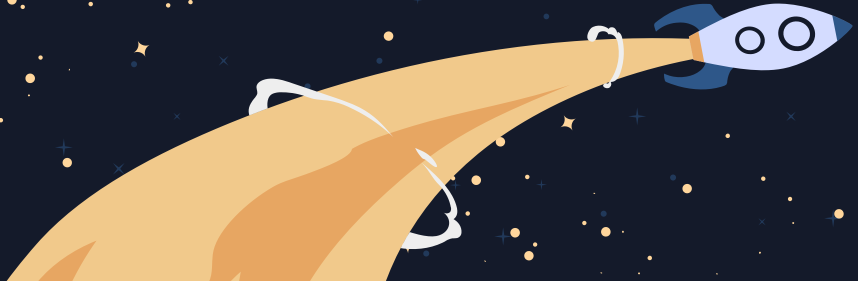
A World Away: Hunting for Exoplanets with AI

Super Code Ninja Girlies

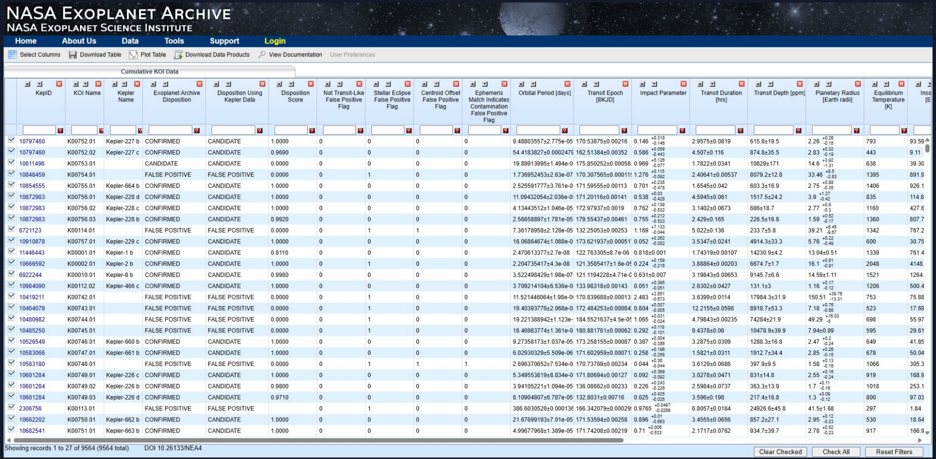
Shalisha Mathew
Ronia Matthews
Breanna Hernandez
Stefany Buccat



CHALLENGE: USING AI FOR DATA ANALYSIS AND EXOPLANET IDENTIFICATION

Thousands of new exoplanets have been discovered manually from data collected through space-based exoplanet surveying missions. To overcome the manual process of exoplanet identification, artificial intelligence and machine learning (AI/ML) can be used to automatically analyze large sets of data and identify exoplanets.

OUR GOAL: create and train an AI/ML model using NASA's open-source exoplanet datasets to analyze data for the purpose of identifying exoplanets



KepID	KOI Name	Kepler Name	Exoplanet Archive Disposition	Disposition Using Kepler Data	Disposition Score	Not Transit-Like False Positive Flag	Stellar Eclipse False Positive Flag	Centroid Offset False Positive Flag	Ephemeris Match Indicates Contamination False Positive Flag
10797460	K00752.01	Kepler-227 b	CONFIRMED	CANDIDATE	1.0000	0	0	0	0
10797460	K00752.02	Kepler-227 c	CONFIRMED	CANDIDATE	0.9690	0	0	0	0
10811496	K00753.01		CANDIDATE	CANDIDATE	0.0000	0	0	0	0
10848459	K00754.01		FALSE POSITIVE	FALSE POSITIVE	0.0000	0	1	0	0
10854555	K00755.01	Kepler-664 b	CONFIRMED	CANDIDATE	1.0000	0	0	0	0
10872983	K00756.01	Kepler-228 d	CONFIRMED	CANDIDATE	1.0000	0	0	0	0
10872983	K00756.02	Kepler-228 c	CONFIRMED	CANDIDATE	1.0000	0	0	0	0
10872983	K00756.03	Kepler-228 b	CONFIRMED	CANDIDATE	0.9920	0	0	0	0
6721123	K00114.01		FALSE POSITIVE	FALSE POSITIVE	0.0000	0	1	1	0
10910878	K00757.01	Kepler-229 c	CONFIRMED	CANDIDATE	1.0000	0	0	0	0
11446443	K00001.01	Kepler-1 b	CONFIRMED	CANDIDATE	0.8110	0	0	0	0
10666592	K00002.01	Kepler-2 b	CONFIRMED	CANDIDATE	1.0000	0	1	0	0
6922244	K00010.01	Kepler-8 b	CONFIRMED	CANDIDATE	0.9980	0	0	0	0
10984090	K00112.02	Kepler-466 c	CONFIRMED	CANDIDATE	1.0000	0	0	0	0
10419211	K00742.01		FALSE POSITIVE	FALSE POSITIVE	0.0000	0	1	0	0
10464078	K00743.01		FALSE POSITIVE	FALSE POSITIVE	0.0000	0	1	0	0
10480982	K00744.01		FALSE POSITIVE	FALSE POSITIVE	0.0000	0	1	0	0
10485250	K00745.01		FALSE POSITIVE	FALSE POSITIVE	0.0000	0	1	0	0
10526549	K00746.01	Kepler-660 b	CONFIRMED	CANDIDATE	1.0000	0	0	0	0
10583066	K00747.01	Kepler-661 b	CONFIRMED	CANDIDATE	1.0000	0	0	0	0
10583180	K00748.01		FALSE POSITIVE	FALSE POSITIVE	0.0000	0	1	1	1
10601284	K00749.01	Kepler-226 c	CONFIRMED	CANDIDATE	1.0000	0	0	0	0
10601284	K00749.02	Kepler-226 b	CONFIRMED	CANDIDATE	0.9800	0	0	0	0
10601284	K00749.03	Kepler-226 d	CONFIRMED	CANDIDATE	0.9710	0	0	0	0
2306756	K00113.01		FALSE POSITIVE	FALSE POSITIVE		0	1	0	0
10662202	K00750.01	Kepler-662 b	CONFIRMED	CANDIDATE	1.0000	0	0	0	0
10682541	K00751.01	Kepler-663 b	CONFIRMED	CANDIDATE	1.0000	0	0	0	0

KepID	Orbital Period [days]	Transit Epoch [BKJD]	Impact Parameter	Transit Duration [hrs]	Transit Depth [ppm]	Planetary Radius [Earth radii]	Equilibrium Temperature [K]	Inso [E
10797460	9.48803557±2.775e-05	170.53875±0.00216	0.146 $^{+0.318}_{-0.146}$	2.9575±0.0819	615.8±19.5	2.26 $^{+0.26}_{-0.15}$	793	93.59
10797460	54.4183827±0.0002479	162.51384±0.00352	0.586 $^{+0.059}_{-0.443}$	4.507±0.116	874.8±35.5	2.83 $^{+0.32}_{-0.19}$	443	9.11
10811496	19.89913995±1.494e-0	175.850252±0.00058	0.969 $^{+5.126}_{-0.077}$	1.7822±0.0341	10829±171	14.6 $^{+3.92}_{-1.31}$	638	39.30
10848459	1.736952453±2.63e-07	170.307565±0.00011	1.276 $^{+0.115}_{-0.092}$	2.40641±0.00537	8079.2±12.8	33.46 $^{+8.5}_{-2.83}$	1395	891.9
10854555	2.525591777±3.761e-0	171.59555±0.00113	0.701 $^{+0.235}_{-0.478}$	1.6545±0.042	603.3±16.9	2.75 $^{+0.88}_{-0.35}$	1406	926.1
10872983	11.09432054±2.036e-0	171.20116±0.00141	0.538 $^{+0.03}_{-0.428}$	4.5945±0.061	1517.5±24.2	3.9 $^{+1.27}_{-0.42}$	835	114.8
10872983	4.13443512±1.046e-05	172.97937±0.0019	0.762 $^{+0.139}_{-0.532}$	3.1402±0.0673	686±18.7	2.77 $^{+0.9}_{-0.3}$	1160	427.6
10872983	2.56658897±1.781e-05	179.55437±0.00461	0.755 $^{+0.212}_{-0.523}$	2.429±0.165	226.5±16.8	1.59 $^{+0.52}_{-0.17}$	1360	807.7
6721123	7.36178958±2.128e-05	132.25053±0.00253	1.169 $^{+7.133}_{-0.044}$	5.022±0.136	233.7±5.8	39.21 $^{+6.45}_{-9.67}$	1342	767.2
10910878	16.06864674±1.088e-0	173.621937±0.00051	0.052 $^{+0.262}_{-0.052}$	3.5347±0.0241	4914.3±33.3	5.76 $^{+0.22}_{-0.49}$	600	30.75
11446443	2.470613377±2.7e-08	122.763305±8.7e-06	0.818±0.001	1.74319±0.00107	14230.9±4.2	13.04±0.51	1339	761.4
10666592	2.204735417±4.3e-08	121.3585417±1.6e-05	0.224 $^{+0.159}_{-0.216}$	3.88864±0.00203	6674.7±1.7	16.1 $^{+0.81}_{-0.91}$	2048	4148.
6922244	3.522498429±1.98e-07	121.1194228±4.71e-0	0.631±0.007	3.19843±0.00653	9145.7±6.6	14.59±1.11	1521	1264.
10984090	3.709214104±6.536e-0	133.98318±0.00143	0.051 $^{+0.395}_{-0.051}$	2.6302±0.0427	131.1±3	1.16 $^{+0.17}_{-0.12}$	1206	500.4
10419211	11.521446064±1.98e-0	170.839688±0.00013	2.483 $^{+2.851}_{-0.673}$	3.6399±0.0114	17984.3±31.9	150.51 $^{+39.76}_{-13.31}$	753	75.88
10464078	19.40393776±2.068e-0	172.484253±0.00084	0.804 $^{+0.007}_{-0.005}$	12.2155±0.0598	8918.7±53.3	7.18 $^{+0.76}_{-0.68}$	523	17.69
10480982	19.221388942±1.123e-	184.5521637±4.5e-05	1.065 $^{+0.031}_{-0.034}$	4.79843±0.00235	74284±21.9	49.29 $^{+16.03}_{-5}$	698	55.97
10485250	16.46983774±1.361e-0	180.881761±0.00062	0.292 $^{+0.118}_{-0.101}$	9.4378±0.06	10478.9±39.9	7.94±0.89	595	29.61
10526549	9.27358173±1.037e-05	173.258155±0.00087	0.387 $^{+0.004}_{-0.386}$	3.2875±0.0309	1288.3±16.8	2.47 $^{+0.2}_{-0.24}$	649	41.85
10583066	6.02930329±5.509e-06	171.602959±0.00071	0.258 $^{+0.196}_{-0.258}$	1.5821±0.0311	1912.7±34.4	2.85 $^{+0.26}_{-0.15}$	678	50.04
10583180	2.696370652±7.534e-0	170.73769±0.00234	0.044 $^{+0.36}_{-0.044}$	3.6129±0.0686	397.9±9.5	1.58 $^{+0.13}_{-0.16}$	1066	305.3
10601284	5.349553819±8.834e-0	171.80694±0.00127	0.092 $^{+0.369}_{-0.092}$	3.0278±0.0471	831±14.8	2.55 $^{+0.16}_{-0.24}$	919	168.9
10601284	3.94105221±1.094e-05	136.08662±0.00233	0.226 $^{+0.243}_{-0.226}$	2.5984±0.0737	363.3±13.9	1.7 $^{+0.11}_{-0.16}$	1018	253.1
10601284	8.10904807±6.787e-05	132.8031±0.00716	0.025 $^{+0.425}_{-0.025}$	3.596±0.198	217.4±16.8	1.3 $^{+0.09}_{-0.12}$	800	97.03
2306756	386.6030528±0.000136	166.342079±0.00029	0.9765 $^{+0.0497}_{-0.0268}$	6.8057±0.0184	24926.6±45.8	41.5±1.68	297	1.84
10662202	21.67699193±7.01e-05	171.53594±0.00258	0.896 $^{+0.01}_{-0.663}$	3.4555±0.0656	857.2±27.1	2.95 $^{+0.12}_{-0.23}$	530	18.64
10682541	4.99677968±1.389e-05	171.74208±0.00219	0.71 $^{+0.006}_{-0.533}$	2.1717±0.0762	934.7±39.7	2.78 $^{+0.52}_{-0.23}$	917	166.9

Showing records 1 to 27 of 9564 (9564 total) DOI 10.26133/NEA4

OUR APPROACH

1 Initial Approach: Random Forest + CNN on Light Curves

- At first, we tried combining **Random Forest models on 4 basic features** with **1D and 2D CNNs** using the raw light curve data (lightkurve)
- The idea was to use the CNN to capture patterns in the light curve shapes, and the Random Forest to combine that with numeric features
- However, this required **folding the light curves for each star and period** using a for loop
- It quickly became **inefficient** because there were thousands of stars to process

- Random Forest was ideal because:
  - It **handles both numeric and categorical features** easily
  - It's **robust to outliers and noisy features**
  - It provides **feature importance** automatically, letting us identify which features contribute most

2 Focus on Outliers and Data Cleaning

- Next, we decided to **handle outliers** in the numeric features (like transit depth, duration, planetary radius)
- We used the **IQR method**: removed values outside 1.5× the interquartile range
- After cleaning, the Random Forest model performed **better on the confusion matrix**:
  - More true positives were correctly classified
  - False positives decreased
  - Overall accuracy improved

```
# Handle Outliers

# List of numeric columns to check for outliers
numeric_cols = ["koi_period", "koi_duration", "koi_prad", "koi_depth"]

# Create copies to avoid modifying original X and y
X_clean = X.copy()
y_clean = y.copy()

for col in numeric_cols:
    # Calculate IQR
    Q1 = X_clean[col].quantile(0.25)
    Q3 = X_clean[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Remove outliers
    mask = (X_clean[col] >= lower_bound) & (X_clean[col] <= upper_bound)
    X_clean = X_clean[mask]
    y_clean = y_clean[X_clean.index]
```

OUR APPROACH

3 Hyperparameter Tuning

- We **increased the number of trees** from 100 → 300 in the Random Forest
  - More trees improve model **stability** and **reduce variance**.
- We also adjusted parameters like max_depth, min_samples_split, and min_samples_leaf to prevent overfitting

```
# Train Random Forest

# Split into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X_clean, y_clean, test_size=0.2, random_state=42)

#rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
#rf_model.fit(X_train, y_train)
rf_model = RandomForestClassifier(
    n_estimators=300,        # more trees = better stability
    max_depth=15,            # prevent overfitting
    min_samples_split=5,     # don't split tiny nodes
    min_samples_leaf=3,      # ensures each leaf has enough data
    random_state=42
)
rf_model.fit(X_train, y_train)
```

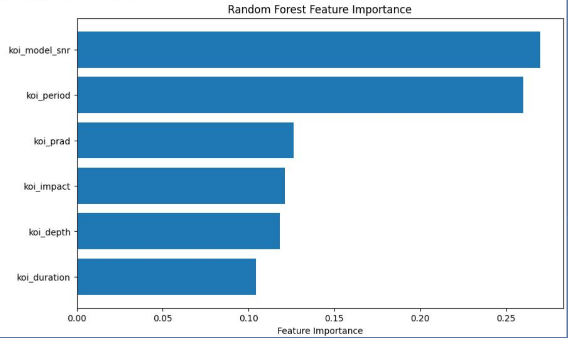
4 Adding More Features

- Initially, we used only 4 features
- We realized that **adding more features could help**, but adding **too many irrelevant features** could actually decrease model performance
- After research, we chose **10 features** that made the most sense physically (transit parameters + stellar properties)

5 Feature Importance Analysis

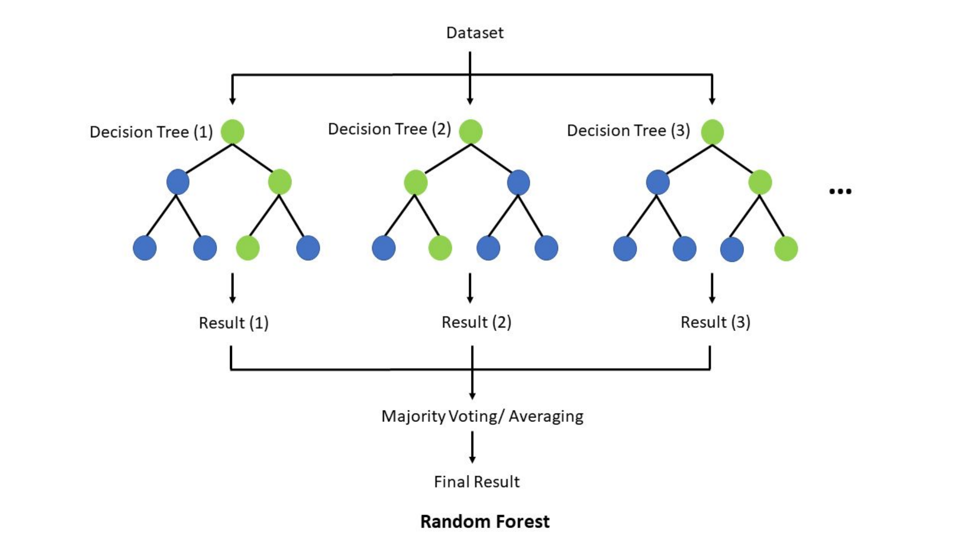
- We used Random Forest's **built-in feature importance**:
  - Each tree in the forest splits data based on features to reduce uncertainty (impurity)
  - Features that **consistently help separate classes** get higher importance scores
- We plotted a **bar chart** to visualize which features mattered most
  - The top features (like koi_model_snr, koi_period) had the highest scores
  - The lowest-ranked features were removed, but the model's performance **didn't change much**, confirming they weren't very useful

	Feature	Importance
5	koi_model_snr	0.269797
0	koi_period	0.260105
2	koi_prad	0.126259
4	koi_impact	0.121304
3	koi_depth	0.118315
1	koi_duration	0.104219



WHAT IS THE RANDOM FOREST CLASSIFIER?

The Random Forest classifier is the AI/ML model's core part. It is an ensemble of 300 decision trees, where each tree will focus and analyze its own data. The trees will then create a prediction (false positive, candidate, or confirmed exoplanet), and the forest will average all the individual predictions for a final prediction. This process was used due to its ability to ensure high accuracy and reliability in determining the most influential features for exoplanet identification, as well as analyzing complex data obtained from space agency open-data, while being robust to noise or outliers.

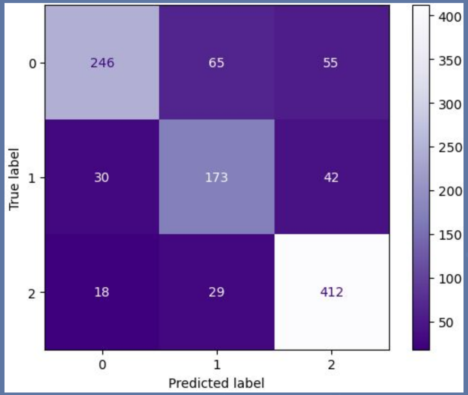


RESULTS AND PERFORMANCE OF THE AI/ML MODEL

- Our AI/ML model analyzed NASA's Kepler exoplanet data using 6 features:
  - koi_period: orbital period
  - koi_duration: transit duration
  - koi_prad: planetary radius
  - koi_depth: transit depth
  - koi_impact: impact parameter
  - koi_model_snr: transit signal-to-noise
- The model removes outliers in koi_period, koi_duration, koi_prad, and koi_depth. With the cleaned data, it will train a Random Forest classifier and evaluate its accuracy. A visual representation of each feature's importance will be displayed through a bar graph, to show which features helped the most to separate classes
- Finally, the model will display:
  - A classification report to summarize the identification of false positives, candidates, and confirmed exoplanets
  - A confusion matrix to compare the predicted labels and the true labels
- After adding more features, the model's accuracy improved from 69% to 78%

```
Classification Report:

              precision    recall  f1-score   support

           0       0.76      0.67      0.71       366
           1       0.50      0.42      0.46       245
           2       0.71      0.84      0.77       459

    accuracy                           0.69      1070
   macro avg       0.66      0.64      0.65      1070
weighted avg       0.68      0.69      0.68      1070

Confusion Matrix:

[[244  55  67]
 [ 50 103  92]
 [ 25  48 386]]
0.6850467289719626
```

```
Classification Report:

              precision    recall  f1-score   support

           0       0.84      0.67      0.75       366
           1       0.65      0.71      0.68       245
           2       0.81      0.90      0.85       459

    accuracy                           0.78      1070
   macro avg       0.76      0.76      0.76      1070
weighted avg       0.78      0.78      0.77      1070

Confusion Matrix:

[[246  65  55]
 [ 30 173  42]
 [ 18  29 412]]
0.7766355140186916
```



USER EXPERIENCE

A web interface platform was created with Streamlit to provide users with accessibility to the AI/ML model. The platform allows users to upload Kepler dataset files from space agency open data. The model will then analyze the uploaded data and display exoplanet identification results. With this interactive platform, scientists, researchers, and students can learn and explore exoplanet data.