosfèrica (*atmospheric retrieval*), i té per objectiu recuperar les proporcions de gasos atmosfèrics, i propietats físiques com ara la temperatura, a partir d'un espectre observat.

### 1.1.2 Metodologies actuals d'extracció atmosfèrica i les seves limitacions

Fins a dia d'avui, l'extracció atmosfèrica ha consistit principalment en ajustar un model atmosfèric (*parametric forward model*) a un espectre observat, i fer una estimació dels paràmetres del model i de les seves incerteses mitjançant un algorisme d'optimització (Madhusudhan 2018) (Figura 1.3). Com a algorisme d'optimització s'utilitza un mètode d'inferència estadística que mostreja l'espai de paràmetres del model, donades les dades observades. Els mètodes més utilitzats es basen en la inferència Bayesiana, la qual permet avaluar les distribucions posteriors dels paràmetres del model. Aquests mètodes inclouen principalment *Nested Sampling* (Madhusudhan 2018), *Markov Chain Monte Carlo* (MCMC), i *Optimal Estimation*.

L'aspecte més important a l'hora d'estimar les abundàncies d'espècies químiques de forma acurada, és la qualitat dels espectres. En especial, la seva precisió i el seu rang espectral. En aquest sentit, les dades disponibles fins a dia d'avui estan fortament limitades en el seu rang




(a) (b)

Figura 1.2: (a) Exemple d'un espectre de transmissió, on s'indiquen les regions on es detecten diferents molècules. Font: Madhusudhan (2019) (b) Senyals espectroscòpiques individuals de les molècules més freqüentment observades. Font: Tinetti et al. (2021)

espectral. Les fonts principals d'espectres de transmissió han estat el Hubble Space Telescope (HST) (Figura 1.3a), amb un rang en l'infrarroig de ∼1.1 – 1.7 µm; i anteriorment el Spitzer Space Telescope, amb un rang de ∼3.6 a 4.5 µm, també en l'infrarroig. Un dels problemes d'observar un rang tan limitat és que diferents conjunts de paràmetres d'un model atmosfèric poden predir un mateix espectre observat (degeneració del model). La principal forma de superar aquesta limitació és mesurar espectres al llarg d'un rang espectral més ampli, idealment des del visible fins a l'infrarroig (Barstow et al. 2016; Madhusudhan 2018).

Tanmateix, espectres més precisos i de major rang espectral també impliquen altres reptes. Principalment, el fet que el model atmosfèric necessari per ajustar l'espectre observat també haurà d'incrementar en complexitat. Això és una limitació important, ja que les tècniques d'inferència Bayesiana impliquen realitzar típicament entre $10^5$ i $10^8$ iteracions del model atmosfèric per assolir convergència i, per tant, només són viables aquells models que tinguin un temps d'execució de l'ordre de segons (Changeat i Yip 2023).

### 1.1.3 Nova generació de telescopis i metodologies

El camp de la caracterització d'atmosferes exopolanetàries està a punt d'experimentar una revolució, amb el llançament de la nova generació de telescopis com el recent James Webb Space Telescope (JWST) de la NASA/ESA/CSA (Greene et al. 2016), i les futures ESA Ariel Mission

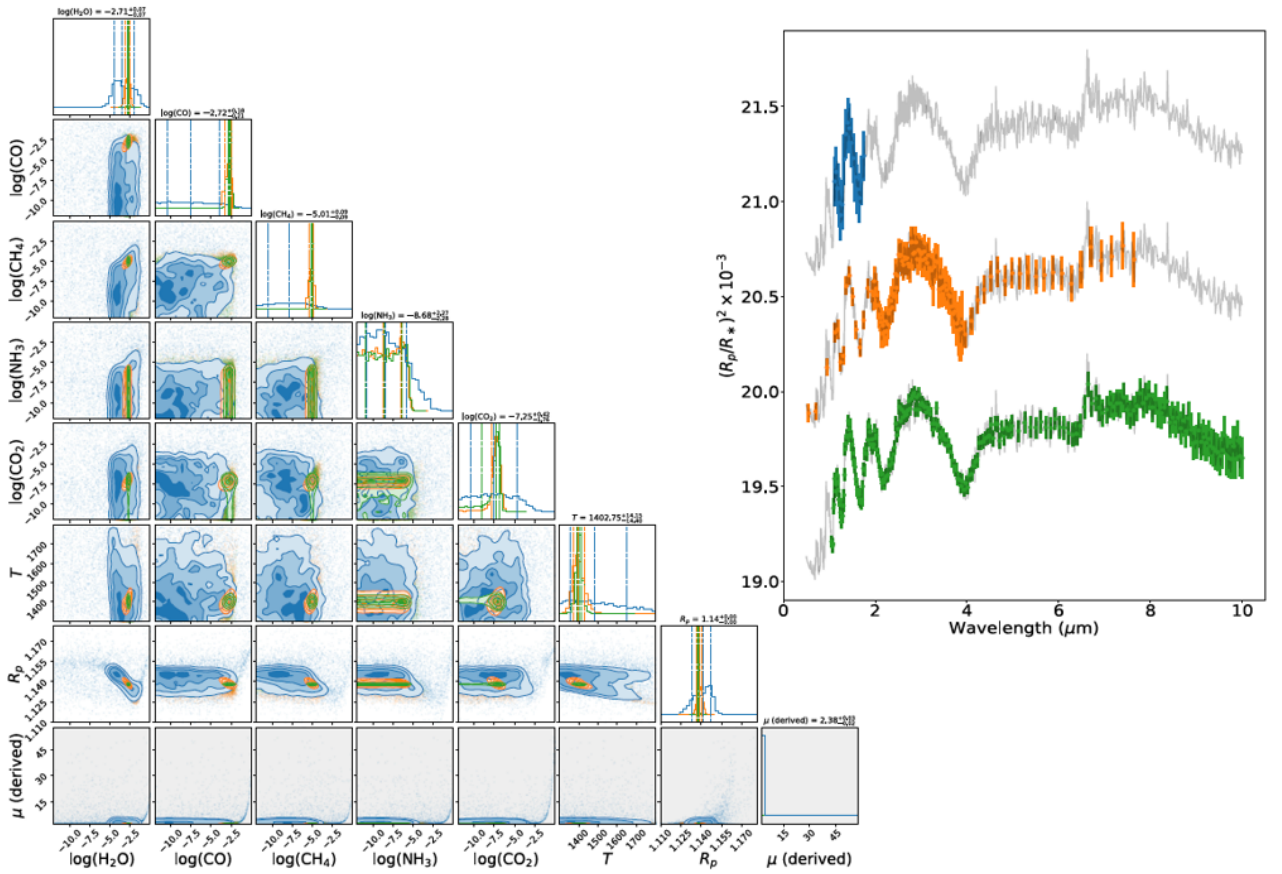



Figura 1.3: Dreta: Espectres simulats pel Hubble Space Telescope (HST, blau), Ariel (taronja) i James Webb Space Telescope (JWST, verd). Es pot veure el gran increment en el rang de longituds d'ona mesurats en Ariel (1.1 - 7.8 µm) i el JWST (0.6 - 28 µm) respecte HST (1.1 - 1.7 µm), la qual cosa implicarà un augment important en el volum de dades i els requeriments computacionals per analitzar-les. Esquerra: Distribucions posteriors dels paràmetres atmosfèrics predits en una extracció atmosfèrica Bayesiana. Es pot observar com el major detall en les dades espectrals d'Ariel i JWST produeixen uns resultats amb menor incertesa que amb HST. Font: Tinetti et al. (2021).

(Tinetti et al. 2021) i BSSL Twinkle Mission (Edwards et al. 2019). Aquests telescopis s'han dissenyat i equipat per mesurar espectres d'alta precisió i amb un ampli rang espectral (Figura 1.3), per centenars d'exoplanetes prèviament confirmats per missions com Kepler i TESS. Així doncs, en els propers anys la qualitat i la quantitat de dades atmosfèriques augmentaran de forma exponencial.

Aquest gran volum de dades representa una oportunitat sense precedents per progressar en la comprensió sobre la composició dels exoplanetes, ja que permetrà determinar un major nombre d'espècies químiques, se superaran les limitacions de la degeneració dels models, i per primera vegada, es podrà generar un inventari complet de composicions atmosfèriques.

Aquesta oportunitat, però, ve acompanyada d'un gran repte, que és precisament com analitzar aquest gran i complex volum de dades espectrals. L'alt cost computacional de les tècniques utilitzades fins ara, basades en tècniques d'inferència Bayesiana, poden representar un coll d'ampolla a l'hora de processar tot el volum previst de dades.

Davant d'aquest escenari, s'ha posat de manifest la urgència de desenvolupar nous mètodes per a l'extracció atmosfèrica (Changeat i Yip 2023; Greene et al. 2016). En aquest sentit, les tècniques d'aprenentatge automàtic (*machine learning*) es presenten com una de les candidates més prometedores. Alguns estudis pioners ja han demostrat la capacitat de tècniques com *Random Forest* (Fisher et al. 2020; Marquez-Neila et al. 2018) i diferents tipus de xarxes neuronals profundes per aconseguir resultats equiparables a les tècniques tradicionals d'inferència Bayesiana (Martínez et al. 2022; Zingales i Waldmann 2018). Malgrat tot, aquests treballs encara es troben en una fase incipient i han tingut accés a dades d'entrenament molt limitades.

Per tal d'incentivar el progrés en el desenvolupament de tècniques d'aprenentatge automàtic aplicades al problema de l'extracció atmosfèrica, Changeat i Yip (2023) han elaborat un conjunt de dades sintètiques simulant els espectres que mesurarà la missió Ariel, tant en quantitat com en qualitat. Les dades han estat pre-processades i documentades per tal de facilitar que científics no experts en l'àmbit de l'astronomia puguin aportar la seva contribució des d'una òptica interdisciplinària i de la ciència de dades.

## 1.2 Objectius i hipòtesi

L'objectiu d'aquest treball és elaborar un model basat en tècniques d'aprenentatge automàtic que sigui capaç de realitzar l'extracció atmosfèrica de dades espectrals obtingudes amb la nova generació de telescopis dedicats a la caracterització atmosfèrica. Per això, el model es desenvoluparà sobre el conjunt de dades sintètiques Atmospheric Big Challenge Database (ABC Database, Changeat i Yip (2023)), que simulen tant la quantitat com la qualitat de les dades que s'obtindran en la futura missió Ariel (Tinetti et al. 2021).

Estudis recents han aplicat satisfactòriament una varietat tècniques com *Random Forest* (Marquez-Neila et al. 2018; Fisher et al. 2020; Nixon i Madhusudhan 2020), xarxes neuronals convolucionals (Zingales i Waldmann 2018; Martínez et al. 2022), xarxes neuronals Bayesianes (Cobb et al. 2019), i altres tipus de xarxes neuronals profundes (Waldmann 2016; Soboczenski et al. 2018) al problema de l'extracció atmosfèrica. D'entre elles, *Random Forest* es presenta com a particularment interessant, ja que en ser un model relativament simple, pot ser competitiu a nivell de requeriments computacionals.

Per tant, en aquest treball es partirà del model *Random Forest* com a model base, i es comprovarà com s'adequa a l'ABC Database, més divers i complex que els conjunts de dades utilitzats en treballs anteriors. A continuació, es proposaran models alternatius per millorar el model base. En concret, s'utilitzaran altres tècniques de conjunt basades en arbres de decisió, com *Gradient Boosting* i *XGBoost*, els quals encara no s'han explorat a la literatura científica en el context de l'extracció atmosfèrica. Aquesta elecció es basa en la hipòtesi que l'aprenentatge seqüencial inherent als mètodes de *boosting* pot produir models amb un major rendiment superior als models de Random Forest, basats en un aprenentatge en paral·lel.

Més concretament, els objectius d'aquest treball són:

- **Realitzar una anàlisi exploratòria de l'ABC Database**, amb la finalitat d'entendre com és la mostra de dades, i esbrinar si existeix algun patró o estructura que pugui ser rellevant a l'hora de construir o interpretar els models d'aprenentatge automàtic en les següents etapes del treball.

- **Desenvolupar el model base amb *Random Forest***. Aquest model estarà basat en Marquez-Neila et al. (2018), i es comprovarà com s'ajusta aquest model al conjunt de dades de l'ABC Database.

- **Desenvolupar models alternatius**. Es desenvoluparan els models de Gradient Boosting basat en histogrames i XGBoost com a candidats a millorar el rendiment del model base.

## 1.3 Metodologia i planificació

### 1.3.1 Cronograma

El treball es preveu desenvolupar en les següents fases (Figura 1.4):

- **Recerca bibliogràfica**

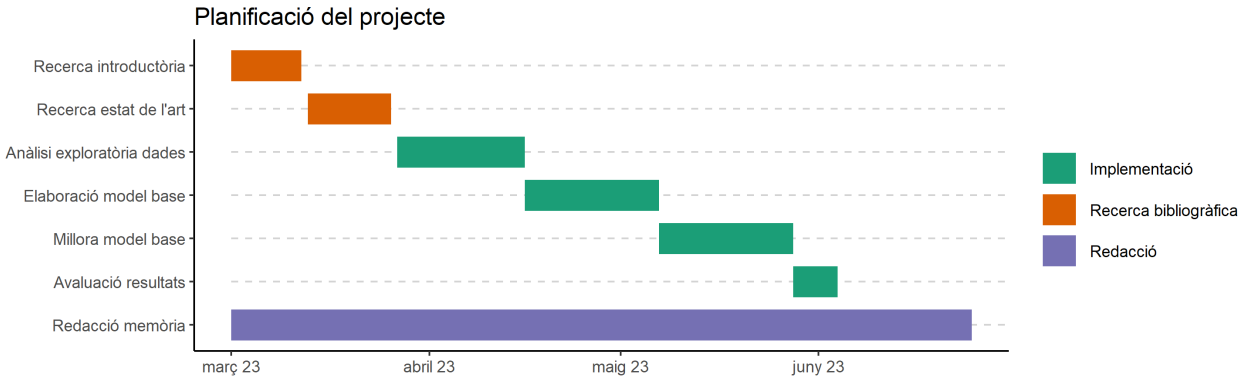



Figura 1.4: Planificació de les diferents etapes del projecte.

Aquesta fase inicial consisteix en fer una revisió bibliogràfica sobre el problema plantejat en aquest treball. En especial, es revisarà l'estat de l'art en l'aplicació de tècniques d'aprenentatge automàtic per a l'extracció atmosfèrica. Aquesta informació s'utilitzarà per a refinar els objectius del treball.

Duració aproximada: 2 setmanes.

- **Anàlisi exploratòria de l'ABC Database**

  En aquesta fase es realitzarà una anàlisi exploratòria per conèixer en detall les característiques del conjunt de dades. S'utilitzaran eines d'estadística exploratòria i tècniques d'agregació no supervisada.

  Duració aproximada: 3 setmanes.

- **Elaboració del model base**

  Aquesta fase consisteix en desenvolupar un model Random Forest basat en Marquez-Neila et al. (2018) a partir de les dades de l'ABC Database.

  Duració aproximada: 3 setmanes.

- **Millora del model base**

  Aquesta fase consisteix en extendre o millorar el model base, mitjançant tècniques basades en Gradient Boosting.

  Duració aproximada: 3 setmanes.

- **Avaluació dels resultats**

  La implementació del treball conclourà amb una comparació dels diferents models realitzats, tant des de la perspectiva de la qualitat dels resultats obtinguts com dels re-

queriments computacionals. En base a això es trauran unes conclusions crítiques i es proposaran vies de millora per a futurs treballs.

Duració aproximada: 1 setmana.

- **Redacció de la memòria**

  Finalment, es redactarà la memòria, que contindrà el resultat de tota la recerca bibliogràfica i les anàlisis realitzades en la fase d'implementació, així com una discussió dels resultats i conclusió final.

  Duració aproximada: 3 setmanes.

### 1.3.2 Recursos

Per a la realització del treball es preveu utilitzar les següents eines i recursos:

- Per la cerca bibliogràfica s'utilitzarà el gestor de referències Zotero.
- Per l'anàlisi exploratòria i la construcció dels models s'utilitzarà el llenguatge Python i l'entorn de desenvolupament Kaggle. Aquesta plataforma ofereix un entorn integrat amb les principals llibreries per al processament de dades i la implementació de models; així com maquinari amb accelerador GPU. A més, també facilita la col·laboració i compartició de codi amb la comunitat.
- Tot el codi generat es farà públic a través del repositori Github https://github.com/EEjarque/TFM_Extraccio_atmosferica.
- Per a redactar la memòria s'utilitzarà el sistema LaTeX.

## 1.4 Competència de compromís ètic i global (CCEG) i Objectius de Desenvolupament Sostenible (ODS)

*Sostenibilitat*: Un dels objectius d'aquest treball és disminuir els costos computacionals de les tècniques actuals per a analitzar dades espectrals. El progrés cap a processos computacionalment més eficients impliquen un ús més racional dels recursos energètics i per tant, una disminució de l'empremta de carboni (Lannelongue, Grealey i Inouye 2021). Per tant, aquesta línia del treball s'enquadra amb els objectius "ODS 12 Responsible consumption and production" i "ODS 13 Climate action".

*Comportament ètic i responsabilitat social*: Els models desenvolupats en aquest treball tenen per objectiu comprendre com són els milers de planetes que s'han descobert els darrers anys

més enllà del nostre sistema solar, i en darrer terme, comprendre millor la singularitat del mon on vivim. Aquesta branca de coneixement adreça curiositats universals entre les cultures, amb el potencial de generar vincles i promoure la pau en regions postconflicte (Fragkoudi 2020). En aquest sentit, el treball s'enquadra amb l'objectiu "ODS 16 – Peace, justice and strong institutions".

# Capítol 2

# Estat de l'Art

## 2.1 Introducció

En els darrers anys han sorgit diversos estudis proposant una varietat de models d'aprenentatge automàtic per extreure informació atmofèrica a partir de dades espectrals. Aquests treballs encara són poc nombrosos, però els seus resultats ja mostren el potencial d'aquestes tècniques per obtenir resultats comparables a les tècniques estàndard actuals basades en inferència Bayesiana.

El precursor d'aquesta línia d'investigació va ser Waldmann (2016), el qual va entrenar una *Deep Believe Network* per fer prediccions qualitatives sobre quines molècules podien estar presents en un espectre. Més específicament, va crear una xarxa neuronal consistent en tres *Restricted Boltzmann Machine* consecutives, les quals feien un aprenentatge no supervisat dels espectres, amb una capa final de regressió logística que assignava etiquetes a les diferents categories identificades. Aquestes categories corresponien a les molècules que previsiblement eren presents en aquell espectre, i aquesta informació s'utilitzava per restringir i acotar els paràmetres d'un procediment estàndard d'extracció atmosfèrica, fent-lo així computacionalment més eficient. Per tant, no es tractava encara d'una solució completa d'extracció atmosfèrica, però va ser una primera demostració que a partir de xarxes neuronals es podien interpretar dades espectrals i extreure'n informació sobre la seva composició.

Aquest treball va ser el precursor d'altres estudis que proposaven solucions completes d'extracció atmosfèrica. A grans trets, aquestes noves propostes es basaven en dues famílies de tècniques. D'una banda, es van proposar models basats en xarxes neuronals profundes, i de l'altra, basats en regressors Random Forest.

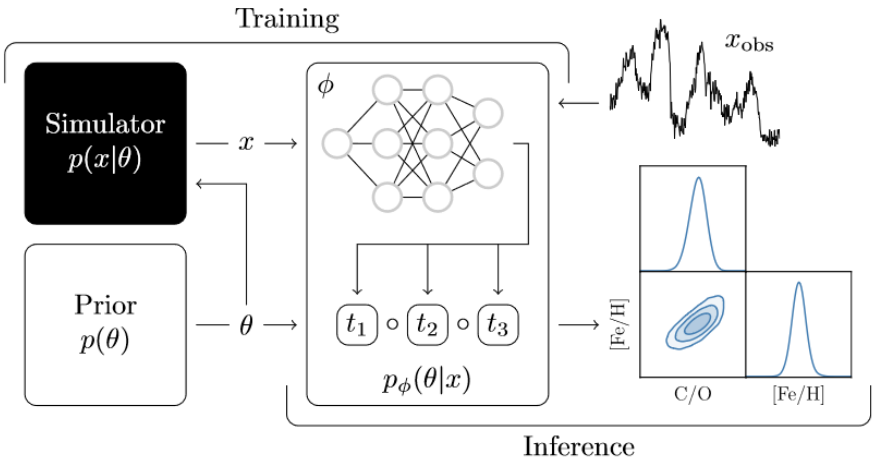



Figura 2.1: Sequència per l'extracció atmosfèrica mitjançant tècniques d'aprenentatge automàtic. S'utilitza un model simulador per crear espectres sintètics a partir dels quals s'entrena un model (ja sigui de tipus xarxa neuronal o Random Forest). Un cop invertit el temps en entrenar aquest model, es pot realitzar l'extracció atmosfèrica de noves observacions en qüestió de segons. Font: Vasist et al. (2023)

## 2.2 Models basats en xarxes neuronals profundes

Zingales i Waldmann (2018) van proposar la primera arquitectura de xarxa neuronal profunda capaç de realitzar una anàlisi completa d'extracció atmosfèrica. Es tractava d'una *Generative Adversarial Network*, un tipus de xarxa neuronal generativa no supervisada. Aquest model implica dues xarxes neuronals, una xarxa generadora que crea dades "falses" a partir de les dades introduïdes, i una segona xarxa discriminadora que classifica aquesta informació com a certa o falsa. L'entrenament dura fins que la segona xarxa és incapaç de distingir entre la informació certa i la falsa, indicant que la xarxa generadora ha après una bona representació de la distribució de probabilitats de les dades d'entrada. Els autors van utilitzar la capacitat d'aquest model de reconstruir informació mancant en el conjunt de dades d'entrada per tal de predir els paràmetres atmosfèrics d'un espectre. Els resultats del model eren a grans trets comparables amb els obtinguts en una extracció Bayesiana. Les majors divergències van ser amb l'estimació d'un paràmetre concret (el CO, pel qual van preveure una abundància més elevada que amb una extracció clàssica), i en les distribucions posteriors, les quals van sortir més disperses que amb una extracció Bayesiana. El model, però, era relativament complex i implicava l'entrenament de dues xarxes neuronals, la qual cosa el deixava en certa desavantatge respecte a altres models més simples proposats més endavant (Martínez et al. 2022).

D'altra banda, Soboczenski et al. (2018) va desenvolupar una xarxa neuronal convolucional unidimensional, utilitzant capes de Monte Carlo *dropout* per tal de generar distribucions de probabilitat sobre els paràmetres atmosfèrics predits. Els resultats obtinguts eren coherents amb

els que s'obtindrien amb una anàlisi tradicional d'extracció atmosfèrica, però no van realitzar una comparació entre els dos mètodes.

En aquest sentit, va ser Cobb et al. (2019) qui va presentar una millora sobre la predicció de les distribucions posteriors dels paràmetres estimats. Van presentar un model basat en un conjunt de xarxes neuronals Bayesianes, les quals extenen la funcionalitat de les xarxes neuronals profundes aportant distribucions d'incertesa sobre les seves prediccions. En el seu treball aporten una comparació acurada entre els seus resultats i els obtinguts amb una extracció tradicional, mostrant una elevada coincidència de les distribucions posteriors entre els dos mètodes.

Finalment, en l'àmbit de les xarxes neuronals, Martínez et al. (2022) van proposar l'ús d'una xarxa neuronal convolucional unidimensional, però diferia respecte Soboczenski et al. (2018) en la forma d'obtenir les incerteses dels paràmetres atmosfèrics estimats. En aquest cas, a partir d'un espectre, la xarxa retornava una distribució Gaussiana multidimensional que representava les distribucions posteriors dels diferents paràmetres atmosfèrics. Comparant els seus resultats amb una extracció Bayesiana, van observar un rendiment similar en l'estimació dels paràmetres. Malgrat tot, aquest mètode està limitat a la predicció de distribucions en forma Gaussiana, i per tant, no és capaç de predir distribucions multimodals o uniformes, molt comunes en extraccions atmosfèriques.

## 2.3 Models basats en Random Forest

A part dels models basats en xarxes neuronals, una altra línia d'investigació s'ha basat en l'ús de regressors *Random Forest*. Comparat amb les xarxes neuronals profundes, el *Random Forest* presenta l'avantatge de ser un tipus de model més simple, fàcil d'interpretar, i generalment més ràpid d'entrenar. És un mètode de conjunts que combina arbres de decisió (en aquest cas també anomenats arbres de regressió, ja que es fan prediccions de variables numèriques continues) i el *bootstrapping* amb substitució, per tal de realitzar nombrosos arbres de regressió sobre subconjunts aleatoris de les dades originals. Malgrat que cada arbre individualment produeix prediccions amb molta incertesa, en conjunt assoleixen un poder predictiu molt més elevat i amb major robustesa (James et al. 2021a).

Marquez-Neila et al. (2018) van presentar la primera extracció atmosfèrica basada en *Random Forest*, el qual s'ha convertit en el model base que posteriorment altres autors han intentat extendre o millorar. El seu anàlisi es va basar en l'ús de 1000 arbres de regressió, amb els quals van assolir convergència, i a partir de tots aquests arbres aleatoris van reconstruir les distribucions de probabilitat de cada paràmetre estimat. A més, van fer ús d'una anàlisi complementària del *Random Forest*, anomenada importància de les característiques (*feature importance*).

Aquesta anàlisi permet determinar quin pes tenen les variables originals en determinar el valor de cada paràmetre estimat. És a dir, en el cas de l'extracció atmosfèrica, quin pes tenen les diferents longituds d'ona de l'espectre en predir l'abundància relativa de cada molècula, o el valor dels paràmetres físics atmosfèrics. Aquesta anàlisi aporta una informació molt interessant, ja que a diferència de les xarxes neuronals, que es comporten més com a caixa negra, permet comprovar si la relació que el model ha establert entre l'espai dimensional original i els paràmetres predits tenen un sentit físic.

Els resultats obtinguts amb aquest model eren coherents amb els que s'obtindrien amb una extracció tradicional, demostrant així com un model de *Random Forest*, malgrat la seva simplicitat, permet obtenir uns resultats similars a tècniques més complexes basades en xarxes neuronals profundes. Més endavant, Nixon i Madhusudhan (2020) van reproduir els resultats de Marquez-Neila et al. (2018), els va comparar amb una extracció amb *Nested Sampling*, i van posar de manifest que malgrat que les prediccions dels paràmetres coincidien, les distribucions de probabilitat eren més disperses en el cas del *Random Forest*. Davant d'això, van proposar una extensió del mètode que millorava la predicció de les distribucions posteriors dels paràmetres predits. Aquesta extensió consistia en comparar els espectres observats amb una sèrie de *forward models* calculats amb paràmetres estimats a partir d'espectres simulats sense soroll. A partir d'aquí extreien una funció de probabilitat que utilitzaven per calcular la incertesa de cada predicció. Amb aquest mètode obtenien distribucions posteriors més similars a les obtingudes amb un *Nested Sampling*, però també afegien molta complexitat al procés.

D'altra banda, Fisher et al. (2020) van proposar un pas previ a l'anàlisi de *Random Forest*, per tal de fer el mètode extensible a espectres d'alta resolució mesurats amb telescopis terrestres. Amb una anàlisi de *cross-correlation*, van aconseguir reduir fins a 10 vegades la dimensionalitat de les dades espectrals. A continuació, aquestes dades es van introduir al model *Random Forest* de Marquez-Neila et al. (2018) per fer la recuperació dels paràmetres atmosfèrics.

## 2.4 Limitació en les dades d'entrenament

En conjunt, aquests treballs representen un seguit de primeres temptatives prometedores pel que fa a l'ús de tècniques d'aprenentatge automàtic per realitzar l'extracció atmosfèrica, ja que tots ells han aconseguit produir uns resultats com a mínim coherents amb els que s'obtindrien amb una extracció clàssica Bayesiana. A més, destaca el fet que l'ús de tècniques i arquitectures de xarxes neuronals molt diferents aconsegueixen produir unes mètriques de qualitat similars, mostrant la versatilitat de les tècniques d'aprenentatge automàtic.

Malgrat tot, la limitació principal que tenen aquests models és que molts d'ells s'han desenvolupat per un planeta en concret, de manera que no són generalitzables a noves observacions

que es facin en el futur. Aquest fet es deu principalment a una manca de disponibilitat de dades, ja que a dia d'avui, existeixen pocs espectres de transmissió observats (estan comptabilitzades 152 exoatmosferes fins a data actual, *Exoplanet atmospheres* (2023)). Davant d'això, l'aproximació que han seguit aquests treballs, ha estat la de crear un conjunt de dades d'entrenament específic pel planeta pel qual volien analitzar l'espectre (Taula 2.1). Per exemple, Cobb et al. (2019), Marquez-Neila et al. (2018) i Nixon i Madhusudhan (2020) s'han centrat en el planeta WASP-12b, i el seu espectre de transmissió capturat per la WFC3 del Hubble Space Telescope (Kreidberg et al. 2015). Fisher et al. (2020) ha desenvolupat els seu model per KELT-9b a partir d'un espectre d'alta resolució (Hoeijmakers et al. 2018; Hoeijmakers et al. 2019), mentre que Zingales i Waldmann (2018) van crear el seu model sobre un espectre del planeta de tipus *hot-Jupiter* HD 189733b (Tsiaras et al. 2018). Tots ells, com a dades d'entrenament, han generat desenes de milers d'espectres simulats, executant cada vegada un *forward model* amb variacions dels diferents paràmetres atmosfèrics, dins d'un rang físicament coherent amb cada planeta en concret.

El fet de desenvolupar un model específic per cada planeta representa una limitació important quant a requeriments computacionals i a l'hora de generalitzar el mètode, ja que d'aquesta manera el procediment per realitzar l'extracció atmosfèrica no consisteix només en entrenar el model i fer prediccions, sinó que s'ha d'afegir a cada cas la generació de les dades d'entrenament, el qual és un procediment molt costós. En conjunt, tot el procés pot arribar a igualar el temps de computació d'una extracció Bayesiana (Nixon i Madhusudhan 2020).

La forma de superar aquesta limitació passa per generar conjunts de dades d'entrenament representatives d'un rang el més ampli possible de planetes. D'aquesta manera, un cop entrenat el model, les extraccions atmosfèriques de noves observacions només implicarien la fase d'inferència (Figura 2.1), i per tant serien només de l'ordre de segons, enlloc de l'ordre de dies. Aquesta aproximació s'ha aplicat en alguns treballs basats en xarxes neuronals (Martínez et al. 2022; Soboczenski et al. 2018) però encara no s'ha aplicat en models de *Random Forest*.

L'ABC database, creada per Changeat i Yip (2023), representa una oportunitat per desenvolupar models més generalitzables. Concretament, aquest conjunt de dades ha estat elaborat a partir de les característiques dels exoplanetes confirmats, i aquells inclosos a la llista de candidats de la missió TESS (Ricker et al. 2014), a data de maig de 2022 en el marc de la *ESA-Ariel Target list initiative* (Edwards et al. 2019; Edwards i Tinetti 2022). Es tracta, per tant, d'un conjunt de dades que inclou les característiques dels planetes que seran mesurats properament per la missió Ariel, i per tant, un model que funcioni bé per aquest conjunt de dades podria servir per fer l'extracció atmosfèrica de les observacions que es facin al llarg d'aquesta missió.

En aquest treball, s'avançarà en la línia d'investigació centrada en l'aplicació de models *Random Forest* per a l'extracció atmosfèrica. A més, s'utilitzarà l'ABC Database com a opor-

tunitat per desenvolupar models que siguin capaços d'incorporar aquesta diversitat de senyals espectrals i que per tant, siguin generalitzables a poblacions diverses d'exoplanetes.

| **Estudi** | **Tècnica** | **Planetes modelitzats** | **Mida conjunt de dades** | **Paràmetres atmosfèrics predits** | **Rang espectral (instrument)** | **Temps d'entrenament** | **Temps de predicció** |
|---|---|---|---|---|---|---|---|
| Waldmann (2016) | Deep Belief Neural Network (RobERt) | WASP-12b, HD189733b, HD209458b, HAT-P-11b, GJ1214b | 85,750 (temps de creació: 3h) | $H_2O$, HCN, $CH_4$, $CO_2$, CO, $NH_3$, NO, SiO, TiO, VO | 1 - 20 $\mu$m (simulat) | $\sim$ 1.5h en 6 nuclis CPU o <10 min. amb GPU. | no mencionat |
| Zingales & Waldmann (2018) | Generative Adversarial Network (ExoGAN) | HD 189733b | $10^7$ (temps de creació no mencionat) | $H_2O$, $CO_2$, $CH_4$, CO, massa del planeta, radi del planeta, temperatura | 0.3 - 50 $\mu$m (simulat) | 9h en GPU o 3 dies en 20 nuclis CPU, per època | $\sim$2 Minuts |
| Marquez-Neila et al. (2018) | Random Forest (HELA) | WASP-12b | 100,000 (temps de creació no mencionat) | $H_2O$, HCN, $NH_3$, opacitat de núvols, temperatura | 0.84 to 1.67 $\mu$m (HST/WFC3) | no mencionat | segons |
| Soboczenski et al. (2018) | Bayesian deep learning model (INARA) | Planetes rocosos; ventall ampli de característiques planetàries | 3,000,000 (temps de creació no mencionat) | $H_2O$, $CO_2$, $O_2$, $N_2$, $CH_4$, $N_2O$, CO, $O_3$, $SO_2$, $NH_3$, $C_2H_6$ i $NO_2$ | No especificat (Large UltraViolet/Optical/InfraRed Surveyor LUVOIR) | no mencionat | segons |
| Cobb et al. (2019) | Bayesian neural networks (plan-net) | WASP-12b | 100,000 (el mateix que HELA) | $H_2O$, HCN, $NH_3$, opacitat de núvols, temperatura | 0.84 to 1.67 $\mu$m (HST/WFC3) | 20 min | $\sim$1.5 segons |
| Fisher et al. (2020) | Cross-correlation + Random Forest (HELA) | KELT-9b | 65,000 | Fe, Ti, $Fe^+$, $Ti^+$, temperatura, metal·licitat | $\sim$ 0,36 - 0,70 $\mu$m (HARPS-N) | no mencionat | $\sim$ 20 segons |
| Nixon & Madhusudhan (2020) | Random Forest (extensió de HELA) | WASP-12b, HD 209458b | 100,000 (el mateix que HELA) | $H_2O$, HCN, $NH_3$, opacitat de núvols, temperatura | 0.84 to 1.67 $\mu$m (HST/WFC3) | 4-80 segons en 4 nuclis CPU | segons |
| Martínez et al. (2022) | Convolutional neural networks | Ventall ampli de característiques planetàries | 200,000 | $H_2O$, CO, $CO_2$, $CH_4$, $NH_3$, HCN, TiO, VO, AlO, FeH, OH, $C_2H_2$, CrH, $H_2S$, MgO, $H^-$, Na, K | 0.84 to 1.67 $\mu$m (HST/WFC3) | 1-3h | segons |

Taula 2.1: Resum d'estudis que han proposat tècniques d'aprenentatge automàtic per a l'extracció atmosfèrica

# Capítol 3

# Anàlisi exploratòria de les dades

## 3.1 L'ABC Database

L'ABC Database (Changeat i Yip 2023; Changeat i Yip 2022) consisteix en un compendi d'espectres atmosfèrics sintètics, acompanyats de les seves respectives composicions i temperatures atmosfèriques, així com de dades auxiliars sobre els exoplanetes que representen. Les dades han estat pre-processades i estructurades per ser utilitzades per entrenar i desenvolupar models d'aprenentatge automàtic per a l'extracció atmosfèrica.

El procés de síntesi dels espectres s'explica de manera resumida a la Figura 3.1. Es parteix d'una selecció d'exoplanetes, elaborada a partir de la llista de planetes confirmats i candidats (Edwards i Tinetti 2022). A cadascun se li ha assignat, d'una banda, els paràmetres estel·lars, orbitals i planetaris que es troben a la literatura. D'altra banda, se li ha assignat una composició atmosfèrica aleatòria, assumint una atmosfera primària de $He/H_2 = 0.17$ (ràtio solar) a la qual s'afegeixen quantitats traça d'aigua ($H_2O$), diòxid de carboni ($CO_2$), metà ($CH_4$), monòxid de carboni (CO) i amoníac ($NH_3$). Les abundàncies de cada gas s'han assignat aleatòriament seguint una llei log-uniforme. Per la temperatura, s'ha assumit un perfil isotèrmic a la temperatura d'equilibri del planeta. A partir de tots aquests paràmetres, s'ha utilitzat un model físic per calcular la contribució dels diferents processos que afecten el pas de la llum a través de l'atmosfera (com l'absorció molecular, l'absorció induïda per col·lisions, i la dispersió de Rayleigh) per generar un espectre teòric d'alta resolució. Finalment, aquest espectre s'ha passat per un simulador que convoluciona l'espectre a la resolució d'Ariel, i simula el soroll en funció de la longitud d'ona, obtenint així un espectre comparable als que s'obtindran amb el telescopi Ariel.

Les dades obtingudes (105887) han passat per un procés de neteja fins a obtenir finalment 91392 espectres simulats, definits per 52 longituds d'ona. Aquests espectres, reescalats entre 0 i 100, representen les variables predictores, i l'objectiu final és utilitzar aquestes dades per predir

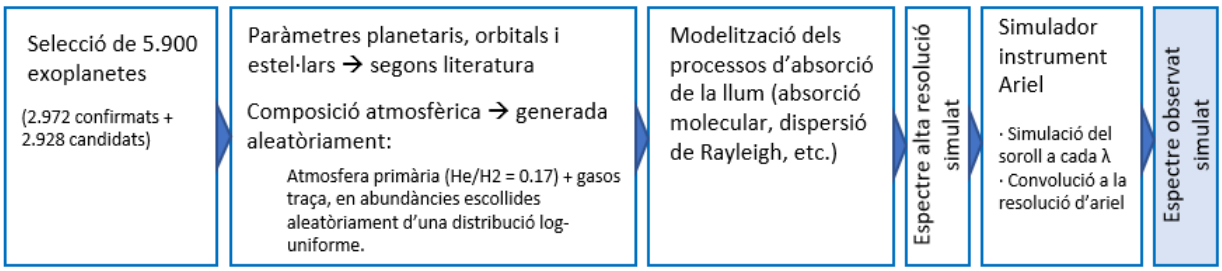



Figura 3.1: Esquema del procés de síntesi dels espectres de transmissió simulats pel telescopi Ariel, els quals formen el conjunt de dades de l'ABC Database.

les concentracions dels gasos ($H_2O$, $CO_2$, $CH_4$, CO, $NH_3$) i la temperatura que han donat lloc a aquests espectres.

Per una fracció dels espectres, la base de dades també inclou els resultats d'una extracció atmosfèrica tradicional basada en inferència Bayesiana. Més concretament, s'ha utilitzat l'algorisme Nested Sampling (Feroz, Hobson i Bridges 2009) com a optimitzador per ajustar els espectres a un model atmosfèric i trobar-ne els paràmetres òptims. Els detalls del procediment es poden consultar a Changeat i Yip (2023). Els resultats d'aquesta extracció permetran comparar els mètodes desenvolupats en aquest treball amb la metodologia estàndard utilitzada fins a dia d'avui per realitzar l'extracció atmosfèrica.

## 3.2 Mètodes de l'anàlisi exploratòria

Per tal d'estudiar la diversitat de senyals espectrals continguda a la base de dades, s'ha realitzat una anàlisi de components principals (PCA) sobre els espectres atmosfèrics. Aquesta anàlisi permet reduir la dimensionalitat del conjunt de dades i identificar els components més significatius que contribueixen a la variabilitat observada. A continuació, s'ha utilitzat l'anàlisi d'agregació basada en densitats DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) sobre l'espai dimensional reduït per agrupar els espectres en base a les seves similituds espectrals. Els agregats s'han identificat ajustant els hiperparàmetres `eps` (distància mínima entre dues mostres per ser considerades veïnes) i `min_samples` (nombre de mostres en un veïnat perquè una mostra es consideri un nucli). Finalment, s'han utilitzat gràfics de *boxplot* per comparar la distribució de les propietats atmosfèriques predominants a cada agregat.

Les anàlisis PCA i DBSCAN s'han realitzat a través de la llibreria `scikit.learn`, mentre que els gràfics *boxplot* s'han realitzat mitjançant la llibreria `seaborn`, ambdues del llenguatge Python.

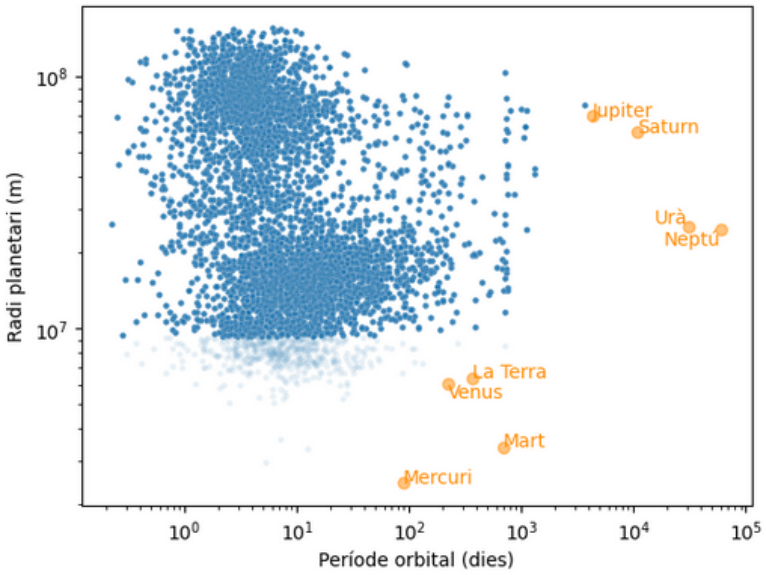



Figura 3.2: Característiques planetàries (radi i període orbital) dels planetes inclosos en aquest estudi, referenciats amb els planetes del nostre sistema solar.

## 3.3 Característiques demogràfiques de la mostra d'exoplanetes

A causa de la gran extensió de l'ABC Database, les característiques dels exoplanetes inclosos a l'estudi reflecteixen els trets demogràfics dels exoplanetes descoberts fins ara. Comparat amb els planetes del nostre sistema solar, presenten uns períodes orbitals molt curts, majoritàriament per sota del període orbital de Mercuri. A més, s'observen dues poblacions en quant a radi planetari: una població es troba en un rang entre la Terra i Urà/Neptú, i una altra per sobre de Júpiter i Saturn (Figura 3.2). Entremig d'aquestes dues poblacions s'observa una menor quantitat de planetes que transicionen entre les dues poblacions. En definitiva, aquestes dades mostren de manera sintètica la diversitat de tipologies d'exoplanetes representades a la base de dades.

## 3.4 Característiques dels espectres atmosfèrics

El conjunt d'espectres atmosfèrics, definits per 52 longituds d'ona, s'ha reduït a dues components principals, les quals capturen respectivament el 46.1% i el 29.0% de la variància total (Figura 3.3A). En aquest espai dimensional reduït, les mostres es troben distribuïdes formant tres agrupacions, indicant per tant, tres grups de mostres amb un conjunt de característiques espectrals més homogènies. Utilitzant l'anàlisi DBSCAN amb uns hiperparàmetres d'`eps` = 6 i `min_samples` = 167, s'han pogut separar aquests grups de mostres (Figura 3.3B). Els tres

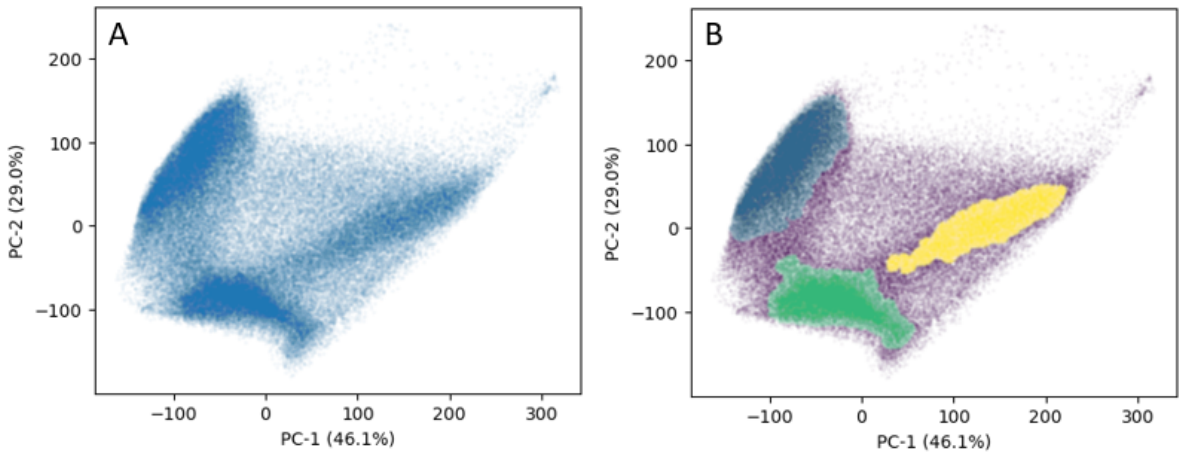



Figura 3.3: A. Distribució dels espectres atmosfèrics en l'espai dimensional reduït a les dues primeres components d'una Anàlisi de Components Principals. B. Mitjançant una anàlisi DBSCAN s'han pogut separar tres agregats, els quals contenen mostres amb unes característiques espectrals més homogènies. La resta de mostres es troben disperses entre els tres agregats, presentant per tant un ventall més divers de característiques espectrals.

agregats principals (grups 0, 1 i 2) contenen el 28.5%, el 23.8% i el 13.7% de les dades respectivament. La resta de les dades (grup -1) forma un quart conjunt de mostres que es troba dispers en tot l'espai i no s'ha pogut assignar a cap agregat.

Cada agregat representa, per tant, un grup de mostres amb unes característiques espectrals similars. Així, cadascun presenta un espectre mitjà amb unes característiques diferenciades (Figura 3.4). En els tres casos la intensitat mínima es produeix per sota d'1 $\mu m$, mentre que la majoria de trets distintius es troben entre els 2 i 8 $\mu m$. El grup 0 presenta una banda d'intensitat característica als 3.3 $\mu m$, mentre que el grup 1 té un mínim d'intensitat en aquesta mateixa regió (3.7 $\mu m$), i intensitats màximes a la part final de l'espectre. El grup 2 també té un mínim d'intensitat als 3.7 $\mu m$, però aquest mínim va seguit d'un pic molt acusat, als 4.31 $\mu m$, i intensitats relativament baixes a la regió final de l'espectre.

Finalment, s'observa que per cada grup espectral predominen unes característiques atmosfèriques diferenciades (Figura 3.5). El grup 0 destaca per tenir una composició més elevada en $CH_4$, mentre que el grup 1 presenta una major concentració d'$H_2O$ i de $NH_3$. El grup 2 destaca per tenir concentracions més elevades de $CO_2$. En canvi, pràcticament no s'observen diferències de temperatura o CO entre els grups espectrals.

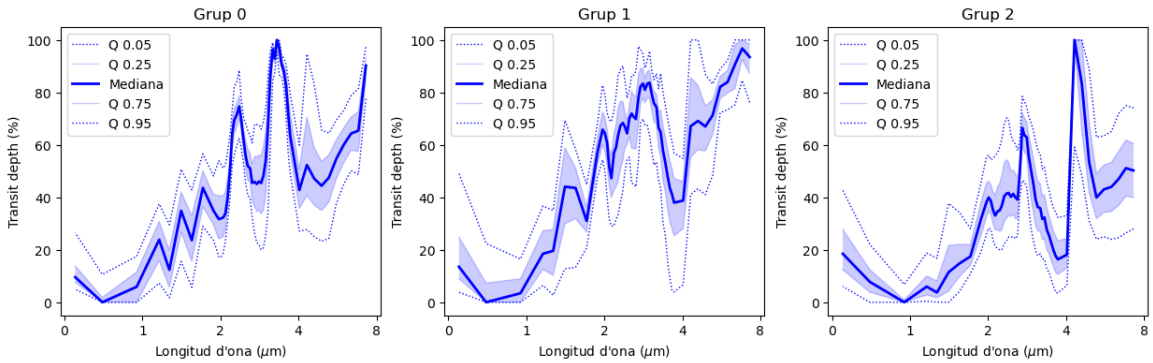



Figura 3.4: Resum estadístic de les característiques espectrals de cada agregat de mostres, identificats mitjançant una Anàlisi de Components Principals i una anàlisi d'agrupació DBSCAN.

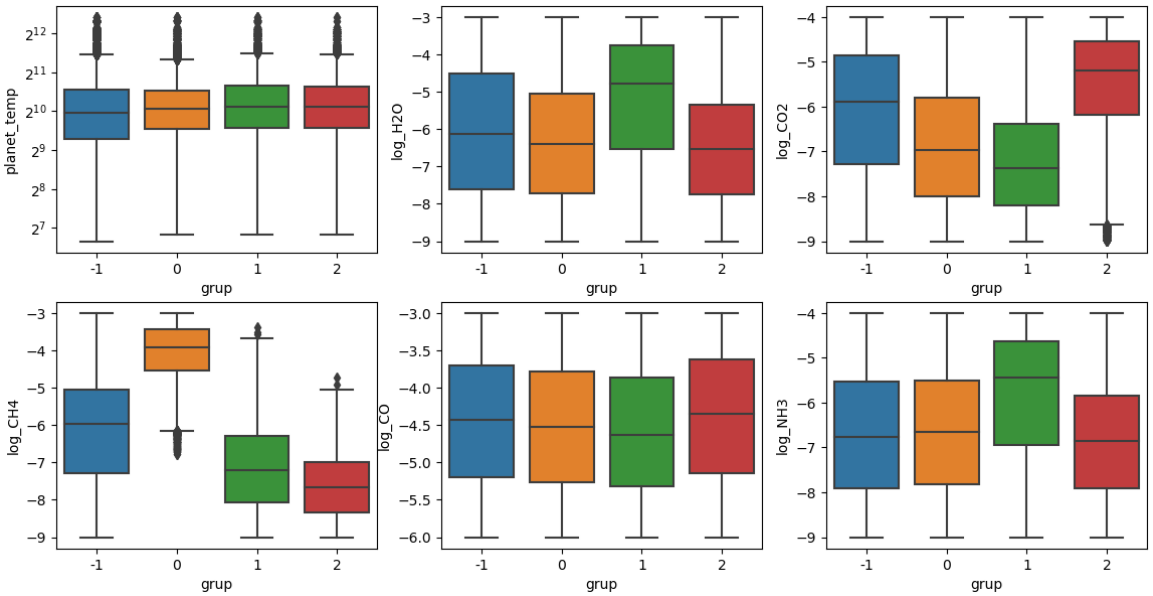



Figura 3.5: Característiques atmosfèriques pels 3 agregats de mostres amb característiques espectrals similars (grups 0, 1 i 2) i per la resta de mostres més diverses i que no corresponen a cap agregat (grup -1).

# Capítol 4

# Modelització

## 4.1 El procés de modelització

Pel desenvolupament dels models s'ha seguit el següent procediment. En primer lloc, les dades s'han dividit aleatòriament en dos subconjunts d'entrenament i test (75% i 25% de les dades, respectivament).

Amb el conjunt d'entrenament s'ha realitzat una cerca i optimització dels hiperparàmetres dels models mitjançant el mètode de la validació creuada. S'han utilitzat 4 plecs de validació i el coeficient de determinació ($R^2$) com a mètrica per avaluar el rendiment dels models. Per limitar les combinacions d'hiperparàmetres, l'estratègia general ha consistit en assignar inicialment els valors per defecte, o bé uns valors relativament laxes; i després anar ajustant els hiperparàmetres un a un, seqüencialment. Primer s'ha ajustat el nombre d'estimadors; a continuació, els hiperparàmetres relacionats amb l'estructura dels estimadors; després, els relacionats amb la regularització del model; i finalment, s'ha acabat d'ajustar la taxa d'aprenentatge; segons corresponia a cada model. A cada cas, s'ha escollit el valor pel qual la $R^2$ era màxima, o bé a partir del qual el model assolia la convergència i ja no millorava de forma rellevant. Com que l'objectiu de predicció inclou 6 variables predictores (temperatura, $H_2O$, $CO_2$, $CH_4$, CO, $NH_3$), s'han ajustat els hiperparàmetres per cadascuna d'aquestes variables de sortida.

Un cop s'han identificat els hiperparàmetres més adequats per a cada model, s'ha procedit a entrenar els models utilitzant les dades del conjunt d'entrenament. Posteriorment, s'ha avaluat el rendiment dels models utilitzant les dades del conjunt de test, consistent en observacions que els models no han vist durant el procés d'entrenament.

En tot el procés, el rendiment dels models s'ha avaluat amb la mètrica $R^2$, ja que és adimensional i permet comparar la qualitat de les prediccions pels diferents paràmetres atmosfèrics, que tenen diferents unitats i escala de variació. La mètrica s'ha calculat amb la implementació de la llibreria `scikit.learn` de Python, la qual està dissenyada per tasques predictives, i es-

pecialment per avaluar l'ajust entre els valors predits i esperats en models de tipus regressió. En aquesta implementació, la $R^2$ pot prendre un valor màxim de 1, indicant una correspondència perfecta entre els valors predits i esperats; en canvi, pot prendre valors negatius si les prediccions són pitjors que la mitjana dels valors esperats.

El rendiment dels models s'ha avaluat tant per tot el conjunt de dades test, com per cadascun dels agregats espectralment homogenis identificats amb l'anàlisi de PCA i DBSCAN, per valorar si hi ha diferències de rendiment entre diferents subconjunts de la mostra de dades. Per avaluar la capacitat de generalització del model i detectar possibles casos de sobreajustament (*overfitting*), s'ha comparat el valor de $R^2$ obtingut en les prediccions utilitzant les dades d'entrenament amb les prediccions utilitzant les dades de test. Grans discrepàncies entre aquestes dues prediccions poden ser indicadores que el model no generalitza bé a noves observacions.

Per tal de determinar quines longituds d'ona han tingut més influència en les prediccions del model, s'ha fet una anàlisi de la importància de les característiques (*feature importance*). Aquests resultats ja venen incorporats en la sortida de l'entrenament dels models. En el Random Forest, aquesta anàlisi es basa en mesurar la disminució de precisió que es produeix quan es permuten les dades d'una característica determinada. Quan una característica és important per a la predicció, la permutació d'aquesta característica provoca una caiguda significativa en la precisió del model. En el cas del Gradient Boosting, l'anàlisi es basa en mesurar la reducció de pèrdua que aporten les diferents variables durant el procés de construcció del model.

Finalment, l'extracció atmosfèrica del millor model obtingut en aquest treball s'ha comparat amb l'extracció feta amb el mètode tradicional Nested Sampling. Amb aquesta finalitat, s'ha comparat la $R^2$ i els gràfics de dispersió dels valors predits i observats dels dos mètodes, per cadascun dels paràmetres atmosfèrics.

Quant a recursos computacionals, els models s'han executat en un sistema amb 4 nuclis de CPU i 30 gigabytes de RAM. D'altra banda, pels models que requerien acceleració, s'ha utilitzat un maquinari equipat amb una GPU d'acceleració (1 Nvidia Telsa P100 GPU), 2 nuclis CPU i 13 gigabytes de RAM.

Tot el codi generat durant el procés de modelització està públicament disponible en un repositori de Github [1].

## 4.2 Model base amb Random Forest

Random Forest es basa en l'ús de múltiples arbres de decisió per realitzar tasques de classificació o, en aquest cas, de regressió. Aquest mètode combina la idea de l'aprenentatge en conjunts (*ensemble learning*) amb la tècnica del *bootstrap aggregation* (o *bagging*) (Criminisi, Shotton i

[1] https://github.com/EEjarque/TFM_Extraccio_atmosferica

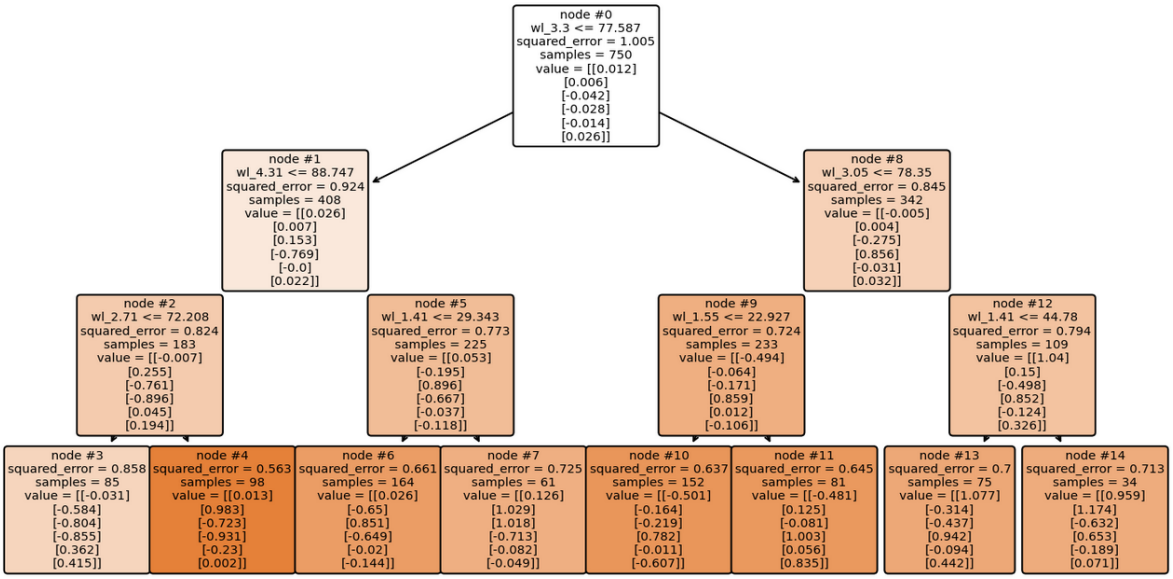



Figura 4.1: Exemple d'arbre de decisió construït a partir de 750 mostres de l'ABC Database. Inicialment totes les mostres pertanyen al mateix node (node #0), i la primera partició ha dividit l'espai entre aquells espectres que tenen una intensitat a $\lambda_{3.3\mu m}$ per sobre (node #1) o per sota (node #2) de 77.587. A continuació, cada node s'ha partit dues vegades més, successivament, fins a obtenir finalment un arbre de 4 nivells de profunditat i 8 nodes finals (fulles). A cada fulla, els valors predits de temperatura, $H_2O$, $CO_2$, $CH_4$, CO i $NH_3$ corresponen a la mitjana de les mostres assignades a aquella fulla.

Konukoglu 2012).

Un arbre de decisió divideix l'espai predictor (és a dir, l'espai definit per les longituds d'ona dels espectres atmosfèrics) en diferents particions binàries, successives i no solapades, en un procés anomenat *recursive binary splitting* (James et al. 2021b). Cada partició divideix una dimensió predictora en un punt de tall, de manera que es minimitzi una funció de cost, com per exemple el guany d'informació (entropia). D'aquesta manera, es van realitzant particions fins que a cada node final (també anomenat fulla) només hi ha una mostra. Altrament, per evitar arbres massa complexos, es pot aturar la creació d'un arbre quan s'assoleixin certs criteris com ara una profunditat màxima de l'arbre o un nombre mínim de mostres en els nodes finals. Finalment, per totes aquelles mostres que pertanyin a un mateix node final se li assigna la mitjana de totes les mostres assignades a aquella fulla (Figura 4.1).

Malgrat que cada arbre de forma individual pugui tenir un poder predictiu molt baix, la combinació de les prediccions de múltiples arbres permet obtenir un resultat molt més robust i precís. En el cas del Random Forest, es combinen múltiples arbres de decisió construïts cadascun a partir d'una submostra aleatòria del conjunt de dades original (*bootstrap aggregation* o *bagging*), i també considerant només un subconjunt aleatori de característiques per a cada

partició en el procés de construcció de l'arbre. Això fa que els arbres siguin independents i poc correlacionats entre ells, disminuint la variància de la mitjana de tots els arbres. Durant la fase de predicció, en el cas de tasques de regressió, es pren la mitjana de les prediccions de tots els arbres per obtenir el valor final de sortida.

En aquest treball s'ha desenvolupat un Random Forest com a model base (en endavant, referit com a RF), utilitzant la implementació de la llibreria `cuML` de Python, la qual permet l'ús de GPU. Els hiperparàmetres s'han optimitzat seqüencialment segons es detalla a la Taula 4.1.

| Hiperparàmetre | Descripció | Valors provats | Valors escollits | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Temp | $H_2O$ | $CO_2$ | $CH_4$ | CO | $NH_3$ |
| n_estimators | Nombre d'arbres. | 50 a 2000 | 100 | 100 | 100 | 100 | 100 | 100 |
| max_depth | Profunditat màxima dels arbres, és a dir, nombre màxim d'iteracions de particions. | 3*, 5, 7, 10, 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| min_samples_split | Mínim nombre de mostres que ha de contenir un node perquè se li pugui aplicar una partició. | 2, 5*, 10, 20, 30, 40, 50, 100 | 10 | 10 | 10 | 10 | 10 | 10 |
| min_samples_leaf | Nombre mínim de mostres que ha de contenir una fulla. | 1, 5, 10* | 1 | 1 | 1 | 1 | 1 | 1 |
| max_features | Nombre màxim de variables a tenir en compte a cada partició. | auto, sqrt*, log2 | auto | auto | auto | auto | auto | auto |

Taula 4.1: Resum del procés d'optimització dels hiperparàmetres del model RF, presentats en l'ordre seqüencial en què s'han anat ajustant. El símbol * indica el valor assignat inicialment, per fer l'ajust dels hiperparàmetres previs.

## 4.3 Models alternatius amb Gradient Boosting

El Gradient Boosting és també un model de conjunts, que combina múltiples estimadors febles per obtenir un predictor més precís i robust, però es diferencia en la forma com es combinen els estimadors (James et al. 2021b). Mentre que el RF combina múltiples arbres de decisió independents, el Gradient Boosting construeix un sol arbre i el va millorant seqüencialment (Figura 4.2). Comença amb un arbre de regressió inicial molt senzill, el qual tindrà uns residuals entre les seves prediccions i les dades reals. A cada iteració, s'utilitza el descens del gradient per ajustar un nou arbre als residuals del model del pas anterior. La predicció del nou model s'afegeix a la predicció del model principal, escalada segons una taxa d'aprenentatge per tal d'evitar el sobreajustament. D'aquesta manera, el model va aprenent lentament i es pot ajustar millor als patrons de les dades.

Aquesta forma d'entrenament seqüencial fa que no sigui possible distribuir la computació

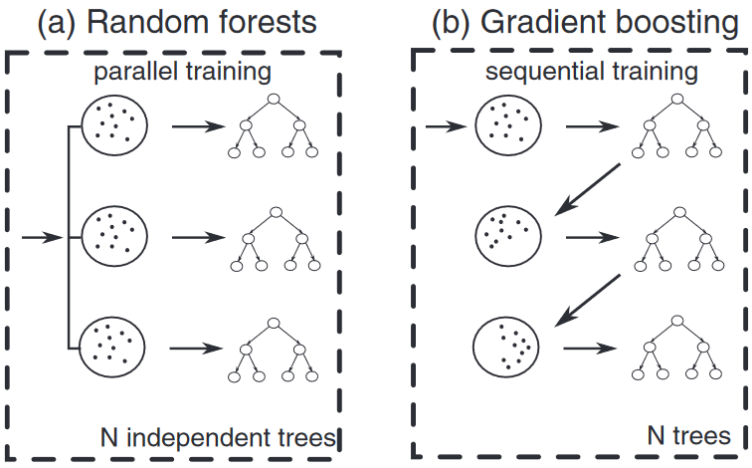



Figura 4.2: Diferències conceptuals entre l'algorisme Random Forest i Gradient Boosting. En el Random Forest, els diferents arbres són independents entre ells i s'entrenen de forma paral·lela, a partir de diferents mostres aleatòries de l'espai de dades original. En canvi, en el Gradient Boosting, s'entrena un sol arbre, el qual es va millorant seqüencialment a partir de la minimització dels seus residuals. Font: Kowalek, Loch-Olszewska i Szwabiński (2019).

de forma paral·lela en diferents nuclis de CPU, com és el cas del RF, i per tant, s'han d'utilitzar altres estratègies per accelerar els temps de càlcul. En aquest sentit, en aquest treball s'han utilitzat i comparat dues implementacions diferents: el Gradient Boosting basat en histogrames, i el XGBoost.

### 4.3.1 Gradient Boosting basat en histogrames

El Gradient Boosting basat en histogrames (HistGB) és una variació de l'algorisme base en què les variables contínues d'entrada són discretitzades mitjançant l'ús d'histogrames. Aquesta variació redueix molt la quantitat de valors que poden prendre les variables d'entrada, la qual cosa fa molt més eficient la posterior creació dels arbres de decisió i permet accelerar el temps de càlcul sense comprometre la qualitat del resultat de sortida (Guryanov 2019). S'ha utilitzat la implementació de la llibreria `scikit.learn` de Python, basada en l'algorisme Light Gradient Boosting Machines (LightGBM, Ke et al. (2017)). Els hiperparàmetres per entrenar el model s'han optimitzat segons es mostra a la taula 4.2.

### 4.3.2 XGBoost

XGBoost és una versió millorada i optimitzada del Gradient Boosting bàsic i que ha esdevingut una de les implementacions més utilitzades en l'actualitat (Chen i Guestrin 2016). En el XGBoost, els arbres de decisió són construïts de forma més regularitzada mitjançant l'ús de

| Hiperparàmetre | Descripció | Valors provats | Valors escollits | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Temp | $H_2O$ | $CO_2$ | $CH_4$ | CO | $NH_3$ |
| max_iter | El nombre màxim d'interacions del procés d'impuls (*boosting*). | 100 a 3800 | 500 | 1500 | 500 | 500 | 2500 | 1500 |
| max_bins | El nombre màxim d'intervals dels histogrames utilitzats per discretitzar les variables d'entrada. | 100, 150, 200, 255* | 150 | 255 | 255 | 255 | 255 | 200 |
| max_depth | Profunditat màxima dels arbres. | 3*, 5, 7, 10, 20 | 10 | 7 | 20 | 20 | 7 | 7 |
| min_samples_leaf | Nombre mínim de mostres que ha de contenir una fulla. | 10, 20*, 50 | 20 | 50 | 20 | 50 | 50 | 20 |
| learning_rate | Taxa d'aprenentatge, és l'escalat realitzat a cada iteració del procés de boosting. | 0.5, 0.1*, 0.01, 0.001 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 | 0,1 |

Taula 4.2: Resum del procés d'optimització dels hiperparàmetres del model HistGB, presentats en l'ordre seqüencial en què s'han anat ajustant. El símbol * indica el valor assignat inicialment, per fer l'ajust dels hiperparàmetres previs.

la tècnica de *split finding* optimitzada. Això permet una millor gestió de la complexitat del model i una reducció del sobreajustament, millorant així la generalització i la precisió. A més, la computació és paral·lelitzable i es pot accelerar mitjançant la utilització de GPU.

En aquest cas, per entrenar el model XGBoost s'ha utilitzat la funció XGBRegressor de l'API de `scikit.learn`. Els hiperparàmetres del model s'han optimitzat segons es detalla a la taula 4.3.

| Hiperparàmetre | Descripció | Valors provats | Valors escollits | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Temp | $H_2O$ | $CO_2$ | $CH_4$ | CO | $NH_3$ |
| n_estimators | El nombre màxim d'interacions del procés de boosting. | 500 a 10000 | 2000 | 2000 | 2000 | 2000 | 6000 | 2000 |
| max_depth | Profunditat màxima dels arbres. | 3*, 5, 10 | 5 | 5 | 5 | 5 | 5 | 5 |
| min_child_weight | Suma mínima dels pesos (o de mostres) requerides per crear un nou node a l'arbre. | 1*, 3, 6 | 1 | 6 | 1 | 1 | 6 | 3 |
| subsample | El percentatge de dades utilitzades per construir els arbres. | 0.8*, 1 | 0,8 | 0,8 | 0,8 | 0,8 | 0,8 | 0,8 |
| colsample_bytree | Submostra de columnes utilitzades per fer cada arbre. | 0.8*, 1 | 0,8 | 1 | 1 | 1 | 1 | 1 |
| gamma | La reducció mínima de la pèrdua per crear una nova partició a l'arbre. | 0*, 0.25, 0.5 | 0 | 0 | 0 | 0 | 0 | 0 |
| learning_rate | Taxa d'aprenentatge, és l'escalat realitzat a cada iteració del procés de boosting. | 0.5, 0.1*, 0.01, 0.001 | 0.1 | 0.1 | 0.1 | 0.1 | 0,1 | 0.1 |

Taula 4.3: Resum del procés d'optimització dels hiperparàmetres del model XGBoost, presentats en l'ordre seqüencial en què s'han anat ajustant. El símbol * indica el valor assignat inicialment, per fer l'ajust dels hiperparàmetres previs.

# Capítol 5

# Resultats de la modelització

## 5.1 Optimització dels hiperparàmetres

Durant el procés d'optimització dels hiperparàmetres s'ha pogut observar successivament com el model s'anava ajustant a la naturalesa de les dades, i quins hiperparàmetres tenien més influència en millorar l'ajust (Figura 5.1). En tots els models, el major increment del rendiment (quantificat mitjançant la $R^2$) s'ha produït en ajustar la profunditat màxima dels arbres (`max_depth`, Figura 5.1). Aquest efecte ha sigut aparent sobretot en el model RF. En canvi, els ajustos en la resta d'hiperparàmetres han estat poc rellevants, i només han suposat millores al tercer decimal de la $R^2$. En el cas dels models basats en Gradient Boosting (HistGB i XGBoost) destaca que des de l'inici, amb l'ajust del nombre d'arbres (o iteracions de *boosting*) s'han obtingut uns valors de $R^2$ molt propers als de l'ajust final (Figura 5.1), indicant per tant la rellevància d'aquest hiperparàmetre en l'ajust final d'aquests models.

Si comparem el nombre d'arbres òptim escollit (Taules 4.1, 4.2 i 4.3), pels diferents paràmetres atmosfèrics, s'observa que el CO seguit de l'$H_2O$ i el $NH_3$, són els que han requerit un valor més elevat per assolir la convergència. En el HistGB el CO ha requerit 2500 iteracions, i l'$H_2O$ i $NH_3$ n'han requerit 1500, mentre que per la resta només n'han calgut 500. Pel XGBoost, el CO ha requerit 6000 iteracions, mentre que la resta de paràmetres han assolit la convergència amb 2000 iteracions. En el RF, tots els paràmetres han assolit la convergència amb tan sols 100 arbres. No obstant, la $R^2$ obtinguda al final del procés d'optimització ha sigut menor pel CO, seguit del $NH_3$ i l'$H_2O$. En conjunt, són indicis que aquests paràmetres atmosfèrics mantenen una relació més complexa amb les variables predictores, de manera que el model té més dificultats per capturar i aprendre aquesta relació.

Amb els hiperparàmetres finalment ajustats, s'han obtingut uns valors de $R^2$ mitjana pels conjunts de validació creuada d'entre 0.851 (CO) i 0.984 ($CO_2$) pel RF; d'entre 0.892 ($NH_3$) i 0.989 ($CO_2$) pel HistGB; i d'entre 0.901 ($NH_3$) i 0.991 ($CO_2$) pel XGBoost. Per tant, XGBoost

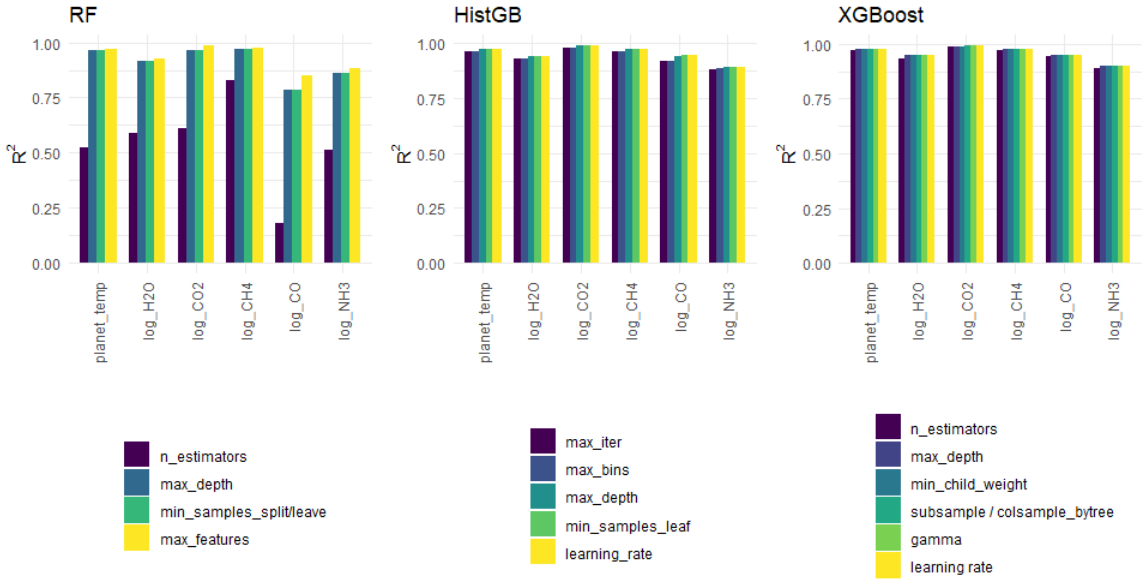



Figura 5.1: Procés d'optimització dels hiperparàmetres dels models Random Forest (RF), Gradient Boosting basat en histogrames (HistGB) i XGBoost. La cerca dels millors hiperparàmetres s'ha realitzant mitjançant la validació creuada en el conjunt de dades d'entrenament i cercant millores en la $R^2$ mitjana.

és el que presenta un millor ajust a les dades d'entrenament.

## 5.2 Capacitat predictiva dels models

Amb els hiperparàmetres optimitzats, s'han entrenat els models RF, HistGB i XGBoost amb el conjunt de dades d'entrenament. Posteriorment, els models s'han utilitzat per fer prediccions amb el conjunt de dades test, les quals són observacions noves que el model no ha vist durant la fase d'entrenament. Amb aquestes noves prediccions s'ha valorat la capacitat predictiva del model mitjançant el coeficient de determinació $R^2$ (Figura 5.2).

En tots els paràmetres atmosfèrics, s'observa que el model XGBoost és el que té un major rendiment, amb una $R^2$ entre 0.905 ($NH_3$) i 0.991 ($CO_2$); seguit del HistGB, amb una $R^2$ entre 0.899 pel $NH_3$ i de 0.987 pel $CO_2$. Finalment, el RF és el que té un rendiment més baix, amb una $R^2$ entre 0.865 pel CO i 0.986 pel $CO_2$. En tots ells, la temperatura i el $CO_2$ són els paràmetres atmosfèrics que han assolit unes $R^2$ més elevades.

Si ens fixem en els diferents agregats de mostres amb característiques espectrals homogènies, identificats mitjançant l'anàlisi PCA i DBSCAN, s'observen variacions de la $R^2$ respecte al que s'obté amb el conjunt de totes les mostres test. En el cas de la temperatura, el $CO_2$ i el CO, aquesta variació és poc perceptible. No obstant, en el cas del $H_2O$, $CH_4$ i el $NH_3$, les diferències són més pronunciades.

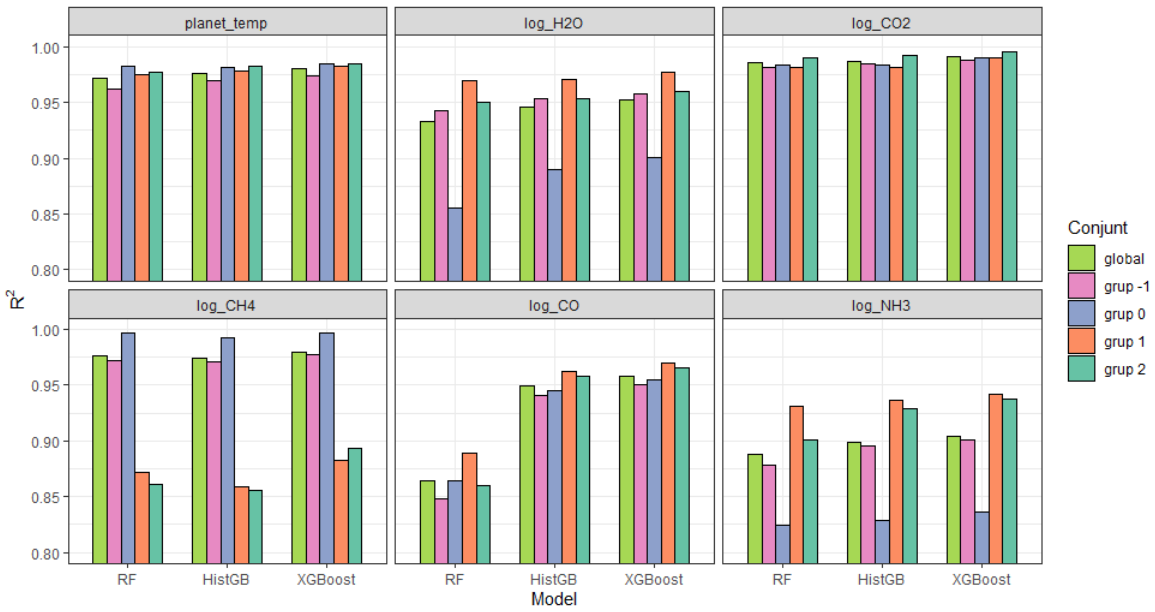



Figura 5.2: Capacitat predictiva dels models segons el coeficient de determinació ($R^2$). Es mostren els resultats pel conjunt de totes les mostres (global), i pels agregats amb característiques espectrals més homogènies identificats mitjançant l'Anàlisi de Components Principals i DBSCAN (grups 0, 1 i 2). Es mostren també els resultats pel grup de mostres no assignades a cap agregat (grup -1).

En el cas del $CH_4$, el grup 0 té una $R^2$ per sobre de la que s'obté amb tot el conjunt de mostres, mentre que pels grups 1 i 2 la $R^2$ disminueix de manera molt acusada, fins situar-se entre 0.856 i 0.894. En canvi, per l'$H_2O$ i el $NH_3$ veiem que el grup 0 és el que es modelitza pitjor ($R^2$ d'entre 0.855 i 0.901 per l'$H_2O$ i de 0.825 i 0.836 pel $NH_3$); mentre que els grups 1 i 2 modelitzen fins i tot millor que per tot el conjunt de dades test.

Finalment, s'ha comparat la $R^2$ obtinguda amb les prediccions del conjunt de dades test, amb la $R^2$ obtinguda amb les prediccions del conjunt de dades d'entrenament, per tal de valorar la magnitud de l'efecte de sobreajustament (*overfitting*) del model. En tots els casos, la $R^2$ del conjunt de test ha sigut inferior a la $R^2$ del conjunt d'entrenament, indicant per tant un cert efecte de sobreajustament. De totes maneres, aquesta diferència es manté molt petita en tots els casos, la qual cosa suggereix que la capacitat de generalització del model es veu poc afectada. Concretament, el RF és el que presenta un major efecte de sobreajustament, amb una variació de la $R^2$ d'entre 0.015 ($CO_2$) i de 0.115 (CO); mentre que el HistGB és el que ha presentat un menor efecte de sobreajustament, amb una disminució de la $R^2$ d'entre 0.007 ($CO_2$) i 0.05 ($NH_3$).

## 5.3 Comparació amb els resultats d'una extracció Bayesiana

Per 5454 mostres del conjunt test, l'ABC Database contenia els resultats d'una extracció atmosfèrica realitzada amb inferència Bayesiana. Les prediccions obtingudes amb aquest mètode s'han comparat amb les obtingudes amb el model XGBoost, model amb el qual s'ha obtingut una $R^2$ més elevada per tots els paràmetres atmosfèrics (Figures 5.3 i 5.4).

Els punts sobre la línia d'igualtat representen una extracció perfecta, és a dir, una coincidència exacta entre el valor el valor predit i el valor esperat. En canvi, els punts fora de la línia d'igualtat indiquen una desviació de la predicció del model respecte al valor esperat, causant una disminució de la $R^2$.

La temperatura és l'únic paràmetre atmosfèric pel qual XGBoost té una capacitat predictiva inferior a la del Nested Sampling. Mentre que Nested Sampling produeix una predicció perfecta, XGBoost presenta una desviació de les prediccions per sobre dels 3000 K.

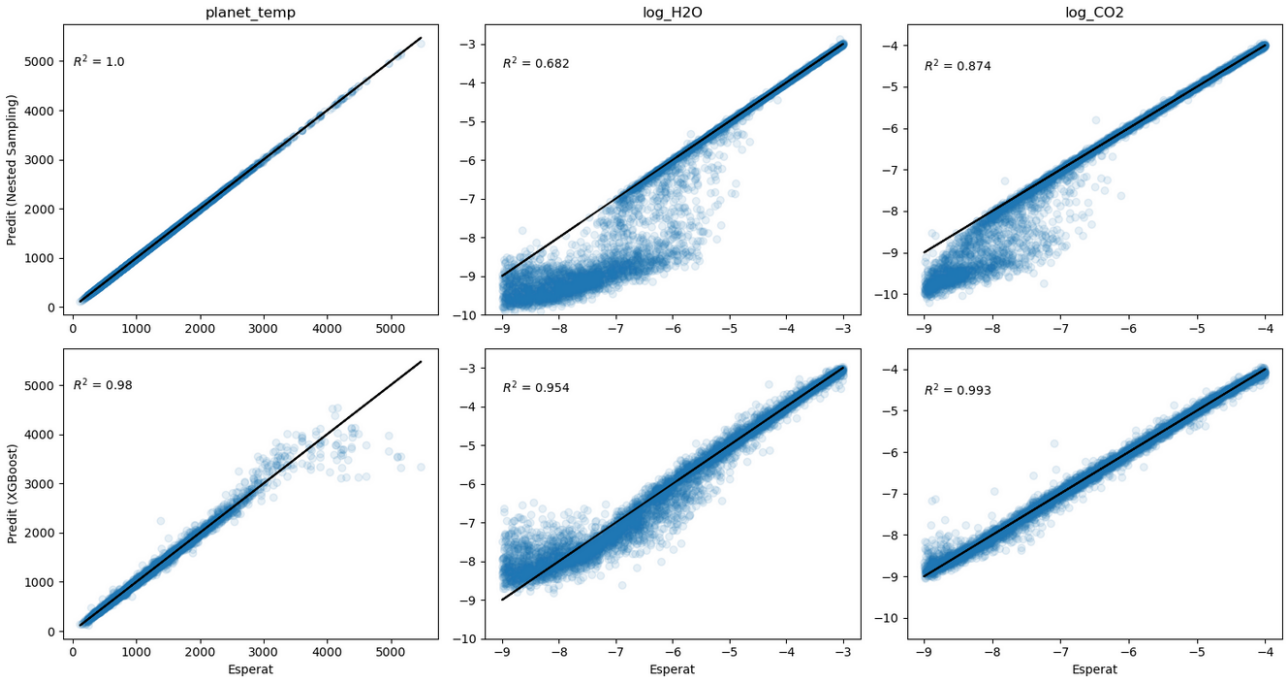



Figura 5.3: Correspondència entre els valors esperats i predits utilitzant una extracció clàssica amb Nested Sampling (fila superior) i una extracció amb XGBoost (fila inferior), pels paràmetres de temperatura, $H_2O$ i $CO_2$.

En canvi, pel que fa als gasos, XGBoost presenta una capacitat predictiva notablement millor. El Nested Sampling presenta una predicció pràcticament perfecta en el rang alt de concentracions. No obstant, a partir d'un cert llindar, apareix una zona de transició a partir de la qual els valors predits s'allunyen de la línia d'igualtat, i la relació entre els valors esperats

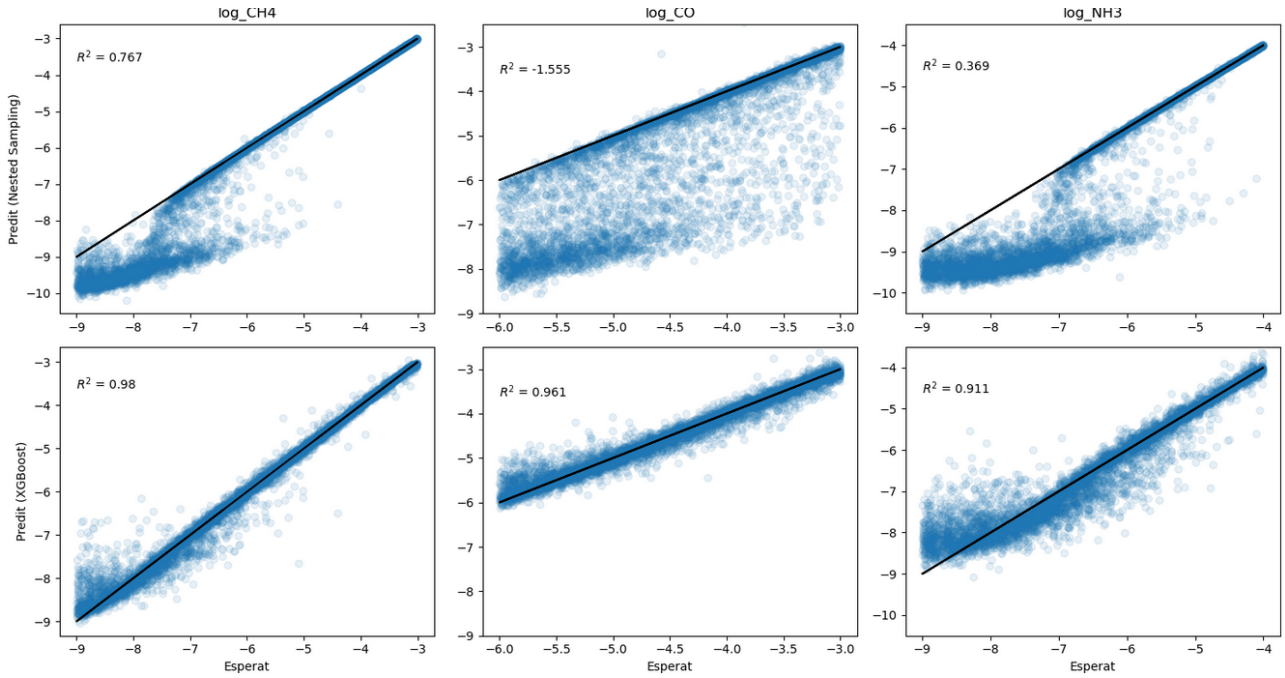



Figura 5.4: Correspondència entre els valors esperats i predits utilitzant una extracció clàssica amb Nested Sampling (fila superior) i una extracció amb XGBoost (fila inferior), pels paràmetres de $CH_4$, CO i $NH_3$.

i predits tendeix a aplanar-se. Amb XGBoost, aquesta zona de transició és molt més reduïda ($H_2O$, $CH_4$ i $NH_3$) i fins i tot pràcticament inapreciable ($CO_2$ i CO), de manera que s'obtenen prediccions més precises al llarg de tot el rang de variació dels gasos. Això es veu reflectit també amb un augment molt important de la $R^2$. El cas més extrem seria el CO, pel qual una $R^2$ de -1.555 amb Nested Sampling esdevé 0.961 amb XGBoost. Tot seguit, les millores més importants s'observen pel $NH_3$ i $H_2O$ (augment de la $R^2$ de 0.369 a 0.911, i de 0.682 a 0.954, respectivament).

## 5.4 Anàlisi de la importància de les característiques

L'anàlisi de la importància de les característiques (*feature importance*) ens permet analitzar quines longituds d'ona han tingut més pes a l'hora de modelitzar els diferents paràmetres atmosfèrics (Figura 5.5). Els resultats pel model XGBoost, que és el que ha presentat un major poder predictiu, mostra que la temperatura i el CO tenen els pesos repartits de forma més homogènia per totes les longituds d'ona, mantenint-se a valors molt baixos, per sota de 0.1. En canvi, la resta de gasos presenten certes longituds d'ona amb un pes més destacat a l'hora de fer les prediccions. Per exemple, l'$H_2O$ presenta un pes molt acusat a les longituds d'ona de 1.41 i 2.22 $\mu m$; mentre que el $CO_2$ presenta el major pes a 4.31 $\mu m$, seguit de 4.91 i 5.24 i

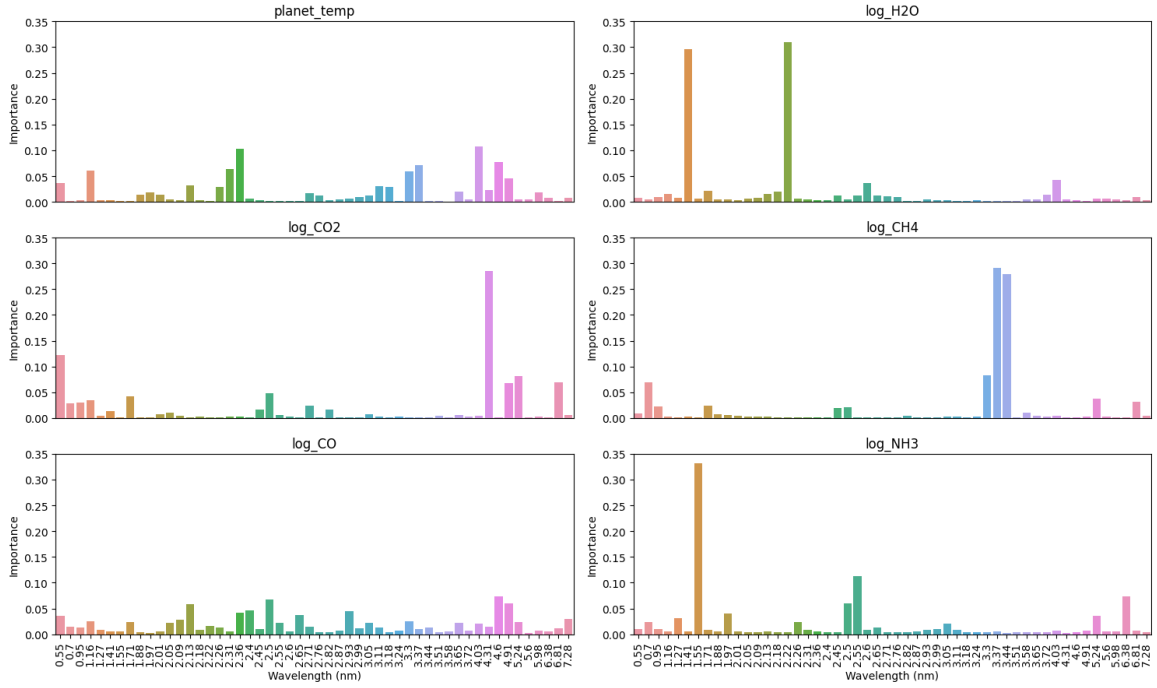



Figura 5.5: Anàlisi de la importància de les característiques, on es mostra el pes que han tingut les diferents longituds d'ona en la modelització de cada paràmetre atmosfèric.

6.81 $\mu m$. El $CH_4$ té els seus majors pesos a unes longituds d'ona una mica inferiors, a la zona de 3.37 i 3.44 $\mu m$. Finalment, l'$NH_3$ té el seu major pes a 1.55 $\mu m$, seguit de la zona de 2.50 i 2.55 $\mu m$ i 6.38 $\mu m$.

## 5.5 Temps d'entrenament dels models

D'entre els tres models desenvolupats, RF i XGBoost són els que han tingut un temps d'entrenament més reduït, reflectint l'ús de la GPU. El model RF ha requerit entre 3 i 5 segons per ajustar cada paràmetre atmosfèric, trigant un temps total de 21 segons per ajustar els 6 paràmetres atmosfèrics. De manera similar, el model XGBoost ha tingut un temps d'entrenament entre 5 i 16 segons per cada paràmetre atmosfèric, amb un temps total de 44 segons.

D'altra banda, el model HistGB, calculat utilitzant CPU, ha requerit un temps d'entrenament comparativament més llarg. Cada paràmetre ha trigat entre 8 i 33 segons, amb un temps total de 84 segons (1.4 minuts). Tot i aquest increment en comparació amb RF i XGBoost, destaca que el temps d'entrenament s'ha mantingut a l'ordre de pocs minuts tot i no utilitzar una GPU com a accelerador.

Finalment, tots tres models han trigat menys d'1 segon per fer la predicció de les 22848 mostres del conjunt de test.

# Capítol 6

# Discussió

En aquest treball s'ha investigat la viabilitat d'utilitzar models de conjunt basats en arbres de decisió per realitzar l'extracció atmosfèrica d'espectres exoplanetaris que s'analitzaran en la propera missió Ariel. Aquest estudi s'emmarca dins els esforços que s'estan fent actualment per trobar alternatives computacionalment més eficients a les tècniques tradicionals basades en inferència Bayesiana. Els resultats obtinguts demostren el potencial de les tècniques basades en Random Forest i Gradient Boosting per analitzar de forma eficient i fiable el gran volum de dades espectrals que s'espera obtenir en els propers anys.

## 6.1 Adequació dels models a l'ABC Database

Aquest treball ha partit del Random Forest com a model base, tècnica per la qual ja hi ha hagut algunes temptatives d'aplicació a l'extracció atmosfèrica (Marquez-Neila et al. 2018; Nixon i Madhusudhan 2020; Fisher et al. 2020). La novetat d'aquest treball ha estat entrenar aquest model amb l'ABC Database, un conjunt de dades espectrals més complex i heterogeni en comparació amb les dades utilitzades en els estudis anteriors. L'objectiu ha estat avaluar si el model manté la seva capacitat predictiva en aquest context més complex. En aquest sentit, amb el nostre model RF s'han obtingut rendiments més elevats en tots aquells paràmetres comuns entre els estudis, és a dir, temperatura, $H_2O$ i $NH_3$ (Taula 6.1). Les millores de rendiment són especialment notables respecte Marquez-Neila et al. (2018) i Nixon i Madhusudhan (2020). Una explicació podria ser que els seus espectres, basats en les característiques del HST, tenien una resolució i rang espectral (13 longituds d'ona en el rang 0.84 - 1.67 $\mu$m) inferior a la dels espectres utilitzats en aquest treball (52 longituds d'ona en el rang 0.55 - 7.28 $\mu$m), simulats segons les característiques d'Ariel. Aquesta hipòtesi encaixa amb el fet que el rendiment del nostre model RF per la temperatura és molt similar a l'obtingut per Fisher et al. (2020), els quals es van basar en espectres d'alta resolució. Aquests resultats posen de manifest com, més

enllà dels mètodes d'extracció, la pròpia naturalesa de les dades reflectint les característiques del telescopi Ariel implica una millora dels resultats obtinguts en l'extracció atmosfèrica.

| Paràmetre | Aquest treball | | | | RF en treballs previs | | |
|---|---|---|---|---|---|---|---|
| | RF | HistGB | XGBoost | Nested Sampling | Marquez-Neila et al. (2018) | Nixon i Madhusudhan (2020) | Fisher et al. (2020) |
| Temperatura | 0.972 | 0.976 | 0.98 | 1.000 | 0.746 | 0.928 | 0.966 - 0.970 |
| H2O | 0.933 | 0.946 | 0.954 | 0.682 | 0.608 | 0.819 | |
| CO2 | 0.986 | 0.987 | 0.993 | 0.874 | | | |
| CH4 | 0.976 | 0.974 | 0.98 | 0.767 | | | |
| CO | 0.865 | 0.950 | 0.961 | -1.555 | | | |
| NH3 | 0.887 | 0.899 | 0.911 | 0.369 | 0.700 | | |

Taula 6.1: Rendiment dels models d'extracció atmosfèrica desenvolupats en aquest estudi, comparat amb Nested Sampling i treballs anteriors basats en Random Forest.

## 6.2 Rendiment dels models i comparació amb una extracció atmosfèrica tradicional

Seguidament, en aquest treball s'han explorat dos models addicionals al RF, basats en Gradient Boosting (HistGB i XGBoost), com a candidats a millorar el rendiment del model base RF. Per tots els paràmetres atmosfèrics, tant HistGB com XGBoost han presentat una millora respecte RF, essent XGBoost el que ha presentat uns valors més elevats de rendiment ($R^2$ d'entre 0.911 i 0.993 pel $NH_3$ i el $CO_2$, respectivament). Seria per tant, el millor model obtingut. No obstant, es destaca que les diferències entre models són força ajustades, i que amb RF s'assoleixen ja unes $R^2$ molt elevades, d'entre 0.865 (CO) i 0.986 ($CO_2$).

Per posar en context la rellevància dels rendiments obtinguts, és necessari comparar els nostres resultats amb els obtinguts en una extracció atmosfèrica tradicional. Pels paràmetres gasosos, en l'extracció amb Nested Sampling s'observa un llindar a partir del qual les observacions es desvien de la línia d'igualtat, i la relació entre els valors predits i observats s'aplana (Figures 5.3 i 5.4, primera fila). Es tracta d'un límit a partir del qual el mètode no és capaç d'extreure les propietats atmosfèriques, i que a efectes pràctics esdevé un límit de detecció per aquella molècula. Aquest llindar afecta un rang de valors més reduït en gasos com el $CO_2$ ($R^2$ = 0.874) o el $CH_4$ (0.767), però arriba a ser molt significatiu en altres gasos com el CO, pel qual el rang complet de valors predits mostra una forta desviació respecte els valors observats. En aquest cas, Nested Sampling no és capaç de realitzar l'extracció per la gran majoria de mostres ($R^2$ = -1.555).

En canvi, en les extraccions realitzades amb XGBoost, l'efecte d'aquest llindar de detecció és molt menys aparent (Figures 5.3 i 5.4, segona fila). Els gasos que més clarament presenten aquest llindar són l'$H_2O$ i el $NH_3$, tot i que apareix a un valor molt inferior ($10^{-7}$ ppmv) que en l'extracció amb Nested Sampling ($10^{-5}$ i $10^{-6}$, respectivament). En canvi, per la resta de gasos, l'efecte d'aquest llindar és molt poc acusat ($CH_4$) o fins i tot pràcticament inapreciable ($CO_2$ i CO). La millora més remarcable es produeix amb el CO, el qual passa de ser pràcticament impredictible amb Nested Sampling, a poder-se predir amb una $R^2$ de 0.961 amb XGBoost, i sense llindar de detecció apreciable. Aquesta capacitat del model de poder extreure el CO és especialment interessant tenint en compte que té unes bandes d'absorció febles i que se solapen amb altres espècies, com el $CO_2$ (Tinetti et al. 2021; Changeat i Yip 2023), i per tant, tradicionalment ha sigut difícil de fer-ne l'extracció. En definitiva, els resultats demostren que, pels paràmetres gasosos, XGBoost és capaç fer prediccions amb una major precisió en tot l'abast de valors en comparació amb el mètode tradicional tradicional Nested Sampling.

La temperatura és l'únic paràmetre pel qual XGBoost assoleix una capacitat de predicció inferior que l'extracció amb Nested Sampling. En concret, s'observa una desviació dels valors predits respecte als observats a valors més alts de 3000 K. Això es pot atribuir al fet que només un 2.2% de les observacions tenen una temperatura per sobre dels 3000 K, i per tant, el model ha tingut molt pocs representats d'aquest tipus d'atmosferes durant el seu entrenament.

## 6.3 Explicabilitat dels models

D'entre les diferents tècniques d'aprenentatge automàtic, els mètodes basats en arbres de decisió tenen l'avantatge de ser intrínsecament interpretables, ja que la pròpia estructura dels arbres dona informació de quines són les variables que tenen més pes a l'hora de fer les prediccions.

En l'anàlisi d'importància de les característiques, s'observen algunes correspondències entre els pesos de les variables i les seves bandes d'absorció. Per exemple, la variable que té més pes a l'hora de modelitzar el $CO_2$ amb XGBoost és la $\lambda$ 4.31 $\mu$m. Aquesta longitud d'ona coincideix amb una de les principals bandes d'absorció del $CO_2$ (Wei et al. 2018). En el nostre conjunt d'espectres sintètics, les mostres del grup 2 (amb predomini de $CO_2$) presenten un pic molt prominent en aquesta mateixa longitud d'ona.

Aquesta coherència també s'observa en el cas del $CH_4$. El model XGBoost per aquest gas presenta els majors pesos a les longituds d'ona entre 3.37 i 3.44 $\mu$m, coincidint amb una de les seves bandes d'absorció (NIST 2023). El grup 0, destacat per una predominància d'aquest gas, presenta un espectre mitjà amb un pic als 3.3 $\mu$m, indicant per tant una relació entre l'estructura apresa pel model, i les bandes d'absorció d'aquest gas.

Per modelitzar l'$H_2O$, les dues longituds d'ona amb major pes han sigut 1.41 i 2.22 $\mu$m, les

quals també s'ajusten molt a les bandes d'absorció teòriques (Wei et al. 2018). Malgrat tot, en cap dels espectres mitjans dels tres grups d'agregats no destaquen pics en aquestes longituds d'ona, la qual cosa s'explica perquè en cap grup no domina aquest gas per sobre dels altres.

El fet que l'$H_2O$ i el $NH_3$ no dominin en cap dels 3 agregats d'espectres indica que no hi ha una quantitat suficient d'espectres on destaqui la senyal d'aquests gasos, i podria ser el motiu pel qual aquests hagin sigut els dos gasos més difícils de modelitzar. Aquesta dificultat de modelització s'ha vist reflectida en la necessitat d'un major nombre d'iteracions per assolir la convergència durant el procés d'aprenentatge dels models HistGB i XGBoost. En el RF, malgrat no haver necessitat un major nombre d'arbres, el model final ha tingut un rendiment inferior per aquests dos gasos. D'altra banda, aquests gasos són els que han mostrat una menor precisió de les seves prediccions en el rang de valors més baixos, en tots els tres models.

## 6.4 Limitacions i vies de millora

Els resultats d'aquest estudi mostren que els models de conjunt basats en arbres de decisió tenen una capacitat elevada per realitzar l'extracció atmosfèrica d'espectres de transmissió, assolint llindars de detecció més baixos que la tècnica d'inferència Bayesiana Nested Sampling, i per tant, és una família de models que és interessant de seguir explorant. Tot i així, encara queda recorregut per tal que aquests models siguin plenament aplicables a l'extracció atmosfèrica.

### 6.4.1 Predicció de les distribucions posteriors

En el camp de l'extracció atmosfèrica, l'interès no es centra només en predir un valor puntual per cada paràmetre atmosfèric, sinó el seu rang complet de valors possibles. En aquest sentit, els mètodes utilitzats tradicionalment, basats en inferència Bayesiana, tenen l'avantatge de proporcionar les distribucions posteriors dels paràmetres extrets. Mètodes com el Random Forest no estan dissenyats per proporcionar directament distribucions posteriors. Per tant, és necessari investigar com extendre els mètodes d'aprenentatge automàtic per tal que incorporin en la seva sortida una estimació de les distribucions posteriors dels paràmetres modelitzats.

En el cas del Random Forest, treballs anteriors (Marquez-Neila et al. 2018; Nixon i Madhusudhan 2020) han utilitzat una aproximació en la qual s'han considerat considerat els valors predictius de tots els arbres aleatoris com a mostres independents i aleatòries per aproximar una distribució posterior. No obstant, aquest enfocament no seria vàlid per al Gradient Boosting, ja que els arbres generats durant el procés d'aprenentatge no són independents entre si, i per tant, no són representatius d'una distribució posterior.

Per tant, per millorar l'aplicabilitat de models de tipus Gradient Boosting en el context de l'extracció atmosfèrica, futurs estudis haurien d'explorar altres possibilitats per estimar aquestes

distribucions. Una opció podria ser aplicar tècniques de *Bootstrap Aggregating (Bagging)* o *Cross-Validation* per obtenir múltiples estimacions del model i, a partir d'elles, construir una aproximació de la distribució posterior.

### 6.4.2 Propostes per millorar el rendiment dels models

Malgrat que els models desenvolupats en aquest treball han presentat una capacitat predictiva molt elevada, encara tenen aspectes que es poden millorar, com per exemple, millorar la precisió en valors baixos per gasos com l'$H_2O$ i el $NH_3$.

Una estratègia podria ser l'augment del conjunt de dades d'entrenament utilitzant informació sobre el soroll a cada longitud d'ona. Aquesta informació, estimada a partir de les especificacions del telescopi Ariel, es troba disponible a l'ABC Database i podria utilitzar-se per crear variacions plausibles dels espectres existents. En concret, es podria aplicar aquesta tècnica per ampliar la representació d'espectres amb temperatures superiors a 3000 K. Aquesta ampliació seria particularment rellevant ja que permetria millorar la capacitat predictiva del model en rangs de temperatura més alts, actualment poc representat en el conjunt de mostres.

Una altra estratègia podria ser ampliar el conjunt de característiques explicatives més enllà de les variables espectrals, i incloure també els paràmetres estel·lars i planetaris disponibles a l'ABC Database com a dades auxiliars. Aquests paràmetres, com ara la distància orbital, la massa del planeta o la temperatura de la seva estrella, tenen una afectació directa sobre la temperatura i pressió atmosfèrica dels exoplanetes. Al seu torn, la temperatura i la pressió són dos factors clau que afecten de forma global la magnitud i escala d'un espectre atmosfèric de transmissió (Madhusudhan 2019). Per tant, incloure aquests paràmetres estel·lars i planetaris podrien donar un context més complet al model i ajudar a millorar les seves prediccions.

# Capítol 7

# Conclusions

Al llarg d'aquest estudi s'han explorat els models RF, HistGB i XGBoost, com a mètodes de conjunt basats en arbres de decisió, per fer l'extracció atmosfèrica del conjunt d'espectres de l'ABC Database[1]. El rendiment d'aquests models s'ha avaluat des de múltiples perspectives, i s'ha arribat a les següents conclusions:

- El model RF aplicat a l'ABC Database ha mostrat uns rendiments notablement més elevats que en estudis previs, els quals havien entrenat els models amb espectres adquirits pel HST. Això reflecteix que el model té una major capacitat de realitzar l'extracció atmosfèrica a partir d'espectres mesurats per Ariel, els quals tenen un major rang espectral i resolució.

- D'entre els models desenvolupats, XGBoost és el que ha presentat una major capacitat de predicció dels paràmetres atmosfèrics. Malgrat tot, tant HistGB com RF han presentat uns rendiments no gaire inferiors, demostrant la capacitat de tota aquesta família de models per fer l'extracció atmosfèrica.

- XGBoost ha mostrat una precisió de les seves prediccions més alta que el mètode tradicional Nested Sampling. Pels paràmetres gasosos, amb Nested Sampling es perd capacitat de predicció a rangs baixos, mentre que XGBoost té una major precisió al llarg de tot el rang de valors.

- Els tres mètodes desenvolupats han presentat uns temps d'entrenament inferiors als 1.5 minuts, la qual cosa els posa en gran avantatge respecte a mètodes com el Nested Sampling, pels quals l'extracció atmosfèrica pot trigar de l'ordre d'hores o dies.

[1] Tot el codi desenvolupat al llarg d'aquest treball està disponible en el repositori https://github.com/EEjarque/TFM_Extraccio_atmosferica

En conjunt, els resultats obtinguts posen de manifest que els models de conjunt basats en arbres de decisió són una alternativa prometedora als mètodes actuals d'extracció atmosfèrica. Per tant, són una línia d'investigació interessant de seguir explorant en futurs estudis sobre l'extracció atmofèrica de dades espectrals massives.

# Glossari

Recull de termes propis del camp de la física i de les ciències exoplanetàries que apareixen al llarg del treball:

**Absorció molecular** Fenomen en què les molècules d'un gas absorbeixen llum en determinades longituds d'ona. Aquesta absorció es deu a les transicions energètiques de les molècules i proporciona informació sobre la composició de l'atmosfera d'un exoplaneta.

**Absorció induïda per col·lisions** Tipus d'absorció en què les col·lisions entre les partícules de l'atmosfera i els fotons de llum causen l'absorció d'energia. Aquest procés pot tenir un impacte significatiu en les regions denses de l'atmosfera, com ara les capes inferiors.

**British Small Satellite Laboratory (BSSL Twinkle)** Missió espacial proposada pel Regne Unit que té com a objectiu estudiar l'atmosfera dels exoplanetes. El llançament de la missió està previst per a l'any 2024.

**Dispersió de Rayleigh** Dispersió de la llum causada per partícules amb una mida molt més petita que la longitud d'ona de la llum incident. Aquest fenomen provoca que la llum blava es dispersi més que la llum vermella.

**Espectre de transmissió** Mesura de com la llum d'una estrella passa a través de l'atmosfera d'un exoplaneta. L'espectre de transmissió es genera analitzant les variacions en la intensitat de la llum en diferents longituds d'ona. Aquest espectre pot revelar informació sobre la composició i estructura de l'atmosfera.

**Extracció atmosfèrica** Processament de dades espectrals per inferir informació sobre la composició i propietats de l'atmosfera d'un exoplaneta.

**Exoplaneta** Planeta que orbita una estrella fora del nostre sistema solar.

**Hubble Space Telescope (HST)** Telescopi espacial de la NASA, llançat el 1990. És conegut per proporcionar imatges i observacions astronòmiques de gran detall i resolució. Malgrat no haver estat dissenyat amb aquest propòsit, des de l'any 2001 ha sigut la principal font de mesures espectrals d'atmosferes exoplanetàries.

**James Webb Space Telescope (JWST)** Telescopi espacial de la NASA llançat el 2021 en col·laboració amb l'Agència Espacial Europea (ESA) i l'Agència Espacial Canadenca (CSA). Un dels seus objectius primaris és fer mesures contínues i d'alta duració d'atmosferes exoplanetàries a través d'espectres d'elevada resolució.

**Atmospheric Remote-sensing InfraRed Large-survey (ARIEL)** : Una missió de l'Agència Espacial Europea (ESA) que té com a missió principal l'estudi de les atmosferes d'exoplanetes. Prevista per al llançament el 2029, durant quatre anys mesurarà fins a un miler d'exoatmosferes dincs un ampli rang espectral que abasta des l'òptic fins a l'infraroig.

**Missió CoRoT (Convection, Rotation and Planetary Transits)** Missió espacial de l'Agència Espacial Europea (ESA) i el Centre Nacional d'Estudis Espacials (CNES) de França, dedicada a la detecció d'exoplanetes i l'estudi de les estrelles variables. CoRoT va ser llançada el 2006 i va ser la primera missió capaç de detectar planetes rocosos, més grans que la Terra.

**Missió Kepler** Missió de la NASA que va funcionar des del 2009 fins al 2018. Kepler utilitzava la tècnica del trànsit per detectar exoplanetes i va ser pionera en la detecció d'exoplanetes rocosos i dins de la zona habitable.

**Missió TESS** Transiting Exoplanet Survey Satellite, una missió de la NASA llançada el 2018. TESS busca exoplanetes mitjançant la tècnica del trànsit, observant més de 200.000 estrelles brillants a tot el cel.

**Spitzer Space Telescope (SST)** Telescopi espacial de la NASA especialitzat en l'observació en infraroig. Va estar operatiu entre el 2003 i el 2020. El SST ha estat crucial per a l'estudi de l'evolució estel·lar, la formació de planetes, i altres fenòmens astronòmics que emeten majoritàriament radiació infraroja.

**Tècnica del trànsit** Mètode de detecció d'exoplanetes que es basa en la observació de la disminució de la intensitat de la llum de l'estrella anfitriona quan l'exoplaneta passa per davant d'ella.

**Velocitat radial** Component de la velocitat d'un objecte astronòmic en la direcció de l'observador. En el context de la detecció d'exoplanetes, la velocitat radial es mesura mitjançant el canvi de longitud d'ona de l'espectre de l'estrella anfitriona causat per l'atracció gravitatòria de l'exoplaneta. Aquesta mesura permet inferir l'existència i les característiques de l'exoplaneta.

# Agraïments

Aquest treball és una aventura que va començar la tardor passada, fent pluja d'idees amb l'Octavi i l'Andrea, en el seu despatx entre chips, Arduinos i columnes de Winogradsky. A tots dos els vull agraïr el seu entusiasme i la seva gran implicació en el treball, però sobretot, que m'hagin donat la oportunitat d'obrir-me una finestra al món de l'astrobiologia.

Un altre agraïment especial el vull dedicar a la Laura, la meva tutora de la UOC, per haver-me ajudat a enfocar el treball dins de la ciència de dades, i que amb els seus comentaris al llarg de les entregues parcials m'ha ajudat a millorar i aconseguir fer un treball del que n'estic molt satisfeta.

Finalment, voldria fer extensius aquests agraïments a la meva família. Als meus pares, per estar sempre al meu costat. I, molt especialment a l'Arnau, per ajudar-me al llarg d'aquests mesos a poder dedicar temps a aquest treball, per fer que tot aquest esforç tingui sentit, i per fer que junts, tot sembli possible.

# Bibliografia


Barstow, Joanna K. et al. (des. de 2016). "A consistent retrieval analysis of 10 Hot Jupiters observed in transmission". A: *The Astrophysical Journal* 834.1. arXiv:1610.01841 [astro-ph], pàg. 50. ISSN: 1538-4357. DOI: 10.3847/1538-4357/834/1/50. URL: http://arxiv.org/abs/1610.01841 (cons. 06-03-2023).

Borucki, William J. et al. (febr. de 2010). "Kepler Planet-Detection Mission: Introduction and First Results". A: *Science* 327.5968. Publisher: American Association for the Advancement of Science, pàg. 977-980. DOI: 10.1126/science.1185402. URL: https://www.science.org/doi/abs/10.1126/science.1185402 (cons. 05-03-2023).

Changeat, Quentin i Kai Hou Yip (juny de 2022). *Ariel Big Challenge (ABC) Database*. eng. Type: dataset. DOI: 10.5281/zenodo.6770103. URL: https://zenodo.org/record/6770103 (cons. 05-04-2023).

— (gen. de 2023). *ESA-Ariel Data Challenge NeurIPS 2022: Introduction to exo-atmospheric studies and presentation of the Atmospheric Big Challenge (ABC) Database*. en. arXiv:2206.14633 [astro-ph, physics:physics]. URL: http://arxiv.org/abs/2206.14633 (cons. 18-02-2023).

Charbonneau, David et al. (gen. de 2000). "Detection of Planetary Transits Across a Sun-like Star". en. A: *The Astrophysical Journal* 529.1, pàg. L45-L48. ISSN: 0004637X. DOI: 10.1086/312457. URL: https://iopscience.iop.org/article/10.1086/312457 (cons. 04-03-2023).

Chen, Tianqi i Carlos Guestrin (ag. de 2016). "XGBoost: A Scalable Tree Boosting System". A: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. arXiv:1603.02754 [cs], pàg. 785-794. DOI: 10.1145/2939672.2939785. URL: http://arxiv.org/abs/1603.02754 (cons. 20-05-2023).

Cobb, Adam D. et al. (juny de 2019). "An Ensemble of Bayesian Neural Networks for Exoplanetary Atmospheric Retrieval". en. A: *The Astronomical Journal* 158.1, pàg. 33. ISSN: 1538-3881. DOI: 10.3847/1538-3881/ab2390. URL: https://iopscience.iop.org/article/10.3847/1538-3881/ab2390 (cons. 18-02-2023).

Criminisi, Antonio, Jamie Shotton i Ender Konukoglu (març de 2012). "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and

Semi-Supervised Learning". English. A: *Foundations and Trends® in Computer Graphics and Vision* 7.2–3. Publisher: Now Publishers, Inc., pàg. 81 - 227. ISSN: 1572-2740, 1572-2759. DOI: 10.1561/0600000035. URL: https://www.nowpublishers.com/article/Details/CGV-035 (cons. 27-05-2023).

Edwards, Billy i Giovanna Tinetti (juny de 2022). "The Ariel Target List: The Impact of TESS and the Potential for Characterizing Multiple Planets within a System". en. A: *The Astronomical Journal* 164.1. Publisher: The American Astronomical Society, pàg. 15. ISSN: 1538-3881. DOI: 10.3847/1538-3881/ac6bf9. URL: https://dx.doi.org/10.3847/1538-3881/ac6bf9 (cons. 25-02-2023).

Edwards, Billy et al. (abr. de 2019). "Exoplanet spectroscopy and photometry with the Twinkle space telescope". en. A: *Experimental Astronomy* 47.1, pàg. 29 - 63. ISSN: 1572-9508. DOI: 10.1007/s10686-018-9611-4. URL: https://doi.org/10.1007/s10686-018-9611-4 (cons. 04-03-2023).

*Exoplanet atmospheres* (2023). URL: http://research.iac.es/proyecto/exoatmospheres/index.php (cons. 23-03-2023).

Feroz, F., M. P. Hobson i M. Bridges (oct. de 2009). "MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics". A: *Monthly Notices of the Royal Astronomical Society* 398.4. arXiv:0809.3437 [astro-ph], pàg. 1601 - 1614. ISSN: 00358711, 13652966. DOI: 10.1111/j.1365-2966.2009.14548.x. URL: http://arxiv.org/abs/0809.3437 (cons. 04-06-2023).

Fisher, Chloe et al. (maig de 2020). "Interpreting High-resolution Spectroscopy of Exoplanets using Cross-correlations and Supervised Machine Learning". A: *The Astronomical Journal* 159. ADS Bibcode: 2020AJ....159..192F, pàg. 192. ISSN: 0004-6256. DOI: 10.3847/1538-3881/ab7a92. URL: https://ui.adsabs.harvard.edu/abs/2020AJ....159..192F (cons. 19-02-2023).

Fortney, Jonathan J., Rebekah I. Dawson i Thaddeus D. Komacek (març de 2021). "Hot Jupiters: Origins, Structure, Atmospheres". en. A: *Journal of Geophysical Research: Planets* 126.3. ISSN: 2169-9097, 2169-9100. DOI: 10.1029/2020JE006629. URL: https://onlinelibrary.wiley.com/doi/10.1029/2020JE006629 (cons. 04-03-2023).

Fragkoudi, F. (març de 2020). "Astronomy as a Tool for Peace and Diplomacy: Experiences from the Columba-Hypatia Project". en. A: *Communicating Astronomy with the Public Journal* 27, pàg. 14. URL: https://ui.adsabs.harvard.edu/abs/2020CAPJ...27...14F/abstract (cons. 10-03-2023).

Gaudi, B. Scott, Jessie L. Christiansen i Michael R. Meyer (oct. de 2021). *The Demographics of Exoplanets*. en. arXiv:2011.04703 [astro-ph]. DOI: 10.1088/2514-3433/abfa8fch2. URL: http://arxiv.org/abs/2011.04703 (cons. 04-03-2023).

Greene, Thomas P. et al. (gen. de 2016). "Characterizing transiting exoplanet atmospheres with JWST". en. A: *The Astrophysical Journal* 817.1, pàg. 17. ISSN: 1538-4357. DOI: 10.3847/0004-637X/817/1/17. URL: https://iopscience.iop.org/article/10.3847/0004-637X/817/1/17 (cons. 04-03-2023).

Guryanov, Aleksei (2019). "Histogram-Based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees". en. A: *Analysis of Images, Social Networks and Texts*. Ed. de Wil M. P. van der Aalst et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, pàg. 39 - 50. ISBN: 978-3-030-37334-4. DOI: 10.1007/978-3-030-37334-4_4.

Haghighipour, Nader (maig de 2013). "The Formation and Dynamics of Super-Earth Planets". en. A: *Annual Review of Earth and Planetary Sciences* 41.1. arXiv:1306.5567 [astro-ph], pàg. 469 - 495. ISSN: 0084-6597, 1545-4495. DOI: 10.1146/annurev-earth-042711-105340. URL: http://arxiv.org/abs/1306.5567 (cons. 05-03-2023).

Hoeijmakers, H. J. et al. (jul. de 2019). "A spectral survey of an ultra-hot Jupiter: Detection of metals in the transmission spectrum of KELT-9 b". en. A: *Astronomy & Astrophysics* 627, A165. ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/201935089. URL: https://www.aanda.org/10.1051/0004-6361/201935089 (cons. 19-03-2023).

Hoeijmakers, H. Jens et al. (ag. de 2018). "Atomic iron and titanium in the atmosphere of the exoplanet KELT-9b". en. A: *Nature* 560.7719. Number: 7719 Publisher: Nature Publishing Group, pàg. 453 - 455. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0401-y. URL: https://www.nature.com/articles/s41586-018-0401-y (cons. 19-03-2023).

James, Gareth et al. (2021a). *An Introduction to Statistical Learning: with Applications in R*. en. Springer Texts in Statistics. New York, NY: Springer US. ISBN: 978-1-07-161417-4 978-1-07-161418-1. DOI: 10.1007/978-1-0716-1418-1. URL: https://link.springer.com/10.1007/978-1-0716-1418-1 (cons. 19-03-2023).

— (2021b). "Tree-Based Methods". en. A: *An Introduction to Statistical Learning: with Applications in R*. Ed. de Gareth James et al. Springer Texts in Statistics. New York, NY: Springer US, pàg. 327 - 365. ISBN: 978-1-07-161418-1. DOI: 10.1007/978-1-0716-1418-1_8. URL: https://doi.org/10.1007/978-1-0716-1418-1_8 (cons. 19-03-2023).

Ke, Guolin et al. (2017). "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". A: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html (cons. 27-05-2023).

Kowalek, Patrycja, Hanna Loch-Olszewska i Janusz Szwabiński (set. de 2019). "Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach". en. A: *Physical Review E* 100.3, pàg. 032410. ISSN: 2470-0045, 2470-0053. DOI:

10.1103/PhysRevE.100.032410. URL: https://link.aps.org/doi/10.1103/PhysRevE.100.032410 (cons. 22-05-2023).

Kreidberg, Laura et al. (nov. de 2015). "A detection of water in the transmission spectrum of the hot Jupiter WASP-12b and implications for its atmospheric composition." en. A: *The Astrophysical Journal* 814.1, pàg. 66. ISSN: 1538-4357. DOI: 10.1088/0004-637X/814/1/66. URL: https://iopscience.iop.org/article/10.1088/0004-637X/814/1/66 (cons. 19-03-2023).

Lannelongue, Loïc, Jason Grealey i Michael Inouye (2021). "Green Algorithms: Quantifying the Carbon Footprint of Computation". en. A: *Advanced Science* 8.12. _eprint: https://onlinelibrary.wile pàg. 2100707. ISSN: 2198-3844. DOI: 10.1002/advs.202100707. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/advs.202100707 (cons. 10-03-2023).

Madhusudhan, Nikku (2018). "Atmospheric Retrieval of Exoplanets". A: arXiv:1808.04824 [astro-ph], pàg. 2153 - 2182. DOI: 10.1007/978-3-319-55333-7_104. URL: http://arxiv.org/abs/1808.04824 (cons. 02-03-2023).

— (ag. de 2019). "Exoplanetary Atmospheres: Key Insights, Challenges and Prospects". A: *Annual Review of Astronomy and Astrophysics* 57.1. arXiv:1904.03190 [astro-ph], pàg. 617 - 663. ISSN: 0066-4146, 1545-4282. DOI: 10.1146/annurev-astro-081817-051846. URL: http://arxiv.org/abs/1904.03190 (cons. 26-02-2023).

Marquez-Neila, Pablo et al. (juny de 2018). *Supervised Machine Learning for Analysing Spectra of Exoplanetary Atmospheres*. en. arXiv:1806.03944 [astro-ph, physics:physics]. URL: http://arxiv.org/abs/1806.03944 (cons. 18-02-2023).

Martínez, F. Ardévol et al. (juny de 2022). "Convolutional neural networks as an alternative to Bayesian retrievals for interpreting exoplanet transmission spectra". en. A: *Astronomy & Astrophysics* 662. Publisher: EDP Sciences, A108. ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/202142976. URL: https://www.aanda.org/articles/aa/abs/2022/06/aa42976-21/aa42976-21.html (cons. 18-02-2023).

Mayor, Michel i Didier Queloz (nov. de 1995). "A Jupiter-mass companion to a solar-type star". en. A: *Nature* 378.6555. Number: 6555 Publisher: Nature Publishing Group, pàg. 355 - 359. ISSN: 1476-4687. DOI: 10.1038/378355a0. URL: https://www.nature.com/articles/378355a0 (cons. 04-03-2023).

NASA Exoplanet Archive, NASA (març de 2023). *A service of NASA Exoplanet Science Institute*. URL: https://exoplanetarchive.ipac.caltech.edu/ (cons. 05-03-2023).

NIST (2023). *Methane*. en. Publisher: National Institute of Standards and Technology. URL: https://webbook.nist.gov/cgi/cbook.cgi?ID=C74828&Type=IR-SPEC&Index=1 (cons. 14-06-2023).

Nixon, Matthew C i Nikku Madhusudhan (jul. de 2020). "Assessment of supervised machine learning for atmospheric retrieval of exoplanets". A: *Monthly Notices of the Royal Astronomical Society* 496.1, pàg. 269 - 281. ISSN: 0035-8711. DOI: 10.1093/mnras/staa1150. URL: https://doi.org/10.1093/mnras/staa1150 (cons. 18-02-2023).

Potthast, Roland (abr. de 2006). "A survey on sampling and probe methods for inverse problems". en. A: *Inverse Problems* 22.2, R1 - R47. ISSN: 0266-5611, 1361-6420. DOI: 10.1088/0266-5611/22/2/R01. URL: https://iopscience.iop.org/article/10.1088/0266-5611/22/2/R01 (cons. 06-03-2023).

Pätzold, M. et al. (set. de 2012). "Transiting exoplanets from the CoRoT space mission - XXIII. CoRoT-21b: a doomed large Jupiter around a faint subgiant star". en. A: *Astronomy & Astrophysics* 545. Publisher: EDP Sciences, A6. ISSN: 0004-6361, 1432-0746. DOI: 10.1051/0004-6361/201118425. URL: https://www.aanda.org/articles/aa/abs/2012/09/aa18425-11/aa18425-11.html (cons. 05-03-2023).

Ricker, George R. et al. (oct. de 2014). "Transiting Exoplanet Survey Satellite". en. A: *Journal of Astronomical Telescopes, Instruments, and Systems* 1.1, pàg. 014003. ISSN: 2329-4124. DOI: 10.1117/1.JATIS.1.1.014003. URL: http://astronomicaltelescopes.spiedigitallibrary.org/article.aspx?doi=10.1117/1.JATIS.1.1.014003 (cons. 13-03-2023).

Soboczenski, Frank et al. (gen. de 2018). "Bayesian Deep Learning for Exoplanet Atmospheric Retrieval". en. A.

Tinetti, Giovanna et al. (abr. de 2021). *Ariel: Enabling planetary science across light-years*. arXiv:2104.04824 [astro-ph]. DOI: 10.48550/arXiv.2104.04824. URL: http://arxiv.org/abs/2104.04824 (cons. 25-02-2023).

Tsiaras, A. et al. (març de 2018). "A Population Study of Gaseous Exoplanets". en. A: *The Astronomical Journal* 155.4, pàg. 156. ISSN: 1538-3881. DOI: 10.3847/1538-3881/aaaf75. URL: https://iopscience.iop.org/article/10.3847/1538-3881/aaaf75 (cons. 23-03-2023).

Vasist, Malavika et al. (febr. de 2023). *Neural posterior estimation for exoplanetary atmospheric retrieval*. en. arXiv:2301.06575 [astro-ph]. URL: http://arxiv.org/abs/2301.06575 (cons. 26-03-2023).

Waldmann, I. P. (març de 2016). "Dreaming of atmospheres". A: *The Astrophysical Journal* 820.2. arXiv:1511.08339 [astro-ph], pàg. 107. ISSN: 1538-4357. DOI: 10.3847/0004-637X/820/2/107. URL: http://arxiv.org/abs/1511.08339 (cons. 02-03-2023).

Wei, Peng-Sheng et al. (oct. de 2018). "Absorption coefficient of carbon dioxide across atmospheric troposphere layer". A: *Heliyon* 4.10, e00785. ISSN: 2405-8440. DOI: 10.1016/j.heliyon.

2018.e00785. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6174548/ (cons. 14-06-2023).

Zingales, Tiziano i Ingo P. Waldmann (nov. de 2018). "ExoGAN: Retrieving Exoplanetary Atmospheres Using Deep Convolutional Generative Adversarial Networks". en. A: *The Astronomical Journal* 156.6, pàg. 268. ISSN: 1538-3881. DOI: 10.3847/1538-3881/aae77c. URL: https://iopscience.iop.org/article/10.3847/1538-3881/aae77c (cons. 18-02-2023).