Small Exoplanet Classification with Machine Learning

Lorraine Nicholson
Term Project, 4/26/2023

Context

- The radius valley separates exoplanets into two classes: super-Earths and sub-Neptunes

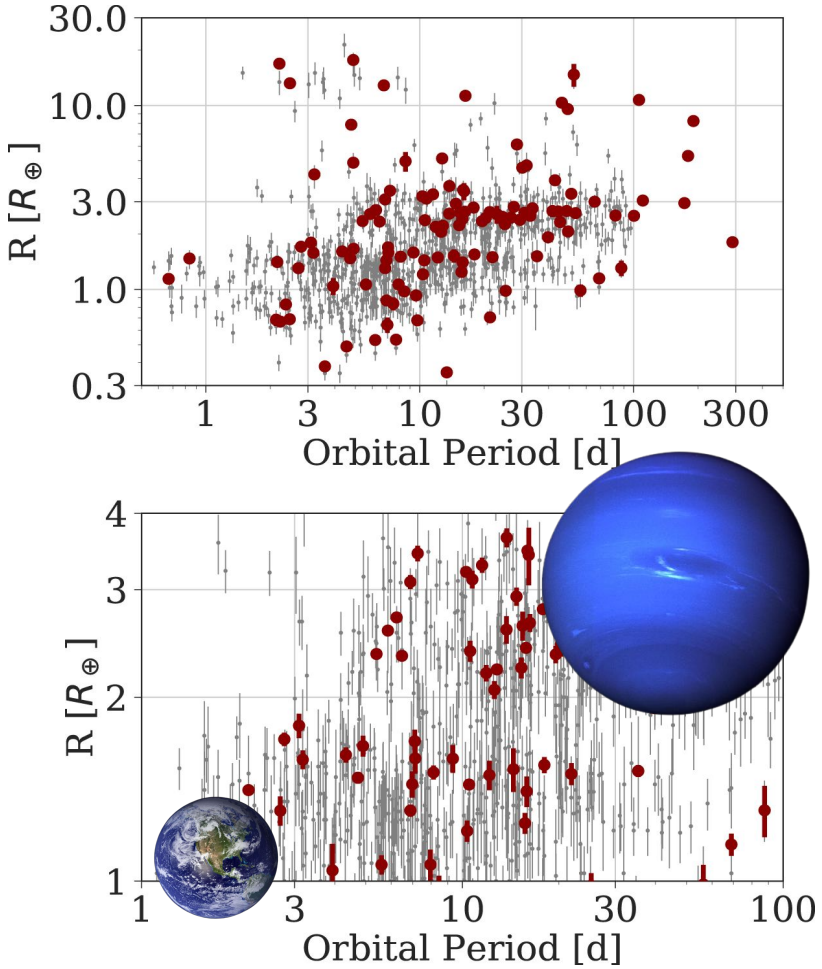


Figure: Van Eylen et. al. 2018

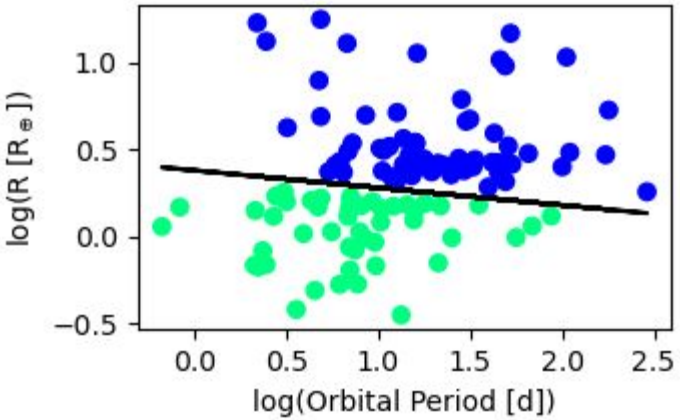
Context

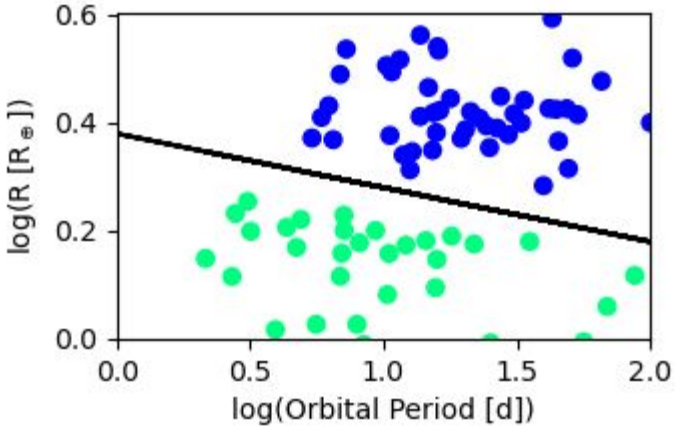
The equation of the radius valley as determined by SVM is:

$$\log_{10}(R_P) = -0.10 \log_{10}(P) + 0.38$$

SVM did a good job at predicting the slope, but **can unsupervised clustering algorithms also solve this problem?**

Spoiler alert: NO





Clustering algorithms set-up & hyperparameters

Training dataset is the planetary observations reported in Van Eylen et. al. (2018).

K-means Clustering

- Number of clusters = 2

Gaussian Mixture Model

Model 1 (terrible):
- Number of clusters = 2

Model 2:
- Number of clusters = 3

DBSCAN

- ε = 0.40
- Min_samples = 5

(trial and error)

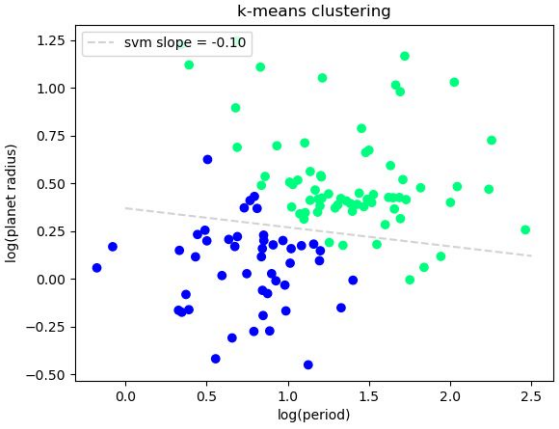
Train-test split to get accuracies

Two criteria to determine the goodness of each model

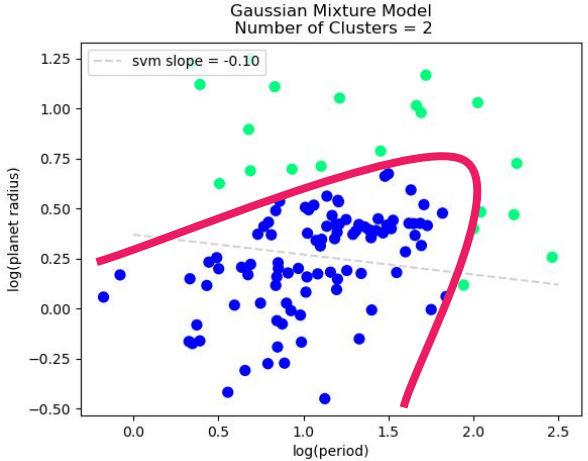
---

1. Test accuracy
2. Shape of the “radius valley” (qualitatively)

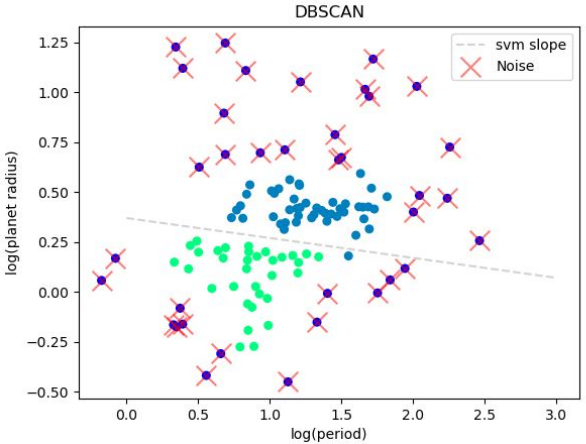
Unsupervised clustering algorithms didn't do a good job



Testing accuracy ~ 96%

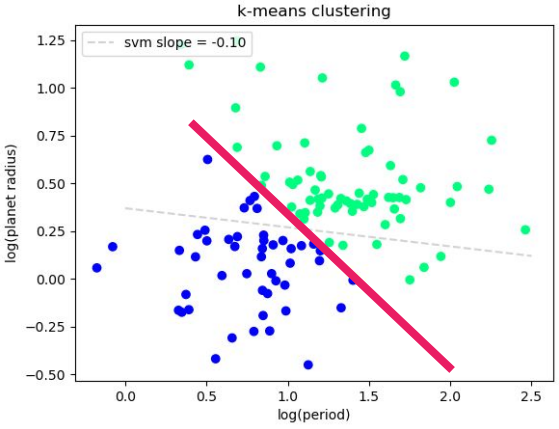


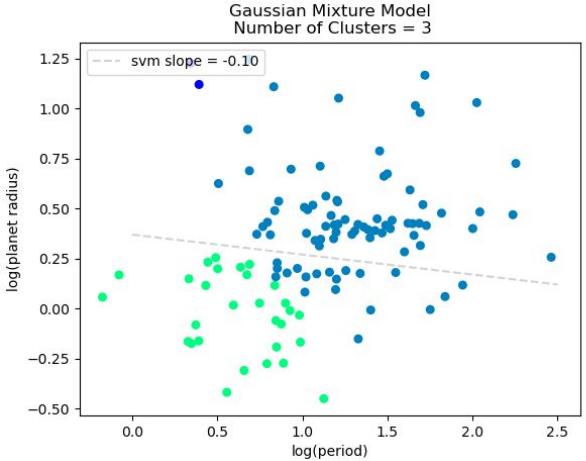
Testing accuracy ~ 55%

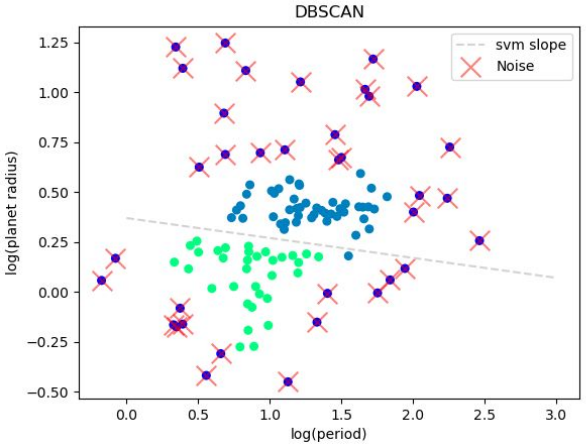


Training accuracy ~ 68%

Unsupervised clustering algorithms didn't do a good job





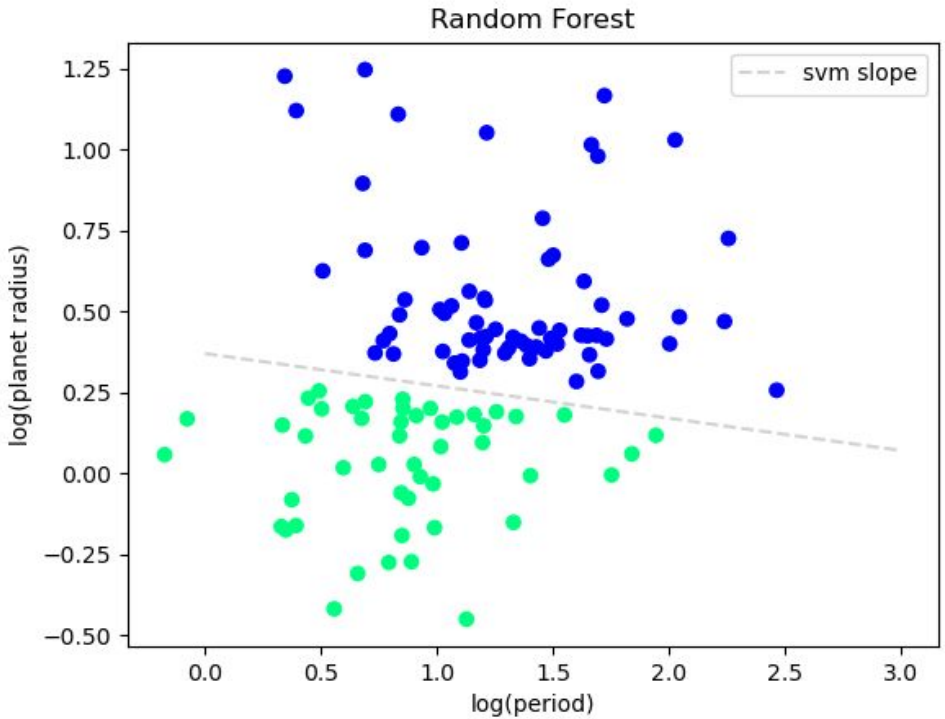


Testing accuracy ~ 96%

Testing accuracy ~ 79%

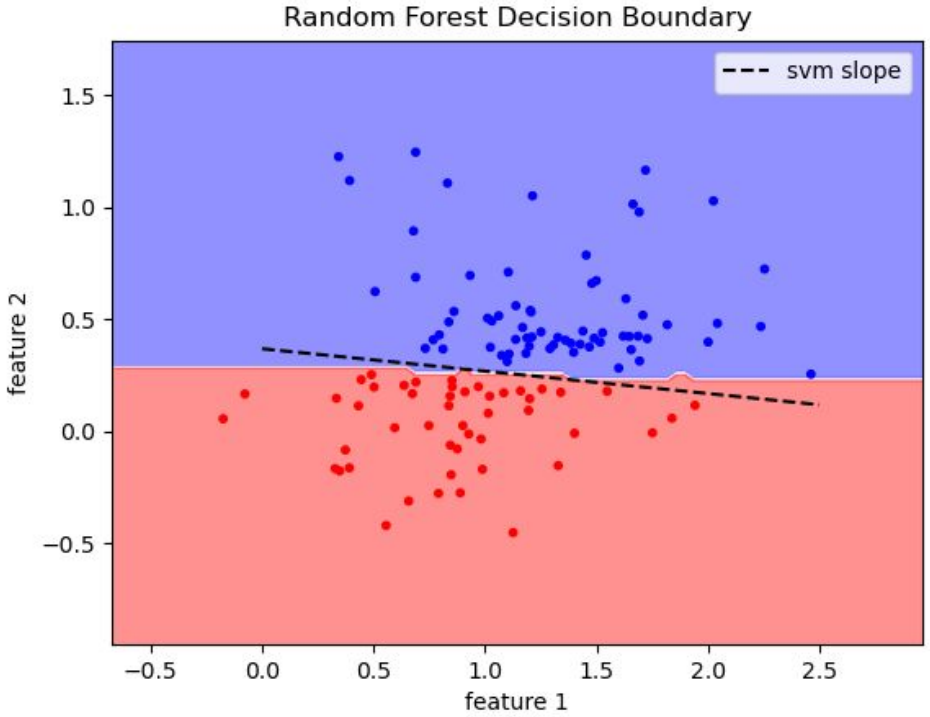
Training accuracy ~ 68%

Random Forest saves the day?




```
Testing accuracy = 100%
Out-of-bag score = 99%
```

Random Forest saves the day?



Testing accuracy = 100%
Out-of-bag score = 99%

Predicts a very shallow radius valley slope.

Things I could've done differently

- Use a larger training dataset: NASA Exoplanet Archive (MacDonald 2019)
  - All known planets with,
    - P < 50 days
    - R < 4.0 $R_\oplus$
- Incorporate bagging into my semi-supervised models?
- I'm curious what would happen if I added some additional features, such as stellar mass

Conclusions

- Machine learning, specifically clustering algorithms, overcomplicate this problem.

- Random forest did a good job at classification but a bad job of recreating the slope of the radius valley.

- Clustering algorithms perform significantly worse than a simple Support Vector Machine method.