Deciphiring the Cosmos:
Applying Machine Learning Techniques to Exoplanetary Data

Ishani Raj

March 14, 2025

1 Abstract

The Kepler Space Telescope's mission has provided a wealth of data in the form of Kepler Objects of Interest (KOI), presenting unique challenges and opportunities for astronomical data analysis. This study utilizes two distinct machine learning techniques to enhance the understanding and classification of exoplanets. Firstly, a t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization incorporating several exoplanet statistics is employed to explore the underlying data structure and visualize the complex relationships within the high-dimensional space of Kepler observations. This method effectively highlights the clustering patterns and distinctions between confirmed exoplanets, candidates, and false positives, for a nuanced depiction of data groupings. Secondly, a Random Forest classifier is built to independently predict the disposition of KOIs based on their physical and orbital characteristics. This approach not only corroborates findings from the t-SNE visualization but also reveals the inherent predictive power of the dataset's raw features. The integration of these methods sheds light on the classification of exoplanets and sets a precedent for the use of advanced machine learning techniques in the interpretation of extensive astronomical data, offering pathways for future explorations and methodological advancements in the field of planetary search.

2 Data

The Kepler Space Telescope was NASA's first planet-hunting mission, assigned to search a portion of the Milky Way galaxy for Earth-sized planets orbiting stars outside our solar system. During nine years in deep space Kepler, and its second act, the extended mission dubbed K2, showed our galaxy contains billions of hidden "exoplanets," many of which could be promising places for life. The dataset used in this study is downloaded from Kaggle [1] and published as-is by NASA [2]. It is a cumulative record of all observed Kepler "objects of interest" (KOI), last updated on October 26, 2017. It includes roughly 10,000 exoplanet candidates identified during the mission encompassing a variety of properties such as planetary size, orbit characteristics, and equilibrium temperature, with each object classified as either a 'Confirmed' planet, a 'Candidate,' or a 'False Positive', which is the disposition of the KOI. Understanding these classifications helps astronomers distinguish between likely planets and other celestial phenomena mimicking planetary transits. It facilitates advancement of our knowledge of the cosmos, particularly in the search for potentially habitable exoplanets.

3 Previous Work

The NASA Exoplanet Archive website [3] features approximately 600 research papers that acknowledge its contributions. One noteworthy study [4] employs machine learning and deep learning algorithms to enhance the detection of transiting exoplanets in K2 data. In this research, two variations of generative adversarial networks are implemented: semi-supervised generative adversarial networks and auxiliary classifier generative adversarial networks. The findings demonstrate that these models significantly improve the classification of stars hosting exoplanets, achieving a recall and precision rate of 1.00 on the test data. Additionally, the semi-supervised approach effectively addresses the challenging task of creating a labeled dataset, streamlining the data preparation process.

4 Exploration and Analysis

Initial Exploration The initial exploratory data analysis of the Kepler Objects of Interest (KOI) dataset involved preprocessing to handle missing values, followed by generating histograms to visually explore the distribution of various planetary characteristics among confirmed exoplanets (Figure 1). A few factors such as planet radius ('koi_prad'), orbital period ('koi_period'), and equilibrium temperature ('koi_teq') were selected to assess their impact on planet confirmation.

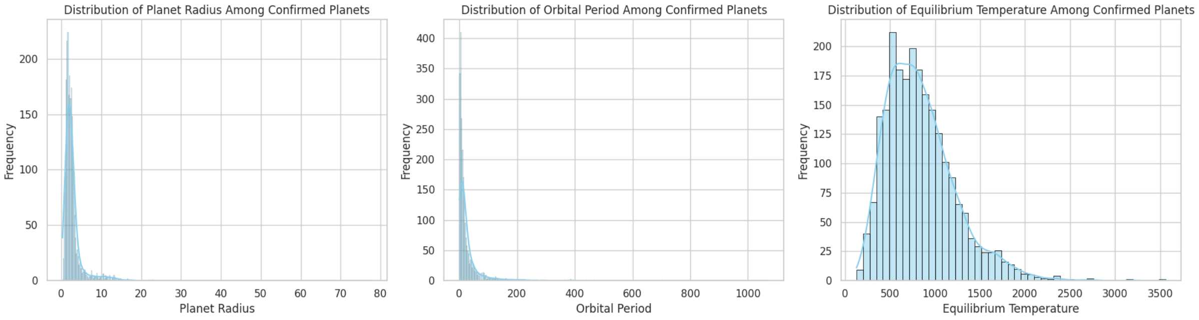

Figure 1: Distribution of planetary characteristics among confirmed exoplanets.

The histograms of confirmed exoplanets within the Kepler dataset reveal distinct patterns in planetary characteristics: a prevalence of smaller radii suggests a higher frequency of smaller, likely terrestrial planets. The orbital period distribution indicates a concentration of planets with shorter orbital periods, suggesting easier detection due to their frequent transits [5]. Lastly, the equilibrium temperature histogram shows a skew towards cooler temperatures, indicating a commonality of cooler planets among those confirmed, which might be conducive to certain types of planetary atmospheres.

This groundwork led to the application of t-SNE for high-dimensional data visualization, which provided an intuitive representation of data clusters and separations not immediately apparent from these traditional plots.

Application of Machine Learning Techniques The t-SNE visualization (Figure 2) effectively captures the complex relationships within the high-dimensional feature space of the KOI dataset. After preprocessing to handle missing values and scaling the selected features including as confidence score, orbital period, transit duration, depth of transit, planetary radius, equilibrium temperature, insolation flux, and stellar parameters including effective temperature, surface gravity, and radius, t-SNE reduces this multidimensional data into a comprehensible two-dimensional format. The dataset's "koi_disposition" is encoded as 'CANDIDATE': 0, 'CONFIRMED': 1, 'FALSE POSITIVE': 2, facilitating a clear visual segmentation in the resulting scatter plot.

The green points, representing false positives, are noticeably clustered towards higher values of both dimensions. The presence of this cluster suggests that there are discernible patterns or feature similarities among false positives. False positives might share characteristics that initially make them appear as valid exoplanet signals but are distinguishable upon closer analysis, such as similarities in how they transit their stars or certain noise patterns in the data.

The blue points, representing confirmed exoplanets , are mostly concentrated towards lower values of both dimensions, distinctly away from false positives. This clustering suggests that confirmed exoplanets

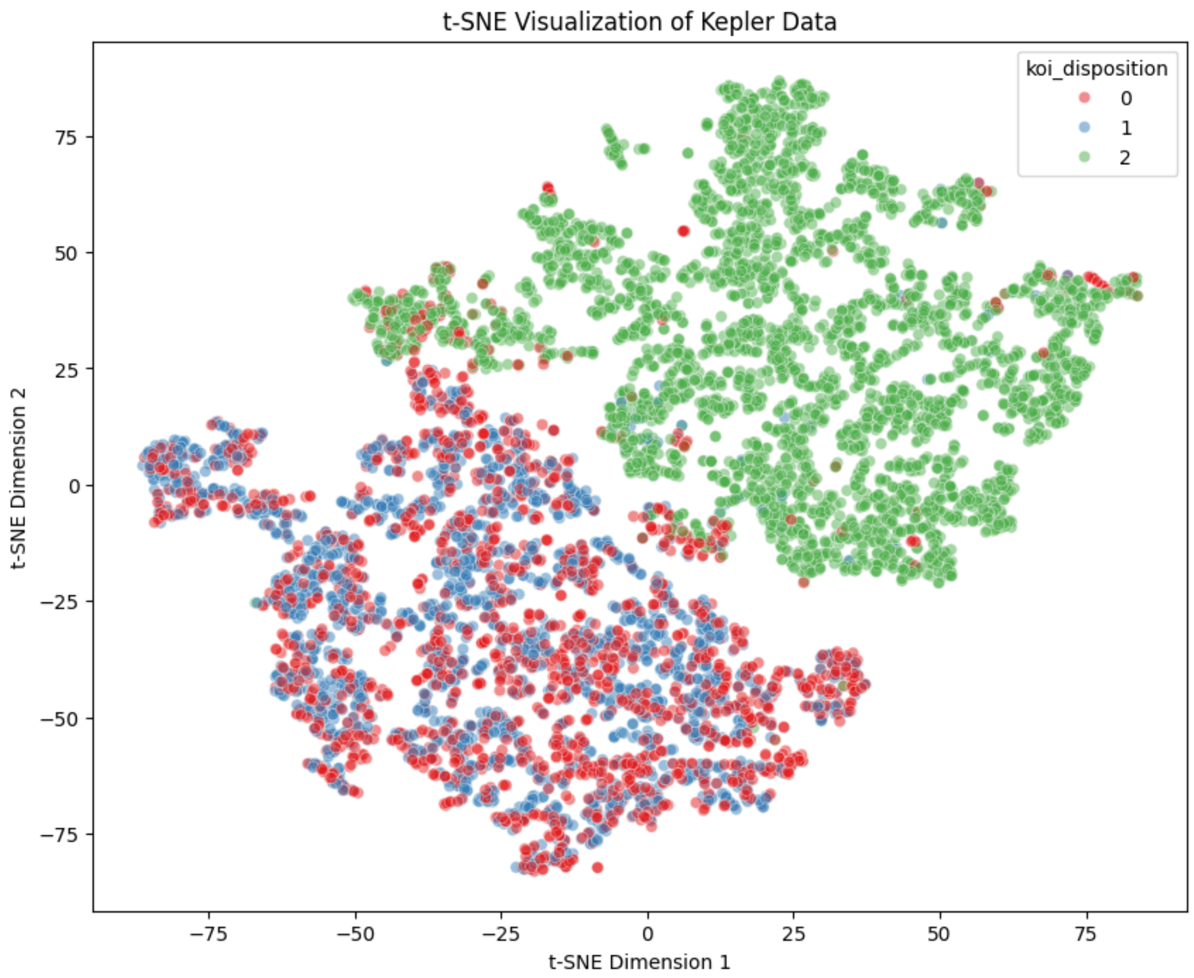

Figure 2: t-SNE Visualization of Kepler Data.

share similar characteristics that are relatively unique to them compared to false positives. It indicates consistency in the data features that characterize confirmed exoplanets, potentially related to their orbital periods, planetary radii, or other measurable astrophysical properties that differentiate them from unconfirmed or falsely identified objects.

Red points are spread widely across the plot, overlapping significantly with the confirmed exoplanets. This spread indicates a variability in the characteristics of candidate exoplanets, reflecting the uncertainty and diversity in this group. Candidates might have features that mostly align closely with confirmed exoplanets and other times with false positives, highlighting the indeterminate nature of their current classification.

Since we now know that each class of the exoplanets shares properties, we implement a Random Forest model for classification. To capitalize on the distinct properties shared by confirmed exoplanets, this model predicts planetary dispositions based on a comprehensive set of features including planet size, orbital data, and host star characteristics. Importantly, the confidence score is excluded from the feature set to ensure that the model's performance is evaluated based solely on the raw predictive power of physical measurements without any pre-calculated confidence scores.

Overall, the model correctly predicts the class 71% of the time across all predictions made. In the analysis of individual classes of the classification model, the results for each class indicate varying levels of effectiveness. For the "Candidate" class, the model achieves a precision of 0.52, meaning it correctly identifies candidates 52% of the time, with a recall of 0.39, indicating that it correctly identifies 39% of all actual candidates. The F1-score of 0.45 suggests moderate effectiveness, potentially hampered by overlapping characteristics with other classes. For the "Confirmed" category, the precision is higher at 0.66, with a recall of 0.75, showing that the model is quite effective at detecting confirmed exoplanets, as reflected by a relatively high F1-score of 0.70. This is highly useful as the presence of confirmed exoplanets can be extended to check for potential habitation. Lastly, the "False Positive" class shows the strongest performance with a precision of 0.80 and a recall of 0.85, indicating the model's strong capability in correctly identifying non-planetary phenomena as false positives. This category's F1-score of 0.82 underscores its robustness, crucial for efficient resource utilization.

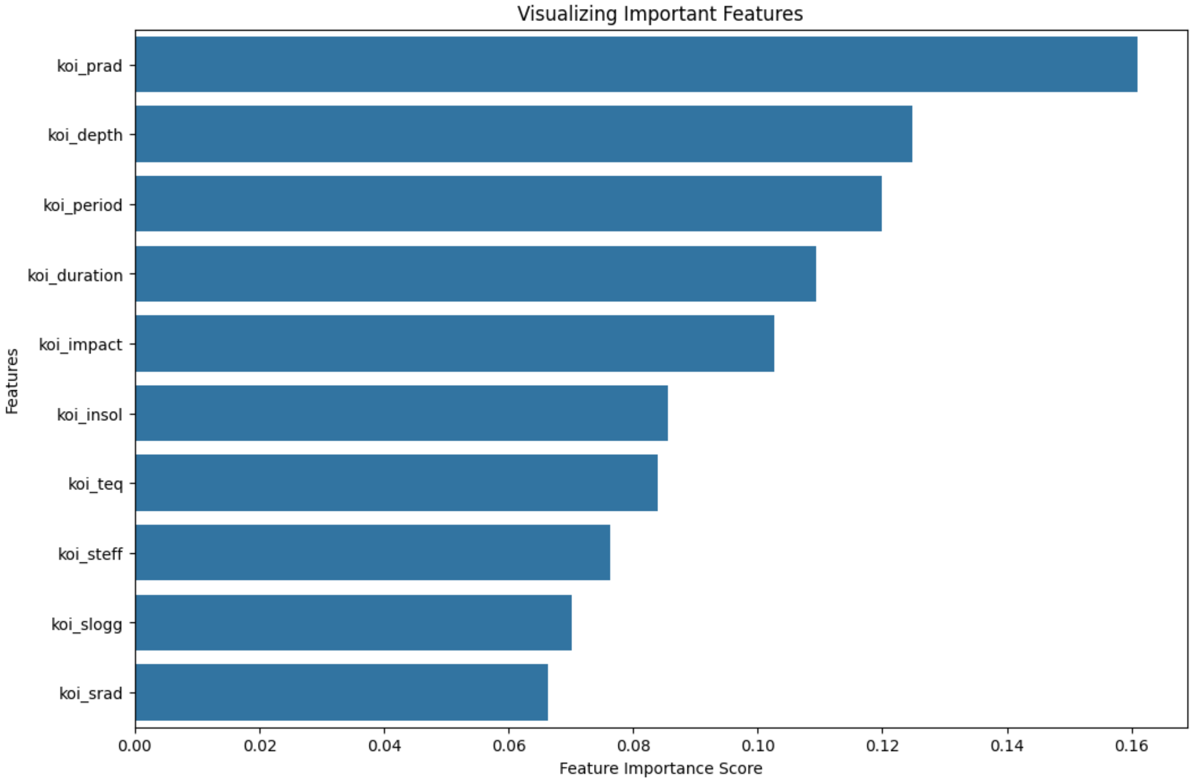

Figure 3: Feature Importance Graph.

The feature importance graph (Figure 3) from the trained model provides valuable insights into which features are most influential in predicting the disposition of the KOI. Planet Radius has the highest importance score, suggesting that the size of the planet is a significant predictor of its classification status. Larger planets might be easier to confirm or distinguish from false positives. The depth of the transit, which measures the amount by which the star's light dims when the planet passes in front of it, is also highly influential. This aligns with expectations, as deeper transits generally indicate larger planets or more significant blocking of the star's light. [5] The time it takes for the planet to complete one orbit around its star is crucial, likely because it relates to the planet's distance from the star and can influence the temperature and environment of the planet. The other features are less critical but still contribute to predictions, likely providing context about the planet's environment and its host star's characteristics.

Evaluating the confusion matrix (Figure 4), the model shows the strongest performance in identifying false positives, indicating effective discrimination of non-planetary phenomena which is crucial for reducing resource expenditure on non-viable targets. It performs relatively well in identifying confirmed exoplanets, but there are still some misclassifications, particularly confusing them with candidates. The candidate class shows the weakest performance in terms of precision and recall, with substantial confusion between the other two classes, reflecting the inherent uncertainty and variability within this group.

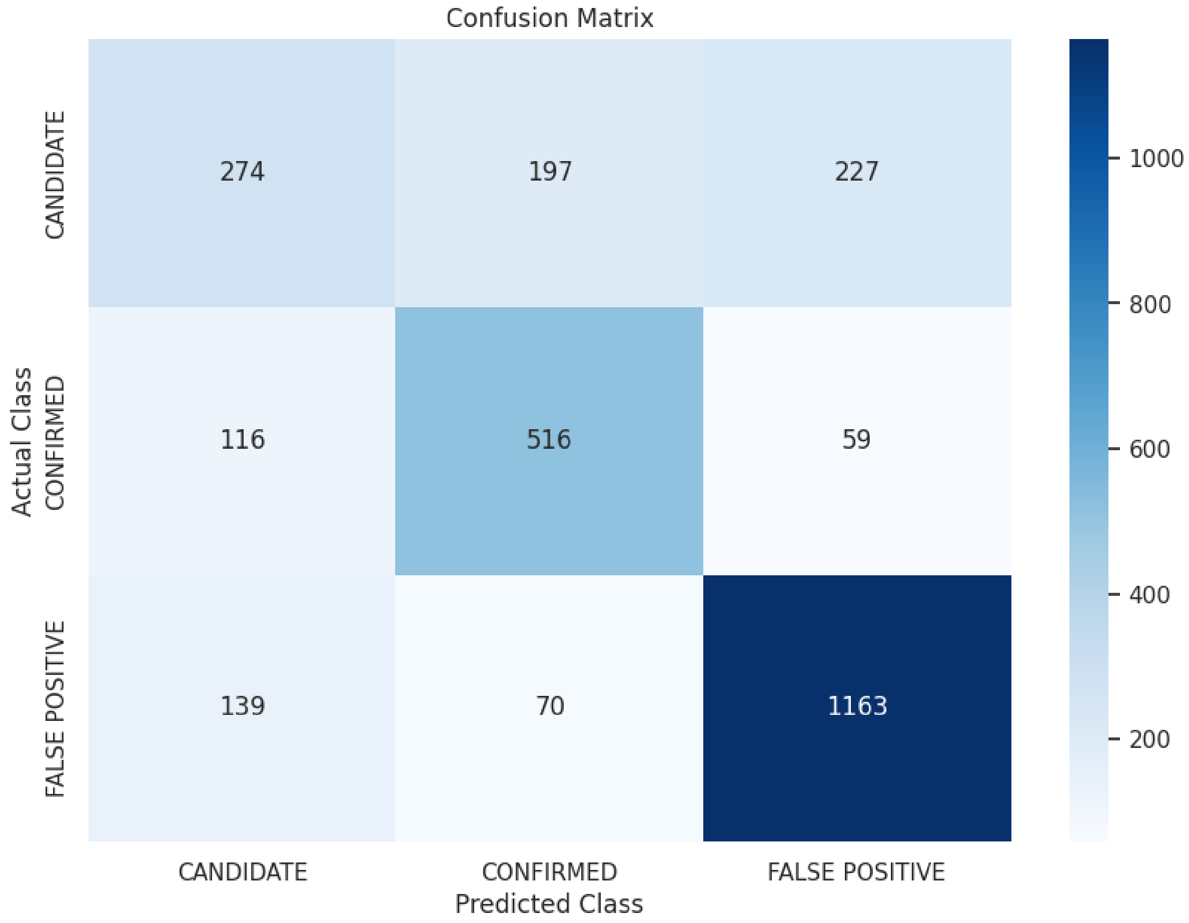

Figure 4: Confusion Matrix.

Comparison of Applications t-SNE is an unsupervised non-linear dimensionality reduction technique used for exploring and interpreting high-dimensional data by reducing it to two or three dimensions, while attempting to preserve the relative distances between points [6]. Its strength lies in its ability to help visually identify clusters and patterns that might not be obvious in the high-dimensional space, making it particularly useful for exploratory data analysis in complex datasets like that of the KOI. This makes t-SNE most suitable for preliminary data exploration where the objective is to generate hypotheses or understand the data structure visually, not for making predictions.

Random Forest, on the other hand, is a supervised machine learning technique used for classification or regression tasks [7]. It uses multiple decision trees to make predictions. It is highly valued for its ability to handle large datasets with higher dimensionality and its ability to provide importance scores for various features, which helps in understanding the features that contribute most to the output. Random Forest is particularly advantageous for its performance in diverse and imbalanced datasets, as it offers detailed performance metrics like accuracy, precision, recall, and F1-score, which are crucial for evaluating model efficiency in real-world scenarios.

The suitability of each technique depends on the specific needs of the analysis: t-SNE is more appropriate for data exploration and visualization to identify inherent data groupings, while Random Forest is better suited for tasks requiring reliable predictions and statistical analyses. In terms of performance, Random Forest generally outperforms other models in predictive tasks due to its ensemble approach, which helps in reducing variance and avoiding overfitting. While t-SNE provides a visual clustering of the KOI dispositions and generates the hypothesis that the class of confirmed exoplanets shares properties, the Random Forest classifier predicts the KOI dispostion with a 71% accuracy.

References

[1] NASA, *Kepler Exoplanet Search Results*, Kaggle, revision 2017, https://www.kaggle.com/datasets/nasa/kepler-exoplanet-search-results/data.

[2] Caltech, *Kepler Objects of Interest Table*, NASA Exoplanet Archive, 2022, https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=koi.

[3] Caltech, *Exoplanet Bibliography*, NASA Exoplanet Archive, 2022, https://exoplanetarchive.ipac.caltech.edu/docs/exobib.html#2019.

[4] ArXiv, *arXiv e-prints*, ADS, 2022, https://ui.adsabs.harvard.edu/abs/2022arXiv220709665A/abstract.

[5] Las Cumbres Observatory, *Exoplanets: The Transit Method*, LCO SpaceBook, 2022, https://lco.global/spacebook/exoplanets/transit-method/#:~:text=Like%20the%20radial%20velocity%20method,big%20planets%20around%20small%20stars.

[6] L.J.P. van der Maaten and G.E. Hinton, *Visualizing Data using t-SNE*, Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008, https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[7] ScienceDirect, *Random Forest*, ScienceDirect Topics in Engineering, 2023, https://www.sciencedirect.com/topics/engineering/random-forest.