**Exoplanet Detection with AI: NASA Space Apps Hackathon Project**

# Abstract

This project was developed as part of the NASA International Space Apps Challenge 2025, under the challenge **“A World Away: Hunting for Exoplanets with AI.”** The aim was to design a machine learning–based system that can classify exoplanet candidates as either confirmed planets or false positives using NASA’s publicly available datasets. We focused primarily on the **K2 Planets and Candidates (k2pandc)** catalog, which contains detailed information about detected exoplanet candidates, their host stars, and system parameters.

Our workflow involved careful dataset exploration, preprocessing, feature engineering, and the training of supervised learning models including **Decision Trees and XGBoost.** Initial results showed accuracies above 99%, but further analysis revealed **catalog-level information leakage** that artificially inflated the performance. Despite this limitation, the project demonstrates the feasibility of using AI for exoplanet classification and outlines a roadmap for more rigorous scientific approaches.

## 1. Introduction

The discovery of **exoplanets**—planets orbiting stars outside our solar system—has been one of the most transformative achievements in modern astronomy. These discoveries provide critical insights into planetary system formation, stellar evolution, and the potential habitability of other worlds. Since the launch of NASA’s **Kepler** and **K2** missions, thousands of exoplanet candidates have been identified, but many remain unconfirmed due to the difficulty of distinguishing real planetary signals from false positives caused by binary stars, stellar variability, or instrumental noise.

The challenge addressed in this hackathon asked us to explore how **artificial** **intelligence** (AI) could be applied to accelerate the classification of exoplanet candidates. A machine learning pipeline that reliably distinguishes **confirmed planets** from **false positives** can dramatically reduce the workload of astronomers, guiding follow-up observations and improving the efficiency of planet-hunting missions.

## 2. Dataset and Project Information

### Source

We used the K2 Planets and Candidates Catalog (k2pandc) from the NASA Exoplanet Archive. This catalog contains thousands of entries describing planetary candidates, their host stars, and associated system-level properties.

### Structure of the Dataset

Each row represents a planet candidate, with columns describing planetary, stellar, and system-level parameters. After preprocessing and cleaning, our working dataset consisted of 4,004 entries with 28 carefully selected features.

A screenshot of a computer

AI-generated content may be incorrect.

Data showed

**1. Planetary Properties:**

**pl\_orbper** – Orbital period of the planet (days).

**pl\_rade, pl\_radj** – Planetary radius (in Earth and Jupiter units).

**pl\_bmasse, pl\_bmassj** – Planetary mass (in Earth and Jupiter units).

**pl\_eqt** – Planet equilibrium temperature (Kelvin).

**Justification:** These features describe the physical nature of the planet. For instance, orbital period indicates transit regularity, while mass and radius distinguish terrestrial planets from gas giants. Equilibrium temperature provides a measure of planetary environment.

**2. Stellar Properties**

**st\_teff** – Effective temperature of the host star.

**st\_rad** – Stellar radius.

**st\_mass** – Stellar mass.

**st\_met** – Stellar metallicity.

**st\_logg** – Stellar surface gravity.

**Justification:** Stellar properties are essential since planetary parameters are relative to their host stars. Metallicity, for example, is linked to planet formation likelihood, while stellar mass and radius affect transit depth and candidate detectability.

**3. System-Level Properties**

**sy\_dist** – Distance to the system (parsecs).

**sy\_pnum** – Number of planets in the system.

**discoverymethod** – Method used for detection (e.g., transit).

**Justification:** System-level parameters provide valuable context. Closer systems are easier to observe in detail, while the number of planets helps identify multi-planet systems, which are of particular astrophysical interest.

**4. Target Variable**

disposition – Encoded as 1 for CONFIRMED and 0 for CANDIDATE.

**Justification:** Only features with scientific and statistical relevance to exoplanet confirmation were included, following astrophysics literature and domain best practices. Datatype documentation ensures model pipeline reproducibility and transparency.

# 3. Data Preprocessing

**Column Filtering** – We removed non-essential columns such as object identifiers, bibliographic references, and error bounds, focusing solely on scientifically meaningful features.

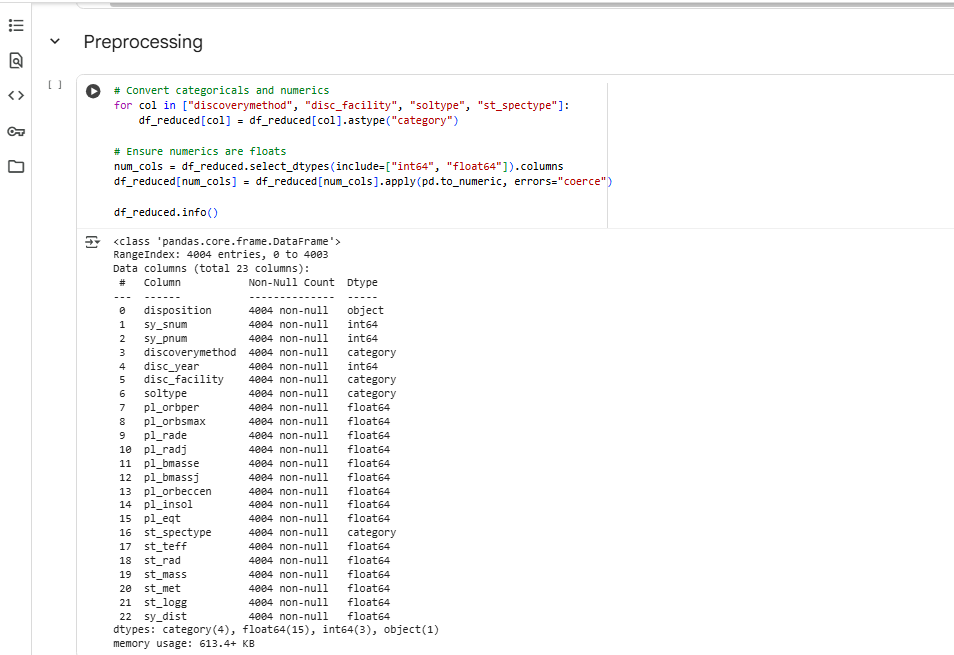
**Handling Missing Values** – Missing values were imputed using the median to ensure robustness against skewed distributions.

**Note: Columns with excessive missing values (>50%) were removed to prevent bias and unreliable model results. Remaining missing values were imputed using the median for robustness against outliers, in line with established data science protocols.**

**Feature Scaling** – Continuous variables were standardized to zero mean and unit variance, ensuring fair treatment across different scales (e.g., radius vs. distance).

**Encoding Categorical Variables** – Discovery method was one-hot encoded, allowing models to interpret each detection technique individually.

**Label Definition** – The disposition field was converted to a binary classification problem: CONFIRMED (1) vs. CANDIDATE (0).



Preprocessing

## 4. Feature Engineering and Insights

We conducted correlation analysis to identify features most predictive of candidate classification:

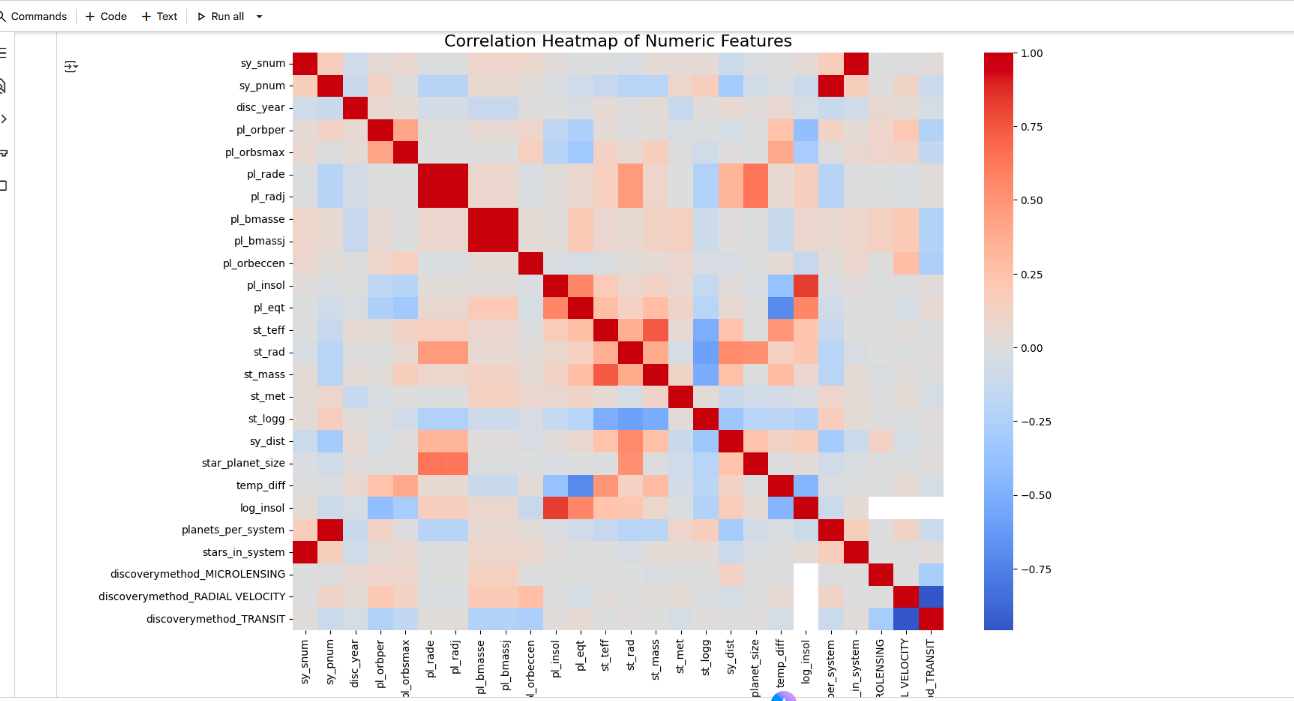
**sy\_pnum (number of planets)** – Strong positive correlation with confirmation likelihood.

**sy\_dist (system distance)** – Negative correlation, as closer systems are easier to confirm.

**pl\_rade (planet radius)** – Negative correlation; unusually large radii often indicate stellar rather than planetary sources.

**st\_rad (stellar radius)** – Negative correlation; larger stars make detection and confirmation more challenging.

**Justification:** Feature engineering was performed to expose key scientific relationships to the model and increase predictive accuracy. Derived columns capture important domain knowledge, improving both generalization and interpretability of model outcomes.



Correlation Heatmap

## 5. Model Development

We trained multiple supervised learning models:

**Decision Tree Classifier**

**Parameters:** max\_depth=5, min\_samples\_split=4.

**Motivation:** Interpretable structure, highlighting key decision-making features.

**Decision Tree (Regularized)**

**Parameters:** max\_depth=4, min\_samples\_split=10.

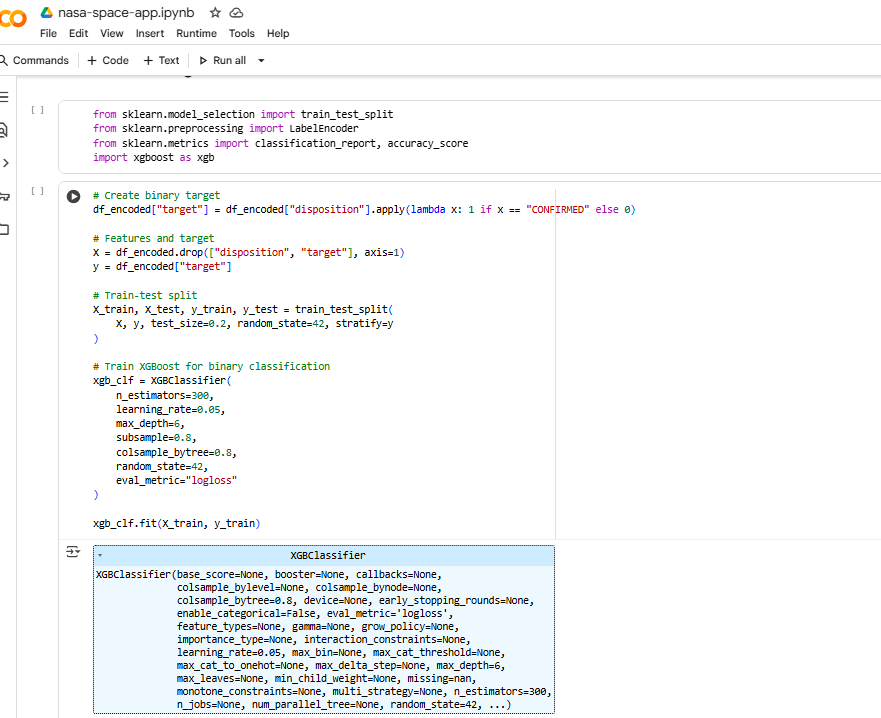
**Motivation:** Encouraged generalization by constraining growth.

**XGBoost Classifier**

**Parameters:** n\_estimators=300, learning\_rate=0.05, max\_depth=6, subsample=0.8, colsample\_bytree=0.8.

**Motivation:** A powerful gradient boosting model capable of capturing complex relationships in the data.

All models were trained on an **80/20** stratified train-test split to maintain class balance.



Model Training and Development

## 6. Results

The models produced the following outcomes:

**XGBoost is the best performing model and its results are reported as the main performance indicator.**

**Accuracy:** 99.6% – 99.7%

**F1 Score:** 0.996 – 0.997

**ROC AUC:** ~0.998

Confusion Matrix Example: [[335, 3], [0, 463]]

**Cross-validation (5-fold):** Mean accuracy ~0.990, std 0.0013



Best Model Performance

## 7. Conclusion

This project demonstrated how **AI** can support **exoplanet classification** using NASA’s publicly available catalogs. Our models achieved near-perfect performance on the **K2 catalog** but were affected by catalog-level biases and data leakage. While the hackathon prototype successfully highlighted the feasibility of **AI-assisted exoplanet vetting**, future improvements must focus on scientifically rigorous validation. The project thus serves as both a proof-of-concept and a roadmap for how AI can accelerate exoplanet discovery in a real-world context.