Classification of TESS data for rocky exoplanets

Luis Andres García López
Facultad de Ciencias
Universidad Autónoma de Baja California (UABC)
Ensenada, Baja California
Email: luis.andres.garcia.lopez@uabc.edu.mx

***Abstract*—This article presents a novel approach to classify rocky exoplanets using data mining and machine learning techniques. It begins by introducing the Transit-Exoplanet Survey Satellite (TESS) mission, which has revolutionized the search for exoplanets. Despite the success of TESS, accurately classifying rocky exoplanets remains a challenge due to the complexity of the data. To address this issue, a data mining methodology is proposed that can efficiently classify exoplanets based on their characteristics. The methodology is described in detail, including data preprocessing steps, feature selection, and the machine learning algorithms used. The results of the project are presented, demonstrating the effectiveness of the proposed methodology in classifying rocky exoplanets. The article concludes by discussing the implications of this work for the search for Earth-like exoplanets and the future of exoplanet research.**

I. Introduction

The search for exoplanets, planets outside our solar system, is an exciting and evolving field in astronomy. The Transiting Exoplanet Survey Satellite (TESS) mission by NASA, launched in 2018, has opened a new window for detecting Earth-like planets. TESS[1] has discovered over 4,000 exoplanet candidates and confirmed the existence of more than 200 exoplanets. However, a concrete method for classifying a rocky planet has not yet been developed. This is because determining its classification accurately requires a lot of telescope observation time, or alternatively, analysis of its atmosphere and comparisons with planets that have been extensively studied[1].

As more exoplanets are discovered, the analysis and classification of data becomes increasingly complex. Data mining can provide a faster and more efficient solution to this problem [6]. This recent project has generated a large amount of data that needs to be analyzed. By identifying exoplanets with Earth-like characteristics, astronomers can focus on studying these planets and searching for signs of life. Our aim is to Provide an automated methodology for classifying exoplanets using data mining and machine learning techniques. The hypothesis that guides our project is that Planets with a density between 3 $\left(\frac{\mathrm{g}}{\mathrm{cm}^3}\right)$ to 6 $\left(\frac{\mathrm{g}}{\mathrm{cm}^3}\right)$ will be classified as rocky regardless of their atmosphere, depending on the criteria applied, also, the ESI (Earth Similarity Index) is used as a comparative measure specifically targeting rocky planets[4].

Nevertheless, Classifying exoplanets is a crucial process for understanding the characteristics and properties of these celestial bodies. Manual classification of exoplanets is a laborious task that requires expert and detailed knowledge of the exoplanet features, but data mining offers a faster and more efficient solution for classifying large exoplanet datasets.

In this project, we use data science techniques to group rocky exoplanets using data from the TESS project. Certain techniques will be applied to different parameters of these planets, such as their densities and distances from their respective stars, among other factors. We explore the public TESS database[2], which contains information on the location, size, temperature, and other characteristics of exoplanets discovered by TESS.

II. Data Description

The data being used is directly extracted from a CALTECH database and provided by NASA. These data have variations in each observation since TESS scans specific areas on a specific date, but the most recent ones have been taken. The variables that characterize the planet, such as its volume, density, stellar flux, among others, will be used for comparative and descriptive studies with respect to Earth and Jupiter. The aim is to obtain results from the perspective of a rocky planet and a gaseous planet. Only numerical data has been taken for this study. In some cases, observations of a planet may not have all their variables recorded, so a value of 0 is assigned to them, which will ultimately remove them from the study.

The extracted data was modified to support the research purposes. These modifications included the removal of irrelevant data, implementation of boundaries or limits for very large data, as well as rounding off specific data as the source includes measurement error margins.

Table I
Characterized data table

dens_pl_ert	dens_pl_jup	densEar_discret	densJup_discret
0.345628	1432.885986	med	med
0.723651	2984.573098	med	med
0.835612	3423.419255	med	med
0.185851	772.029713	med	med
0.030080	124.642288	low	med
0.025251	104.631284	low	med

[1]https://tess.mit.edu/

[2]https://exoplanetarchive.ipac.caltech.edu/

III. Methodology

The following points present the methodology used and the rationale behind its application in the project

1) Data Selection
   The selected data from the TESS database were filtered as suggested in [4], aiming to have a solid comparison base like Earth. Therefore, we expect that the obtained similarity values will provide us with better indicators for classification.
2) Preprocessing.
   Due to the nature of the data being handled, there were initially missing values, which were removed along with the unused data. Additionally, noise reduction has been performed regarding error margins in certain values. These error margins, resulting from computational calculations, error propagation, or the precision of planet observations, would not be taken into account. After applying preprocessing techniques, a subset of the data was obtained, which is presented in Table 1
3) Data transformation
   Once the data has been cleaned, a feature engineering process will be performed, including data discretization and normalization. This treatment aims to create new meaningful features from the existing data and ensure that the data is standardized for further analysis. While normalization ensures that the data is scaled to a consistent range to eliminate biases caused by different units or scales. These steps will enhance the quality and compatibility of the data for subsequent analysis in the project.
4) Experimentation
   We will apply the respective attributes to the ESI[4] index expression, aiming to emulate it in a machine learning environment. This approach will enable us to cluster planets based on their similarities.
5) Characterization
   To visualize the distribution of data based on a variable, characterization techniques are employed. These techniques help gain insights and understand the distribution patterns of the data. Prior to applying various tools such as statistical correlation models and classification models like KNN, the data underwent preprocessing to calculate density values for each planet relative to Earth and Jupiter. Following that, a nominal segmentation was performed to validate the correlation between planetary density and radius, providing graphical evidence of accurate density calculations. It was assumed that planets with larger radius would typically exhibit lower densities, as shown in the table. I.

IV. Related Work

Due to the recent nature of TESS data, most research conducted on these topics and data science is still being published or in the process of being developed. As a result, only a few recent articles have been identified. One of the noteworthy articles used as a reference citation covers the following topics. [5]

- The article focuses on the use of deep learning techniques, specifically neural networks, to classify light curves and detect exoplanetary transit signals in the data generated by the TESS mission.
- A high-quality dataset is presented, containing light curves from full-frame images of both the Primary Mission and the First Extended Mission of TESS.
- The Box Least Squares method is employed to detect periodic signals in the light curves.
- The dataset is curated through a thorough manual review process before being used to train the Astronet-Triage-v2 neural network.
- High levels of precision and recall are achieved in the detection of transit/eclipse events in the test set.

Another article[1] makes significant progress in the classification of rocky exoplanets by examining a dataset of 720 previously identified rocky planets from a source other than TESS. The authors employ an analog approach based on our solar system to assess the degree of similarity to Mars and Earth among these exoplanets. They utilize a Monte Carlo model for this analysis, aiming to determine which exoplanets merit focused investigation in the search for potential biological signals.

V. Experiments

As initial steps in the data exploration, the most general values are plotted and compared to those of Earth. In this case, as shown in Figure 1, a comparison of the density variation between Earth and the studied exoplanets is depicted.

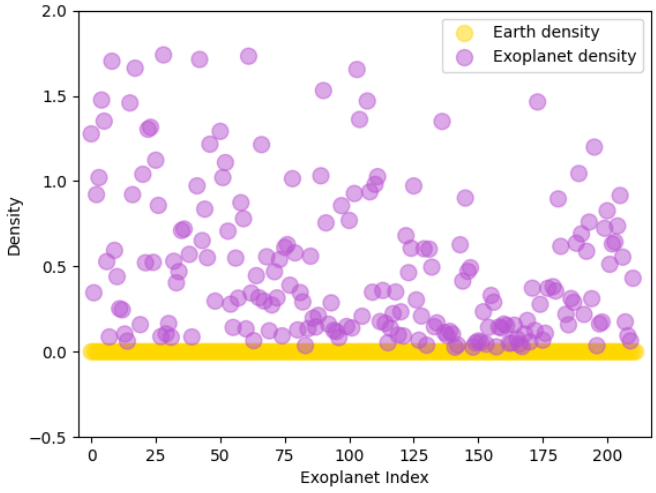

Figure 1. Density comparison

In the following experiment, we implemented the KNeighborsRegressor algorithm with the aim of creating a trained model that, based on the equilibrium temperature, stellar

temperature, radius, and density values, can determine how similar an exoplanet is to Earth.

To use the KNeighborsRegressor model, the training data is first prepared by selecting relevant features such as planet density, radius, stellar temperature, and equilibrium temperature. Next, the data is divided into training and test sets, where the training set is used to train the model and the test set is used to evaluate its performance. An instance of the KNeighborsRegressor model is created, specifying the number of neighbors to consider for making predictions. Then, the model is trained using the training data by finding the nearest neighbors to an input data point and using their similarity values to make predictions. Once trained, the model can make predictions on new data by finding the nearest neighbors in the training set and using their similarity values to predict the similarity of the new point to Earth. Finally, the model's performance is evaluated by comparing the predictions to the actual values in the test set using metrics such as mean squared error (MSE), which quantifies the average squared difference between the predictions and the actual values. In summary, the KNeighborsRegressor is used to estimate the similarity of exoplanets to Earth based on relevant features and similarity to the nearest neighbors, through the process of training, prediction, and model evaluation.

As depicted in Figure **??**, the model exhibits commendable learning and functionality. Subsequently, we proceed to assess the model's performance by subjecting it to testing using TESS data from unidentified celestial objects, which have yet to be classified as exoplanets, as well as the planets within our solar system. This endeavor aims to provide a more comprehensive evaluation of the model's efficacy

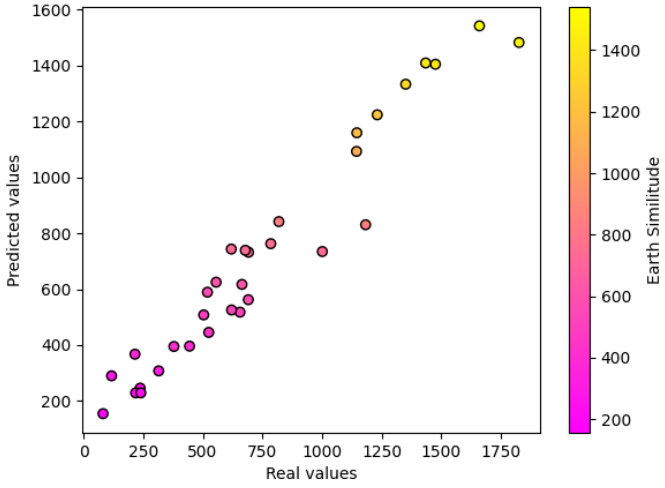

Figure 2. KNRegressor Earth similitude

The regression model provides insights suggesting that values closer to 0 exhibit similarities to Earth, while the dissimilarity increases as the similarity value rises. Furthermore, the aforementioned process was applied utilizing density, radius, equilibrium temperature (K), and stellar temperature (K) as reference values. The outcomes, specifically for Jupiter, are elaborated upon in Figure 3.

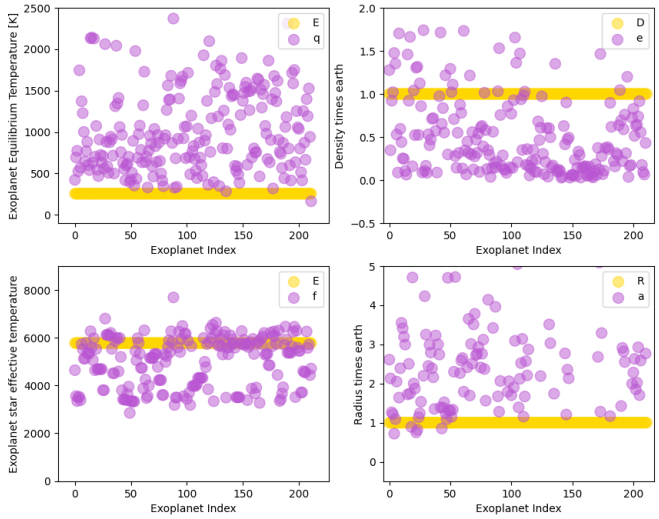

Figure 3. Variables visualization

Table II
SIMILITUDE COMPARISON

pl_name	SimilSolar	ExoPlName	ExoSimilEarth
Venus	**347.3039505767211**	LP 890-9 b	167.986768
Mars	**347.3039505767211**	TOI-1696 b	240.148491
Jupiter	**330.4411580141149**	TOI-1452 b	72.292377
saturno	**330.4411580141149**	TOI-3884 b	218.531977
Uranus	**330.4411580141149**	LTT 3780 c	101.341191
Neptuno	**330.4411580141149**	TOI-163 b	1434.874870

Table 2 presents the similarity results of a data sample that underwent the KNregressor model, compared to the values derived from certain planets within the solar system. These findings provide clear evidence that the temperature of the star is one of the significant factors in the model, as the planets within the solar system exhibit similar similarities. However, when examining the similarity values of exoplanets from different solar systems, significant variations arise.[3]

Furthermore, by comparing these similarity values with those obtained using the Earth Similarity Index (ESI), we can assert that the model is robust. For instance, when considering Venus with an ESI of 0.78, our comparison yields a normalized similarity of 0.8, based on our similarity values ranging from 0 to 1.[2]

VI. CONCLUSION

As shown in Fig.4 the model is robust enough to make accurate predictions regarding similarity to Earth.Applying a machine learning model is not a trivial task, and it is necessary to conduct utility tests to decide which methods to use based

on the problem or objective. Based on the acquired knowledge, it is possible to carry out data analysis with enough value to determine if a planet is similar to Earth, according to the variables considered in the study.

pl_rade	dens_pl_ert	st_teff	pl_eqt	pl_name	Prediction KNR
6046.0790	5.239459	5773	226	Venus	0.80
3389.3720	3.933038	5772	210	Mars	0.83
69910.8943	1.329949	5774	110	Jupiter	0.20
58231.5771	0.699770	5777	63	saturno	0.24
25361.6768	19.049723	5778	64	Uranus	0.44
24617.5440	1.639776	5780	51	Neptuno	0.44

Figure 4. Variables visualization

Table III
STUDIED DATA

pl_rade	pl_radj	pl_bmasse	pl_bmassj	st_teff	pl_eqt	Pl_eh_vol	Pl_J_vol	dens_pl_ert	dens_pl_jup
2.61	0.233	22.7	0.07142	4640	422	17.779581	0.012649337	1.276749132	5279.852677
2.13	0.19	3.34	0.01051	3556	693	9.663597	0.006859	0.34562798	1432.885986
1.264	0.113	1.86	0.00585	3347	773	2.019487744	0.001442897	0.921028266	3791.318391
1.217	0.109	1.84	0.00579	3505	525	1.802485313	0.001295029	1.020815662	4180.890486
0.718	0.064	0.546	0.00172	3522	1745	0.370146232	0.000262144	1.475096886	6135.617067
1.09	0.097	1.75	0.00551	3384	568	1.295029	0.000912673	1.351324937	5645.548178
2.05	0.183	4.55	0.01432	5732	1371	8.615125	0.006128487	0.528142418	2185.040397
11.5	1.026	131.89945	0.415	5521	1228	1520.875	1.080045576	0.086726276	359.3153215
1.64	0.146	7.51	0.02363	5125	1001	4.410944	0.003112136	1.702588239	7100.270202

REFERENCES

[1] Sarah R N McIntyre, Penelope L King, and Franklin P Mills. A rocky exoplanet classification method and its application to calculating surface pressure and surface temperature. *Monthly Notices of the Royal Astronomical Society*, 519(4):6210–6221, 01 2023.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] Dirk Schulze-Makuch, Abel Méndez, Alberto G. Fairén, Philip von Paris, Carol Turse, Grayson Boyer, Alfonso F. Davila, Marina Resendes de Sousa António, David Catling, and Louis N. Irwin. A two-tiered approach to assessing the habitability of exoplanets. *Astrobiology*, 11(10):1041–1052, 2011. PMID: 22017274.

[4] Edward W. Schwieterman, Nancy Y. Kiang, Mary N. Parenteau, Chester E. Harman, Shiladitya DasSarma, Theresa M. Fisher, Giada N. Arney, Hilairy E. Hartnett, Christopher T. Reinhard, Stephanie L. Olson, Victoria S. Meadows, Charles S. Cockell, Sara I. Walker, John Lee Grenfell, Siddharth Hegde, Sarah Rugheimer, Renyu Hu, and Timothy W. Lyons. Exoplanet biosignatures: A review of remotely detectable signs of life. *Astrobiology*, 18(6):663–708, 2018. PMID: 29727196.

[5] Evan Tey, Dan Moldovan, Michelle Kunimoto, Chelsea X. Huang, Avi Shporer, Tansu Daylan, Daniel Muthukrishna, Andrew Vanderburg, Anne Dattilo, George R. Ricker, and S. Seager. Identifying exoplanets with deep learning. v. improved light-curve classification for TESS full-frame image observations. *The Astronomical Journal*, 165(3):95, feb 2023.

[6] Yanxia Zhang, Hongwen Zheng, and Yongheng Zhao. Knowledge discovery in astronomical data. In Alan Bridger and Nicole M. Radziwill, editors, *Advanced Software and Control for Astronomy II*, volume 7019, page 701938. International Society for Optics and Photonics, SPIE, 2008.