Description of Data Preprocessing and Machine Learning Model Training

1. Data Preprocessing

Data preprocessing is the process of converting raw data into a clean and structured format suitable for machine learning models. Real-world datasets, especially astronomical data, often contain missing values, noise, duplicates, and inconsistent units. If these issues are not handled properly, they can reduce model accuracy and reliability.

The dataset used in this project contains planetary and stellar features such as planet radius, planet mass, orbital period, equilibrium temperature, star temperature, luminosity, metallicity, and star type. These features are important for determining the habitability potential of exoplanets.

The first step in preprocessing is **data quality assessment**, where missing values, null values, duplicate rows, and inconsistent units are identified. Summary statistics are generated to understand data distribution, and visual techniques like missing value heatmaps help detect data gaps.

Missing data handling depends on the type of feature. Numerical planetary features are filled using mean or median imputation, while star temperature is handled using median to reduce the effect of outliers. Categorical features such as star type are filled using the mode. Rows with completely missing critical information are removed to avoid misleading patterns.

Outlier detection is performed using statistical methods like Z-score and Interquartile Range (IQR). Physically impossible values such as negative planet radius or unrealistic temperatures are removed, while extreme but realistic values may be capped.

To maintain consistency, **unit standardization** is applied. All measurements are converted into standard astronomical units such as Earth radii, Earth masses, Astronomical Units (AU), and Kelvin. This prevents scale-related bias during model training.

Feature engineering is used to improve model performance by creating new meaningful features. These include a Habitability Score Index, Stellar Compatibility Index, and Orbital Stability Factor, which convert scientific knowledge into machine-readable signals.

Categorical features like star type are converted into numerical form using **One-Hot Encoding**, and numerical features are scaled using StandardScaler or

MinMaxScaler. Finally, a **target variable** is created to represent habitability as either binary (habitable or non-habitable) or multi-class. The cleaned dataset is saved as preprocessed.csv.

---

2. Machine Learning Model Training

Machine learning model training is the process of teaching an algorithm to learn patterns from preprocessed data and make predictions. In this project, the problem is defined as a **supervised classification task**, where the model predicts the habitability class of an exoplanet.

The dataset is split into **features (X)** and **target (y)**, followed by an **80–20 train-test split**. This ensures that the model is evaluated on unseen data and avoids overfitting.

A **baseline model** such as Logistic Regression or a shallow Decision Tree is first trained to establish reference performance. This helps in comparing the effectiveness of advanced models.

Advanced models like **Random Forest** and **XGBoost** are then trained. Random Forest handles non-linear relationships well and provides feature importance, while XGBoost is highly efficient for structured tabular data and captures complex feature interactions.

To avoid data leakage, **machine learning pipelines** are used to combine preprocessing steps such as scaling, encoding, and model training into a single workflow.

Model performance is evaluated using metrics such as **Accuracy, Precision, Recall, F1-score, and ROC-AUC**. Confusion matrices and ROC curves are used for deeper analysis.

After evaluation, **hyperparameter tuning** is performed using GridSearchCV or RandomizedSearchCV to optimize model parameters. The best model is selected based on high recall, F1-score, and stable performance.

Finally, the selected model generates **habitability probability scores**, which are used to rank exoplanets from most to least habitable. Feature importance analysis is performed to interpret results and provide scientific justification.