# Kepler Exoplanet Analysis

Aditeya Baral

Ameya Bhamare

Saarthak Agarwal

# NASA Exoplanet Exploration

For several decades, planet identification has been a task performed by specialized astronomers and domain experts.

With the advent of computational methods and access to satellite data from space missions, this trend has changed.

NASA's Exoplanet Exploration program has provided us with vast amounts of data on celestial objects to assist in space exploration.

The Kepler Mission has identified over 4000 transiting exoplanets since the commencement of the mission in 2007

# Objective

- NASA’s  focus lay on exploring planets and planetary systems.
- It has traditionally been a time-intensive task reserved for domain experts with access to specialized equipment.
- These experts study images collected by satellite-based telescopes like the Hubble.
- A new generation of modern satellites such as Kepler has opened the door for exoplanet exploration
- New technologies are trying to partially automate scientific observations.

# Dataset

NASA has provided us with a catalog of discoveries that help in computing planet occurrence rates as a function of a star's properties.

This information is catalogued in the Cumulative Kepler Object of Information table available for public domain use on NASA's exoplanet archive.

These satellites not only take pictures but also process those images using astronomical techniques to produce data with a variety of features for identifying these exoplanets.

This data has democratized the once arduous task of exoplanet exploration among data scientists, engineers and statisticians alike.

# Approach

- Machine learning algorithms can be applied to exoplanet data to detect overlooked exoplanets in data archives or automate the classification of objects of interest.
- The KCOI data has been cleaned to retain the most important features
- Finally, models have been fit on this data and eventually used to assign a probability for an observation in the KCOI table being an exoplanet
- The models explored include Support Vector Machines, Random Forest Classification, AdaBoost and Deep Neural Networks.
- The Random Forest Classifier was selected as the optimum machine learning model and returned an F-1 score of 98% on the dataset.

# Evaluation of Solution

Considering the imbalance of the dataset, we propose to use measures that will suit our problem.

We use the F-1 score as the base statistic to measure our model's performance.

Additionally, we propose to use cross-validation with a fold size of 10 across our entire dataset

We estimate our model on these folds to analyse how the model performs across the entire set of observations.

This helps us to prevent any inherent bias the model could have learnt.

# Observations

We observe that there is significant overlap between the different classes of exoplanets, making it increasingly difficult for scientists to predict their habitability.

We also observe that most of the exoplanet characteristics are independent of each other, with very few attributes having significant correlation.

Subsequently, this leads to high variance contributions from all the features

A difference in feature rank importance was observed across the different algorithms, showing the differences in the working of each model.

Additionally, we see that machine learning algorithms prefer categorical variables for classification as it allows them to form decisions faster and reduce entropy quicker.

# Team Contribution

Aditeya created the data preprocessing pipeline which allowed for removal of noise and redundant features in the dataset. Additionally, he worked on classification models based on traditional Machine Learning approaches.

Ameya worked on the literature review to provide us with background information about exoplanets. He also produced the various hypotheses to be tested and worked on Deep Learning based approaches.

Saarthak provided us with the necessary visualisations to help us understand the data better. Additionally, he worked on the analysis of the models themselves, including the ranking of feature importance to give us greater insights on our classification.


# Thank You