Work Flow :

**1. Understand the Dataset**

* **Explore Data Fields**: Begin by analyzing the columns in the CSV file. Typical fields may include:
  + Exoplanet mass, radius, semi-major axis, orbital period, etc.
  + Host star properties like temperature, luminosity, metallicity.
  + Detection methods (e.g., transit, radial velocity).
* **Check Data Quality**: Look for missing, incomplete, or noisy data. Understanding data limitations helps design preprocessing steps.

**2. Define Research Objectives**

Align your research with the dataset's content. Examples:

* **Detection**: Predict whether an observation corresponds to an exoplanet.
* **Characterization**: Predict physical or orbital properties (e.g., mass, radius) based on stellar and observational data.
* **Classification**: Identify patterns related to detection methods (e.g., which features predict success for transit vs. radial velocity).

**3. Preprocess the Data**

* **Handle Missing Values**: Impute or drop rows/columns with excessive missing data.
* **Normalize/Standardize Features**: Scale continuous variables to ensure ML models perform optimally.
* **Feature Engineering**: Create derived features like star-planet ratios, or binary flags for detection techniques.
* **Dimensionality Reduction**: Use techniques like PCA if you have many features.

**4. Choose a Machine Learning Approach**

**Supervised Learning**

* **Objective**: Build predictive models using labeled data.
  + **Binary Classification**: Classify whether an object is a planet.
  + **Regression**: Predict continuous characteristics (e.g., mass, radius).
* **Example Algorithms**:
  + Random Forests, XGBoost, or LightGBM for structured tabular data.
  + Neural Networks if the dataset is large and complex.

**Unsupervised Learning**

* **Objective**: Discover patterns without explicit labels.
  + **Clustering**: Group exoplanets based on similar properties.
  + **Dimensionality Reduction**: Understand feature importance and correlations.

**5. Split Data into Train and Test Sets**

* Use a **train-test split (80-20)** or **cross-validation** to assess model performance.

**6. Train and Evaluate Models**

* **Model Training**:
  + Experiment with multiple algorithms to identify the best-performing one.
  + Use metrics like **accuracy, precision-recall (for classification)** or **RMSE (for regression)**.
* **Hyperparameter Tuning**:
  + Optimize models using techniques like Grid Search or Bayesian Optimization.
* **Feature Importance**:
  + Analyze which features are most relevant for detection or characterization.

**7. Validate Insights**

* Cross-check machine learning findings with astrophysical theories to ensure interpretability.
* Use visualizations to validate:
  + Correlations (e.g., planet mass vs. detection method).
  + Model decision boundaries or feature contributions.

**8. Write Your Research Paper**

Organize your research into the following sections:

1. **Introduction**: Overview of machine learning applications in exoplanet research.
2. **Literature Review**: Discuss related studies.
3. **Methodology**:
   * Data preprocessing steps.
   * Algorithms and models used.
   * Validation strategies.
4. **Results**:
   * Model performance (tables, graphs).
   * Key insights derived from the data.
5. **Discussion**:
   * Compare your findings with established knowledge.
   * Limitations and future directions.
6. **Conclusion**: Summarize findings and contributions.

**9. Possible Extensions**

* Investigate **deep learning** models for feature extraction.
* Apply transfer learning using pre-trained models from astronomy datasets (e.g., Kepler or TESS data).
* Explore automated feature selection or neural architecture search.