Predicting Earth-Similarity and Habitability of Exoplanets

Ciro Zhang
Riley Gaddis
Sofia Tkachenko
Zheng Lu

1 Introduction

As human development accelerates throughout the 21st century, so do the challenges facing our planet. Climate change, depletion of natural resources, and global tension threaten to make Earth increasingly inhospitable, forcing us to look beyond our world for solutions. One possibility lies in the search for Earth-like exoplanets that not only serve as a potential escape route, but also as a way to better understand the conditions that make a planet habitable. Could there be another world out there that can sustain life, or perhaps already does? In this project, our team wishes to quantify the rarity of Earth-like and habitable exoplanets, as well as to analyze general trends to optimize the search for them. What factors pertaining to exoplanet characteristics and host star characteristics could have strong associations with habitability and similarity to Earth?

The Earth Similarity Index (ESI) will be the primary metric for Earth-likeness. This is a quantitative measure used to compare the similarity of exoplanets and other celestial bodies to Earth. It ranges from 0 (no similarity) to 1 (identical to Earth) and is computed based on key planetary properties such as radius, density, surface temperature, and atmospheric pressure. The Habitable Worlds Catalog, which is a dataset being referenced for this project, computes ESI from the planet's stellar flux, radius, or mass, in response to limited information on planet surface temperature. This is the version of ESI to be used as a reference in this project.

Our main goals for this project are to (1) quantify the relationship between ESI and exoplanet habitability by producing a logistic regression model to predict habitability; (2) identify variables with the highest predictive power for ESI, beyond those used in the current formula, to make ESI prediction more accessible when conducting habitability analysis; (3) forecast ESI values for planets with limited data through regression and direct application of astrophysical results to further optimize computation of ESI. All of these steps unite to answer the question: how can we statistically optimize our ability to classify habitable and Earth-similar exoplanets?

2 Dataset Overview

Our two main sources of data are the NASA Exoplanet Archive and the Habitable Worlds Catalog from the University of Puerto Rico at Arecibo. The NASA Exoplanet Archive contains information on various planetary and stellar attributes for every discovered exoplanet. While this dataset is comprehensive in terms of accounting for every known exoplanet, it does not always contain relevant attribute information and thus has significantly more missing data than is feasible to work with at times. As a result, much of our actual analysis will be conducted with the Habitable Worlds Catalog. The HWC provides two main datasets. The first is a subset of NASA's archive, which contains most of those planets for which we actually have attribute information, instead of majority missing values. The second is a further reduced subset that only contains those planets the HWC has identified to be potentially habitable, whether by conservative or optimistic estimates. In this project, we will not be differentiating between conservative and optimistic predictions for habitability. Both the HWC datasets contain the Earth Similarity Index as an additional feature, which is not captured in the NASA dataset. As explained earlier, the ESI will serve as the primary focus of this research. Missing data was inevitable, and we opted to handle it by dropping it from our analysis. The exception for this was Section 7, which sought to directly process missing data. In this section, some noncritical missing values were imputed via median imputation. The primary datasets used were from the HWC. While we did not extensively use NASA's data in our analyses, it is still crucial to mention it as the main source of the HWC data.

Our main dataset from HWC contains the following attributes

- P_NAME: Planet Name
- P_DETECTION: Detection Method (e.g. transit)
- P_DISCOVERY_FACILITY: Discovery Facility
- P_YEAR: Discovery Year
- P_PERIOD: Orbital period (planet's year)
- P_MASS Planet's Mass, relative to Earth
- P_RADIUS Planet's radius, relative to Earth
- P_FLUX Stellar energy received by planet
- P_TEMP_SURF The surface temperature
- S_TIDAL_LOCK: Tidal locking likelihood
- P_HABZONE_OPT: Fit within habitable zone
- P_ECCENTRICITY: Orbit shape deviation
- S_LUMINOSITY: Star's brightness level
- S_TYPE_TEMP: Star's temperature level
- S_METALLICITY: Star's heavy element content
- P_HABITABLE: If the Planet is Habitable
- P_ESI: Earth Similarity Index

3 Exploratory Data Analysis

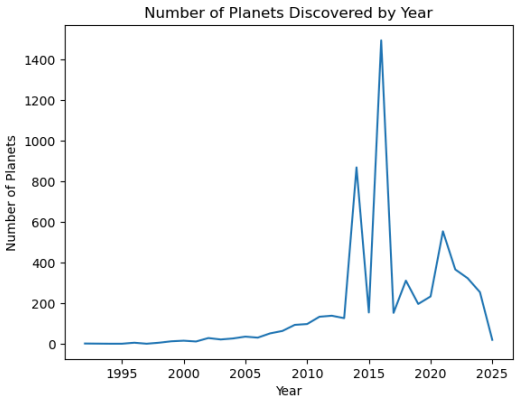


Figure 3.1: Number of Discovered Planets Across Years

Figure 3.1 shows the trend of exoplanet discovery by year. The spike from 2013 to 2018 is due in part to the contributions of the NASA Kepler telescope, which operated directly in space to discover over 28,000 exoplanets in its lifetime. Since its retirement in 2018, discovery has since dropped off but remains on an increasing trend since the early 2000s.

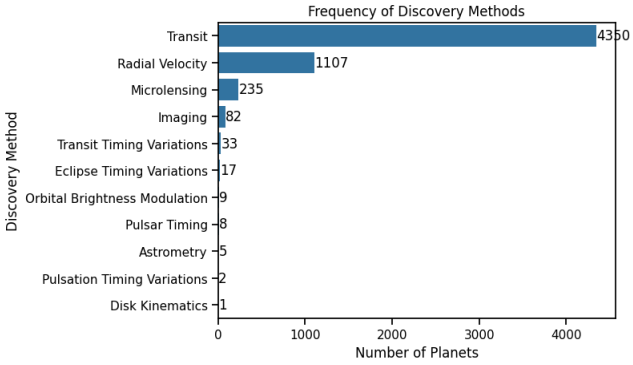


Figure 3.2: Distribution of Discovery Methods

The above figure summarizes which methods are most common in discovering exoplanets. The transit method is the most common way to discover exoplanets, and it works by detecting slight dimming of a star as a planet passes in front of it. The radial velocity method, though less common, tracks a star's "wobble" caused by orbiting planets, measuring shifts in its light wavelengths. This technique is particularly effective for detecting massive, close-orbiting planets like "hot Jupiters," making this method less ideal for finding Earth-like planets.

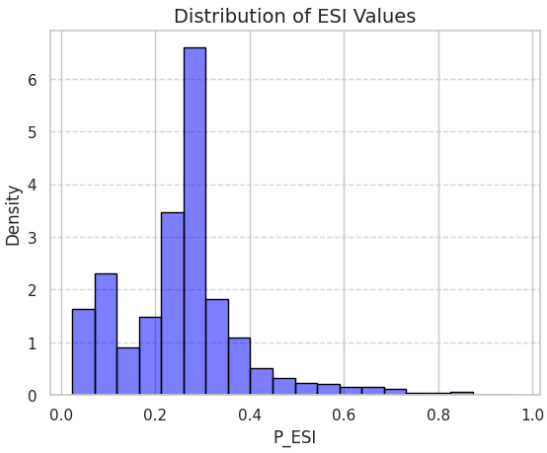


Figure 3.3: Distribution of P_ESI

Figure 3.3. indicates that the ESI rating distribution is highly right-skewed, with most planets receiving an ESI score at around 0.3. This indicates that planets similar to Earth, as defined by the Habitable Worlds Catalog (ESI > 0.8), tend to be extremely rare.

3.1 P_ESI to Single Feature Correlations Analysis

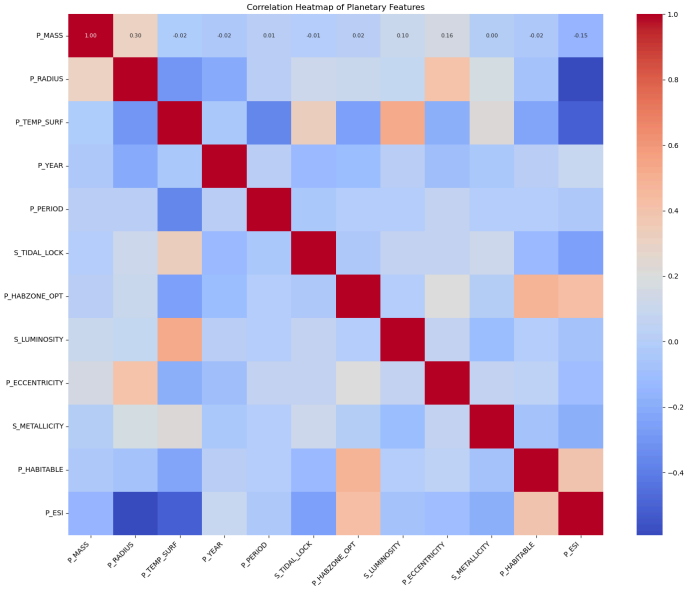


Figure 3.4: Heat Map of Every Numerical Feature vs P_ESI

From Figure 3.4, we can observe that features such as P_RADIUS and P_TEMP_SURF have a strong negative correlation to P_ESI. This is expected since the planet's radius and surface temperatures are often used in calculating planet ESI.

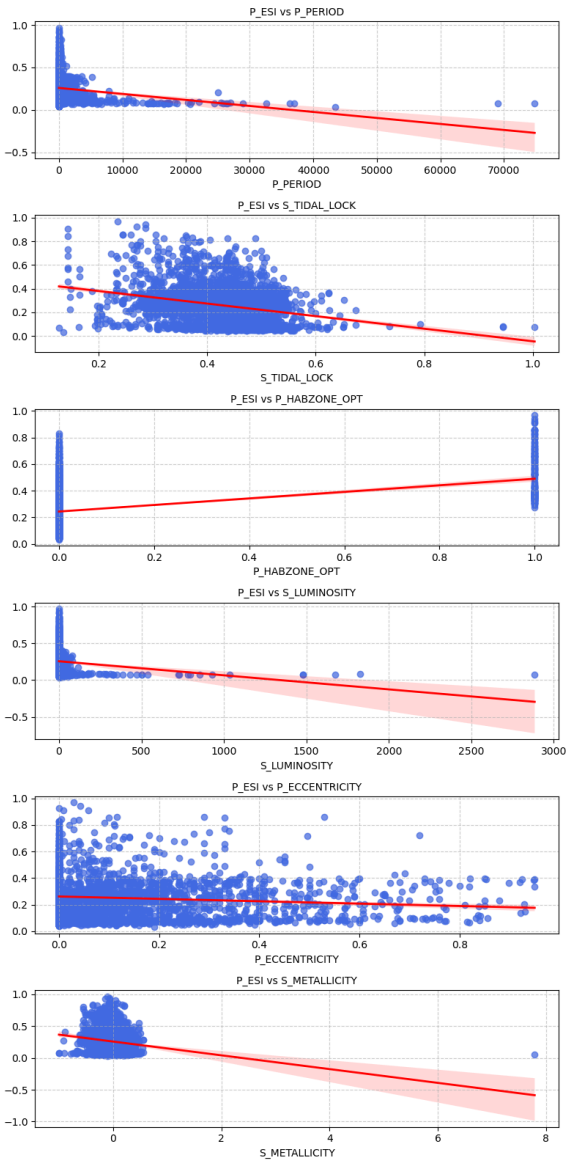


Figure 3.5: Scatter_plot of Every Numerical Feature vs P_ESI

Figure 3.5 provides scatter plots of various numerical features against P_ESI, allowing us to examine their individual relationships. When excluding surface temperature from our features, we observe that the remaining features exhibit weaker predictive power. Some features, such as P_PERIOD and P_ECCENTRICITY, show mild trends, but their ability to independently predict P_ESI is limited. Several features, such as star metallicity, star luminosity, and tidal lock have outliers that influence the data considerably.

While no single feature strongly predicts P_ESI on its own, a combination of multiple features might still capture useful patterns. Further analysis using feature selection techniques, such as Principal Component Analysis (PCA) could help improve the results.

4 Clustering

In this section, we employ clustering techniques to analyze how different features influence P_ESI as a group. Since we know that individual features are correlated with ESI but tend to be weaker. In addition, since we know that mass, radius, and flux contribute directly to ESI calculations, we will exclude these features and perform clustering to assess the impact of the remaining variables.

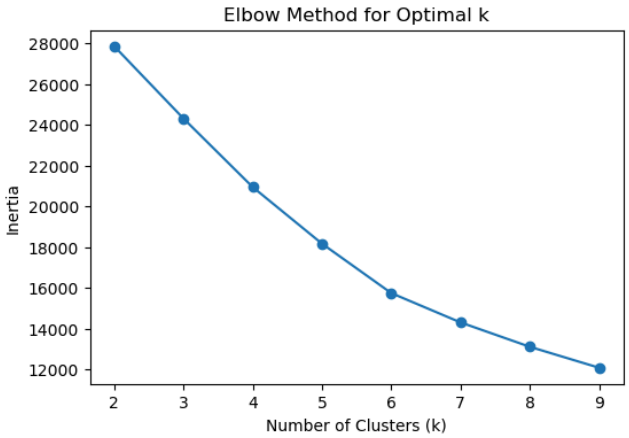


Figure 4.1: Line plot of Inertia vs Number of Clusters

Using the elbow method, we determine the optimal number of clusters by plotting inertia against the number of clusters. As shown in Figure 4.1, the elbow point occurs at k = 4, suggesting that four clusters provide a balance between compactness and variance.

4.1 P_ESI to Multi-Feature Correlations Analysis

Cluster	P_YEAR	P_PERIOD	S_TIDAL_LOCK	P_HABZONE_OPT	S_LUMINOSITY	P_ECCENTRICITY	S_METALLICITY	P_HABITABLE	P_ESI
2	2013.366379	626.502039	0.428267	1.0	2.088940	0.210386	0.026922	0.362069	0.489312
0	2016.037332	93.976181	0.432519	0.0	3.597534	0.028788	0.013037	0.000000	0.250859
1	2014.248244	3489.369435	0.456514	0.0	12.768695	0.415369	0.093208	0.000000	0.172303
3	2016.166667	680.580799	0.569965	0.0	1265.665406	0.155167	-0.395833	0.000000	0.076204

Figure 4.2: Mean of every feature in each cluster

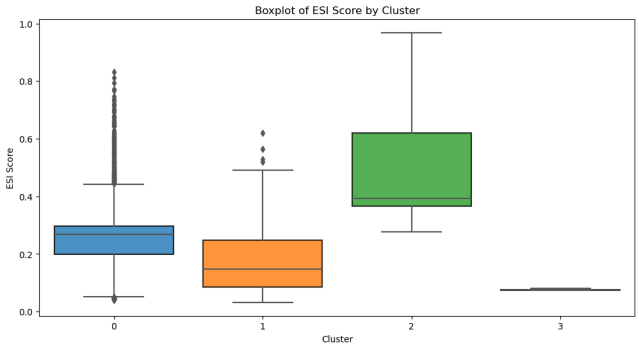


Figure 4.3: Distribution of ESI in each cluster

After determining k, we apply clustering to our dataset and visualize the results. The distributions of feature values across different clusters (shown in figure 4.2 and 4.3) suggest that the selected features together still contribute meaningfully to the clustering structure. If these features had no relationship with P_ESI, we would expect more random or overlapping cluster distributions.

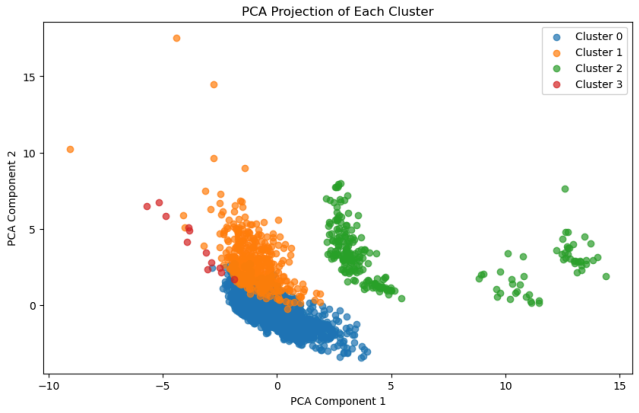


Figure 4.4: Scatter plot of the PCA projection of each cluster

The PCA projection in Figure 4.4 visually reinforces the separation between clusters. Notably, Cluster 2 (green) stands out, with its planets positioned far from the others—likely due to a significantly higher mean P_ESI of 0.489 (Figure 4.2). This suggests that though we did not include mass, radius, flux, and surface temperature in our input features, other features when put together still had the distinct characteristics to separate the planets with high ESI values apart from the rest.

4.2 Quantifying Rarity of Earth like Planets

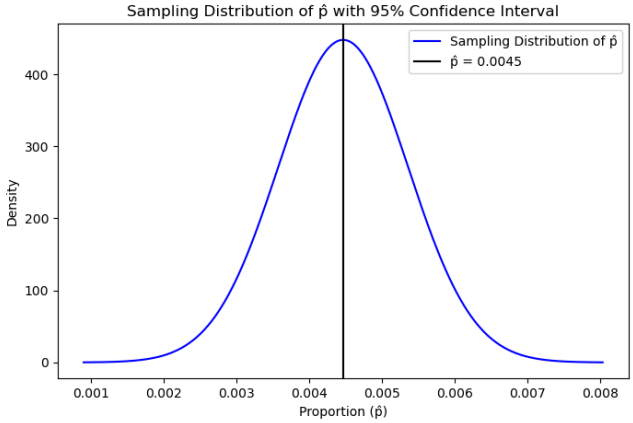


Figure 4.5: 95% confidence interval of Earth-like planets

Our clustering results show a group with relatively high ESI scores, but none stand out as truly Earth-like. The largest average ESI of any cluster is only about 0.489. This indicates how rare a planet similar to Earth is. Only 0.45% of the planets in our dataset have an ESI $\geq$ 0.8, and a 95% confidence interval of [0.00271, 0.00621] (Figure 4.5) supports this scarcity.

5 Habitability and ESI

In addition to ESI scores, the Habitable Worlds Catalog (HWC) provides data on 70 (possibly) habitable exoplanets. According to the Planetary Habitability Laboratory (PHL), ESI scores are independent of habitability criteria. Hence, the focus of this section will be to differentiate habitable exoplanets from those that have higher ESI but are not habitable. In this section, we will shorten the latter description and refer to such planets as "promising" (which is not to be equated with "similar to Earth").

The general threshold for habitable exoplanets is ESI > 0.5, as suggested by the PHL. However, there are some exceptions, with one habitable planet demonstrating an ESI of around 0.47. The threshold used in this section will be determined by this minimum value. Planets whose ESI >= 0.47 will be considered "promising" exoplanets per our definition.

The justification for this threshold is to maintain a consistent cutoff for comparison of habitable to non-habitable planets. A major drawback to the data used in this project is its low number of habitable planets relative to all discovered exoplanets. By employing a cutoff equal to the smallest ESI value of all habitable planets, we ensure that the planets studied here are, to some degree, generally comparable to each other. This allows us to make more meaningful, relative comparisons. This also helps to ensure that statistical significance does not occur

purely as a result of there being a disproportionate number of non-habitable planets in the entire dataset.

One major goal of our research was to quantify the rarity of Earth-like exoplanets. We can extend this goal to habitable exoplanets, and in particular, the rarity of habitable exoplanets amongst planets with higher ESI values. Using the data from HWC, we found that there were 296 planets whose ESI scores were at least 0.47. Of those 296 planets, exactly 70 were predicted to be habitable by the HWC. The observed proportion of habitable exoplanets out of the sample of promising exoplanets is thus around 0.2365.

It is reasonable to assume that exoplanet observations are independent. Also note that 70 >= 10 and (296-70) >= 10, which indicates that our sample is sufficiently large. The conditions to produce a 95% CLT-based confidence interval for the true proportion of habitable exoplanets out of all promising exoplanets are satisfied. The procedure yields a confidence interval as follows:

[0.18807886, 0.28489411]

With 95% confidence, we approximate around 19% to 28% of promising exoplanets (ESI >= 0.47) could be potentially habitable, as defined by the HWC.

This indicates that habitable planets are moderately rare, even among planets with already higher ESI values. Hence there is reason to suspect that ESI score is not the sole predictor of planet habitability. We can further compare habitable and non-habitable exoplanets with ESI >= 0.47 by exploring their numerical attribute averages.

is_habitable	P_ESI	P_MASS	P_RADIUS	P_FLUX
False	0.575213	7.216232	2.363757	2.721243
True	0.738725	5.288748	1.770514	0.866255

Figure 5.1: Summary of key ESI attributes v.s. habitability.

Summarized in the table above, we can see that habitable exoplanets have 0.16 more in ESI score, 2 less in mass, 0.6 less in radius, and 1.9 less in flux, compared to non-habitable exoplanets, on average. We chose these variables because the latter three are used in computing HWC's ESI, and we believe that individual analysis could yield more nuanced insight into which aspects of ESI are most significant to predict habitability. Several two-sample *t*-tests and nonparametric tests were performed to verify if these differences were statistically significant. For all tests, the chosen significance level was α=0.01.

5.1 Planet ESI

One way to distinguish habitable from non-habitable promising exoplanets is to compare ESI values between the two groups. The ESI distributions by group are summarized below.

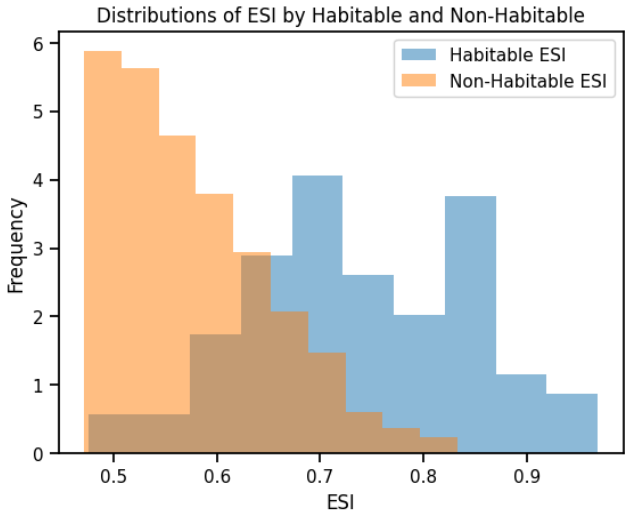


Figure 5.2: ESI Distribution by Habitability Status.

The histogram conveys several observations. The distribution of non-habitable ESI is strongly right-skewed while the distribution of habitable ESI is comparably more normal. Consequently, a nonparametric test is preferable to determine if ESI scores differ between groups. A one-sided two-sample Kolmogorov-Smirnov test was performed on the two samples, which we can reasonably assume are independent of each other. The ECDF plots are shown.

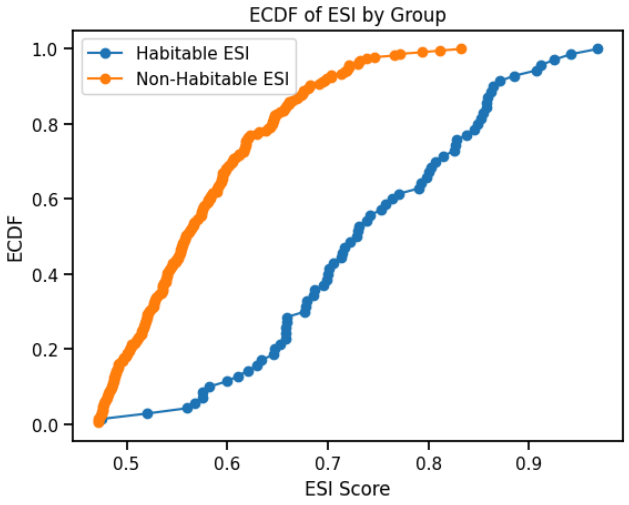


Figure 5.3: ECDFs of ESI by Habitability Status.

The documentation for the two-sample K-S test in SciPy elucidates how to adapt the K-S test to a one-sided alternative hypothesis. Using this reference and the plots above, we propose the hypotheses as follows:

- **H_0:** The ECDF of ESI for habitable planets is greater than or equal to the ECDF of ESI for non-habitable planets.
- **H_a:** The ECDF of ESI for habitable planets is less than the ECDF of ESI for non habitable planets for at least one observation.

The *p*-value of the test was around 1.7e-21, which is highly statistically significant against the threshold α=0.01. This implies that ESI values are almost consistently higher for habitable exoplanets compared to non-habitable exoplanets.

This test shows that ESI could be a strong predictor of habitability, even though ESI and habitability are functionally independent from each other. Hence, there is motivation to include ESI in a logistic regression model to predict habitability.

5.2 Planet Mass

The ESI used in this project is computed from some combination of three factors: radius, mass, and stellar flux. We can further assess statistical significance of these attributes directly. The distributions of mass are right-skewed for both habitable and non-habitable exoplanets. Applying an identical log transformation (base 2.3) to both distributions yields roughly normal data, according to the following histograms and QQ-plots.

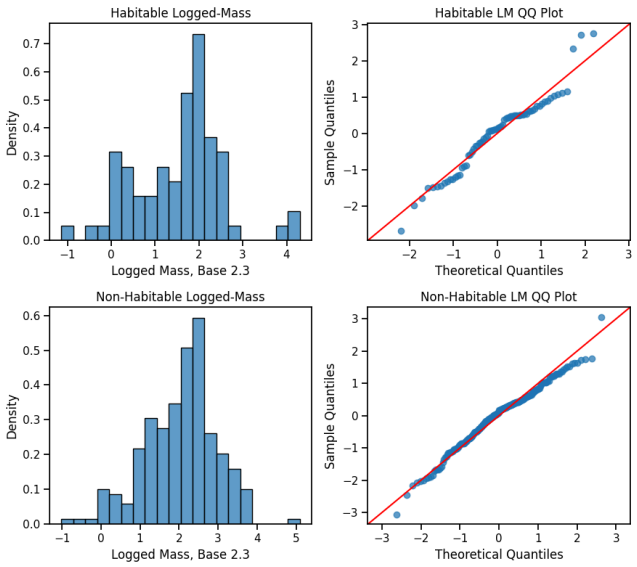


Figure 5.4: Histograms and QQ-plots of logged-mass, base 2.3.

Since the number of habitable exoplanets is limited, the assumptions of normality may be somewhat uncertain. However, the data appears to be normal enough with the transformation. We will note that results should be interpreted with caution and proceed with the test. The hypotheses are as follows:

- **H_0:** Habitable and non-habitable logged-masses are equal, on average.
- **H_a:** Non-habitable logged-masses are larger than habitable logged-masses, on average.

A Welch's two-sample *t*-test was conducted using the SciPy Stats module, setting `equal_var` to False and `alternative` to "greater" in accordance with the one-sided hypothesis test. The `ttest_ind` function produced a *p*-value of 2.1e-4, which is highly statistically significant under the chosen threshold α=0.01. There is evidence that promising but non-habitable exoplanets have a higher logged-mass than habitable exoplanets, on average. Note that interpretability of results is limited to the log-scale, as these are the values that satisfy the conditions of the *t*-test. As a result, significance in the original scale may not occur.

5.3 Planet Radius

A similar analysis as in Section 5.1 may be made for planet radii, which is one of the measures used to compute ESI in the HWC. As with mass, the radii distributions for both habitable and non-habitable exoplanets are right-skewed, requiring a log-transformation to enact a two-sample *t*-test to compare means. Applying an identical log transformation (base 1.5) to both distributions yields roughly normal data, according to the following histograms and QQ-plots.

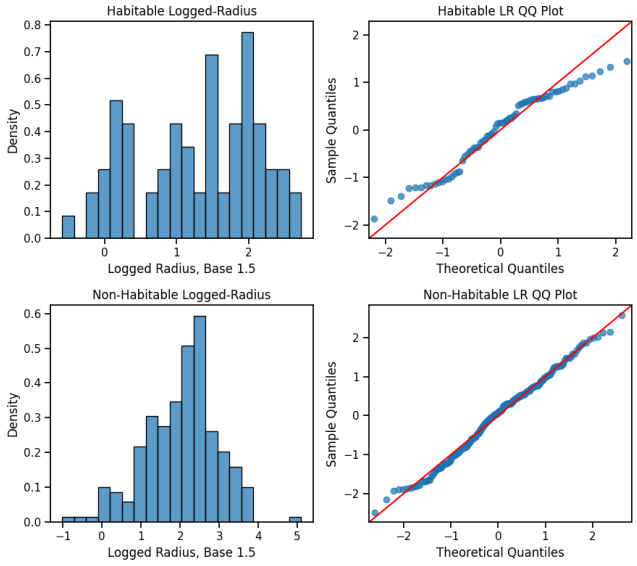


Figure 5.5: Histograms and QQ-plots of logged-radius, base 1.5.

Normality for the logged-radius of non-habitable exoplanets appears to be satisfied. However, we are limited by the small number of habitable exoplanets. It is not entirely clear if the logged-radius of habitable exoplanets can follow a normal

distribution, at least when an identical transformation is used. It is necessary to obtain more samples in order to see more of the true distribution. As a result, a nonparametric test may be preferable given the constraints of our data. As with ESI values, a one-sided two-sample K-S will be performed to compare distributions of planet radii by habitability status.

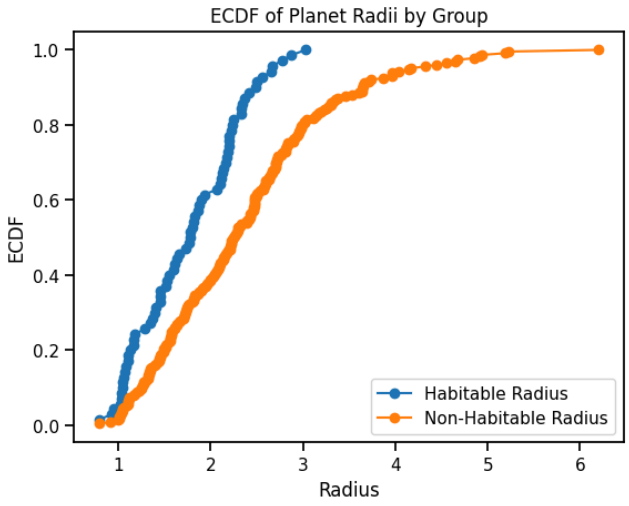


Figure 5.6: ECDFs of Radii by Habitability Status.

Using the plots above, we propose the hypotheses as follows:

- **H_0:** The ECDF of radii for habitable planets is less than or equal to the ECDF of radii for non-habitable planets.
- **H_a:** The ECDF of radii for habitable planets is greater than the ECDF of radii for non habitable planets for at least one observation.

The *p*-value of the test was around 3.3e-6, which is highly statistically significant against the threshold α=0.01. This suggests that radii of habitable exoplanets are generally smaller than radii for non-habitable, promising exoplanets. Such a result is unsurprising, as the largest habitable planet radius is around half the size of the largest radius among all promising exoplanets.

5.4 Stellar Flux

Stellar flux is roughly characterized by the amount of energy received by a planet from its host star. It is also a key value used in computing ESI, according to HWC.

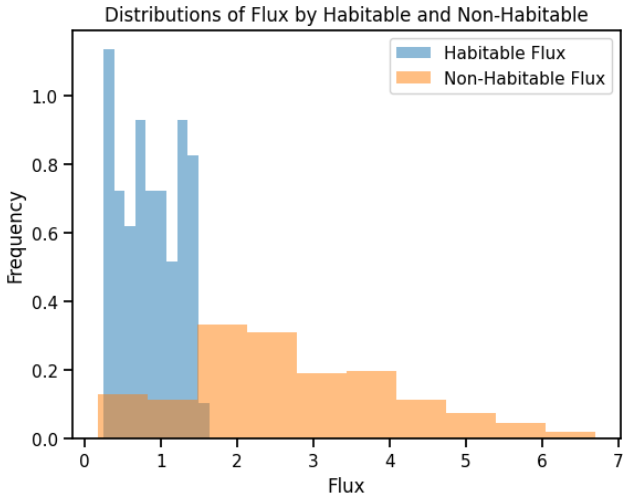


Figure 5.7: Flux Distribution by Habitability Status.

There is immediate suspicion that habitable flux is generally lower than non-habitable flux. A one-sided two-sample K-S test was performed to statistically confirm this.

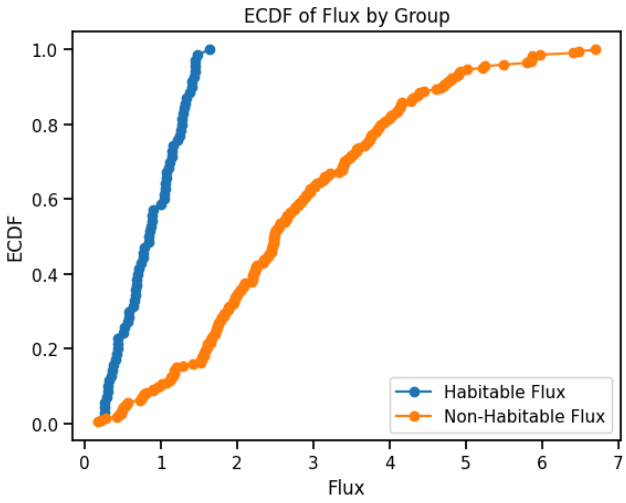


Figure 5.8: ECDFs of Flux by Habitability Status.

Using the plots above, we propose the hypotheses as follows:

- **H_0:** The ECDF of flux for habitable planets is less than or equal to the ECDF of flux for non-habitable planets.
- **H_a:** The ECDF of flux for habitable planets is greater than the ECDF of flux for non habitable planets for at least one observation.

The *p*-value of the test is 1.5e-38, which is highly statistically significant against the threshold α=0.01. This suggests that flux of habitable exoplanets is generally smaller than flux for non-habitable, promising exoplanets. Again, such a result is

unsurprising, given that the largest flux of a habitable planet is considerably smaller than the largest flux of all given promising exoplanets.

5.5 Star Type, Temperature

While Sections 5.2 to 5.4 cover quantities directly related to ESI, there may be additional factors needed to classify a planet as actually being habitable. This section explores whether exoplanet habitability is independent of host star temperature type, furthering the comparison of habitable exoplanets against non-habitable exoplanets. While a planet may be similar to Earth in terms of radius, mass, or flux, the temperature of the star it orbits can completely negate any chance of habitability. This pairing is intended to provide nuance not directly sourced from ESI, which could result in better model performance.

S_TYPE_TEMP	F	G	K	M
is_habitable				
False	2	46	71	107
True	0	4	14	52

Figure 5.9: Contingency table by Star Temperature Type.

The contingency table above summarizes the total number of exoplanets of each habitability status and star-temperature pairing. Notably, there are very few promising planets that orbit a star of temperature type *F*. This temperature type is the hottest of all those visualized in the table. By comparison, Earth's Sun is of type *G*, which is one classification level below *F*. It appears that most habitable planets orbit stars of type *G* or cooler.

The assumptions of the chi-square test of independence require that expected counts are sufficiently large (>= 5) in most cells. Due to how few planets orbit stars of type *F*, it is preferable to drop this column in the contingency table before proceeding with the test. After the drop is performed, the chi-square test for independence yields a *p*-value of 2.8e-4, which is highly statistically significant relative to the threshold α=0.01. This implies that star temperature and habitability are not independent in the context of promising exoplanets.

5.6 Predicting Habitability

Among planets that are already deemed promising (ESI >= 0.47), what additional factors determine habitability? In this subsection, a process fitting several logistic regressions will be followed, and an analysis of these models will be made to understand which factors are most predictive of habitability.

5.6.1 Baseline

The baseline model solely relies on ESI as the predictor for habitability. Since ESI is independent of habitability criteria, there is justification in using ESI to predict habitability.

Logit Regression Results

Dep. Variable:	is_habitable	**No. Observations:**	296
Model:	Logit	**Df Residuals:**	294
Method:	MLE	**Df Model:**	1
Date:	Sun, 16 Mar 2025	**Pseudo R-squ.:**	0.3815
Time:	12:25:31	**Log-Likelihood:**	-100.15
converged:	True	**LL-Null:**	-161.91
Covariance Type:	nonrobust	**LLR p-value:**	1.068e-28

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-12.3480	1.422	-8.681	0.000	-15.136	-9.560
P_ESI	17.2533	2.125	8.119	0.000	13.088	21.418

Figure 5.10: Results of Baseline Logistic Regression.

The results of the model are shown. The ROC curve is pictured below, with an AUC score of around 0.88.

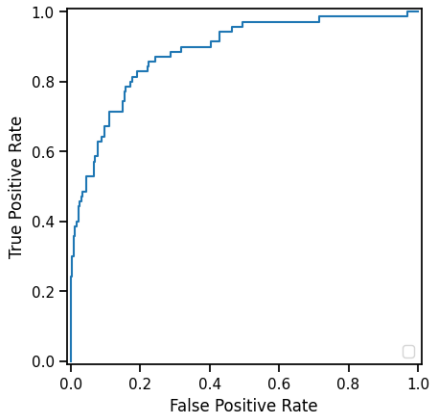


Figure 5.11: ROC Curve of Baseline Model.

The high AUC score indicates that ESI already has strong predictive power for habitability. The results suggest that ESI alone may have success in predicting habitability. Any subsequent models will be evaluated in reference to this baseline model, as it is already quite powerful.

5.6.2 Planet Radius, Mass, Flux

The following logistic regression model will assess whether a "dissected" version of ESI is more effective in predicting habitability. We've shown that radius, mass, and flux may differ significantly between habitable and non-habitable promising exoplanets. Since ESI is already a powerful predictor of habitability, will extracting the values used to compute ESI explain more variance in the data? To answer this question, we fit

a new logistic regression model to predict habitability solely from planet radius, mass, and flux.

Logit Regression Results

Dep. Variable:	is_habitable	**No. Observations:**	296
Model:	Logit	**Df Residuals:**	292
Method:	MLE	**Df Model:**	3
Date:	Mon, 17 Mar 2025	**Pseudo R-squ.:**	0.6468
Time:	00:22:44	**Log-Likelihood:**	-57.192
converged:	True	**LL-Null:**	-161.91
Covariance Type:	nonrobust	**LLR p-value:**	3.851e-45

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	8.9077	1.379	6.459	0.000	6.205	11.611
P_MASS	0.0025	0.060	0.042	0.967	-0.115	0.120
P_RADIUS	-2.2753	0.448	-5.079	0.000	-3.153	-1.397
P_FLUX	-3.4844	0.531	-6.558	0.000	-4.526	-2.443

Figure 5.12: Results of Second Logistic Regression.

The results of the model are shown. The ROC curve is pictured below, with an AUC score of around 0.972.

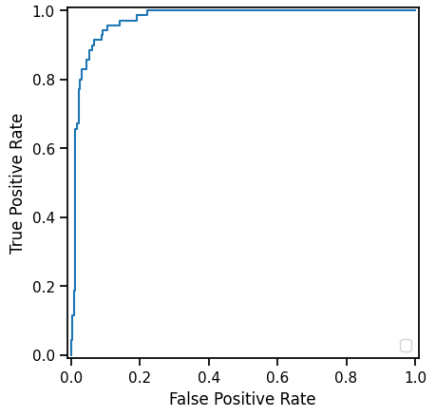


Figure 5.13: ROC Curve of Second Model.

The improvement of the second model was quite significant compared to the first attempt, suggesting outstanding ability to discriminate between habitable and non-habitable (promising) exoplanets. However, as is evident from the summary results, mass was insignificant as a predictor. After checking VIF scores, we concluded that none were exceptionally large enough to suggest multicollinearity:

- VIF: P_MASS: 1.550
- VIF: P_RADIUS: 1.632
- VIF: P_FLUX: 1.084

However, we decided to remove mass anyway so that it would not contribute noise to the model. The resulting model did not suffer from any major differences as a result. The model in Figure 5.14 demonstrates the same pseudo-R-squared and log-likelihood value as the model in Figure 5.12.

Logit Regression Results

Dep. Variable:	is_habitable	**No. Observations:**	296
Model:	Logit	**Df Residuals:**	293
Method:	MLE	**Df Model:**	2
Date:	Mon, 17 Mar 2025	**Pseudo R-squ.:**	0.6468
Time:	00:35:34	**Log-Likelihood:**	-57.192
converged:	True	**LL-Null:**	-161.91
Covariance Type:	nonrobust	**LLR p-value:**	3.322e-46

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	8.9071	1.379	6.459	0.000	6.204	11.610
P_RADIUS	-2.2667	0.398	-5.697	0.000	-3.047	-1.487
P_FLUX	-3.4859	0.530	-6.574	0.000	-4.525	-2.447

Figure 5.14: Results of No-Mass Model.

5.6.3 Planet Radius, Flux, Star Type

The final model we considered combined planet radius, flux, and star type as a categorical variable to predict habitability. Due to low counts of the *F* star type, the model failed to converge without data manipulation. Because of this, we decided to merge *F* (yellow-white) and *G* (yellow) star type counts together to avoid such issues. While not ideal, we believe these categories are similar enough to combine. Star types *F* and *G* are adjacent in terms of the ranges of temperature they represent, so from an ordinal perspective, *F* and *G* are adjacent categories to one another. The results of the model are summarized below.

Logit Regression Results

Dep. Variable:	is_habitable	**No. Observations:**	296
Model:	Logit	**Df Residuals:**	291
Method:	MLE	**Df Model:**	4
Date:	Mon, 17 Mar 2025	**Pseudo R-squ.:**	0.6589
Time:	00:53:08	**Log-Likelihood:**	-55.222
converged:	True	**LL-Null:**	-161.91
Covariance Type:	nonrobust	**LLR p-value:**	4.987e-45

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	7.6673	1.611	4.759	0.000	4.509	10.825
S_TYPE_TEMP[T.K]	0.1638	0.799	0.205	0.838	-1.402	1.730
S_TYPE_TEMP[T.M]	1.1665	0.797	1.464	0.143	-0.396	2.729
P_RADIUS	-1.9811	0.407	-4.868	0.000	-2.779	-1.183
P_FLUX	-3.5152	0.546	-6.438	0.000	-4.585	-2.445

Figure 5.15: Results of Star Type Model.

The ROC curve is pictured below, with an AUC score of 0.973.

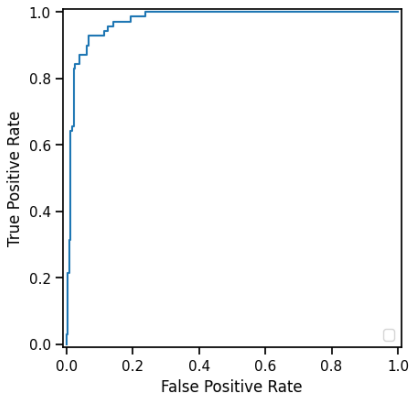


Figure 5.16: ROC Curve of Star Type Model.

There was a very slight improvement in AUC score upon inclusion of star type in the model. However, most of the star type coefficients are insignificant under the α=0.01 threshold. In order to balance model complexity with model fit, we decided to drop star type from the model altogether. It is possible that flux already accounts for stellar properties, including the temperature of a star, hence making the inclusion of star type redundant.

5.6.4 Final Model

To the extent of our analysis, the final model we propose is the model summarized in Figure 5.14, which we shall characterize as the "deconstructed-ESI" model. The only two predictors used are planet radius and flux. It appears that planet radius and flux are strong predictors of planet habitability. We can observe the confusion matrix for this model, using a threshold of 0.5 to make predictions.

	Predicted: 0	Predicted: 1
Actual: 0	217	9
Actual 1:	12	58

The precision is 0.8657, the recall is 0.8286, and the F1-score is 0.8467. It is clear that the model is not exploiting class imbalance of habitable and non-habitable exoplanets, since all scores are generally quite high. The model does suffer slightly in identifying habitable planets, meaning that it has a slightly higher False Negative rate. For this reason, there are improvements that could be made on this model, perhaps in future implementations of this project. There may be other predictors that we did not consider that might have even more predictive power than what we have currently included.

5.7 Takeaways

We conclude this section by commenting on the relationship between habitability and ESI, as well as highlighting some limitations of our findings. Even though habitability criteria are independent of ESI, as emphasized by the HWC, the process utilized in Section 5 has shown that there exists a strong relationship between ESI, variables used to compute ESI (such as radius and flux), and habitability status. Therefore—at least within the context of planets that are already of higher ESI value—habitability can more or less be accurately predicted from a small subset of variables. However, it is crucial to note that our analysis is restricted to those planets whose ESI is at least 0.47. If we had considered all planets, there would likely be significantly more nuance and variance in numerical attributes, as well as outliers and more significant skew. Maintaining an ESI threshold allows for convenient and comparable analysis, but if habitable planets are to be understood in reference to all other exoplanets, a more intensive analysis will be required beyond what has been done here.

6 Alternative Predictors for ESI

Due to the nature of astrophysical data, information on radius, mass, and flux may not always be available. This would make it difficult to perform habitability analysis as outlined in Section 5. In this section, we analyzed which other variables could be used to predict ESI. To avoid using physical parameters (such as mass, radius, flux, and surface temperature) that directly contribute to the ESI calculation, this analysis selected features related to planetary environment and stellar properties that do not directly determine the ESI. These features include:

- **P_PERIOD**: Planetary orbital period
- **S_TIDAL_LOCK**: Tidal locking status
- **P_HABZONE_OPT**: Indicator of being in the optimal habitable zone
- **S_LUMINOSITY**: Stellar luminosity
- **P_ECCENTRICITY**: Orbital eccentricity
- **S_METALLICITY**: Stellar metallicity
- Along with one-hot encoded variables for **P_DETECTION**, **P_TYPE_TEMP**, and **P_DISCOVERY_FACILITY**.

After removing missing values and applying the necessary encoding, the dataset was split into an 80/20 training/testing set. A linear regression model was then built to predict the continuous P_ESI values.

Analysis Results

- **Mean Squared Error (MSE):** 0.007503
- **Coefficient of Determination (R^2):** 0.607353

This indicates that the current model is able to explain approximately 60.7% of the variance in the ESI, reflecting a moderate level of explanatory power for these environmental and stellar features.

Example Regression Coefficients :

- **P_PERIOD:** 0.026232
- **S_TIDAL_LOCK:** -0.207784
- **P_HABZONE_OPT:** +0.174986
- **S_LUMINOSITY:** -0.063514
- **P_ECCENTRICITY:** -0.043872
- **S_METALLICITY:** -0.149656

To further evaluate the suitability of the linear regression model, we examined the residuals. Figure 6.2 shows the Residuals vs. Predicted Values plot, while Figure 6.1 displays the QQ Plot of Residuals.

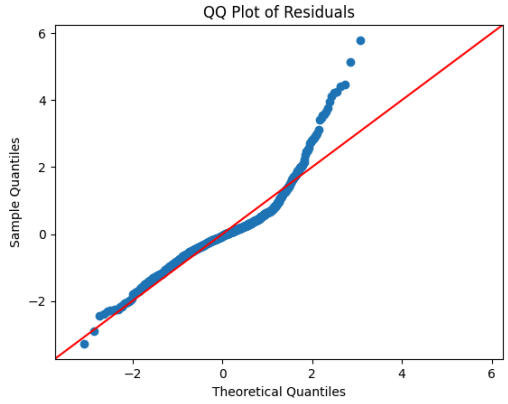


Figure 6.1: QQ plot of model Residuals

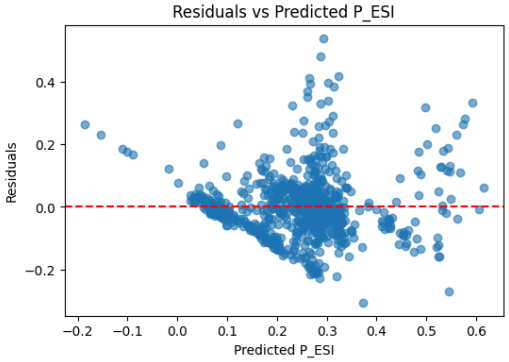


Figure 6.2: Scatter plot of Residual vs Prediction

The coefficients for the different detection methods, temperature types, and discovery facility variables vary, reflecting the statistical relationships between these observation methods and the ESI.

6.1 Residual and QQ Plot Analysis:

The residual plot (Figure 6.2) shows that most residuals from the predicted values are clustered around 0, indicating minimal overall bias; however, there are some localized deviations suggesting that a portion of the data is not fully captured by the current model.

6.2 Conclusion (Linear Regression):

The current linear regression results indicate that the selected environmental and stellar features—along with a log transformation to reduce skewness—have a greater explanatory power for the ESI than before. Although the model only explains about 60.7% of the variance, the results reflect the statistical association between these factors and the ESI. Negative coefficients (such as for S_TIDAL_LOCK, S_LUMINOSITY, P_ECCENTRICITY, and S_METALLICITY) suggest that planets with longer orbital periods and those in the optimal habitable zone tend to have higher ESI values, while higher stellar luminosity, greater orbital eccentricity, and higher stellar metallicity are associated with lower ESI values. Additionally, the coefficients for various detection methods indicate differing impacts on the ESI. Please note that the same feature set was used for both regression and classification; the regression model achieved a decent R^2 for these variables, suggesting they have potential in predicting Earth similarity, which we further evaluated in our classification analysis.

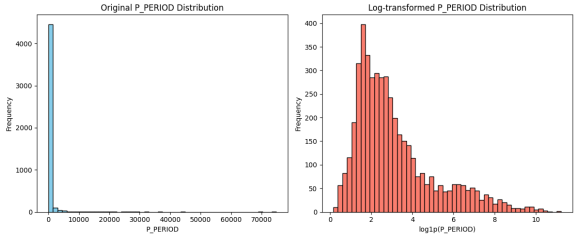


Figure 6.3: Histogram of P_PERIOD and log P_PERIOD

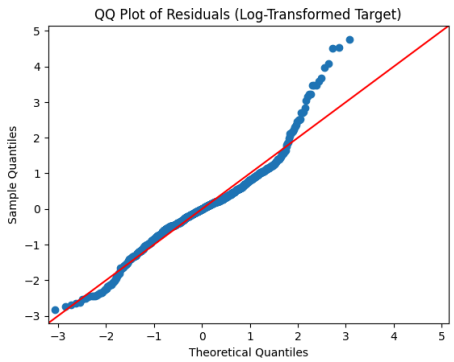


Figure 6.4: QQ plot of model Residuals

Analysis Results

- **Mean Squared Error (MSE):** 0.004151
- **Coefficient of Determination (R^2):** 0.631831

The QQ plot shows that while the residuals have improved in normality (compared to before the log transformation), there is still a noticeable tail at the upper end. However, it more closely follows a normal distribution than prior to the transformation.

Although the QQ plot shows that the residuals do not strictly follow the diagonal line, indicating some deviation from normality, this is a common phenomenon in complex astrophysical data, primarily due to inherent non-linear relationships and measurement complexities in the data. The current model demonstrates strong explanatory power ($R^2 \approx 0.63$), with key features (e.g., the negative effect of tidal locking) aligning with established scientific understanding. Therefore, minor deviations from normality do not diminish the model's practical value in planetary habitability research.

Binary Classification Analysis (ESI > 0.8 as Earth-like)

Do the variables used in the previous linear regression successfully predict whether a planet is similar to Earth? That is, can we successfully predict Earth-similarity without relying on variables typically used to compute ESI? In this subsection, we produced a logistic regression model to answer this question. Planets with a P_ESI greater than 0.8 are considered "Earth-like" in this analysis. A binary classification model was constructed to determine whether a planet has high Earth similarity. During data processing, missing values were removed and categorical variables were one-hot encoded. We also used SMOTE to address class imbalance, then built a logistic regression model using an 80/20 training/testing split. The model accuracy was 99.52%.

6.3 Confusion Matrix:

	Predicted: 0	Predicted: 1
Actual: 0	929	10
Actual: 1	16	913

Figure 6.5: Confusion matrix of model predictions

For both class 0 (ESI $\leq$ 0.8) and class 1 (ESI > 0.8), Precision, Recall, and F1-Score are above 0.98.

6.4 Supplement on SMOTE:

SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the dataset by oversampling the minority class (Earth-like planets in this case). This method synthesizes new examples for the minority class rather than simply duplicating existing ones. By doing so, SMOTE helps ensure that the classifier receives a balanced training set, which in turn improves its ability to distinguish between Earth-like and non-Earth-like planets. In our analysis, using SMOTE led to a robust model with high classification accuracy, even though the original dataset had an imbalanced distribution.

6.5 Conclusion (Binary Classification):

The binary classification results show that, given the selected features and data processing approach (including SMOTE for balancing), the model effectively distinguishes between high ESI (Earth-like) and low ESI planets. The high accuracy (~98.61%) and robust confusion matrix indicate that there are distinct feature differences in the current dataset that allow for effective differentiation between the two categories.

Based on the current analysis results, we conclude that (1) after applying log transformations to certain skewed variables, the model explains approximately **63.2%** of the variance in P_ESI, which is an improvement over previous runs without the transformation. (2) Negative coefficients indicate that higher tidal lock values, higher stellar luminosity, greater orbital eccentricity, and higher stellar metallicity are associated with lower ESI, while longer orbital periods and an optimal habitable zone indicator correlate with higher ESI. (3) The residual and QQ plots show a partial improvement in normality and variance, suggesting the log transformation helped, although some non-linearity remains.

6.6 Limitations

1. Data Quality and Missing Values:

Despite cleaning, some features may still have underlying measurement errors or biases that could affect the model's performance.

2. Model Assumptions:

Although log transformations improved linearity, residual analysis indicates that some non-linearity and heteroscedasticity may persist. Thus it is not recommended to use this as a final model until all assumptions have been satisfied.

3. Feature Selection:

The analysis intentionally excluded physical parameters such as planetary mass, radius, and surface temperature (which directly determine ESI). However, this limits the amount of information available to the model, and other relevant features (e.g., stellar radius, effective temperature) might further improve predictions.

6.7 Relation/Changes Since Project Proposal

Since the proposal, we've focused on environmental and stellar features instead of physical parameters that directly determine the ESI. We applied a log transformation (clipping negatives to zero) to selected predictors (P_PERIOD, S_LUMINOSITY, P_ECCENTRICITY, S_METALLICITY), improving R^2 from ~54.6% to ~60.7%. We also used SMOTE to balance the classes for better Earth-like planet detection and added residual analysis, QQ plots, and distribution visualizations to further assess model performance.

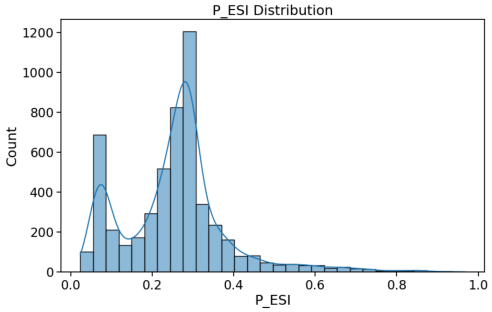
7 Estimating Unknown ESI Values Using a Hybrid Model Approach

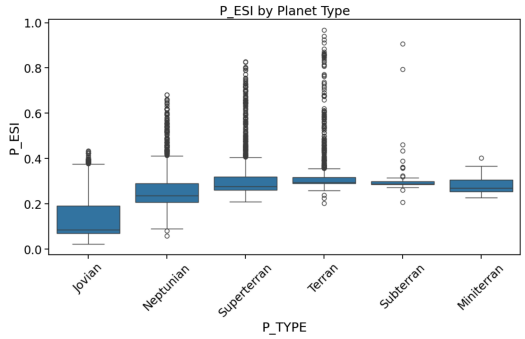
The typical formula used to compute ESI requires measurements for density, radius, surface temperature, and escape velocity. The HWC data, however, uses a different formula relying on radius, mass, and flux. Due to the fact that the HWC measurement does not include surface temperature, we do not have all of the necessary measurements available to compute ESI with total accuracy. The HWC tries to get around this by using flux in their alternative formula, but the formula is arbitrary in that it sometimes uses mass in its calculations and sometimes uses radius. Not to mention the fact that the true ESI formula does not even use flux. So rather than using an arbitrary formula to compute missing ESI values, we will use a hybrid model approach to analyze all of the variables available to us and use those to make a less arbitrary weighted calculation of the missing ESI values.

By investigating the HWC dataset, we hoped that analyzing variables that are correlated with ESI could help us predict some of the ESI values missing from our dataset. Our plan was to utilize a combination of astrophysical reasoning with statistical modeling techniques. The statistical methods used were linear ElasticNet regression and nonlinear Random Forests.

After reading the HWC dataset, we extracted the most relevant information, like observed ESI, mass, radius, equilibrium temperature, and other parameters as well. Then we proceeded to analyze that information for consistency. Knowing that astronomical datasets often have some missing entries, we implemented strategies to help maintain the integrity of our data.

Variables for which more than half of the exoplanets were missing a measurement were automatically ignored. This helped to ensure that our model was only trained on features with a sufficient amount of data. For other numerical variables, median value imputation was often applied during preprocessing to fill in missing data. However, if an entry was missing a measurement for a crucial variable like mass, radius, or gravity then that entry was excluded from the prediction process. We had to do this because mass, radius, and gravity are absolutely necessary for the accurate calculation of predicted ESI.





To fill in some of the gaps in our data, we used measurements that were given to us to make sensible physics-based calculations, and

use those calculations in place of the non-existing data. For instance, when estimating planetary density, we used the formula:

$$\rho = m / ((4/3)\, \pi\, r^3)$$

and for escape velocity, we used the formula:

$$v_{esc} = \sqrt{2GM/r}$$

Since these formulas are based on proven astrophysical principles, we were able to maintain data integrity, even though we decided to substitute missing data with our own calculations.

We were able to refine our dataset even further by consolidating infrequent variables, like detection method and planet type. This helped us reduce sparsity in our data and improve the robustness of our analysis.

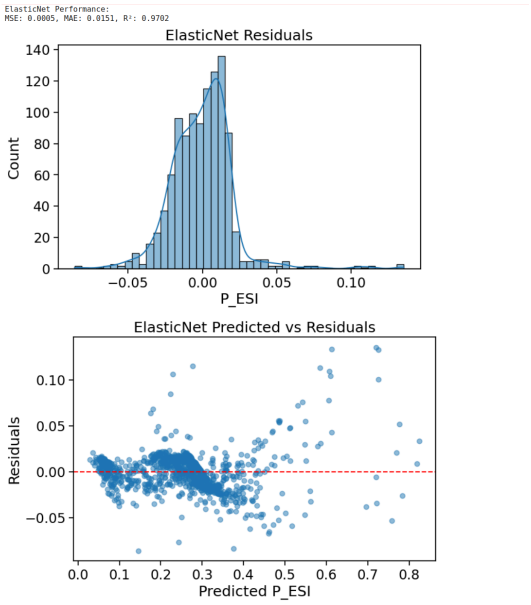
One of the most important parts of our analysis was feature engineering, and a major part of that process was the utilization of similarity scores. Having similarity scores for each of our most influential variables, helped in the quantification of how Earth-like each planet was. For a parameter x, its corresponding similarity score was calculated using the formula:

$$S_x = 1 - | x - x_{earth} | / (x + x_{earth})$$

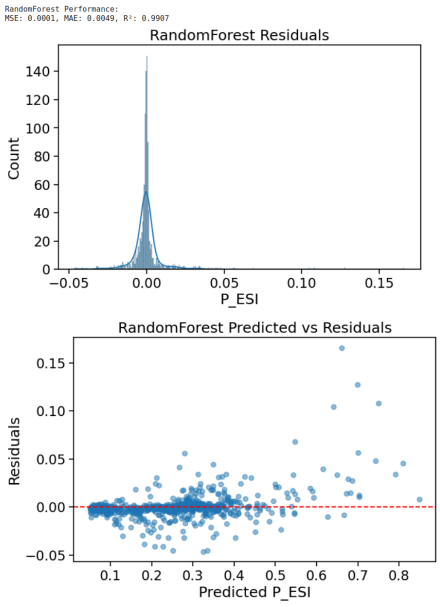
where x_{earth} represents Earth's corresponding measurement. The use of similarity scores provided further clarity by reflecting the similarity between the characteristics of an exoplanet and those of Earth.

Then we proceeded to split the data into training data and testing data, while also making sure to have a balanced representation of the different exoplanet types. This allowed us to generalize our model's performance across a diverse range of conditions.

We further enhanced our generalization by combining calculations from multiple models. This made our predictions much more accurate and robust. The first model we used, the ElasticNet regression model, was able to analyze linear trends in our similarity scores and the exoplanets' data. ElasticNet regression is a method that linearly combines the L1 (Lasso) regression and L2 (Ridge) regression. This technique improved our model's interpretability and aided in the handling of multicollinearity.



Our second model, the Random Forest model, was essentially a combination of numerous decision trees that tried to analyze complex non-linear trends in our data. Leveraging multiple trees and averaging their predictions, helped our model become more resistant to overfitting, which was a major advantage given all of the intricacies of astrophysical data.

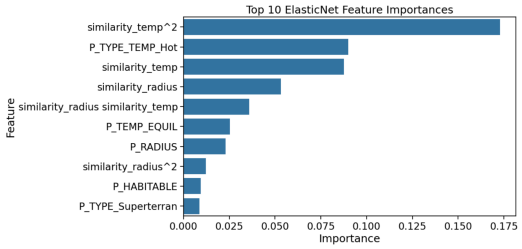


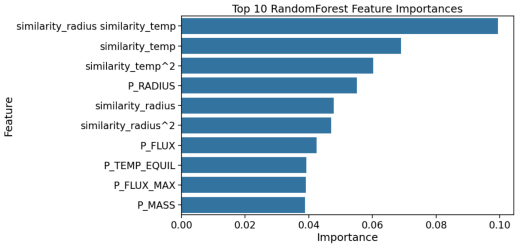
Once our models were trained, their outputs were aggregated using a weighted approach. The weights were based off of each

model's normalized R^2 score, which made it so that our final prediction

$$y_{pred} = w_1 (y_{pred1}) + w_2 (y_{pred2})$$

was within the range [0, 1]. This method ensured that both our linear model and our non-linear model provided an appropriate contribution to our predicted ESI values.





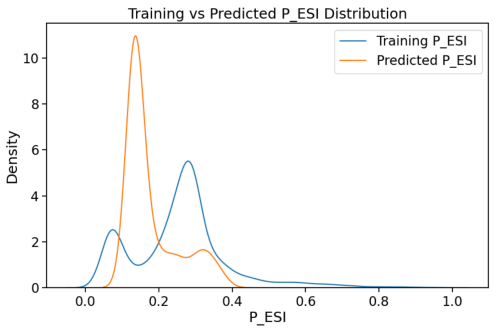
Model Performance:

ElasticNet Performance:

MSE: 0.0005, MAE: 0.0151, R^2: 0.9702

RandomForest Performance:

MSE: 0.0001, MAE: 0.0049, R^2: 0.9907



Both of our models achieved remarkably low MSE scores and remarkably high R^2 scores. This may be a result of the use of polynomial features and similarity scores capturing redundant correlations rather than true causality. However, it is also likely that our methodology made it inherently easy to approximate ESI for entries where only the ESI value was missing. Since ESI is traditionally calculated using a weighted product of radius, surface temperature, escape velocity, and density, our decision to train our models exclusively on entries where most of these measurements were available could have made it easier than expected to predict ESI.

A part of our analysis that remains ambiguous is the effect of our use of equilibrium temperature rather than surface temperature in our predicted ESI calculations. We decided to use equilibrium temperature as a proxy for surface temperature because the HWC dataset does not include measurements of the exoplanets' surface temperatures. While Earth's equilibrium temperature and surface temperature are nearly identical, that is not the case for every celestial body. Atmosphere plays a major role in that discrepancy, but since the atmosphere of these exoplanets is unknown, we had to rely on equilibrium temperature in our calculations. This may have caused an underestimation or overestimation in some of our results.

Truthfully it is hard to tell whether our models' excellent performances can be attributed to a genuine understanding of the underlying astrophysical relationships or a fallacious understanding rooted in redundancy. Whatever the case our framework gave us a much better understanding of planetary habitability. Our integrated approach succeeded in using a combination of astrophysical insights and advanced statistical techniques as a way to predict unknown ESI values. We hope that this framework can help others better understand some of the factors that drive planetary habitability. We also hope that others can build upon these foundations by incorporating additional data and exploring more sophisticated statistical methods.

Rows with missing P_ESI but P_MASS, P_RADIUS, and P_GRAVITY present: 234

Shortened List of Predicted ESI Values:

Predicted P_ESI Entries (Ordered Largest to Smallest):

P_NAME	P_TYPE	S_NAME	P_DETECTION	Guessed P_ESI
KMT-2020-BLG-0414L b	Terran	KMT-2020-BLG-0414L	Microlensing	0.377853
OGLE-2016-BLG-1195L b	Terran	OGLE-2016-BLG-1195L	Microlensing	0.374637
OGLE-2013-BLG-0341L B b	Terran	OGLE-2013-BLG-0341L B	Microlensing	0.367767
PSR B0329+54 b	Terran	PSR B0329+54	Other	0.362550
KMT-2016-BLG-1105L b	Terran	KMT-2016-BLG-1105L	Microlensing	0.362127
OGLE-2019-BLG-1053L b	Terran	OGLE-2019-BLG-1053L	Microlensing	0.360467
OGLE-2019-BLG-0960L b	Terran	OGLE-2019-BLG-0960L	Microlensing	0.355080
KMT-2021-BLG-0912L b	Terran	KMT-2021-BLG-0912L	Microlensing	0.352591
OGLE-2017-BLG-0173L b	Superterran	OGLE-2017-BLG-0173L	Microlensing	0.346483
MOA-2013-BLG-605L b	Superterran	MOA-2013-BLG-605L	Microlensing	0.341785
KMT-2017-BLG-1194L b	Superterran	KMT-2017-BLG-1194L	Microlensing	0.341073
MOA-2007-BLG-192L b	Superterran	MOA-2007-BLG-192L	Microlensing	0.339792
OGLE-2018-BLG-0677L b	Superterran	OGLE-2018-BLG-0677L	Microlensing	0.335707
KMT-2019-BLG-1367L b	Superterran	KMT-2019-BLG-1367L	Microlensing	0.333708
PSR B1257+12 d	Superterran	PSR B1257+12	Other	0.328919
KMT-2021-BLG-1391L b	Superterran	KMT-2021-BLG-1391L	Microlensing	0.326881
KMT-2019-BLG-1806L b	Superterran	KMT-2019-BLG-1806L	Microlensing	0.325932
OGLE-2017-BLG-1434L b	Superterran	OGLE-2017-BLG-1434L	Microlensing	0.325272
PSR B1257+12 c	Superterran	PSR B1257+12	Other	0.322461
MOA-2022-BLG-249L b	Superterran	MOA-2022-BLG-249L	Microlensing	0.319507
KMT-2017-BLG-1003L b	Superterran	KMT-2017-BLG-1003L	Microlensing	0.319425
OGLE-2017-BLG-1806L b	Superterran	OGLE-2017-BLG-1806L	Microlensing	0.318029
OGLE-2005-BLG-390L b	Superterran	OGLE-2005-BLG-390L	Microlensing	0.315532
KMT-2017-BLG-0428L b	Superterran	KMT-2017-BLG-0428L	Microlensing	0.313470
KMT-2018-BLG-1988L b	Superterran	KMT-2018-BLG-1988L	Microlensing	0.312817
KMT-2018-BLG-1025L b	Superterran	KMT-2018-BLG-1025L	Microlensing	0.308873
OGLE-2018-BLG-0383L b	Superterran	OGLE-2018-BLG-0383L	Microlensing	0.307133
OGLE-2018-BLG-0977L b	Superterran	OGLE-2018-BLG-0977L	Microlensing	0.305994
MOA-2012-BLG-505L b	Superterran	MOA-2012-BLG-505L	Microlensing	0.302834
OGLE-2018-BLG-0532L b	Superterran	OGLE-2018-BLG-0532L	Microlensing	0.302329
KMT-2018-BLG-0029L b	Superterran	KMT-2018-BLG-0029L	Microlensing	0.291230
OGLE-2018-BLG-1185L b	Superterran	OGLE-2018-BLG-1185L	Microlensing	0.289262
PSR B1257+12 b	Other	PSR B1257+12	Other	0.289043
OGLE-2017-BLG-0482L b	Superterran	OGLE-2017-BLG-0482L	Microlensing	0.282440
KMT-2019-BLG-0253L b	Superterran	KMT-2019-BLG-0253L	Microlensing	0.279961
MOA-2010-BLG-328L b	Superterran	MOA-2010-BLG-328L	Microlensing	0.276695
MOA-2009-BLG-266L b	Neptunian	MOA-2009-BLG-266L	Microlensing	0.266825
KMT-2019-BLG-0842L b	Neptunian	KMT-2019-BLG-0842L	Microlensing	0.266389
MOA-2020-BLG-135L b	Neptunian	MOA-2020-BLG-135L	Microlensing	0.264774
KMT-2021-BLG-0171L b	Neptunian	KMT-2021-BLG-0171L	Microlensing	0.261771
OGLE-2018-BLG-0596L b	Neptunian	OGLE-2018-BLG-0596L	Microlensing	0.252033
OGLE-2017-BLG-1691L b	Neptunian	OGLE-2017-BLG-1691L	Microlensing	0.251080
OGLE-2005-BLG-169L b	Neptunian	OGLE-2005-BLG-169L	Microlensing	0.249448
KMT-2019-BLG-0953L b	Neptunian	KMT-2019-BLG-0953L	Microlensing	0.246123
KMT-2022-BLG-0440L b	Neptunian	KMT-2022-BLG-0440L	Microlensing	0.243497
OGLE-2018-BLG-0506L b	Neptunian	OGLE-2018-BLG-0506L	Microlensing	0.241012
MOA-2011-BLG-262L b	Neptunian	MOA-2011-BLG-262L	Microlensing	0.239842
OGLE-2015-BLG-1670L b	Neptunian	OGLE-2015-BLG-1670L	Microlensing	0.237170
OGLE-2018-BLG-1126L b	Neptunian	OGLE-2018-BLG-1126L	Microlensing	0.236592
MOA-2011-BLG-291L b	Neptunian	MOA-2011-BLG-291L	Microlensing	0.234454
KMT-2021-BLG-1253L b	Neptunian	KMT-2021-BLG-1253L	Microlensing	0.232893
OGLE-2018-BLG-0516L b	Neptunian	OGLE-2018-BLG-0516L	Microlensing	0.230823
OGLE-2007-BLG-368L b	Neptunian	OGLE-2007-BLG-368L	Microlensing	0.230376
OGLE-2015-BLG-0966L b	Neptunian	OGLE-2015-BLG-0966L	Microlensing	0.225134
KMT-2021-BLG-2294L b	Neptunian	KMT-2021-BLG-2294L	Microlensing	0.221541
MOA-2008-BLG-310L b	Neptunian	MOA-2008-BLG-310L	Microlensing	0.218914
MOA-2011-BLG-028L b	Neptunian	MOA-2011-BLG-028L	Microlensing	0.206273
KMT-2019-BLG-1216L b	Neptunian	KMT-2019-BLG-1216L	Microlensing	0.202164
KMT-2021-BLG-0320L b	Neptunian	KMT-2021-BLG-0320L	Microlensing	0.200538
KMT-2017-BLG-0165L b	Neptunian	KMT-2017-BLG-0165L	Microlensing	0.196762
MOA-2015-BLG-337L b	Neptunian	MOA-2015-BLG-337L	Microlensing	0.196254
KMT-2021-BLG-0192L b	Neptunian	KMT-2021-BLG-0192L	Microlensing	0.195362
OGLE-2019-BLG-1492L b	Neptunian	OGLE-2019-BLG-1492L	Microlensing	0.193230
OGLE-2012-BLG-0950L b	Neptunian	OGLE-2012-BLG-0950L	Microlensing	0.192803
KMT-2021-BLG-1554L b	Neptunian	KMT-2021-BLG-1554L	Microlensing	0.192582
KMT-2021-BLG-0712L b	Neptunian	KMT-2021-BLG-0712L	Microlensing	0.189507
MOA-2020-BLG-208L b	Neptunian	MOA-2020-BLG-208L	Microlensing	0.186588
OGLE-2018-BLG-0298L b	Neptunian	OGLE-2018-BLG-0298L	Microlensing	0.186331
KMT-2021-BLG-1689L b	Neptunian	KMT-2021-BLG-1689L	Microlensing	0.184674
OGLE-2012-BLG-0026L b	Neptunian	OGLE-2012-BLG-0026L	Microlensing	0.182758
OGLE-2012-BLG-0838L b	Jovian	OGLE-2012-BLG-0838L	Microlensing	0.176865
OGLE-2008-BLG-092L b	Jovian	OGLE-2008-BLG-092L	Microlensing	0.174038
OGLE-2014-BLG-1722L b	Jovian	OGLE-2014-BLG-1722L	Microlensing	0.172954
OGLE-2011-BLG-0173L b	Jovian	OGLE-2011-BLG-0173L	Microlensing	0.170457
KMT-2018-BLG-0748L b	Jovian	KMT-2018-BLG-0748L	Microlensing	0.168854
KMT-2019-BLG-1042L b	Jovian	KMT-2019-BLG-1042L	Microlensing	0.168626

8 Limitations

There are several limitations to address for our analysis as a whole. Firstly, we are assuming the data provided to us is accurate. Without assurances, the validity of astrophysical measurements is difficult to gauge. The process of taking these kinds of measurements is often prone to errors, sometimes in the form of instrumental errors, and sometimes in the form of observational errors due to environmental factors. Unfortunately there is no way to confirm whether the data we are using is definitively well-measured. Such factors are out of our control, and to account for this we simply note that our results should be interpreted within the context of the data we studied. We would also like to note that in our efforts to perform habitability and ESI analysis, there were very few habitable planets to work with, which raised issues with the sample size, generalizability, and the integrity of our models and statistical tests. In most cases, the sample size was generally sufficient per typical statistical requirements. However, normality assumptions were difficult to verify as a result of there being so few data points.

As described in Section 7, ESI has varying definitions. It is important to clarify that our project mainly used the definition of ESI supplied by the Habitable Worlds Catalog (HWC). Other sources may use different formulae and thus different predictors to generate ESI scores. Thus, our results may not be generalizable to other kinds of ESI scores.

Another major limitation was the nature of the data itself. Astrophysical data is prone to skewness and outliers. In some models, we found it difficult to produce linear relationships, even with log-scaling and transformation. Naturally, this yields models that do not satisfy the assumptions of linear regression, which raises concerns about the validity of their predictions. Consequently, the models made in this project should be interpreted with heavy caution and refined before practical application. This project still needs more refinement, mainly to resolve issues with linearity in models and further feature exploration to produce even stronger relationships to ESI and habitability.

9 Relation to Project Proposal

In our project proposal, our goal was to create one multivariate model to predict ESI scores from the data available to us. However, we realized that ESI is directly computed from specific predictors, such as mass, radius, and flux. This encouraged us to think beyond simple regression, and motivated us to pursue what is, in our opinion, a more intriguing question: how can we optimize classification of habitable exoplanets? By refocusing our project around habitable exoplanets and the relationship habitability has with ESI, we found a new direction by way of predicting ESI from new variables, which yielded more interesting results for ESI prediction.

10 Conclusion

In this project, we organized our goals around one central question: how can we statistically optimize our ability to classify habitable and Earth-similar exoplanets? This question produced several avenues of exploration, resulting in analysis of what indicates habitability, two logistic regression models to predict both habitability and similarity to Earth (ESI $\geq$ 0.8), and a discussion of how to potentially predict ESI values from other predictors in the event that the usual predictors are unavailable.

Climate change, depletion of natural resources, and global tension continue to threaten the habitability of Earth, forcing us to look beyond our world for solutions. Because of this predicament, we set out to optimize the search for habitable Earth-like exoplanets by harnessing data from the NASA Exoplanet Archive and the Habitable Worlds Catalog. We started by using exploratory data analysis to find the relationship between habitability and ESI, but in order to fully understand this relationship, we needed to better understand ESI and find ways to predict its value in exoplanets with an incomplete set of measurements. Our analysis eventually led us to the conclusion that only 0.45% of the exoplanets in our datasets had an ESI $\geq$ 0.8, and only 19-28% of the exoplanets with an ESI $\geq$ 0.47 demonstrated habitability conditions. This showed that there is an utter lack of abundance of Earth-like exoplanets in our universe. We proceeded to utilize methods like clustering, regression, and hybrid models, all carried out under a different set of constraints, to get a more complete picture of our data. While there were certainly limitations to what we could do, it became evident that our decision to take multiple different approaches expanded the scope and clarity of our analysis. We were able to predict ESI values under a number of different conditions, giving us a much better foundation to analyze habitability in the future. In its quest to shed light on the true scarcity of habitable worlds our project attempted to make astrophysical insights through sheer statistical rigor. We hope that our framework can guide future discovery efforts and push humanity closer to a deeper understanding of our place in the universe. For future projects, we encourage improvements to the models produced here, especially to ensure statistical integrity and satisfaction of all model assumptions.

References

[1] NASA Exoplanet Science Institute. (2020). Planetary Systems Table. IPAC. https://doi.org/10.26133/NEA12

[2] PHL @ UPR Arecibo - Earth Similarity Index (ESI). (n.d.). https://phl.upr.edu/projects/earth-similarity-index-esi