MILESTONE - 2

Theory for Data Processing and ML

1. Data Preprocessing & Management

The preprocessing pipeline was designed to transform raw telescope data into a high-quality dataset suitable for machine learning.

Data Acquisition

- **Source**: Data was programmatically fetched from the **NASA Exoplanet Archive** using the astroquery library.
- **Table Selection:** The pscomppars (Planetary Systems Composite Parameters) table was selected to ensure the most complete set of physical and stellar features.

Data Quality Assessment

- **Summary Statistics**: Used df.describe() to identify range, mean, and standard deviation.
- **Null Identification:** Identified significant missing data, particularly in pl_eqt (Equilibrium Temperature) and st_met (Stellar Metallicity).
- **Visualization:** Generated a **Missing Value Heatmap** using Seaborn to visualize gaps across the planetary features.

Handling Missing Data

- **Median Imputation**: Applied to planetary physical values (radius, mass, density, orbital period) to handle skewed distributions without introducing outlier bias.
- **Stellar Features:** Missing host star temperatures and luminosities were filled using the median value of the existing stellar population.
- **Categorical Mode:** Missing st_spectype (Star Type) entries were filled using the most frequent occurrence (Mode).

Outlier Detection & Physical Validation

- **IQR Method:** Calculated the Inter-Quartile Range (IQR) to identify statistical outliers.
- **Hard Constraints:** Implemented filters to remove physically impossible data points, specifically ensuring pl_rade (Radius) and pl_eqt (Temperature) are greater than zero.

Feature Engineering (Custom Indices)

To convert raw physics into machine-readable signals, three custom indices were calculated:

1. **Habitability Score Index:** A combined metric measuring how close a planet's temperature is to 288K and its radius is to 1.0 Earth radii.

2. **Stellar Compatibility Index:** Measures the similarity of the host star to the Sun (5778K), favoring stable G-type stars.
3. **Orbital Stability Factor:** A ratio of the semi-major axis to the orbital period based on Kepler’s Third Law to assess the long-term stability of the planet's path.

Categorical Encoding & Target Creation

- **One-Hot Encoding**: Converted the categorical st_spectype into multiple binary columns (e.g., star_G, star_M).
- **Target Variable:** Created a binary target column where **1** represents the top 5% of planets by habitability score, defining the classification task.

2. AI Model for Habitability Prediction

Following the "Model Tournament," **XGBoost** was selected as the optimal algorithm for this dataset.

Model Architecture: XGBoost Classifier

- **Type**: Gradient Boosted Decision Trees.
- **Reasoning:** XGBoost was chosen because it excels at identifying non-linear relationships in tabular data and includes internal handling for imbalanced datasets.

Training Strategy

- **Data Splitting**: The dataset was divided into an **80% Training Set** and a **20% Testing Set**.
- **Feature Scaling:** Applied StandardScaler to normalize features, ensuring that large-scale values (like orbital period in days) did not overwhelm small-scale values (like mass in Earth masses).
- **Class Imbalance:** Leveraged **SMOTE** (Synthetic Minority Over-sampling Technique) to synthetically balance the rare habitable candidates with the non-habitable majority.

Performance Metrics

- **Winning Score**: The model achieved an **F1-Score of 0.9256**.
- **Validation:** Beyond simple accuracy, the model was evaluated using **ROC-AUC** to confirm its ability to distinguish between classes regardless of the probability threshold.

3. Milestone 2 Deliverables Summary

- **Preprocessed Dataset:** Exported as preprocessed.csv.
- **Pickle Model:** The final XGBoost model and Scaler were serialized into .pkl format for the Flask API.
- **Analysis:** Feature importance plots revealed that the engineered **Habitability Score** and **Stellar Compatibility** were the strongest predictors of the model's decisions.