Using Machine Learning to identify the Habitability of Exoplanets from TESS Transit Data

Eve Belding

October 23, 2022

Abstract

With so many new tools and methods to detect exoplanets, it is time to start determining which are deserving of more research in the hunt for other habitable planets. We began our research with exoplanets that are similar to Earth as the most likely to house life. We used transit measurements such as an exoplanet's size, temperature, and relative distance from its star as well as machine learning to develop two models: one to predict if an object of interest was an exoplanet and another to cluster our data to determine its similarity to Earth. We achieved an 88.7% accuracy for the first model, and our clustering algorithm found interesting results when compared to a graph showing the principal components of the data. We found that our Earth data point fell right on the origin of our 2 dimensional PCA graph and was placed in our largest cluster. Both of our models showed that machine learning can be useful in the search for habitable exoplanets and is deserving of more research.

1. Introduction

With the growing amount of tools to search for exoplanets, the need for efficient research has become more prevalent. We have data for many potential exoplanets, but going through them one by one is time consuming. Machine learning is how we can quickly determine which exoplanets are most likely to be habitable and therefore most warrant further research. We worked with the TESS Objects of Interest data set from the NASA Exoplanet Archives [11], which consists of numerical and categorical data that stems from transit information. Transit Photometry is the study of far away planets based on the fluctuations in light coming from the star it is orbiting[3]. This data was used to create two supervised machine learning models. The first was a classification model to determine if the object of interest was an exoplanet by outputting a true or false label, and the second to cluster the data to determine those most like Earth.

2. Background

While the hunt for exoplanets is relatively new, multiple methods have been tested in determining exoplanets and their possible habitability. Past work discussed how the presence of water could be used to determine if an exoplanet is habitable[7]. Researchers were able to use spectroscopy and hypothesized data to determine if an exoplanet had water. While water is an important feature for a habitable planet, the data they used to train their model was mostly hypothesized and they were unable to test it due to a lack of data. Past work has also used the same transit data used in this paper to determine scores for the likeness of an exoplanet to Earth[6]. They gave a strong focus to the temperature of the exoplanets as that is something that would greatly affect the possibility of human life. Our method in determining similarity to Earth differs in that we included Earth as a row in our dataset and clustered with it.

3. Dataset

The exoplanet dataset used in this work[11] includes 5908 objects of interest ranging from confirmed planets to planetary candidates to false positives. Each object of interest has 65 dimensions including the transit depth, the equilibrium temperature of the object, the radius of its star in relation to the sun, and more as well as the uncertainties and limits[10] to the majority of these measurements.

In our initial cleaning steps, we eliminated all of the columns that were unnecessary for this project or were solely null values. Columns that were used to title the data and describe the time at which they were discovered were unnecessary to our purposes. There were also pairs of columns that included sexigismal time and decimal time. We eliminated the sexigismal time columns as they were redundant time information. Finally, the columns that only had null values, which included many of the uncertainty and limit columns, were eliminated. The columns that had a few null values also had to be dealt with. Objects with null values were eliminated from the dataset as proper replacements were not apparent. This eliminated 1863 of our 5809 objects. The only numerical transformation needing to be performed was on our column that described whether the object was: a known planet, a confirmed planet, a planetary candidate, an ambiguous planetary candidate, a false positive, or a false alarm. This column became very important for the Random Forest Classifier discussed in the Methods/Methodology section.

To better visualize our data, we performed a PCA transformation. This reduced our data's dimensions with the most important factors being the transit depth and the positive uncertainty of the transit depth. We color coded the PCA graph based on the classifications, as shown in figure 1, but observed no visible correlation.

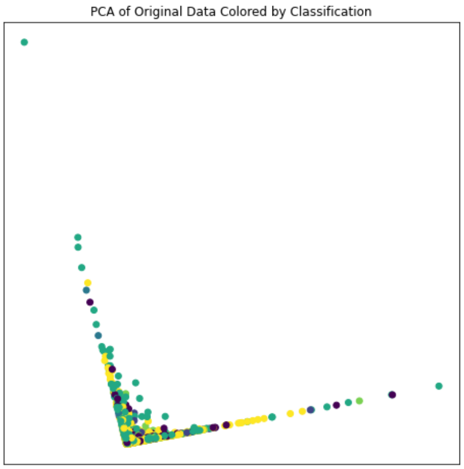

Figure 1: PCA colored by classifications

4. Methodology/Models

We were interested in a two step project. The first part would determine if an object of interest was a planet or not by splitting the data into trues and falses based on the classifications given by our dataset. The second part would then cluster the objects with Earth data to determine those most similar to Earth.

For step one of the project to determine whether or not an object of interest was an exoplanet, we chose a Random Forest classifier. This classifier works by creating lots of different decision trees and then averaging their results. To begin, our data was reconfigured to suit our model. We focused on the points from our dataset that were classified as either “false alarms” and “false positives”, or “known planets” and “confirmed planets". The objects that were not classified in any of these categories were omitted for the time being. These points were then redefined as either true or false. This column of trues and falses became our y values while the other data became our X set. The data was split 20% for testing vs. 80% for training data.

With our data now split, we moved on to maximizing the accuracy of our Random Forest Classifier. To tune the hyperparameters of the model, grid searches were performed. As seen in figure 2, these grid searches were performed for the max depth, random state, and n estimators of the model. The first instance on each of these graphs in which the accuracy was at its highest level were chosen to maximize the total accuracy, so the max depth = 13, random state = 8, and n estimators = 134. We did experiment with both an SVM classifier and a Logistic Regression model, but a higher accuracy could be achieved with the Random Forest Classifier.

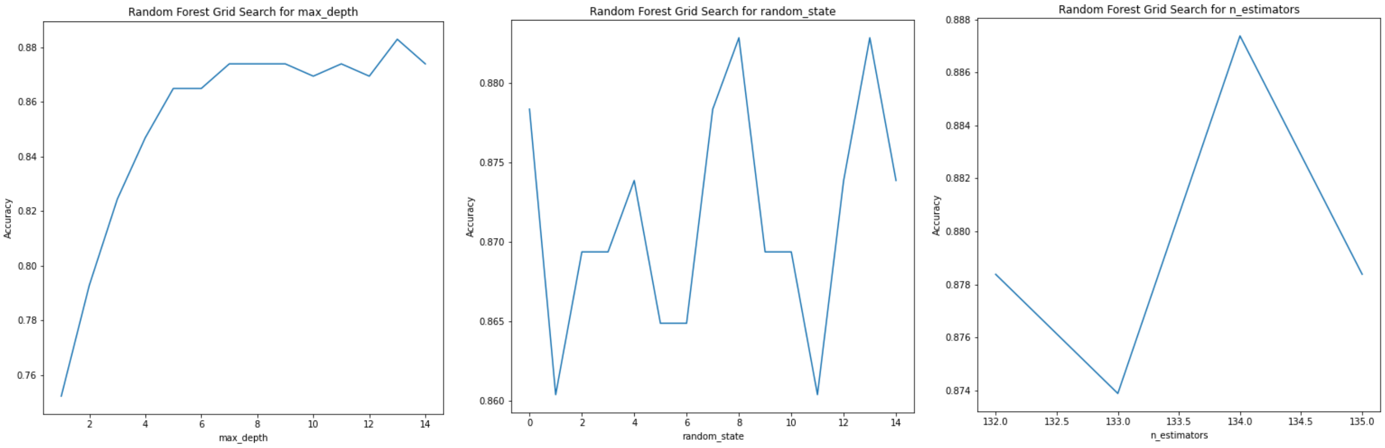

Figure 2: Grid searches for max_depth, random_state, and n_estimators; For max_depth, random_state and n_stimators were their defaults of 0 and 100 respectively. For random_state, max_depth = 13 and n_estimators = 100. Finally, for n_estimators, max_depth = 13 and random_state = 8

For step 2 of the project, we wanted to cluster the exoplanets to determine the exoplanets most like Earth. We predicted that the closer an exoplanet's characteristics were to Earth, the more likely it would be that said exoplanet could be habitable or able to sustain some sort of life.

To create the dataset for our clustering algorithm, all of the error and lim columns[10] from our data as well as columns like the distance from our planetary system to the exoplanet and the proper motion of the right ascension of the planet were eliminated. The corresponding data for Earth was then found and imported into a new row at the end of the dataset.

Once all of the data was organized, we performed a k-means algorithm. We chose to use 5 clusters because it was simple while easy to read and understand. As seen in figure 3, one of our clusters was far larger than the other. This cluster is the one that Earth was sorted into. This makes sense as Earth was used as the baseline for many of the measurements[10]. A couple of different cluster amounts were tried, and all of which had one far larger cluster that included Earth.

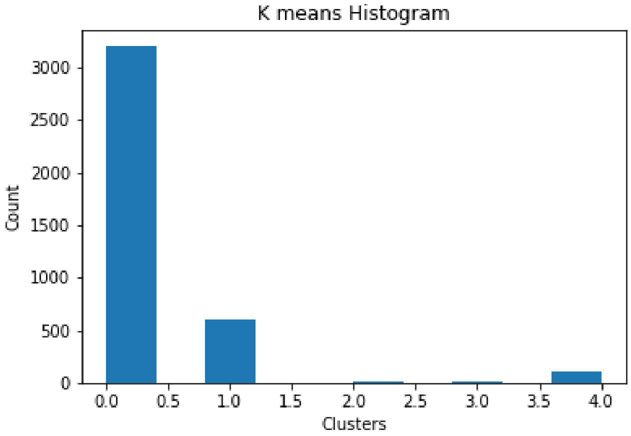

Figure 3: Histogram of the Clustering algorithm. Earth is in the largest column

5. Results and Discussion

As stated earlier, the Random Forest Classifier had the best accuracy of any of the Classification models that were tried. Figure 4 shows the Confusion Matrix and ROC Curve of our model. It had an accuracy score of 88.7% and a precision score of 93.4% . The model was then tested on the undetermined planetary candidates that were separated from the data earlier. The model predicted that 1044 were planets and that 1548 were not. Scientists who work on discovering exoplanets could use our model to determine which candidates to start with in their search.

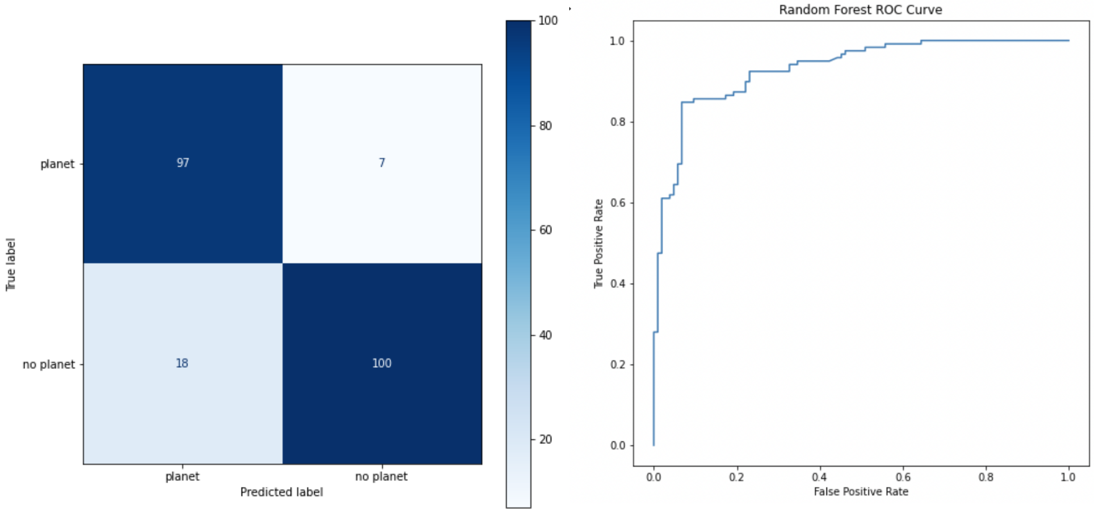

Figure 4: Confusion Matrix and ROC Curve for the Random Forest Classifier

To further explore our model, a PCA decomposition was performed on our model which is pictured in figure 5. The graph was colored based on whether the object of interest was a planet or not. There was some correlation along the axes of the PCA graph, but the trends were not particularly distinct.

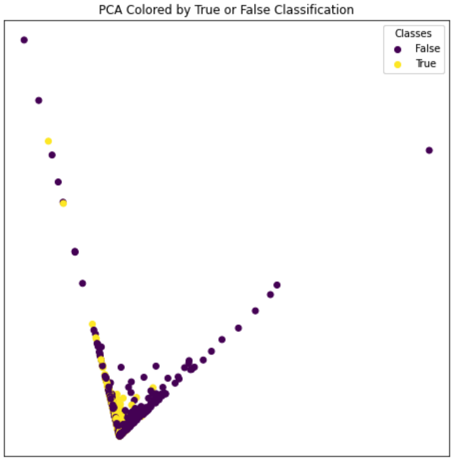

Figure 5: PCA graph colored by True or False Classifications

The results from our k-means clustering model were very interesting when used in relation to their PCA decomposition. These results are pictured in figure 6. There are very distinct sections on the graph that are color coded by the clusters the points were put in. We can also see that Earth is very close to where the origin of the PCA graph is. There are two possibilities to explain this observation. One is that scientists have focused on researching the objects of interest that appear similar to Earth. There is more data for these types of objects, so they carry more weight for the PCA. Option number two is that because much of our data is centered around Earth,[10] it appears near the center of the graph.

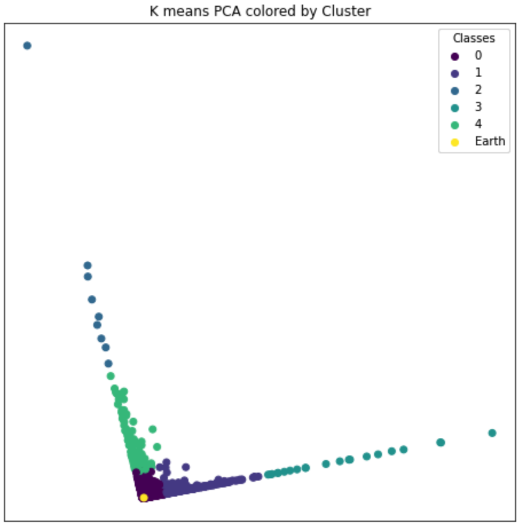

Figure 6: PCA graph colored by cluster

These models are meant to be used in tandem. The first can determine where there are exoplanets, and the second can determine which planets are the most likely to be habitable. However, this method is a preliminary exploration into the data and this concept of determining habitability. Whether the correlations found represent true habitable characteristics has yet to be determined.

6. Conclusion

In this project, TESS Transit data was used to predict whether or not an object of interest was a true exoplanet with a Random Forest model. The likelihood of these exoplanets to be habitable was also predicted based on a k-means clustering algorithm that included the Earth. Machine Learning models like the ones we have come up with can significantly improve the hunt for other life in the universe. This research, however, is far from finished. With further research, higher accuracies and better models can be achieved. We can also better determine the accuracy of our models as more data comes out about these objects of interest. We can especially look towards the future data of the James Webb telescope[2] which will gather the exact data needed to improve our understanding of what a habitable exoplanet could look like. In conclusion, while this research is a great starting point for the benefits of machine learning in this field, there is far more work to be done.

Acknowledgments

We would like to thank the In-Spirit AI scholars program for introducing us to this field and sparking our interest as well as the In-Spirit AI + X program for all of the wonderful resources. We would also like to thank Tomer Arnon for guiding us through this process and all the friends and teachers who helped check programs and facts or just listened to us talk about this project.

Bibliography

1. Basak, S., Mathur, A., Theophilus, A. J., Deshpande, G., & Murthy, J. (2021). Habitability classification of exoplanets: a machine learning insight. *The European Physical Journal Special Topics*, *230*(10), 2221–2251. https://doi.org/10.1140/epjs/s11734-021-00203-z
2. *Exoplanets: Detection, Habitability, Biosignatures – uwastrobiology*. (n.d.). https://depts.washington.edu/astrobio/wordpress/research-areas/exoplanets-detection-habtability-biosignatures/
3. Gould, A. (2016, October 4). *About transits*. NASA. Retrieved October 5, 2022, from https://www.nasa.gov/kepler/overview/abouttransits
4. Kuiper, G. P. (1938). The Magnitude of the Sun, The Stellar Temperature Scale, and Bolometric Corrections. In *Astrophysical Journal* (Vol. 88, pp. 429–471). essay.
5. Lecture 1: Introduction to astronomy 250. (n.d.). Retrieved October 5, 2022, from http://ircamera.as.arizona.edu/astr_250/Lectures/LECTURE_01.HTM#:~:text=The%20total%20range%20of%20right,deg%20%2F%2015%20deg%2Fhr
6. Pratyush, P., & Gangrade, A. (n.d.). *Automation Of Transiting Exoplanet Detection, Identification and Habitability Assessment Using Machine Learning Approaches*. Retrieved September 11, 2022, from https://arxiv.org/pdf/2112.03298.pdf
7. Pham, D., & Kaltenegger, L. (2022). Follow the water: finding water, snow, and clouds on terrestrial exoplanets with photometry and machine learning. *Monthly Notices of the Royal Astronomical Society: Letters*, *513*(1), L72–L77. https://doi.org/10.1093/mnrasl/slac025
8. (2021, November 12). Space Chemistry [Review of *Space Chemistry*]. *Oyla*, 16–21.
9. Sustainable by design :: Declination. (n.d.). Retrieved October 5, 2022, from https://susdesign.com/popups/sunangle/declination.php#:~:text=The%20earth%27s%20equator%20is%20tilted,gives%20rise%20to%20the%20seasons
10. *TESS objects of interest table data columns*. TESS Objects of Interest (TOI) Table Data Column Definitions. (n.d.). Retrieved October 8, 2022, from https://exoplanetarchive.ipac.caltech.edu/docs/API_TOI_columns.html
11. *TESS Project Candidates*. (n.d.). Exoplanetarchive.ipac.caltech.edu. Retrieved September 11, 2022, from https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=TOI
12. The Trustees of Princeton University. (n.d.). Princeton University. Retrieved October 5, 2022, from https://www.astro.princeton.edu/~strauss/FRS113/writeup3/

Our Google Collab Notebook:
https://colab.research.google.com/drive/1cQ2HUtklfbl-f8N-kVqxHIt1-8EVdEQN?usp=sharing