**PROJECT PROPOSAL**

**FALL 2023**

**Unlocking the Secrets of Exoplanets: A Comprehensive Analysis**

**DATS6103 : Introduction to Data Mining**

**Team members**

1. **Apoorva Reddy Bagepalli**
2. **Pratiksha Dange**
3. **Tulasi**
4. **Saniya Samir Shinde**

**Research Proposal**

**Abstract**

This project delves into the rich dataset provided by the NASA Exoplanet Archive, which contains comprehensive information on exoplanets discovered through various methods and across different years. The aim is to uncover patterns, trends, and insights that can enhance our understanding of exoplanetary systems. Several SMART questions guide our exploration, including investigating the historical trends in exoplanet discoveries, the significance of orbital periods in relation to the discovery year, and the relationships between planetary characteristics, host star distances, and orbital dynamics. By rigorously analyzing this dataset, we aspire to shed light on the progress and limitations of observational techniques, ultimately contributing to the advancement of exoplanet research and planetary formation theories.

**SMART Questions**

1. Can we determine any patterns in the detection methods used for exoplanets and their respective discovery years? Are certain methods more effective at finding specific types of planets?
2. What is the historical trend in the discovery of exoplanets based on their detection methods, and what can this tell us about the progress and limitations of our observational techniques over time?
3. How does the discovery year of exoplanets impact the significance of their orbital periods in the context of planetary formation theories?
4. Are there any discernible patterns or trends in the orbital characteristics of planets, such as orbital radius, orbital period, and eccentricity, and how might these relate to their host stars?
5. What is the relationship between the distance of exoplanets from their host stars and their planetary characteristics, such as mass and radius?
6. How do the orbital radii and orbital periods of exoplanets relate to their eccentricity, and how might this information aid in uncovering patterns in orbital dynamics within planetary systems?
7. What are the potential characteristics of a new exoplanet based on features such as distance, stellar magnitude, planet type, discovery year, mass, radius, orbital radius, orbital period, eccentricity, and detection method, in comparison to the existing exoplanets in the dataset?

**Dataset Description**

The NASA Exoplanet dataset contains information on all known exoplanets (planets outside our solar system) discovered by NASA's various space missions, ground-based observatories, and other sources. The dataset includes information such as the planet's name, mass, radius, distance from its host star, orbital period, and other physical characteristics. The dataset also includes information on the host star, such as its name, mass, and radius. We aim to study the properties and distribution of exoplanets in our galaxy. You can find our dataset [here](https://www.kaggle.com/datasets/adityamishraml/nasaexoplanets). It consists of 13 columns and 5250 rows.

**mass of the planet = mass\_multiplier \* mass\_wrt(planet)**

**radius of the planet = radius\_multiplier \* radius\_wrt(planet)**

**Model Methods**

In the upcoming analysis, a decision tree will be deployed to identify patterns in the detection methods used for different types of planets and their respective discovery years. `1A bar chart will be created to display the distribution of discovered exoplanets based on different detection methods such as Radial Velocity, Direct Imaging, and Eclipse Timing Variations. An ANOVA test will be conducted to determine whether statistically significant differences exist in the mean orbital periods of exoplanets discovered in different years. Additionally, a heatmap will be generated to display the correlation coefficients between orbital parameters, including orbital radius, period, and eccentricity. A linear regression model will be utilized to analyze the relationship between the distance of exoplanets from their host stars and their planetary characteristics such as mass and radius.

Furthermore, a KNN model will be deployed to predict the potential characteristics of a new exoplanet based on features like distance, stellar magnitude, planet type, discovery year, mass, radius, orbital radius, orbital period, eccentricity, and detection method. This prediction will be made in comparison to the existing exoplanets in the dataset.

**Github repo :** <https://github.com/apoorvareddy612/DATS6103_Project>