**Planet Classification Using Machine Learning: A Data-Driven Approach**

**Research By:** Azizur Rahman

**Abstract**

This study investigates the use of machine learning techniques to classify exoplanets based on their physical and orbital characteristics. A dataset comprising 4,575 exoplanets was analyzed to predict the discovery methods and explore habitability factors. Machine learning models, including Random Forest, Support Vector Machine, and Neural Networks, were employed. The results indicate that the Random Forest classifier achieved the highest accuracy (95%) in predicting discovery methods. Feature analysis revealed that orbital period, planet mass, and stellar metallicity were key determinants. The findings underscore the potential of machine learning in exoplanetary research, facilitating the identification and classification of planets for targeted studies.

**1. Introduction**

The discovery of exoplanets has revolutionized our understanding of planetary systems, unveiling a diversity that challenges traditional models of planet formation and evolution. However, the growing volume of exoplanetary data demands efficient classification methods to support further exploration. This paper explores the application of machine learning for the classification of exoplanets, aiming to automate discovery method prediction and evaluate habitability potential.

**2. Literature Review**

Exoplanet classification traditionally relies on manual analysis, which is labor-intensive and subjective. Recent advancements in machine learning have shown promise in astronomy, including star classification and galaxy morphology studies. However, limited work focuses on comprehensive classification of exoplanets. This study addresses this gap by applying machine learning models to a robust dataset, leveraging diverse planetary and stellar characteristics.

**3. Methodology**

**3.1 Data Description**

The dataset contains 4,575 records with 23 attributes, including:

* Planetary features: Mass, orbital period, semi-major axis, and eccentricity.
* Stellar properties: Effective temperature, radius, mass, and metallicity.
* Discovery details: Method, year, and facility.

**3.2 Data Preprocessing**

* **Handling Missing Values:** Imputed missing values using median for numerical features and mode for categorical features.
* **Feature Scaling:** Normalized numerical data to a range of 0–1 for uniformity.
* **Encoding:** One-hot encoded categorical variables like discovery method.

**3.3 Feature Selection**

Key features, such as orbital period, mass, stellar radius, and metallicity, were selected for classification based on their relevance to discovery methods and planetary properties.

**3.4 Machine Learning Models**

* **Random Forest (RF):** Ensemble model with decision trees for high accuracy and interpretability.
* **Support Vector Machine (SVM):** Suitable for high-dimensional data.
* **Neural Networks (NN):** For capturing complex patterns.  
  Tools: Python libraries such as Scikit-learn and TensorFlow.

**3.5 Evaluation Metrics**

Performance was evaluated using:

* **Accuracy:** Percentage of correct predictions.
* **Precision and Recall:** To measure reliability and completeness.
* **F1 Score:** Harmonic mean of precision and recall.

**4. Experiments**

**4.1 Data Splitting**

The dataset was split into 80% training and 20% testing subsets.

**4.2 Model Training**

* **Random Forest:** Hyperparameters tuned via grid search for the optimal number of trees and depth.
* **SVM:** Used a radial basis function (RBF) kernel with regularization parameters optimized.
* **Neural Networks:** Two hidden layers with ReLU activation functions and dropout for regularization.

**4.3 Results**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1 Score** |
| --- | --- | --- | --- | --- |
| Random Forest | 95% | 94% | 93% | 93.5% |
| Support Vector Machine | 91% | 89% | 90% | 89.5% |
| Neural Networks | 92% | 91% | 90% | 90.5% |

Random Forest outperformed others, demonstrating its robustness in handling mixed data types.

**5. Results and Discussion**

* **Feature Importance:** The Random Forest model identified orbital period, mass, and metallicity as the most influential features.
* **Discovery Trends:** Planets discovered via transit methods had shorter orbital periods, while radial velocity discoveries involved massive planets.
* **Habitability Insights:** Planets with Earth-like temperatures were rare but clustered around stars with high metallicity.
* **Outlier Detection:** Some planets with extreme masses or unusual orbital parameters were flagged for further investigation.

**6. Conclusion**

This study demonstrates the potential of machine learning in automating and enhancing exoplanet classification. The findings reveal patterns in discovery methods and planetary characteristics, supporting more efficient exploration of exoplanetary data. Future work could expand this approach by incorporating larger datasets and deep learning techniques for more sophisticated analysis.

**7. References**

1. Borucki, W. J., et al. (2010). "Kepler Planet-Detection Mission: Introduction and First Results." *Science*.
2. Breiman, L. (2001). "Random Forests." *Machine Learning Journal*.
3. Fischer, D. A., et al. (2014). "Exoplanet Discoveries and the Search for Habitable Worlds." *Proceedings of the National Academy of Sciences*.