DATA ANALYTICS: Unveiling the Cosmos

<u>Overview</u>

Embark on a celestial journey to distant realms – exoplanets, planets that circle stars beyond our solar confines. These alien worlds captivate scientists and stargazers alike, unraveling the tapestry of our universe.

Imagine diverse planets, some gaseous giants, others rocky landscapes. Their sizes, compositions, and orbits vary, each shaped by cosmic forces. We explore planetary systems, where planets dance around stars, tracing out intricate patterns over time.

These systems reveal mysteries. Do certain planets favor specific types of stars? What makes a planet habitable? Through spectroscopy, we decipher atmospheres, hinting at potential life-sustaining conditions.

Kepler's legacy enriched our understanding. By observing dips in starlight caused by transiting exoplanets, we estimate their sizes and orbits. Radial velocity helps detect wobbles as planets tug stars in orbit, revealing hidden planets.

Exoplanets challenge our assumptions, expanding our cosmic perspective. Some orbit binary stars, while others exhibit exotic traits defying categorization. They rewrite planetary formation theories, nudging us to question and refine our models.

In our quest to understand these enigmas, we stand humbled by the vastness of space. Exoplanets remind us of the countless worlds waiting to be discovered, igniting our curiosity to explore, question, and redefine our place in the universe

General Instructions

1) This is the problem statement for the event Data Analytics.

2) The dataset for the problem can be found here: ✛ phl_exoplanet_catalog_2019

3) The field descriptions can be found here:
   https://phl.upr.edu/projects/habitable-exoplanets-catalog

4) The problem statement consists of 4 problems and each problem has few subparts.

5) These problems are based on the Exoplanet data captured by NASA.

6) Students can participate in teams of size 3-4.

7) The weightage of each question is written in front of the question.

8) Participants are free to use any programming language, environment and library. However, it is preferable to use Google Colab, Jupyter Notebook as the programming environments.

9) For submission, each team should make a PDF report which should contain a detailed solution and approach. All the plots and outputs need to be shown in the report along with proper explanations and descriptions. The final answers should be highlighted.

10) All code snippets should be attached to the report. The snippets should be well commented on and should convey a clear idea of their work.

11) Only 15 minutes will be allocated for each team to present their solutions. Hence, all the relevant conclusions must be presented promptly.

12) Participants will be evaluated on the clarity and applicability of the solution, their innovation and the feasibility of their conclusions.

Problem Statement

1. **Visualization and Analysis of Dataset: (15 points)**

   1.1 Inspect the data type and Extract basic statistics:

       1.2.1 Print Range, Mean, Median, and Standard Deviation of the dataset.

       1.2.2 Does the Dataset require Normalisation?

       1.2.3 What are your inferences based on the above results?

   1.2 Using the Seaborn module, plot a heatmap to explore the various planetary detection methods used over the years

       1.2.1 What do you infer from the above heatmap?

   1.3 Identify the planetary detection methods that have identified the most: i) Uninhabitable planets (0)

       ii) Conservatively habitable planets (1)

       iii) Optimistically habitable planets (2)

   1.4 Determine the Interquartile Range and the Skewness of the Dataset

   1.5 How would you tackle the classification bias (class imbalance) of the Dataset?

   1.6 Report the ratio of the Number of Iron atoms to the Number of Hydrogen atoms for the exoplanets in the dataset by creating a separate feature titled 'Fe to H Ratio'

2. **Interpretation and calculation of physical parameters: (15 points)**

   2.1 Calculate the escape velocities of exoplanets and compare them to their estimated temperatures. Present a plot of estimated temperature with escape velocity. Explain the nature of the plot obtained.

2.1.1 Analyze whether these velocities are sufficient to retain their atmospheres, drawing connections to atmospheric escape processes.

2.2 How does the distribution of host star ages correlate with the metallicity of their associated exoplanets?

2.2.1 Identify any patterns that align with our understanding of stellar evolution and planet formation

2.3 Examine the correlation between host star magnetic field strength and exoplanet atmospheric properties.

2.3.1 Explore how magnetospheric interactions might impact the composition and stability of exoplanet atmospheres.

2.4 What are Spectral Types? How many major star Spectral Types exist?

2.4.1 Present a categorical plot representing various Spectral Types and the habitability associated with it.

2.4.2 Give reasons for the plot obtained by using the knowledge gained regarding star Spectral Types

2.4.3 Is there a relationship between the size and density of exoplanets and the specific spectral characteristics of their host stars?

3. **Feature Engineering: (10 points)**

3.1 What percentage of null values exists in each feature? Visualise these percentages to identify which features have the most missing data.

3.2 Feature Reduction

3.2.1 Given a large number of features, identify redundant or highly correlated features

3.2.2 Choose an appropriate feature reduction method for this

dataset. Expatiate the premise behind choosing that method

        3.2.3 Describe the relationship between various features before and after feature reduction by creating scatter plots and identifying changes in the distribution of data, patterns and clusters

    3.3 Apply a suitable imputation technique to fill the null values of the dataset. Clarify your choice of technique used

4. **Habitability classification: (20 points)**

    4.1 Build a robust and efficient classifier for classifying a new exoplanet into the three classes of habitability namely,

        i) Uninhabitable planets (0)

        ii) Conservatively habitable planets (1)

        iii) Optimistically habitable planets (2)

    by utilizing the target features.

    Implement K-Fold Cross Validation for training. Train the dataset for all values of K from 2-10. Plot the loss and accuracy versus epochs for these K values.

    4.2 Plot the ROC curve and Confusion Matrix to quantify the performance of your classifier.

    4.3 Optimize all the hyperparameters used in the classifier by choosing an appropriate optimization method. Also, explain the premise behind choosing the optimization method.