MACHINE LEARNING FOR THE CLASSIFICATION OF PLANETS

Presented by: Paris Martínez
Last Updated: May 23rd, 2022

INDEX

- Objective
- Data Acquisition
- Data Exploration
- Data Preparation
- Feature engineering
- Selection Model
- Results
- Feature Importance
- Conclusion

OBJECTIVE

Search for habitable planets:

- Habitable Zone
- Planet Type
- Geology
- Temperature
- Atmosphere
- Orbit and Rotation

Problem

Machine Learning

- Supervised
  - Regression
  - Classification
- Unsupervised

DATA ACQUISITION

KEPLER MISSION

During just over nine and a half years in orbit, the Kepler space telescope observed more than half a million stars and discovered more than 2,600 planets.

WEB SCRAPING

Data:

- NASA Exoplanet Archive

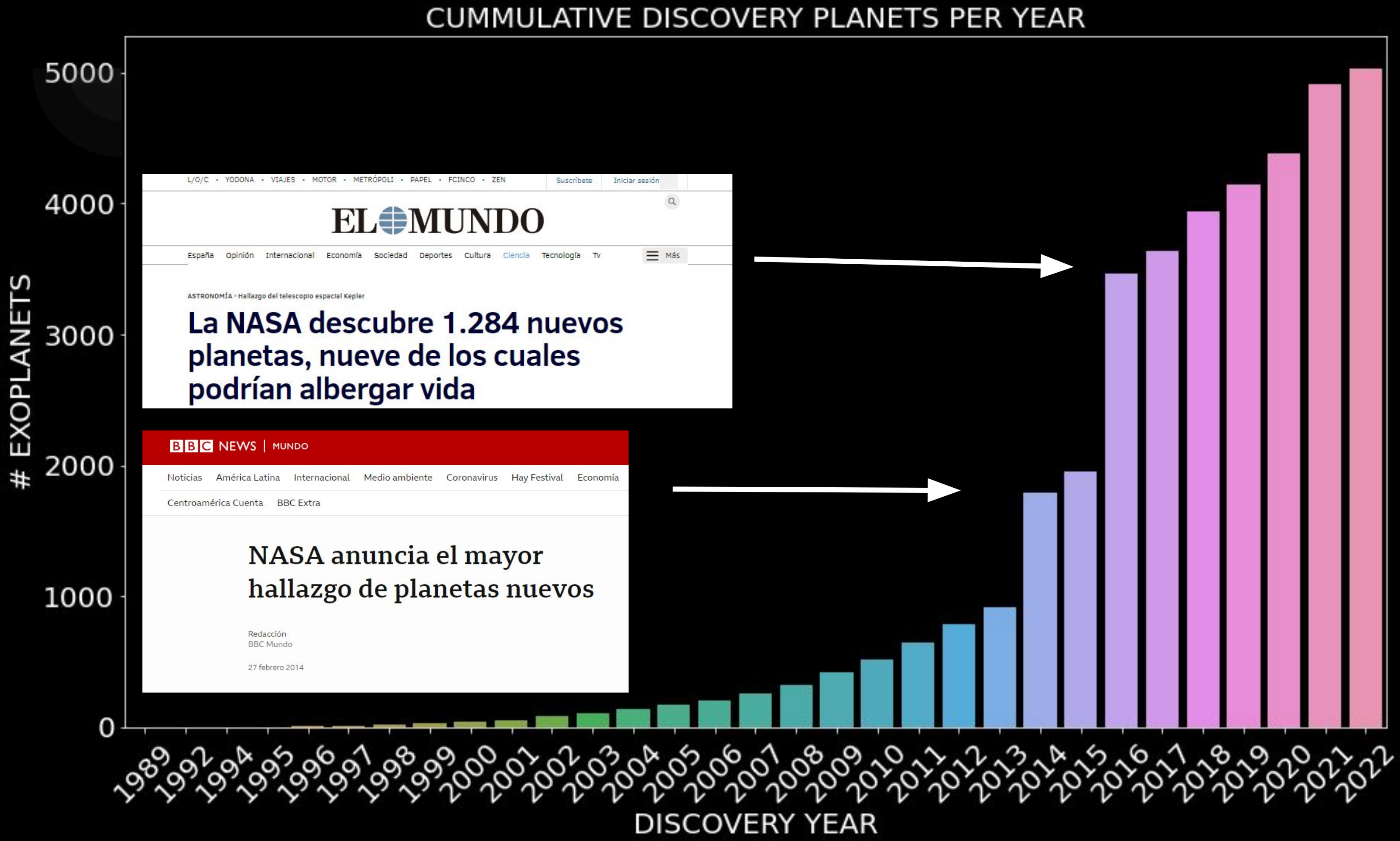
Modules:

- Selenium
- Beautiful Soup

URL:

- https://exoplanets.nasa.gov/discovery/exoplanet-catalog/

DATA EXPLORATION



ASTRONOMÍA • Hallazgo del telescopio espacial Kepler

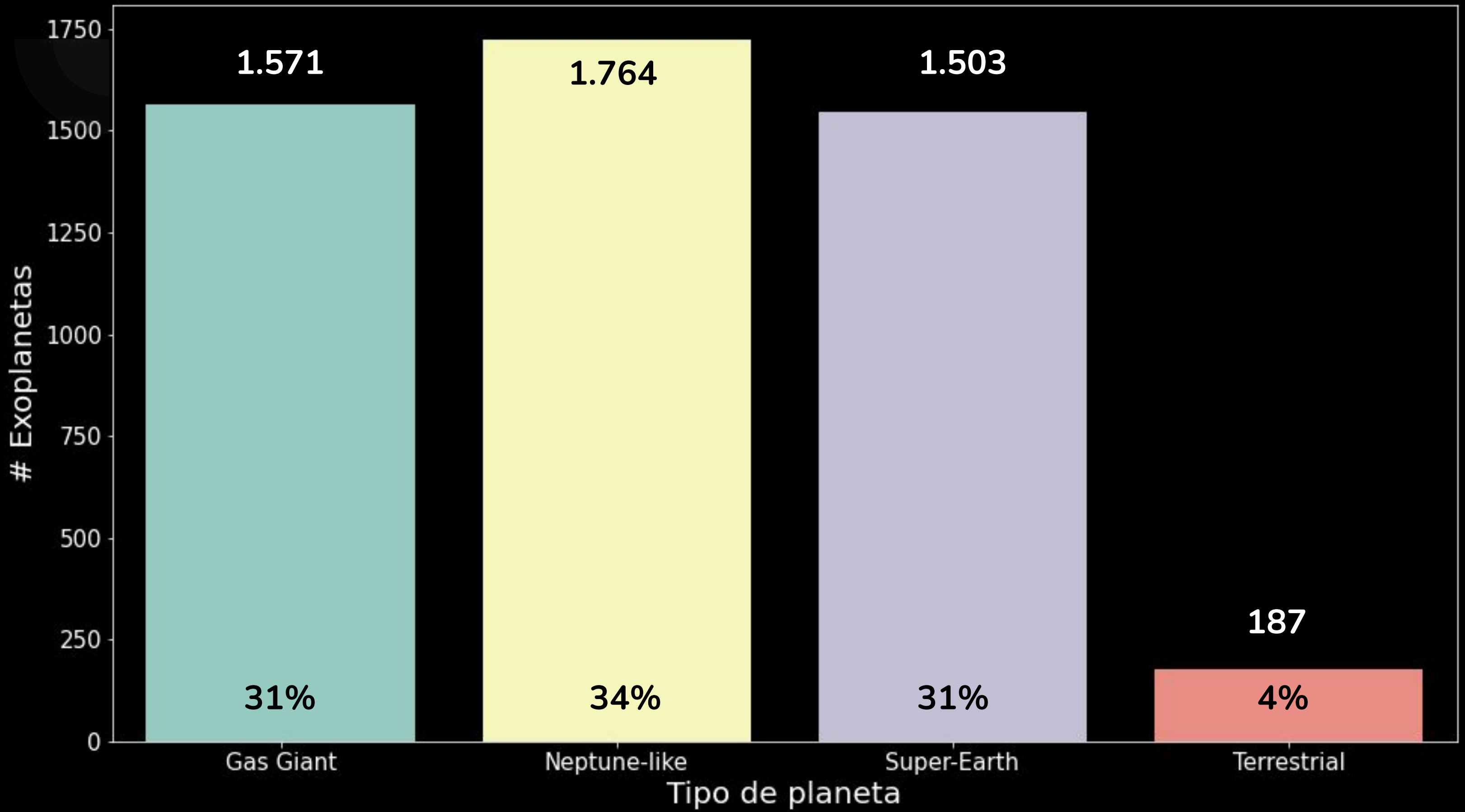
La NASA descubre 1.284 nuevos planetas, nueve de los cuales podrían albergar vida

NASA anuncia el mayor hallazgo de planetas nuevos

Redacción
BBC Mundo

27 febrero 2014

DISCOVERED EXOPLANETS



DATA PREPARATION

- Transform Jupiter mass, radius to Earth
- Drop Categorical columns: names, detection_method
- Drop Unknowns targets: 2 stars, 5 planets
- Encode target column
- Change Infinite with mode in orbital days: 1 planet
- Fill NaN with KNNImpute: 337 NaN

177 eccentricity, 117 stellar_magnitude, 18 mass, 16 distance, 9 radius

FEATURE ENGINEERING

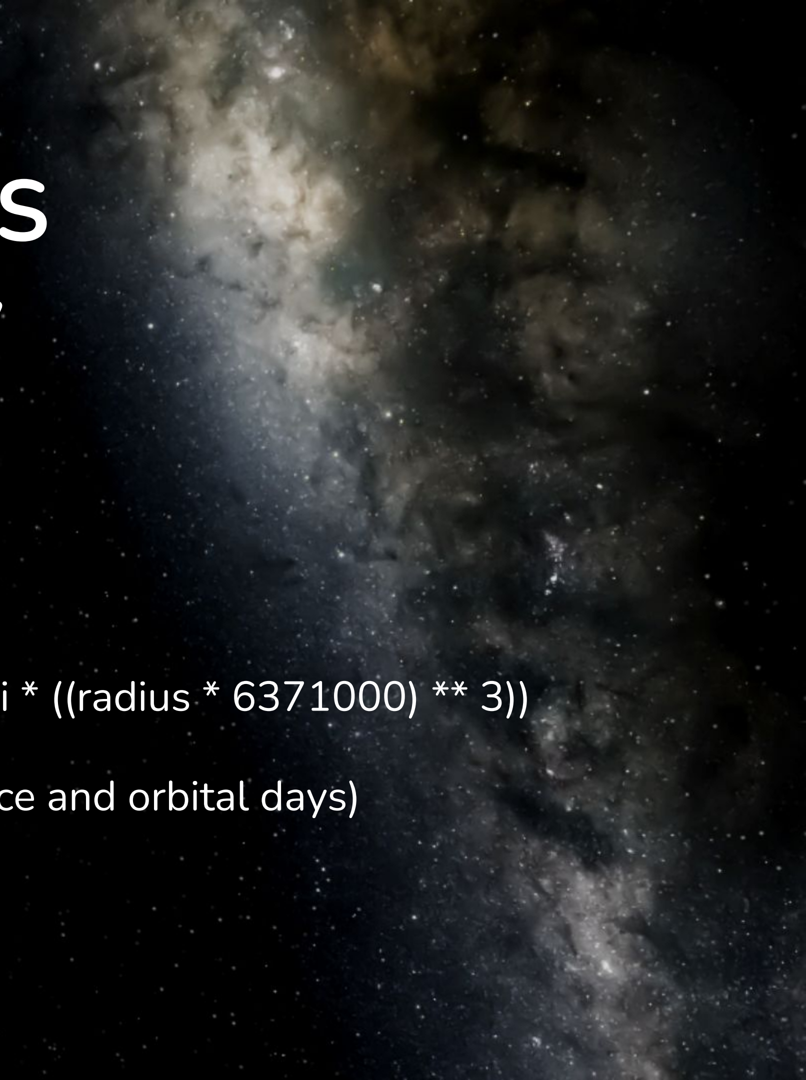
CREATE NEW FEATURES
Density and Log Density

Masa de la Tierra: 5,972 × 10^24 kg

Radio de la Tierra: 6.371 km

Density (Kg/m3) = (mass * (5.972*10**24)) / (4/3 * pi * ((radius * 6371000) ** 3))

Log transform features (mass, radius, density, distance and orbital days)



Features	Log Features
distance	log_distance
mass_E	log_mass
radius_E	log_radius
density	log_density
orbital_days	log_orbital
discovery_year	
stellar_magnitude	
eccentricity	

Target

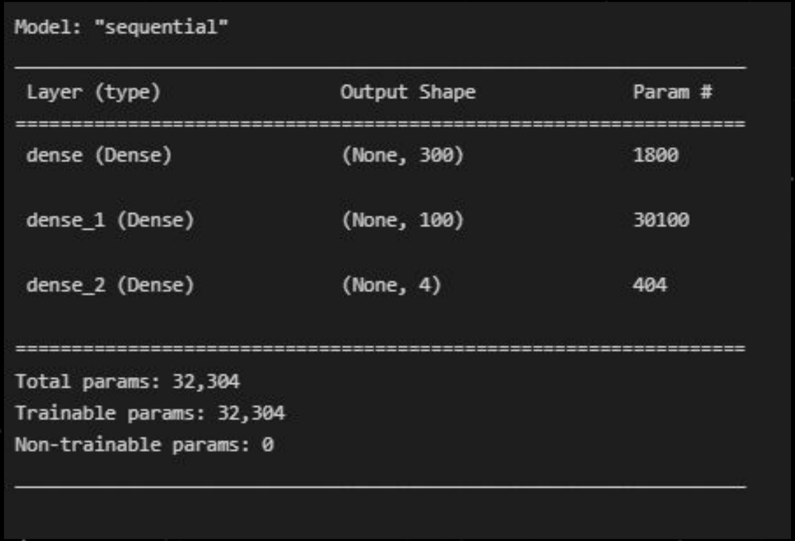
Planet Encode:

- Gas Giant: 3
- Neptune-like: 2
- Super Earth: 1
- Terrestrial: 0

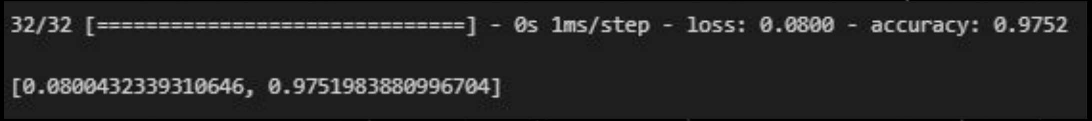
SELECTION MODEL

- Random Forest
- Cat Boost
- Decision Tree Classifier
- Lightgbm
- Multi Layer Perceptron
- Voting(RF+XGB)
- XGBoost
- Ada Boost
- Logistic Regression
- Gradient Boosting
- Deep Learning Model

Deep Learning
Accuracy: 97,5%

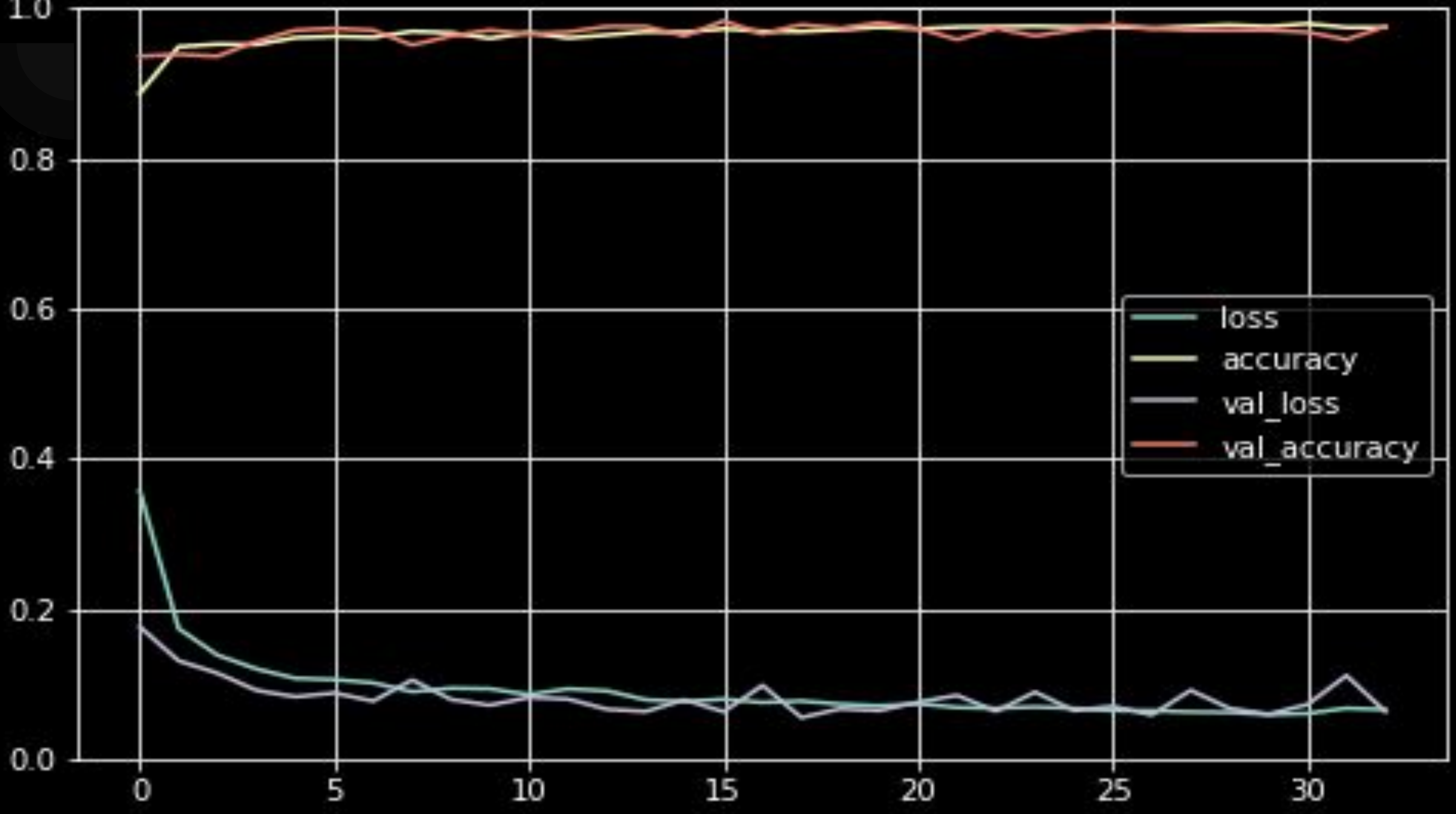


```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 300)               1800

 dense_1 (Dense)             (None, 100)               30100

 dense_2 (Dense)             (None, 4)                 404

=================================================================
Total params: 32,304
Trainable params: 32,304
Non-trainable params: 0
_________________________________________________________________
```



```
32/32 [==============================] - 0s 1ms/step - loss: 0.0800 - accuracy: 0.9752

[0.0800432339310646, 0.9751983880996704]
```

TRAINING


1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
loss
accuracy
val_loss
val_accuracy

VOTING (RF+XGB)

<u>Random Forest</u>

random_state = 42

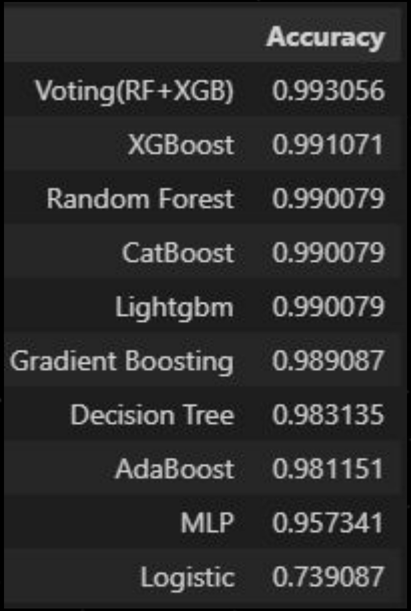
n_estimators = 500

max_leaf_nodes = 16

<u>XGBoost</u>

random_state = 42

Results



	Accuracy
Voting(RF+XGB)	0.993056
XGBoost	0.991071
Random Forest	0.990079
CatBoost	0.990079
Lightgbm	0.990079
Gradient Boosting	0.989087
Decision Tree	0.983135
AdaBoost	0.981151
MLP	0.957341
Logistic	0.739087

Terrestrial Planets

Total: 42

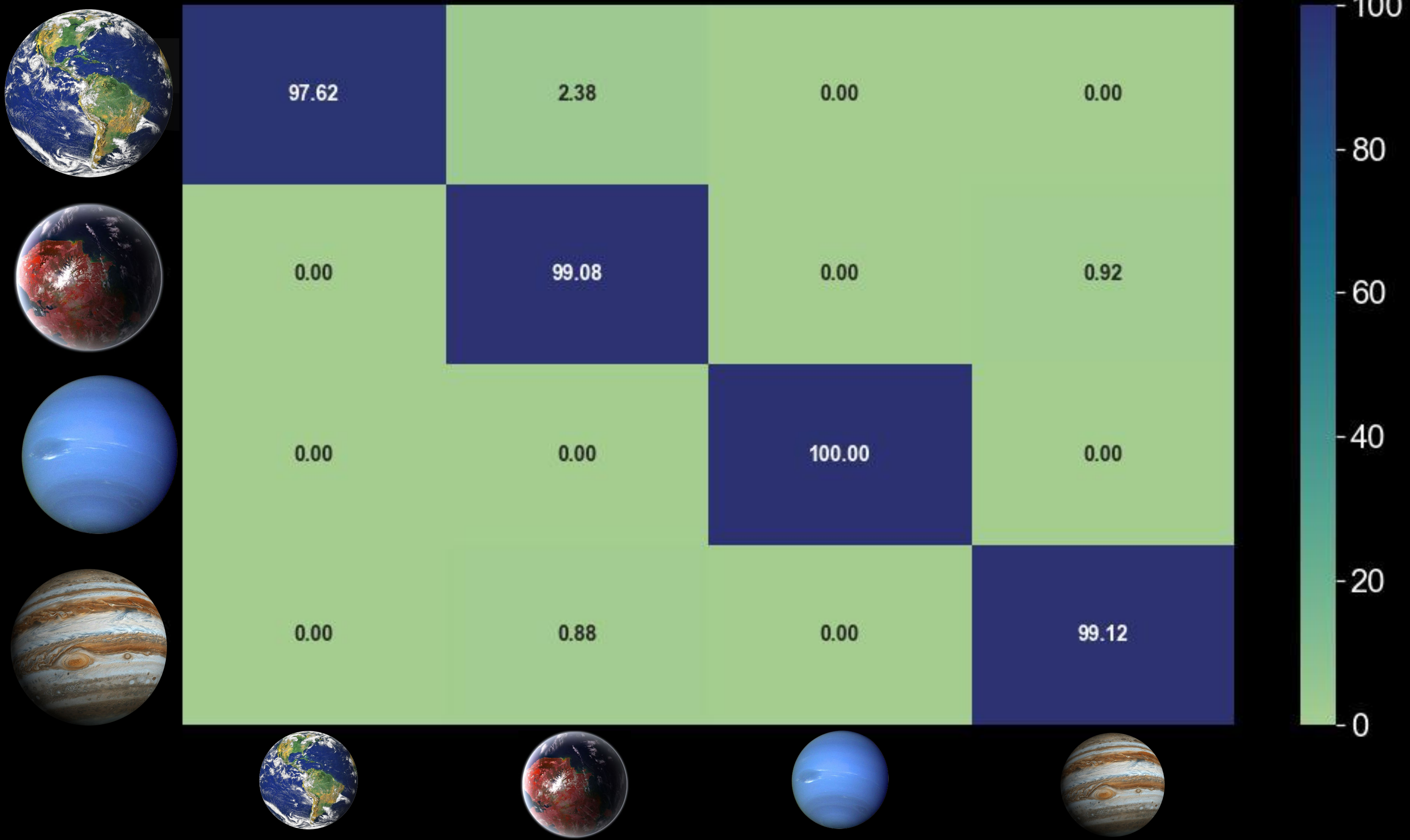
Classified as Terrestrial: 41

Classified as Super-Earth: 1

Classified as Neptune-Like: 0

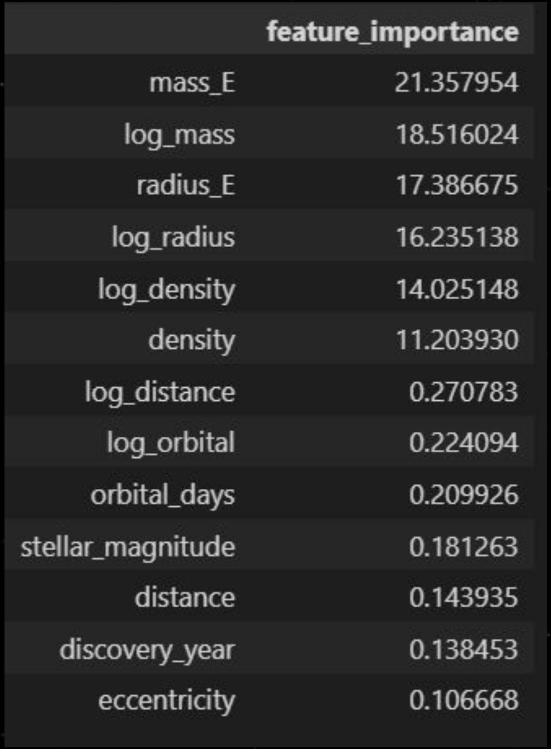
Classified as Gas Giant: 0

CONFUSION MATRIX



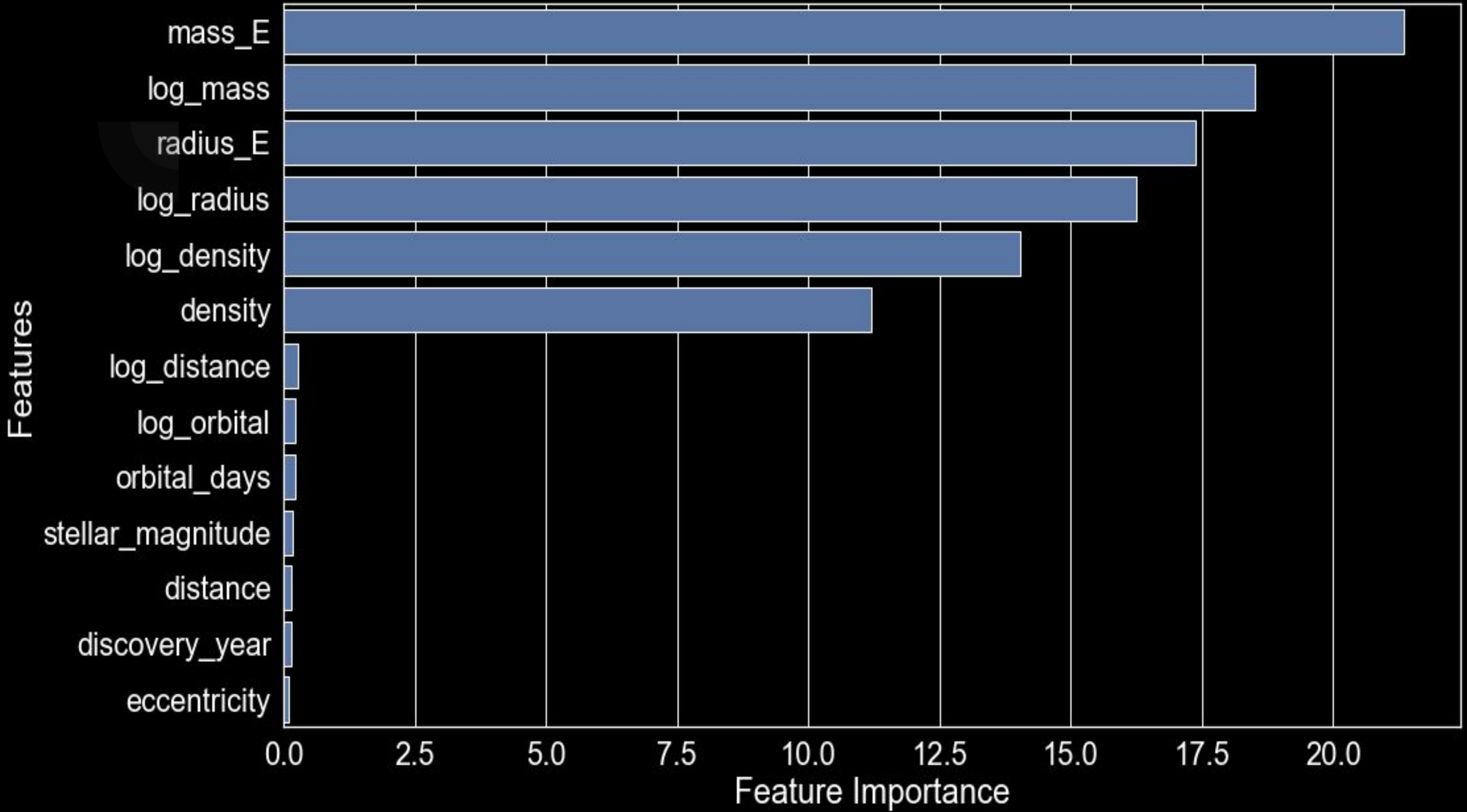
97.62	2.38	0.00	0.00
0.00	99.08	0.00	0.92
0.00	0.00	100.00	0.00
0.00	0.88	0.00	99.12

Feature importance

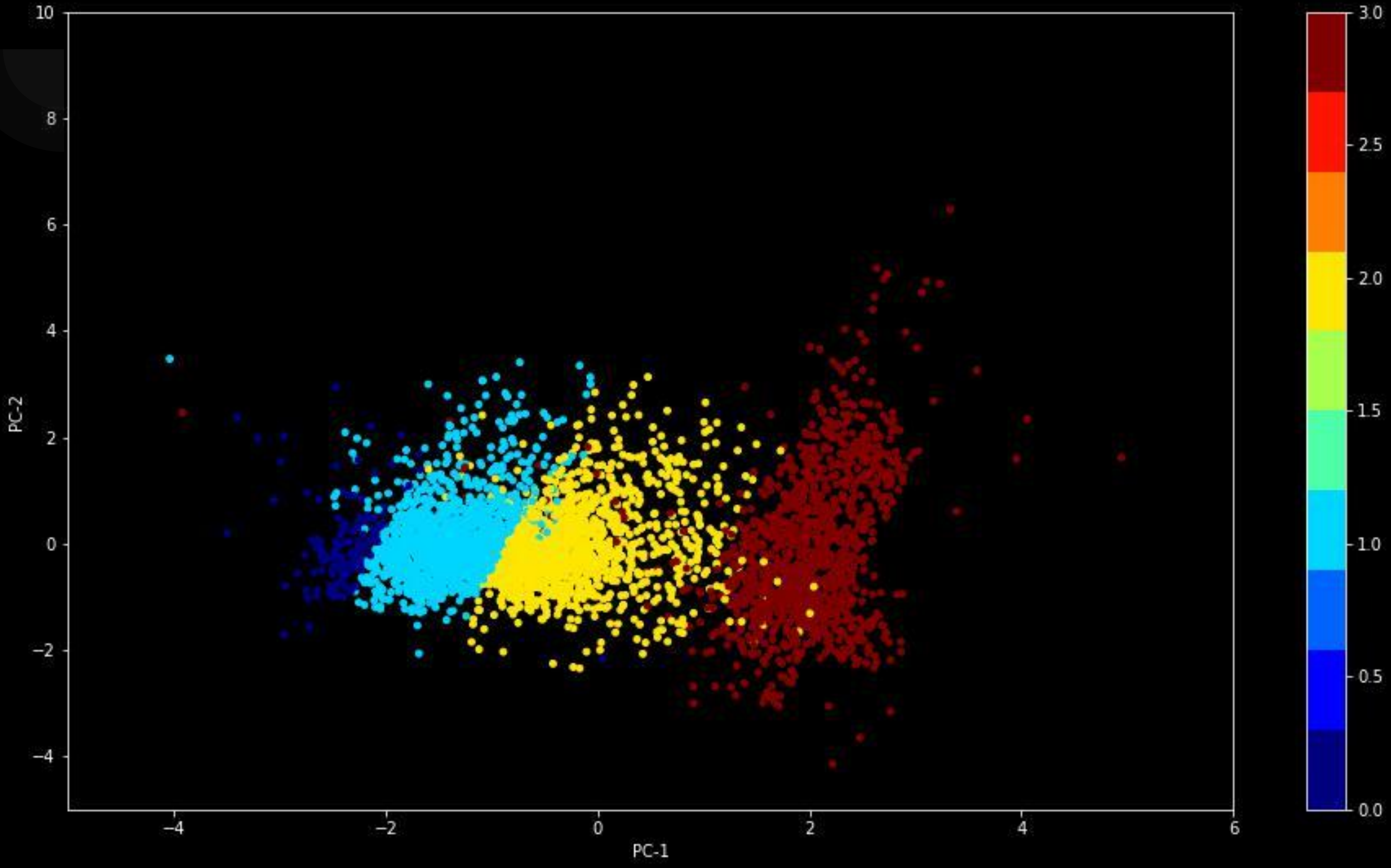


	feature_importance
mass_E	21.357954
log_mass	18.516024
radius_E	17.386675
log_radius	16.235138
log_density	14.025148
density	11.203930
log_distance	0.270783
log_orbital	0.224094
orbital_days	0.209926
stellar_magnitude	0.181263
distance	0.143935
discovery_year	0.138453
eccentricity	0.106668

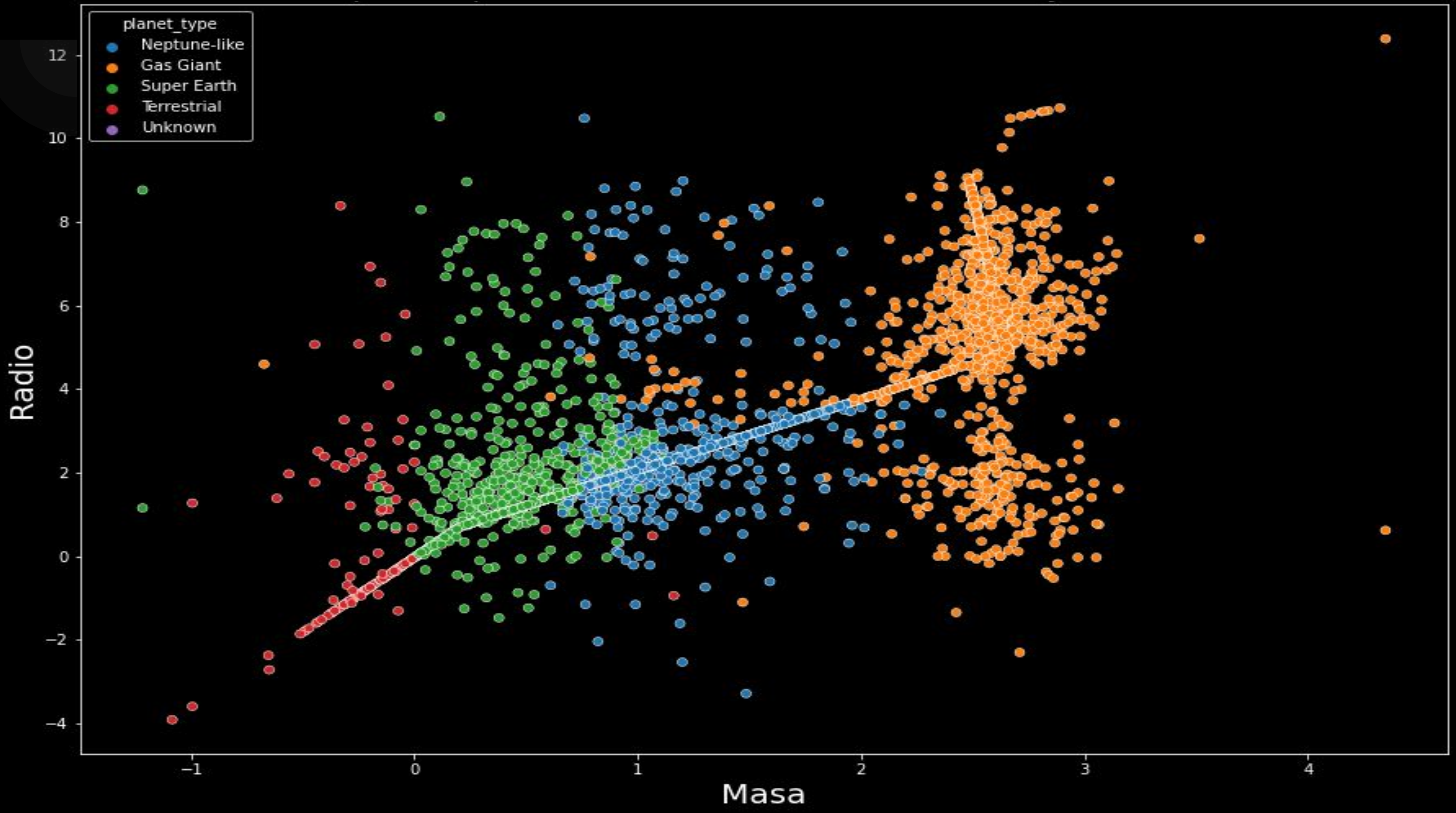
1. Mass
2. Radius
3. Density


mass_E
log_mass
radius_E
log_radius
log_density
density
log_distance
log_orbital
orbital_days
stellar_magnitude
distance
discovery_year
eccentricity
Features
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Feature Importance

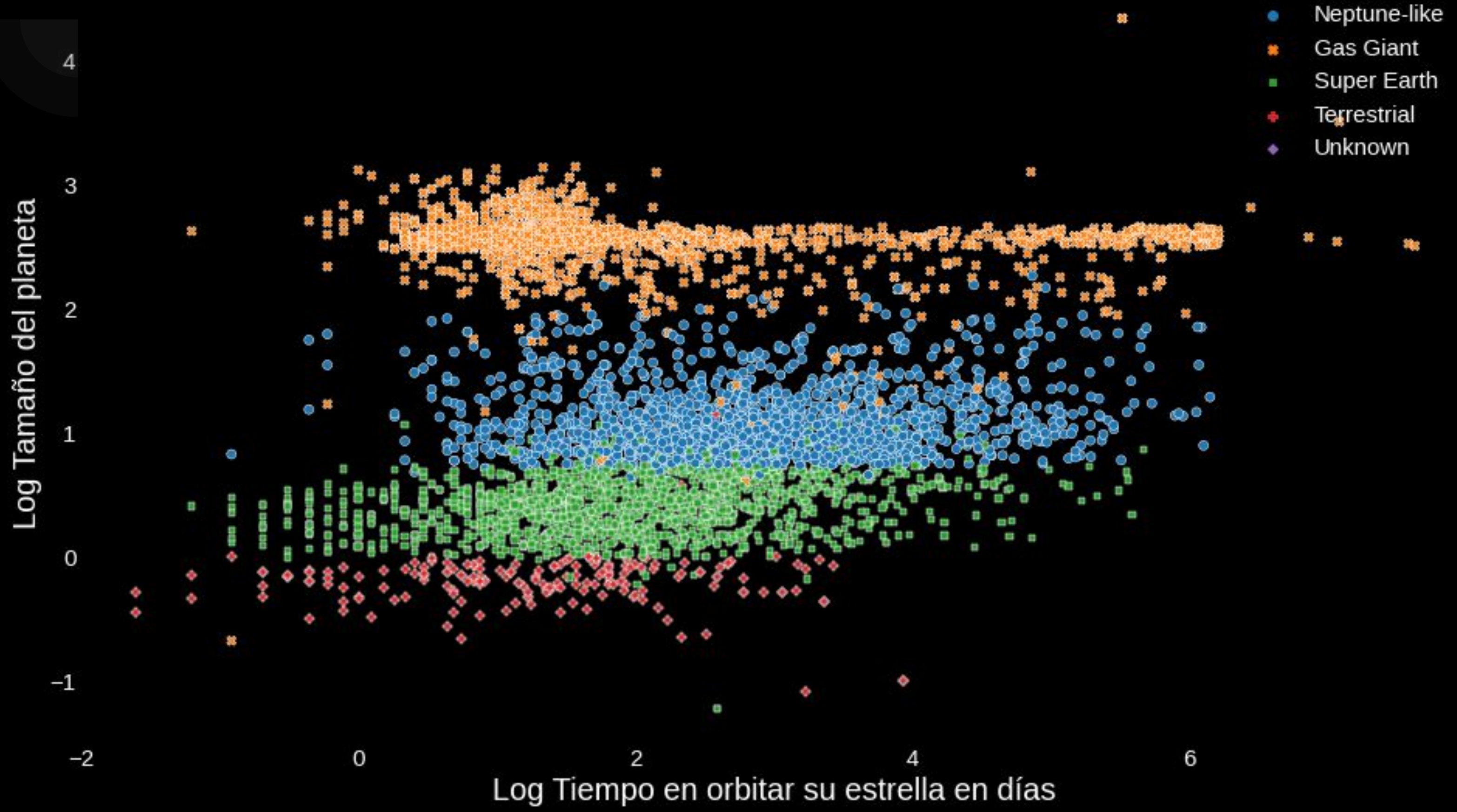
PRINCIPAL COMPONENT ANALYSIS



MASS

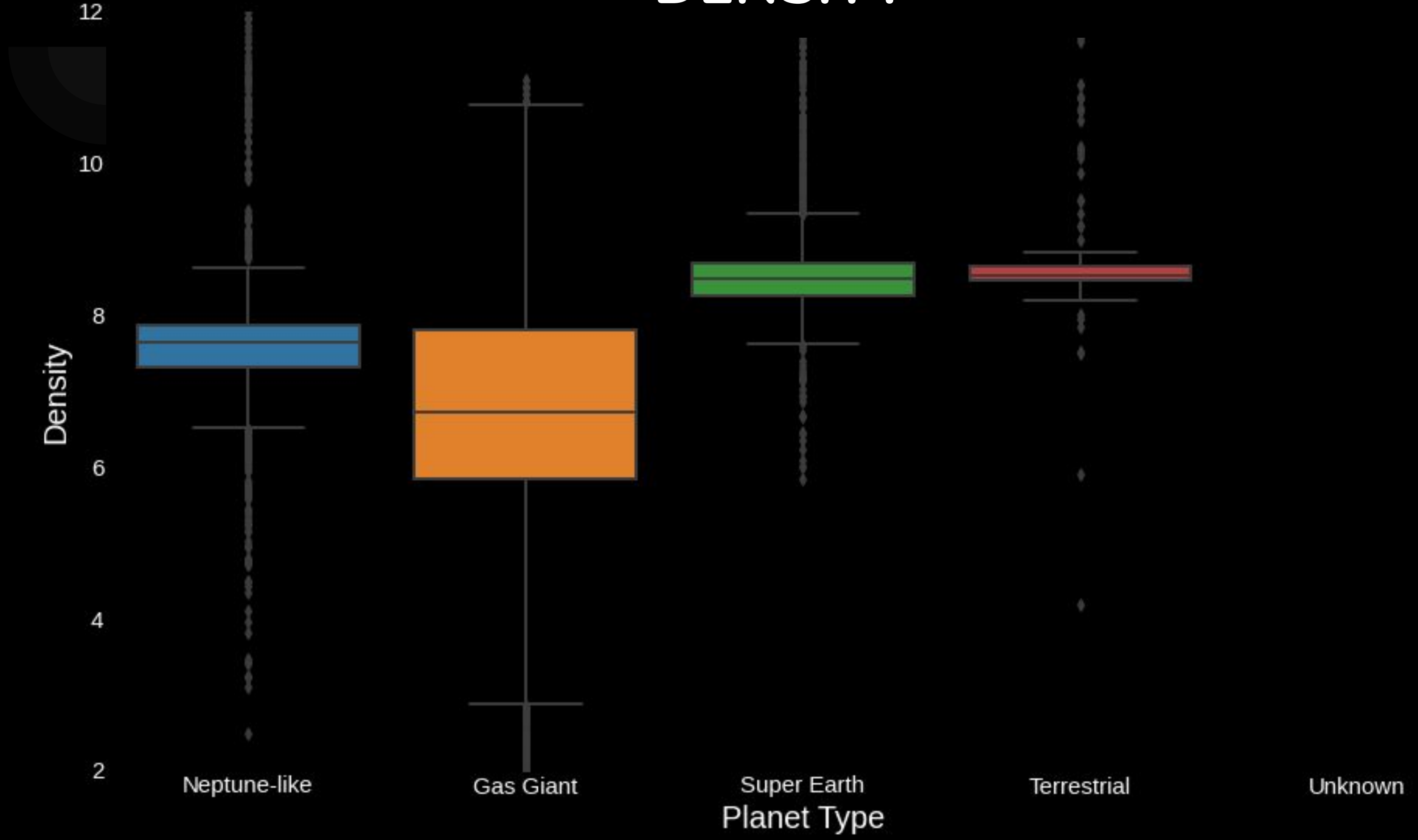

planet_type
Neptune-like
Gas Giant
Super Earth
Terrestrial
Unknown
Radio
12
10
8
6
4
2
0
−2
−4
Masa
−1
0
1
2
3
4

RADIUS

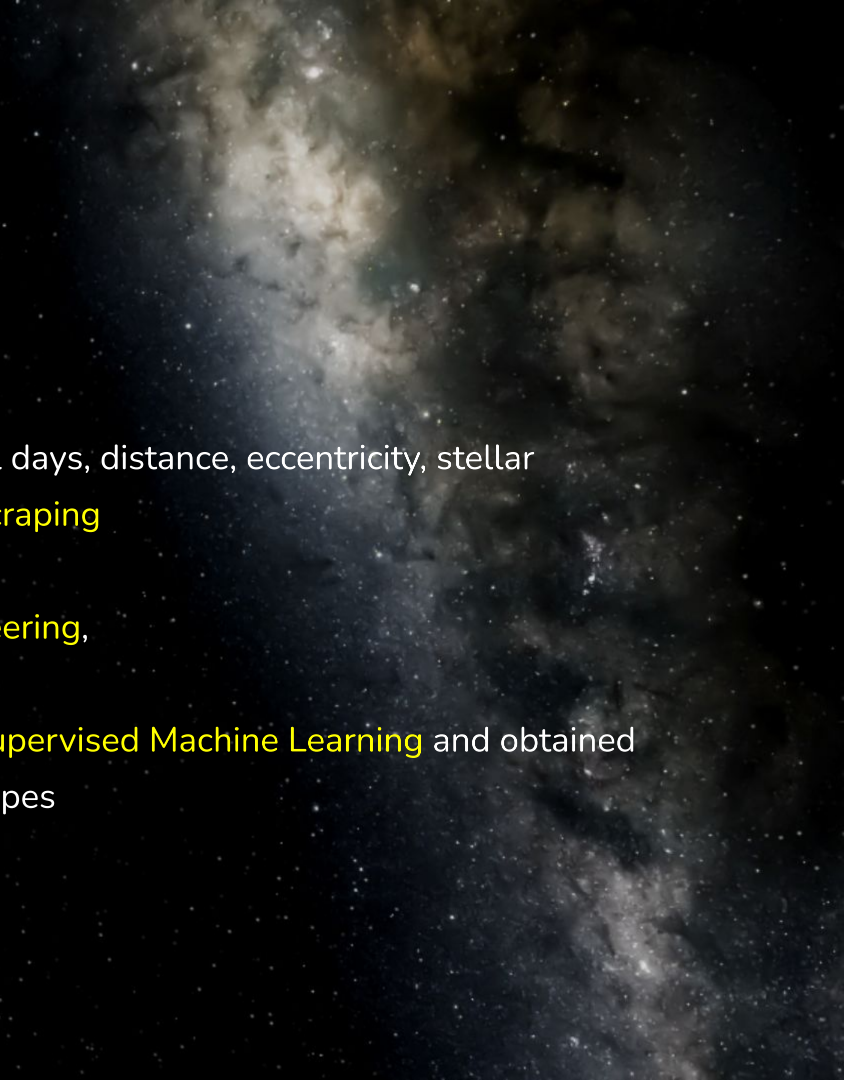


Neptune-like
Gas Giant
Super Earth
Terrestrial
Unknown
Log Tamaño del planeta
4
3
2
1
0
−1
−2
0
2
4
6
8
Log Tiempo en orbitar su estrella en días

DENSITY


12
10
8
6
4
2
Density
Neptune-like
Gas Giant
Super Earth
Terrestrial
Unknown
Planet Type

CONCLUSION



Starting from a few features (massa, radius, orbital days, distance, eccentricity, stellar magnitude and discovery year) obtained by web scraping

Preparing the data and doing some Feature Engineering,

We have solved the Classification Problem with Supervised Machine Learning and obtained an Accuracy of 99.3% in the prediction of planet types

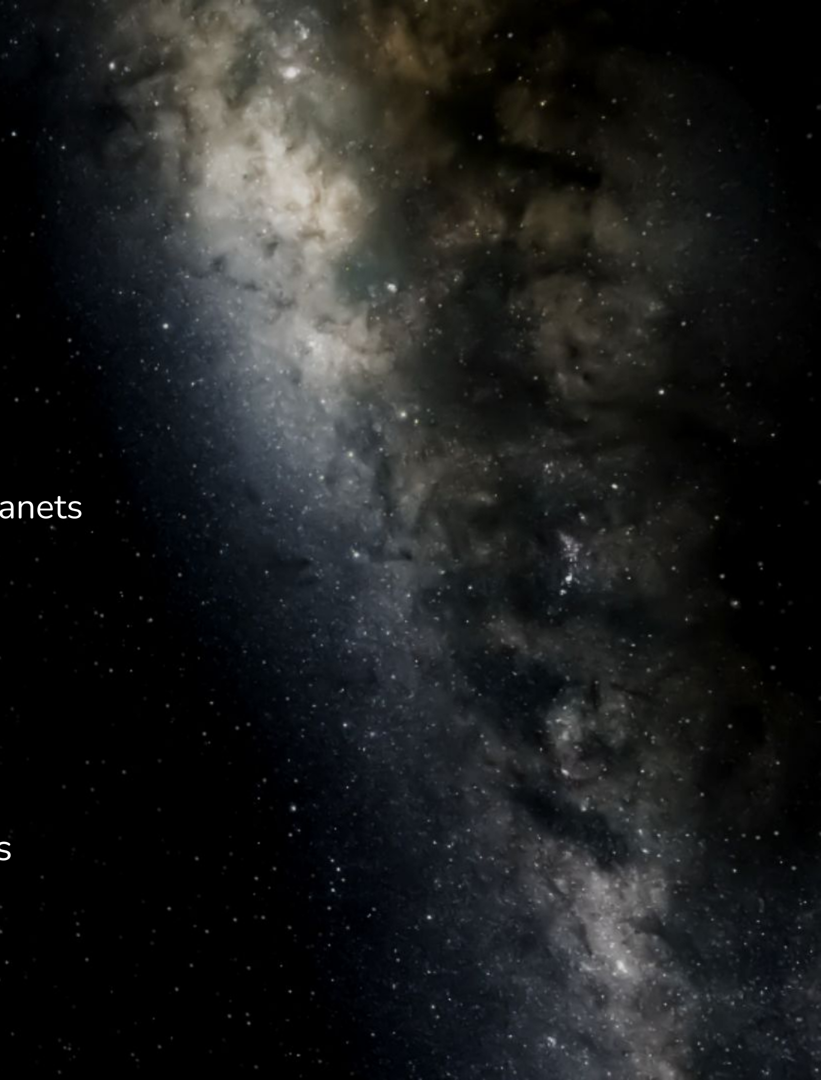
COMPARATIVE

Works of Rincón / Johans González:

- Models: Decision Tree
- Accuracy: 94,83% / 97,31%
- Observations: 3.286 exoplanets / 1.672 exoplanets

My results:

- Model: Voting(RF+XGB)
- Accuracy: 99,31%
- Observations: 5039 exoplanets



FUTURE LINES OF DEVELOPMENT

- Search and detection of exoplanets.
- Classification of habitable and superhabitable planets

SUPERHABITABLE PLANETS

Comparison between the size of Kepler-442b (1.34 $R_{\oplus}$) and Earth (right).