Classifying Exoplanet Types Through Machine Learning

Johan Schoultz

December 5, 2023

1 Introduction

This project explores the use of machine learning techniques to classify the planet type for exoplanets based on their recorded features. The ability to accurately identify what type an exoplanet is with machine learning models would provide an immediate deeper understanding into the composition of a newly discovered planet, depicting whether it is mainly gas-based or of a terrestrial type similar to Earth. Discovering this planet type for newly found exoplanets would enable researchers within Astronomy to determine the relevance of that particular planet within current ongoing research.

2 Exoplanet Data and Preprocessing

To address this challenge, this project makes use of a dataset by NASA consisting of 5250 exoplanets [3]. For each planet, the dataset includes: its distance from Earth, stellar magnitude (brightness of the planet), planet type, year of discovery, mass, radius, orbital radius, orbital period, eccentricity and the method of detection with which the planet was found. Out of these attributes the planet type classifies each of the planets, which is of interest to this project. Scientists from NASA have determined four categories that planet types fall under: gas giants, Neptune-like, super-Earths and terrestrial [2]. Of the 5250 planets in the dataset, there are 1825 Neptune-like, 1630 gas giants, 1595 super-Earths and 195 terrestrial planets. For the remaining five planets their category is currently unknown.

The planet data must be preprocessed to be properly prepared before a machine learning model can be applied. Initially, the mass and radius of each planet is recorded as a factor of either Earth's or Jupiter's mass or radius respectively. To ensure the units of measurement is consistent for all entries, these columns are converted such that all planet masses and radii are listed as multiples of the Earth's mass and radius. This is done by using Jupiter's mass being 317.83 times that of Earth, and Jupiter's radius being larger by a factor of 11.209 [4]. The names and discovery years of the planets are then removed as a potential point of bias in the dataset which could risk mistraining the model. Next, K-Nearest Neighbour (KNN) imputation using K=*5* is used to fill any entries in a planet's data which may be empty by using the five planets that are most similar. This allows all planets where the type is known to be preserved in the dataset and assist in training the model. However, the five planets where the classification is not yet known are removed.

3 Supervised Machine Learning

The problem faced encompasses creating a model that classifies planets from a dataset where the planet type is included. Therefore, supervised learning models would be most effective to employ in such situations where the model would learn from the provided labelled data. Specifically, classification models which aim to classify data points into distinct classes are ideal. Hence, a random forest classifier is selected for this project. Random forests have become popular due to their adaptability and ability to achieve highly accurate results [1]. A random forest attempts to separate data entries into their target classes through creating a diverse set of decision trees that each attempt to separate the data. The random forest then searches for a combination that takes into account the classifications made by these trees that best separate the data into the expected classifications. The use of a high number of decision trees allows the model to avoid overfitting the data [1].

4 Classification Performance and Results

After the data has been preprocessed, the random forest model can be applied and its results evaluated. 80% of the dataset is used to train the model and the model is then tested on the remaining 20%. The first model includes all of the planetary features mentioned in section 2, excluding those removed in preprocessing. This model yields significantly well-performing results, achieving an overall accuracy of 98% and high levels of precision and recall, with F1-scores ranging between 0.94-1.00 for classifying each of the planet types in the test set. The radius and mass are evident to be the most important features for this model, together contributing to approximately 80% of the prediction made.

Taking this significant reliance on radius and mass into account, a second random forest model is deployed which only contains these two features as input parameters. This model is an improvement from the previous, achieving an overall 99% accuracy and nearly perfect F1-scores in the range 0.99-1.00. These results suggest that given the radius and mass of a planet, a random forest classifier is highly effective in predicting the type of a planet.

Building on the highly successful results shown from the first two models, a model is then created that includes all features except mass and radius as input. This would show how effectively the planets can be classified where these two parameters are not known. The results prove to be a significant drop in performance. The accuracy of this model is 64%, and only manages to predict 2 of 37 terrestrial planets in the test set correctly. The F1-scores are noticeably lower than the previous models with results of 0.61 and 0.54 for Neptune-like planets and super-Earths respectively. The score is further reduced for terrestrial planets at 0.09, which could be explained by the significantly lower number of terrestrial planets in the dataset. However, for gas giants the model manages to retain an F1-score of 0.81. This suggests an ability to predict planets as being gas giants more effectively than the other planet types.

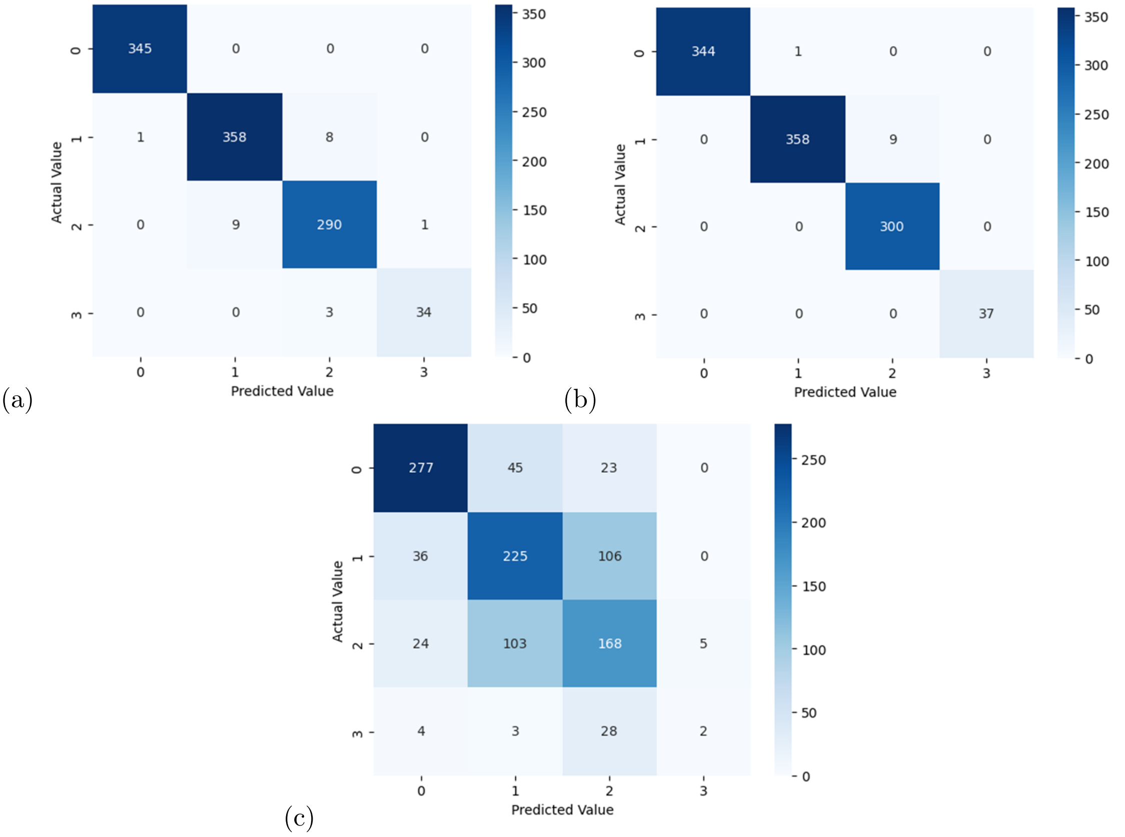
These models can then be used to predict the planet types of the five planets that are unknown in the dataset. The results vary between either Neptune-like and gas giants for each of the planets in each model. However, the classifications from the first two models should be considered less reliable, due to their heavy dependence on the mass and radius of a planet. The mass and radius of these five planets are unknown — which could likely be the reason they are not currently classified — and therefore the first two models make the planet type predictions using imputed values of radius and mass.

However, the results of the third model are more useful due to its independence from planet mass and radius. It predicts all five planets as gas giants, with a probability of 43-44% for three planets and 95% for the two remaining. The lower probability for the first three is likely caused by an additional unknown parameter. The orbital radius being unknown for these also leads to imputed values being used which impact the prediction. In contrast, this model presents an approximate 80% likelihood for the final two being gas giants, due to its higher predictive confidence for these where all parameters are present and its increased performance in classifying gas giants.

5 Limitations and Conclusion

The project is limited by a lower number of terrestrial planets in the dataset which hinders performance in terrestrial classifications. In addition, accurate models being heavily dependent on knowing planet radius and mass limits the ability to classify planets where this information is not known. The use of different machine learning models, such as multi-layer perceptrons, could also be explored further since this project has focused entirely on using random forests with varied input parameters. The reason for this focus is due to the initial success that was found with random forests.

Conclusively, this project shows random forest classifiers are highly accurate in predicting exoplanet types where the mass and radius are known. Without these values, predictions have an increased difficulty, especially for terrestrial planets. On the other hand, gas giants are more reliably classified without mass and radius than other planets and allows us to claim two currently unknown exoplanets to potentially be gas giants with an approximate 80% confidence.



(a)

(b)

(c)

Figure 1: This figure shows the confusion matrices for each of the random forest models employed. The classification results for the initial model that includes all parameters are shown in (a), whereas (b) shows those of the model that only relies on mass and radius as input parameters. The results of the third model which includes all but mass and radius as input are displayed in (c).

References

[1] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

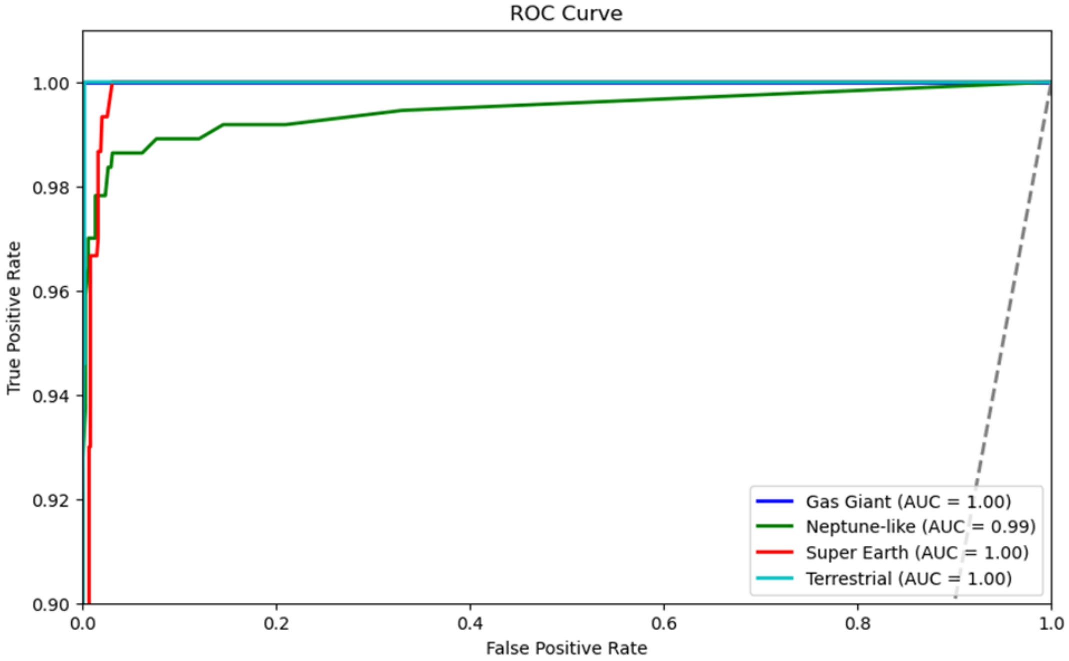
[2] Pat Brennan. What is an exoplanet? - planet types. *NASA*, Apr 2022.

[3] Aditya Mishra. Nasa exoplanets. *Kaggle*, Feb 2023.

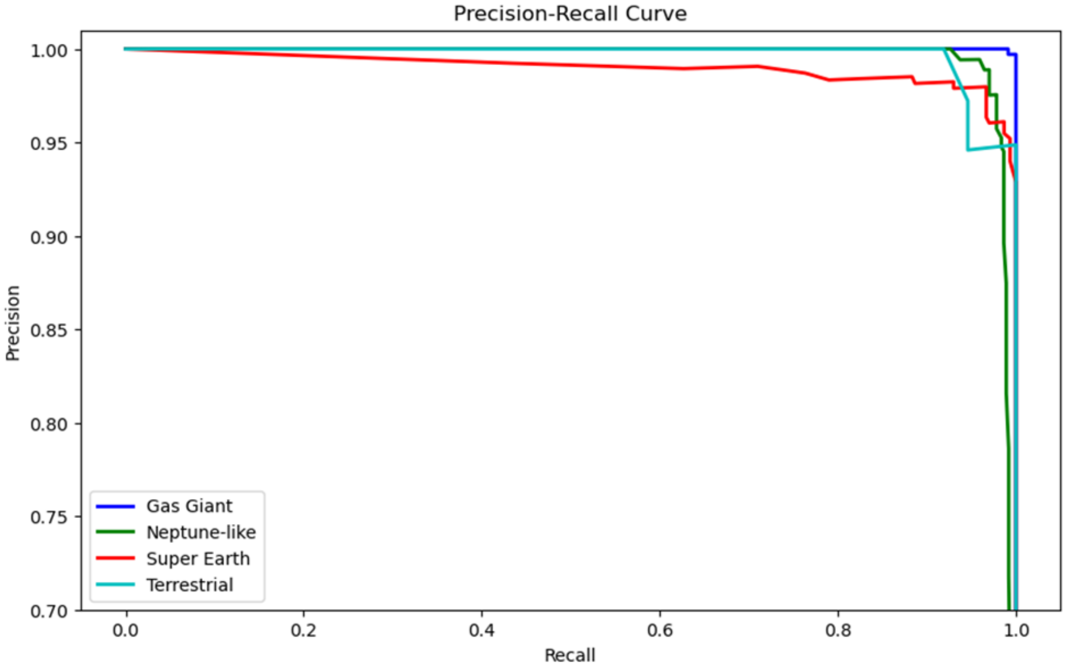
[4] David R. Williams. Jupiter fact sheet. *NASA*, May 2023.

Appendices

Appendix A: ROC curve (a) and precision-recall curve (b) for the first model that includes all features as input parameters.

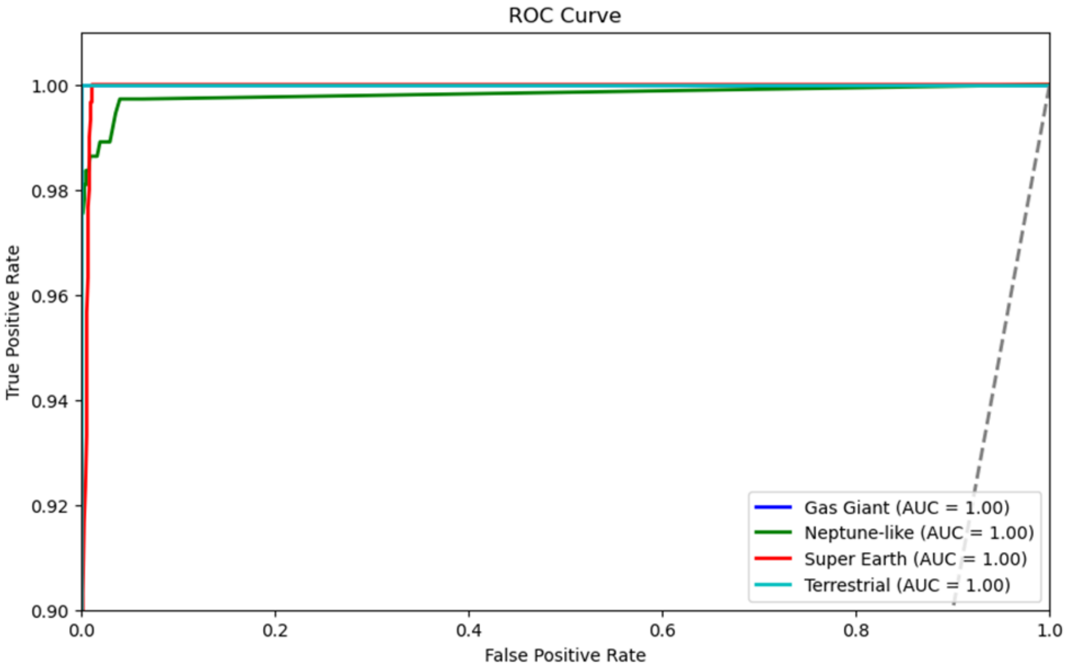


(a)

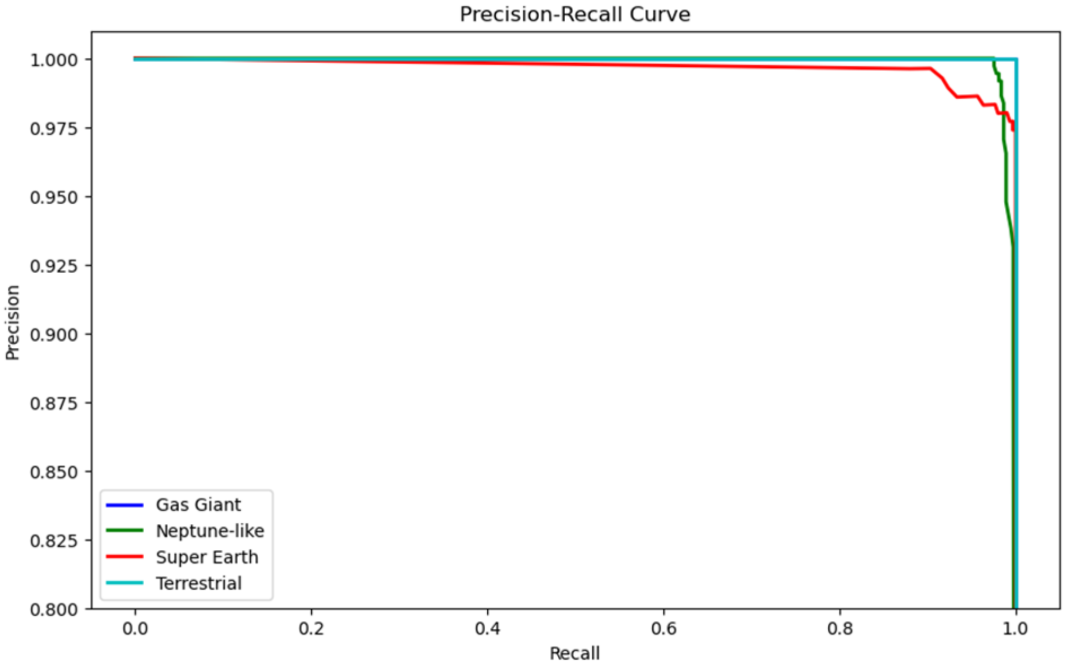


(b)

Appendix B: ROC curve (a) and precision-recall curve (b) for the second model using only mass and radius as input parameters.

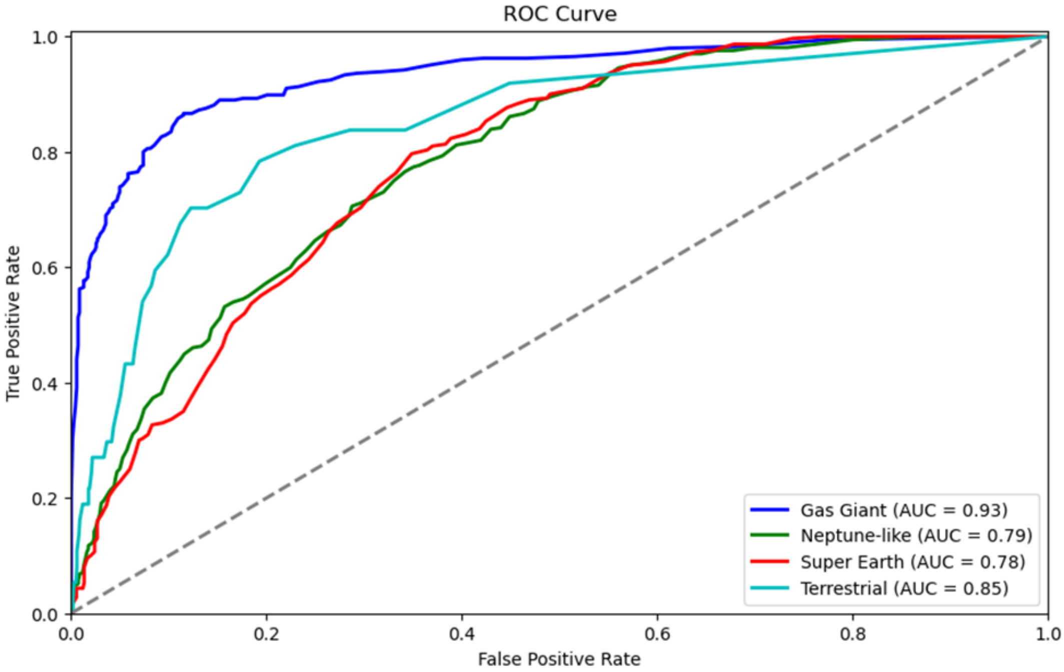


(a)

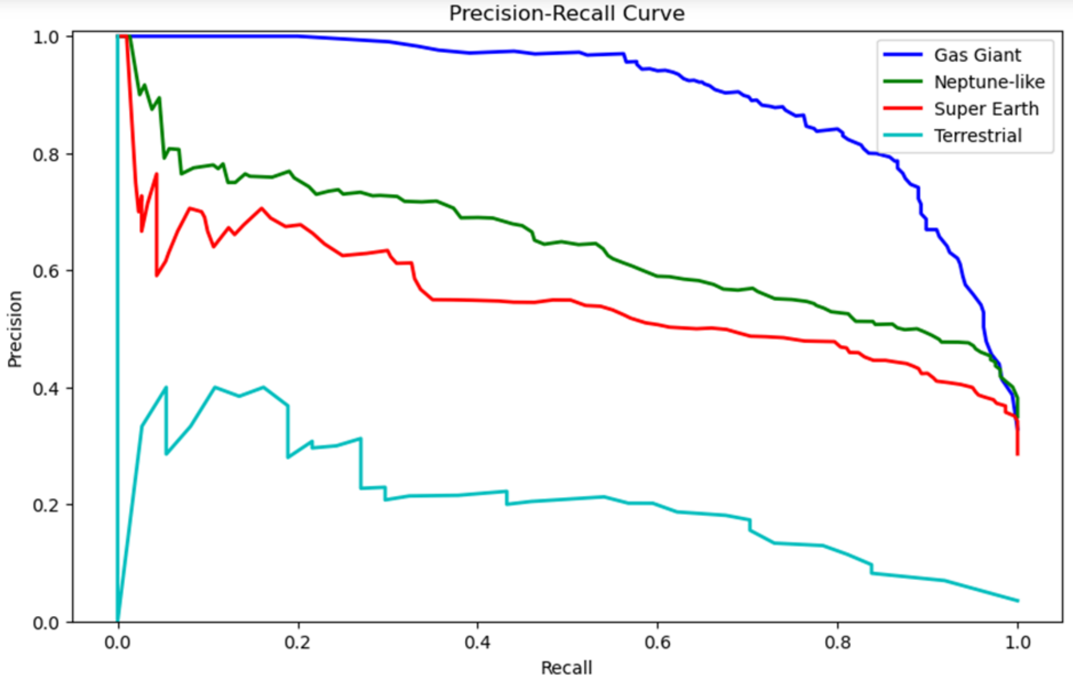


(b)

Appendix C: ROC curve (a) and precision-recall curve (b) for the third model using all input parameters excluding mass and radius.



(a)



(b)