Exoplanet Analysis - NASA Kepler Mission

Introduction:

The detection and classification of exoplanets are critical to understanding planetary systems beyond our own. The Kepler mission, launched by NASA, has provided extensive data on potential exoplanets known as Kepler Objects of Interest (KOIs). However, distinguishing confirmed exoplanets from false positives remains challenging due to variations in host star characteristics, measurement uncertainties, and the complexity of planetary signals.

With a background in physics, including experience in spectroscopy and planetary orbits, I am particularly motivated to explore the factors that influence exoplanet classification accuracy. Understanding these factors can help refine the search criteria for potentially habitable planets, contributing to the broader goal of identifying Earth-like worlds. This research not only addresses scientific curiosity but also has practical applications for future missions like the Transiting Exoplanet Survey Satellite (TESS) and the James Webb Space Telescope (JWST).

This project investigates the influence of host star photometric properties on detection outcomes, identifies key factors that increase the likelihood of exoplanet confirmation, examines clustering patterns among KOIs, and assesses the reliability of the koi_score confidence metric. These insights can refine detection methodologies, reduce false positives, and improve future surveys. The results are actionable for optimizing target selection in these upcoming missions.

The benefits of this work extend to improving our understanding of planetary formation and potentially discovering habitable worlds. Prior studies (e.g., Thompson et al., 2018[1]; Mullally et al., 2016[2]) have shown that host star properties, such as brightness and metallicity, significantly impact detection outcomes. This analysis applies sophisticated machine learning and clustering techniques to dig deeper and try to find hidden patterns from the data.

Analytical Questions and Data:

1. How do variations in the host star's photometric properties across different wavelength bands influence exoplanet detection and classification outcomes?

   Objective: Investigate whether specific photometric properties, such as brightness in different color bands (e.g., red, green, blue), affect KOI classification outcomes. This can help refine star selection criteria for future observations.

2. Which characteristics most influence the likelihood of an exoplanet candidate's confirmation, and how can these insights refine detection methods?

   Objective: Identify key factors distinguishing confirmed exoplanets from false positives and candidates. Try to build a refined, simplified model.

3. Can we identify distinct clusters of KOIs based on physical and orbital characteristics, and do these clusters show similar confirmation outcomes?

   Objective: Discover natural groupings within the KOI dataset based on planetary and orbital characteristics. Determine if certain clusters are more likely to yield confirmed planets.

4. Which columns/features determine the koi_score, and how reliable is

this confidence value compared to classified stars?

Objective: Identify factors influencing the koi_score confidence metric and evaluate its reliability for predicting confirmed exoplanets versus false positives.

The dataset[3] contains 9564 rows (entries) and 141 columns, encompassing planetary attributes, stellar properties, threshold flags, confidence scores, and final classifications. It is extensive and updated daily with approximately 10,000 new entries, making it the most comprehensive resource available for exoplanet research.

A key feature is the presence of eight wavelength bands, enabling detailed photometric analysis. The dataset also includes measurements like transit depth and duration, adding richness and complexity to the analysis. These features enable sophisticated questions about the interplay between photometric and stellar properties in exoplanet detection.

The fourth analytical question, focusing on the koi_score (confidence metric), emerged from initial dataset exploration, underscoring the need to understand its reliability and influencing factors.

Analysis:

Initial Data Processing and Exploration

The dataset underwent comprehensive preprocessing[4] to ensure relevance and reduce noise for the analysis:

- Dropped pixel-based vetting statistics as they focus on contamination detection, which is irrelevant to the physical factors driving planet confirmation. Their inclusion would unnecessarily complicate the analysis without providing meaningful insights.
- Eliminated metadata columns such as planet names and dataset origins, as they offer no predictive or explanatory value for the research questions.
- Derived columns (e.g., error columns) were excluded to avoid introducing confounding factors, as they do not represent independent features.
- Columns with >50% missing values were dropped since imputing them risked biasing the dataset without adding significant analytical value.

The dataset was further pre-processed in each question, to better align with the examined aspect needs.

Question 1: Photometric Properties and Exoplanet Detection

Data Preparation and Derivation:

- Analyzed koi_kepmag along with the other bands to compare its unique contribution to exoplanet detection, given its optimized for this purpose.
- Dropped koi_zmag because its high missing values and similarity to koi_jmag wavelength make them nearly indistinguishable.
- Grouping visible and infrared bands help distinguish and differentiate the two bands, while differences between these groups can amplify subtle variations caused by planetary transits. These derived features added complexity without introducing redundancy.

Methods and Techniques

1. **Distribution Analysis:**

   Initial histograms revealed a consistent left skewness across all wavelength bands, reflecting the higher likelihood of detecting brighter stars (lower magnitudes) and the exclusion of faint stars below detection thresholds. When separated by classification, the distributions showed confirmed KOIs more concentrated around the peak with smaller variability compared to false positives. Box plots further emphasized these trends, highlighting subtle brightness differences between classifications, though no distinct separation was evident.

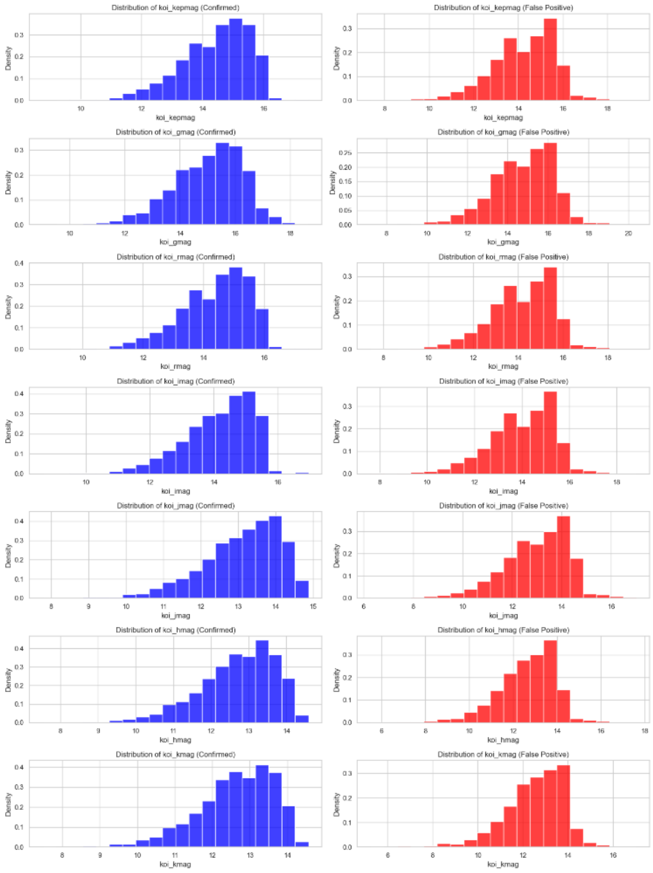

Figure 1: Distribution of different wavelength bands grouped by final classification

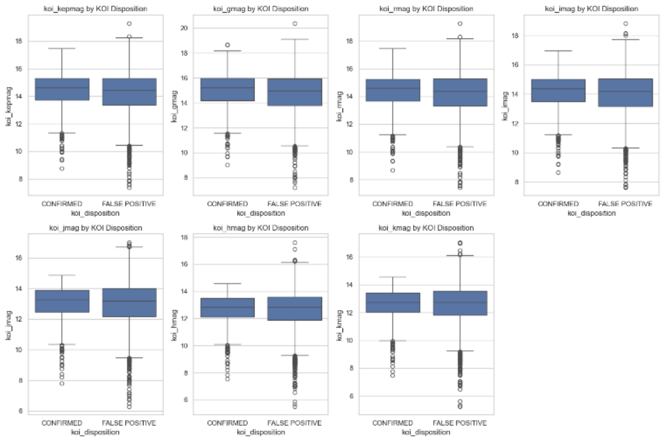

Figure 2: Boxplots of different wavelength bands grouped by final classification

2. **Correlation Matrix and PCA Analysis:**

   ○ The correlation matrix revealed high correlations between wavelength bands but low correlation with the target variable (confirmation status).

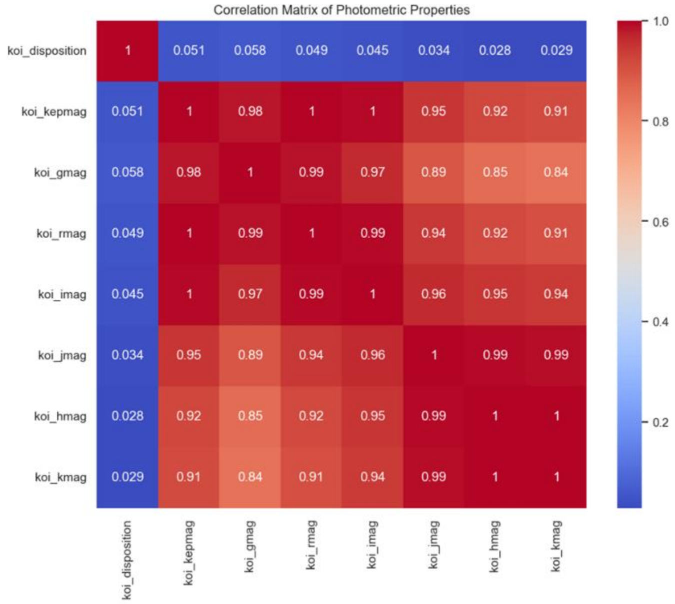

Figure 3: Correlation matrix between wavelength bands and classification

   ○ PCA reduced dimensionality but did not yield clear separation between classifications, indicating overlapping class distributions.

3. **T-Statistic and P-Value Analysis:**
   - Conducted t-tests to assess statistical significance of differences between confirmed planets and false positives across magnitude bands.
   - Validated results with p-values, confirming statistically significant differences.

Band	**t-stat**	**p-value**
koi_kepmag	6.666	2.85E-11
koi_gmag	7.984	1.66E-15
koi_rmag	6.509	8.13E-11
koi_imag	5.856	4.97E-09
koi_jmag	3.902	9.64E-05
koi_hmag	3.13	1.76E-03
koi_kmag	3.089	2.02E-03

Figure 4: Table of Welch's t-test results

Question 2: Key Features Influencing Confirmation

Data Preparation:

- **Prioritized koi_model_snr (signal-noise-ratio):** Dropped rows with missing values, as this is a primary signal indicator.
- **Addressed High Missing Values:** Imputed zeros for koi_max_sngle_ev, koi_max_mult_ev, and koi_num_transits, justified by domain knowledge regarding weak signals as shown in notebook.
- **Feature Selection:** Retained koi_kepmag while dropping other redundant wavelength bands. Removed confidence columns (e.g., probabilities) to focus on independent predictors.

Methods and Techniques

Random Forest Classifier:

- Trained a robust model to handle noise and high-dimensional data.
- Used built-in feature importance metrics to prioritize variables for a reduced model.

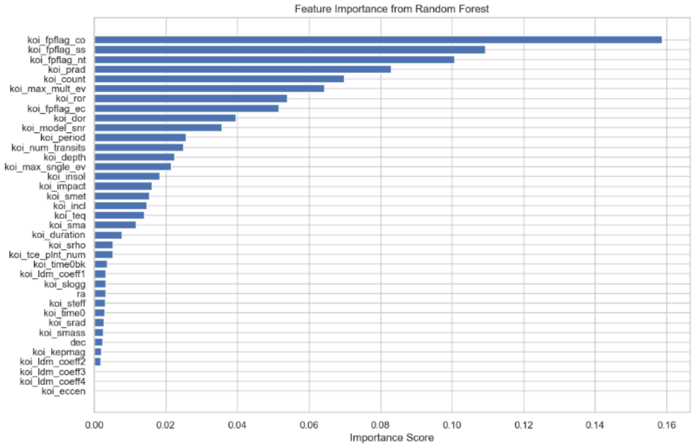

Figure 5: Feature importance bar chart from the Random Forest model.

Cross-Validation (5-Fold):

- Validated model generalizability across multiple data splits, accounting for slight target imbalance.

Dimensionality Reduction and Visualization:

- Applied PCA and t-SNE to identify feature clusters and visualize classification separations.
- t-SNE captured non-linear structures and highlighted subtle distinctions missed by PCA.

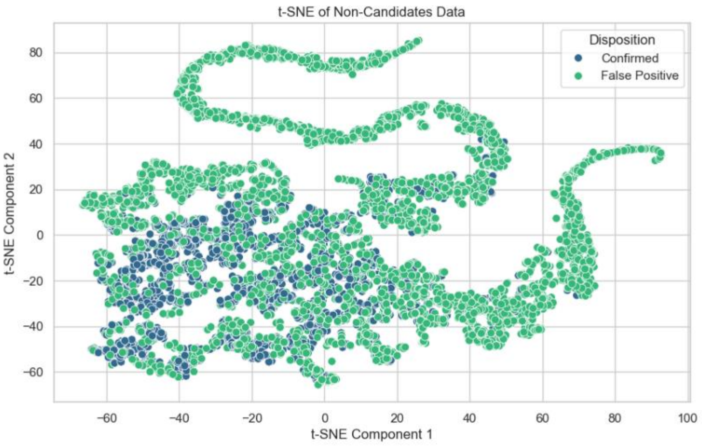

Figure 6: 2D t-SNE plot with differentiating target classes.

Model Simplification:

- Trained a reduced Random Forest model using the top 10 features for improved interpretability without sacrificing accuracy.

Question 3: Clustering KOIs by Characteristics

Data Preprocessing:

- Dropped rows with missing values in crucial features, guided by the earlier finding that missingness correlated with low koi_model_snr.
- Retained only the top 99% quantile to minimize noise from extreme values while preserving the dataset's core structure.
- Since our main goal is clustering, those values would really affect our visuals, especially the more time-efficient PCA that we preferred for this task.

Methods and Techniques

K-Means Algorithm:

Chosen for its simplicity and efficiency in partitioning data into clusters. The elbow method determined the optimal number of clusters, ensuring simpler and more interpretable results.

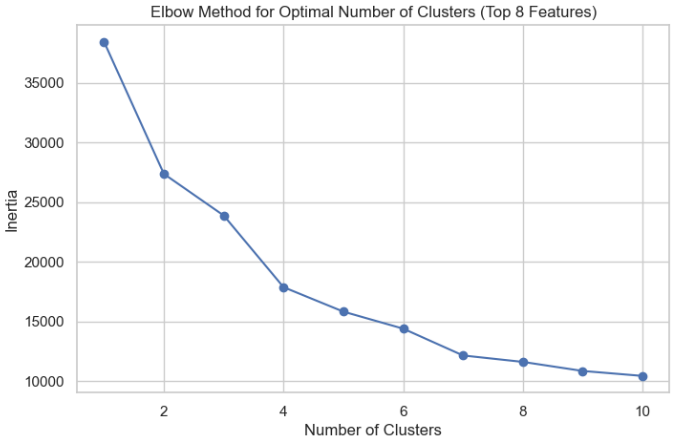

Figure 7: Visual of the elbow method (determining 4 clusters as the "elbow point")

Dimensionality Reduction:

PCA was used for visualization and to minimize feature space, facilitating a clearer understanding of cluster separations.

Refinement:

Reduced features to the top 8 (from Question 2's importance rankings) and repeated the clustering process. This refinement led to fewer, more distinct clusters, further enhancing interpretability.

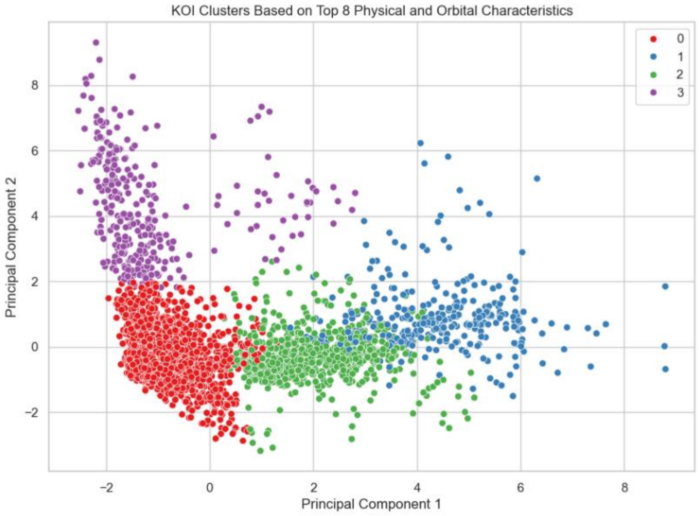

Figure 8: Predicted clusters after elbow method

Visualizations of the distributions within each class was the way we evaluated the clustering success.

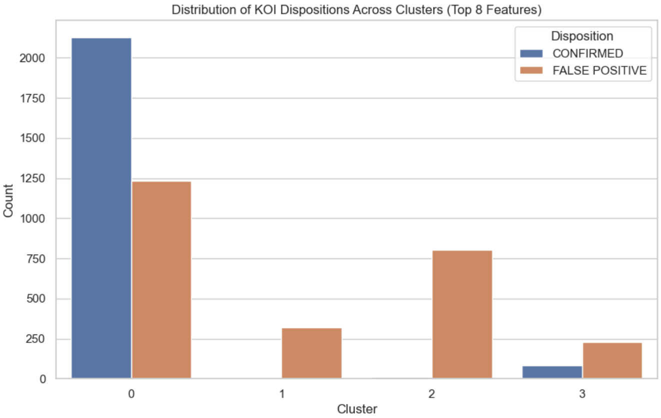

Figure 9: Distribution of KOI classes across the clusters to evaluate the separation

Question 4: Evaluating koi_score Confidence Metric

Data Preprocessing:

Identified a perfect correlation between missing koi_score values and koi_bin_oedp_sig, which measures false-positive signals. Rows with missing data were dropped, as imputing the target variable was infeasible.

Methods and Techniques

Correlation Analysis:

Evaluated Pearson correlations between koi_score and features to identify the top 10 influential predictors. This analysis informed subsequent modeling steps.

Regression Model/Random Forest Regressor:

Chosen for its ability to model non-linear relationships and handle high-dimensional datasets. The model's low Mean Squared Error (MSE) and high R² score indicated strong predictive performance. Calculated koi_scores variance as well to compare with mse and get meaningful insights.

Cross-Validation:

A 5-fold approach ensured the model's generalizability across subsets of the data.

Threshold Optimization:

Analyzed accuracy across a range of thresholds for koi_score, enabling identification of the optimal range for predictions. This step provided actionable insights into the sensitivity of this confidence metric. Graphicly representing it afterwards helped us understand the how directly it impacts the final classification.

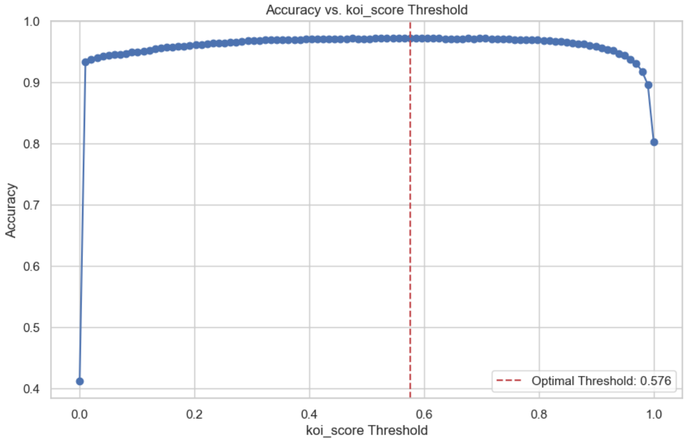

Figure 10: Threshold versus classification accuracy curve (showing optimal line as red)

Further Validation:

Confusion matrices and classification reports (precision, recall, F1-score) evaluated the accuracy of threshold-based classifications.

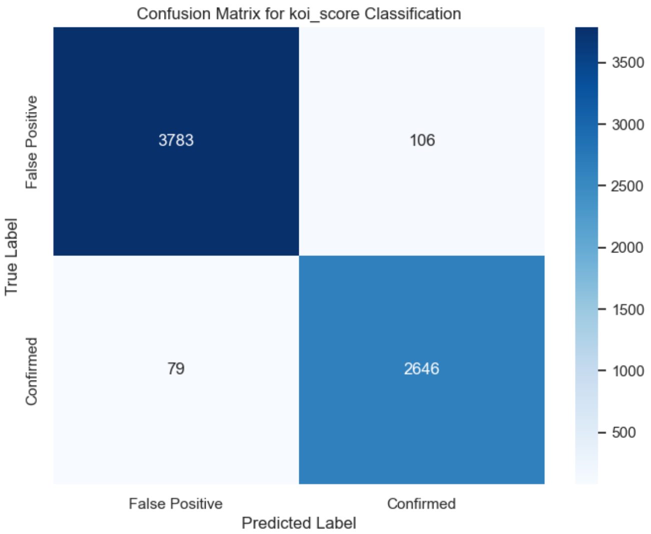

Figure 11: Confusion matrix for koi_score Classification

Findings, Reflections, and Further Work

Key Findings

1. Photometric Properties and Exoplanet Detection: The dataset showed an imbalance, with false positives nearly twice as frequent as confirmed exoplanets. T-test results indicated that brightness in different wavelength bands significantly differentiates confirmed planets from false positives. While histograms did not show a clear distinction, confirmed KOIs were more concentrated around the peak values with smaller spreads. Although the green and red bands showed stronger associations with confirmed KOIs, the correlation matrices and PCA did not reveal any meaningful separation, suggesting no clear benefit in focusing on a specific wavelength band.

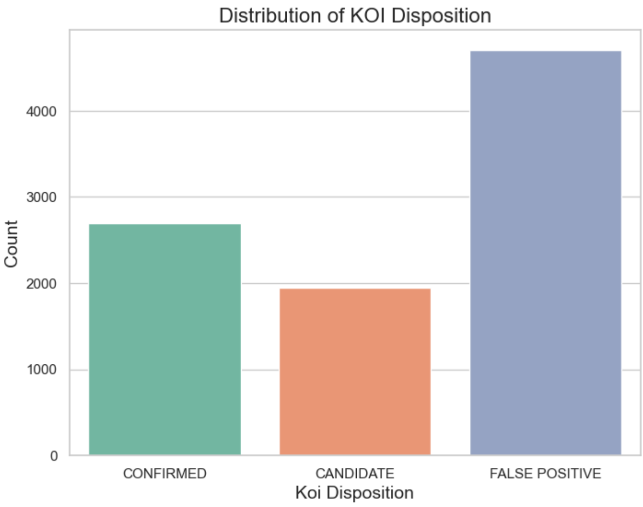

Figure 12: Distribution of KOI Classification

2. Factors Influencing Exoplanet Confirmation: The Random Forest model achieved an initial accuracy of 99.3%, with a 1% average cross-validation error (results similar to this paper[8]). Simplifying the model by retaining the top 10 predictors increased accuracy to 99.5% and reduced the cross-validation error to 0.6%. The model simplification allowed for 3D visualization using t-SNE dimension reduction and plotly library[5], which helped in better understanding the relationships between features and the target variable.

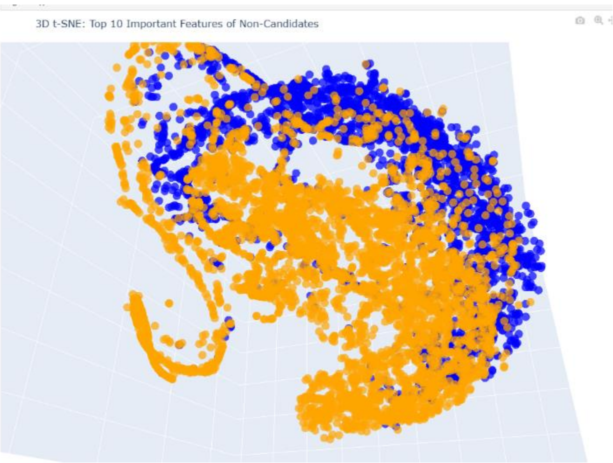

Figure 13: Screenshot of 3D t-SNE representation after feature selection

3. Clustering of KOIs: Initial clustering with 7 clusters did not provide clear separations. After refining the model with the top 8 features, reducing the number of clusters to 4 improved separation. This indicated that KOIs can indeed be grouped based on their physical and orbital characteristics, although further refinement could be done by isolating and splitting the 0 cluster for clearer separation.

4. Evaluation of the koi_score Confidence Metric: Missing koi_score values correlated with the koi_bin_oedp_sig column, a key metric for identifying false positives (Used missingno library[6]). Since imputing these values was not feasible, rows with missing data were dropped. The Random Forest regression model showed that the koi_flags features were highly significant, with validation results showing an MSE of 0.0264 and an R² score of 0.886. The accuracy vs. threshold graph indicated that a threshold of 0.576 provided the best accuracy (97.2%), while the metric is very sensitive meaning with even 0.02 threshold to produce more than 90% accuracy.

Reflections and Limitations

The findings offer valuable insights into exoplanet classification, but several challenges were encountered. The high-dimensionality of the dataset, with over 140 features, complicated the analysis despite dimensionality reduction methods like PCA and t-SNE. These methods sometimes failed to reveal clear patterns, suggesting the need for further refinement. The dataset imbalance, with more false positives than confirmed exoplanets, may have affected model performance and introduced bias. Techniques such as oversampling or rebalancing could help address this issue.

Handling missing data, particularly in the koi_score column, was another challenge. Since imputing values was not feasible, rows with missing data were dropped, which may have led to the loss of valuable information. Future work could benefit from incorporating additional data sources, such as spectroscopic or radial velocity measurements, to improve model generalizability. While the Random Forest model performed well, further model improvements and additional data could enhance its predictive power. The clustering model could also be refined by isolating specific clusters for better separation and deeper insights.

Further Work

Within the timeframe of this project, the four analytical questions were thoroughly explored. However, these questions could be expanded upon further with additional time[7]. For instance, analyzing the dataset over a longer period, such as a month or a year, would allow for the incorporation of external factors like daily weather patterns, seasonal changes, or even material-equipment wear over time. These factors could influence the accuracy of exoplanet detection and

classification, providing deeper insights into the reliability of the findings.
Expanding the analysis with temporal factors would contribute to a more comprehensive understanding of the challenges in exoplanet classification and detection.
Further improvements could include exploring other machine learning models, incorporating more diverse datasets (such as radial velocity data), and investigating alternative clustering techniques to enhance the interpretability of groupings.

Wordcount

Section	Actual	Max Count
Introduction	262	300
Analytical questions and data	287	300
Analysis	984	1000
Findings, reflections and further work	617	600
Total	2150	2200

References

1. Twicken, J. D., Catanzarite, J. H., Clarke, B. D., Girouard, F., Jenkins, J. M., Klaus, T. C., Li, J., McCauliff, S. D., Seader, S. E., Tenenbaum, P., Wohler, B., Bryson, S. T., Burke, C. J., Caldwell, D. A., Haas, M. R., Henze, C. E., & Sanderfer, D. T. (2018). KeplerData Validation I—Architecture, diagnostic tests, and data products for vetting transiting planet candidates. *Publications of the Astronomical Society of the Pacific*, *130*(988), 064502. https://doi.org/10.1088/1538-3873/aab694

2. Coughlin, J. L., Mullally, F., Thompson, S. E., Rowe, J. F., Burke, C. J., Latham, D. W., Batalha, N. M., Ofir, A., Quarles, B. L., Henze, C. E., Wolfgang, A., Caldwell, D. A., Bryson, S. T., Shporer, A., Catanzarite, J., Akeson, R., Barclay, T., Borucki, W. J., Boyajian, T. S., . . . Zamudio, K. A. (2016). PLANETARY CANDIDATES OBSERVED BY KEPLER. VII. THE FIRST FULLY UNIFORM CATALOG BASED ON THE ENTIRE 48-MONTH DATA SET (Q1–Q17 DR24). The Astrophysical Journal Supplement Series, 224(1), 12. https://doi.org/10.3847/0067-0049/224/1/12

3. NASA Exoplanet Archive. (2024, November 8). Kepler Objects of Interest (KOI) cumulative table. https://doi.org/10.26133/NEA4

4. Logan. (2022, January 6). Identifying Exoplanets using Multiple Classification Models. Medium. https://medium.com/@scheidlogan/identifying-exoplanets-using-multiple-classification-models-7ee48024d7fd

5. *plotly.express.scatter_3d — 5.24.1 documentation*. (n.d.). https://plotly.com/python-api-reference/generated/plotly.express.scatter_3d

6. ResidentMario. (n.d.). *GitHub - ResidentMario/missingno: Missing data visualization module for Python*. GitHub. https://github.com/ResidentMario/missingno

7. *The NASA Exoplanet Archive: Data and Tools for Exoplanet Research*. (n.d.). Ar5iv. https://ar5iv.labs.arxiv.org/html/1307.2944

8. Srivathsa, V., & Assaf, R. (2022). Using Machine Learning to determine the most important features in exoplanet verification. *Journal of Student Research*, *11*(3). https://doi.org/10.47611/jsrhs.v11i3.2821