# Can we interpret machine learning? An analysis of exoplanet detection problem


## Abstract

The exoplanet detection problem - planets that orbit a star outside our Solar System - has focused on the use of time-consuming manual process. Now, the promising techniques are machine learning methods. However, the lack of interpretability in order to understand what the models does, has avoided the improvement and development of the models. In this work, we study the use of classical machine learning methods for detecting confirmed objects on the Kepler mission. Using metadata from the objects and hand-crafted features from the *light curves*, our study shows that approximately 93% of the data is correctly detected. The extreme behavior of non-exoplanet objects facilitate the recovery of mostly all these objects (high *recall*), however our work presents difficulties with confirmed objects overlapping with the non-exoplanet objects (low *precision*). Because of this, we provide some insights about where the error could be in order to interpret the learning process of our proposal.


## 1. The exoplanet detection problem

- Challenging task! Exoplanets emit or reflect very dim magnitudes of light compared to their host star and they are very near to them.
- Fine-grained and time-consuming analysis techniques are needed.
- Automatic methods such us machine learning has emerged.
- Good performance but they have some limitations → progress stagnated.

**Our focus:** Machine learning techniques lack of interpretability so we explore on what the model learns and does.

## 2. Data

Given the documentation and effectiveness of the mission, we used the Kepler dataset and focused on the transit method. This is the largest labeled dataset on exoplanet detection which is composed of *Kepler Object of Interest* (KOI). A KOI - candidate object - can be confirmed as exoplanet, rejected based on additional evidence or still be under study. We only used the KOI objects that have their *light curves* - light intensity measurements of a celestial object - available i.e. 2281 *Confirmed*, 3976 *False Positive* and 1797 *Candidate*. Besides, each KOI has metadata related to the study itself.

## 3. Models and Methods

Knowing some metadata of the transit object, a *light curve* fit can be done to get a smooth version of it. We focus on generate different representations based on the Mandel-Agol [1] fit on a *light curve*:

- We applied a *Fourier Transform* to the long *light curves* and extract automatic learning features with two techniques: *PCA* and *ICA*.
- We use hand-crafted techniques (summary statistics as [2, 3]). These features were concatenated to the metadata.
- To reduce and simplify the manual features, we generate a new representation based on a feature selection technique called FSS [4].

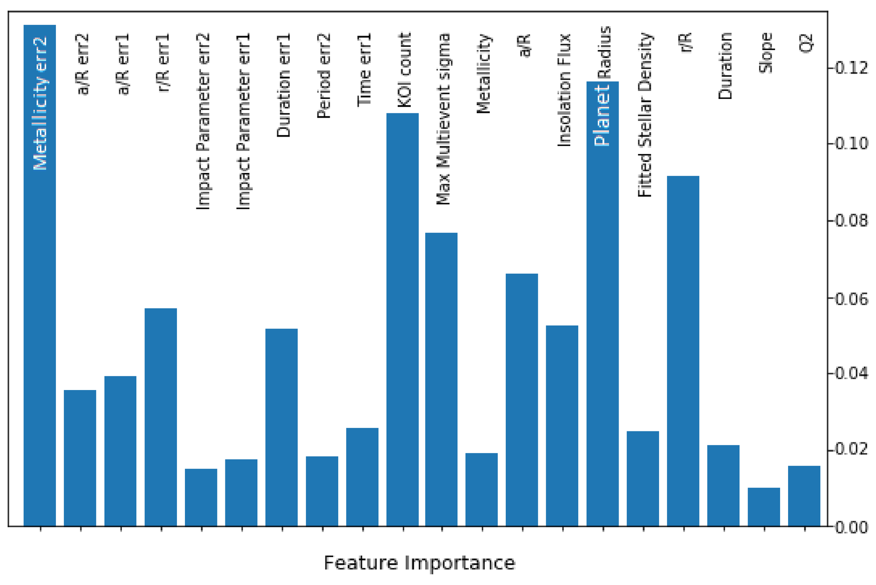
We use a Random Forest model (RFM), as previous work used [5, 6].

## 4. Results

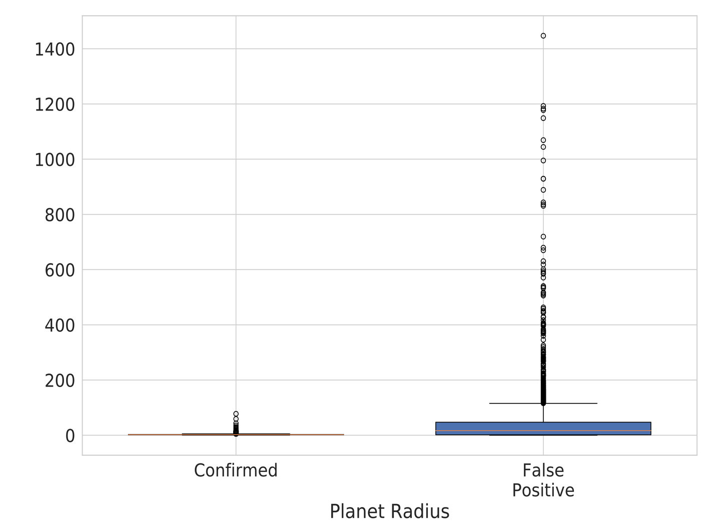
- Unbalanced binary classification problem!
- Metrics: *precision* (P), *recall* (R) and *f1 score* (F1) for each class. F1 is also averaged at macro, micro and weighted.
- We create a test set for evaluation and use cross-validation for model selection.
- Select 20 features over the manual representation is quite meaningful in order to recognize patterns effectively, reaching the best values on results.

| | | *Confirmed* | | | *False Positive* | | | *Global* F1 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Representation** | *d* | P | R | F1 | P | R | F1 | Macro | Micro | Weight |
| Fourier-PCA on Mandel-Agol | 25 | 63.04 | 80.40 | 70.57 | 89.25 | 77.55 | 82.99 | 76.83 | 78.47 | 79.01 |
| Fourier-ICA on Mandel-Agol | 25 | 64.57 | 77.28 | 70.33 | 88.03 | 79.81 | 83.72 | 77.02 | 78.98 | 79.40 |
| Metadata + Manual features | 72 | 86.46 | 92.28 | 89.27 | 96.20 | 93.11 | 94.63 | 91.96 | 92.84 | 92.90 |
| (FSS) Metadata + Manual features | 20 | **87.78** | **92.48** | **90.07** | **96.32** | **93.87** | **95.08** | **92.57** | **93.42** | **93.46** |

**Table:** Different evaluation metrics on the classification of test set by the RFM. The results on all the generated representations are shown, where *d* represent the number of features.



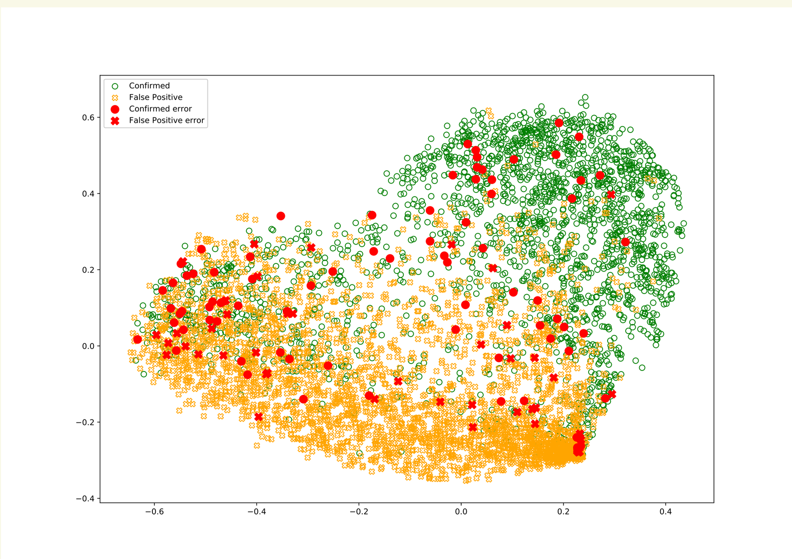

**(a)** Feature importance score, based on RFM ensemble.



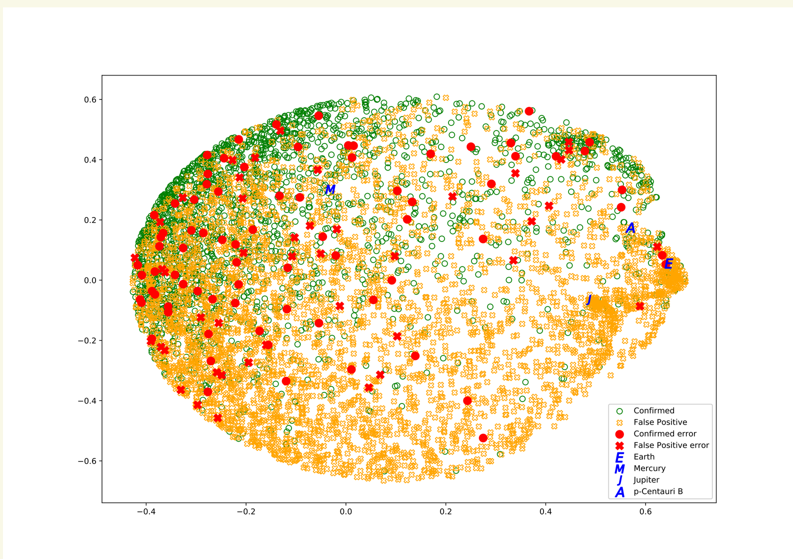

**(b)** Box-plot of planet radius feature.

**Figure:** Feature Analysis.

## 5. Interpretation



**(a)** Best representation features (FSS).



**(b)** Available features of our solar system planets, like radius and temperature.

**Figure:** Data visualization (Kernel PCA). Predictions errors are highlighted in red.

**Based on metrics:** *False Positive* is easier to detect than *Confirmed* over all representations, having always high P and F1. We also detect that is difficult to reach a good P on *Confirmed* class; predictions of this class are usually contaminated.

**Based on feature analysis:** The most important features are related to measurement errors. *False Positive* class has quite extreme values and high variability on planet radius with respect to *Confirmed* class, that clarifies the good detection of this class by the model and the impact if this feature by itself.

**Based on data visualization:** We can see that in our best representation there is still an overlap among classes that makes our model fail. On other side, several exoplanets similar to Mercury has been discovered thanks to the closeness to her star. However, there is still a gap of discovered object on Earth-like and Jupyter-like planets, i.e. relatively big and far from its host star.

## 6. Conclusion

- Careful selection over the manual features is the best representation for the learning task.
- There is still a gap on improve the learnable extracted information from the raw *light curves*.
- Our model fails in the region close to the decision boundaries, showing that there is still some overlap among classes that could be improved.
- The discovery of more exoplanets far of its host star could better explain the class behavior and improve the results.


**Acknowledgements**. The authors would like to thanks the *Programa de Iniciación Científica* PIIC-DGIP of the Federico Santa María University for funding this work.


## 7. References


- K. Mandel and E. Agol, “Analytic light curves for planetary transit searches,” *The Astrophysical Journal Letters*, vol. 580, no. 2, p. L171, 2002.
- J. W. Richards *et al.*, “On machine-learned classification of variable stars with sparse and noisy time-series data,” *The Astrophysical Journal*, vol. 733, no. 1, p. 10, 2011.
- C. Donalek *et al.*, “Feature selection strategies for classifying high dimensional astronomical data sets,” in *Big Data, IEEE International Conference on*, pp. 35–41, 2013.
- R. Caruana *et al.*, “Greedy attribute selection,” in *Machine Learning Proceedings*, pp. 28–36, Elsevier, 1994.
- S. D. McCauliff *et al.*, “Automatic classification of kepler planetary transit candidates,” *The Astrophysical Journal*, vol. 806, no. 1, p. 6, 2015.
- M. Bugueno, F. Mena, and M. Araya, “Refining exoplanet detection using supervised learning and feature engineering,” in *XLIV Latin American Computer Conference (CLEI)*, pp. 278–287, IEEE, 2018.



**Gabriel Molina, Francisco Mena, and Margarita Bugueño**

{gabriel.molina.12, francisco.mena.13, margarita.bugueno.13}@sansano.usm.cl - https://github.com/FMena14/ExoplanetDetection

Universidad Técnica Federico Santa María, Chile


ADASS XXIX