48. **Exploring Planetary Habitability Using Logistic Regression and Predictive Modeling**

**Abstract**

This research paper investigates the habitability of planets by employing various statistical models to analyze their planetary attributes, such as Solar Radiation, Atmospheric Composition, and Distance from Star. The study utilizes Logistic Regression, Linear Discriminant Analysis (LDA), and simpler models to predict whether a planet is habitable or not. The findings provide insights into the predictive capabilities of different models and their performance in classifying planets based on their key features. The analysis highlights the limitations and potential improvements for future research in the field of exoplanet classification.

**1. Introduction**

Predicting planetary habitability is a crucial challenge in exoplanet research. It involves understanding the conditions that allow a planet to support life. In this study, we analyze various planetary attributes, such as Solar Radiation, Atmospheric Composition, and Distance from Star, using Logistic Regression and other statistical models. We aim to evaluate these models' predictive accuracy and understand their strengths and weaknesses in classifying planets as habitable or non-habitable.

**2. Data Overview and Statistical Modeling**

The dataset contains 200 observations across five key features: Solar Radiation, Atmospheric Composition, Distance from Star, Planet, and Habitability. The primary goal is to use these features to predict the binary outcome of habitability (0 for non-habitable, 1 for habitable).

**2.1 Planetary Features**

1. **Solar Radiation**: Measures the solar energy received by each planet.
2. **Atmospheric Composition**: The makeup of the planet's atmosphere, including essential gases and harmful elements.
3. **Distance from Star**: The distance of the planet from its star, affecting its temperature and habitability potential.
4. **Habitability**: A binary classification indicating whether a planet is habitable (1) or not (0).

**3. Model Interpretations**

**3.1 Logistic Regression Model for Planetary Habitability**

The logistic regression model was applied to predict habitability using Solar Radiation, Atmospheric Composition, and Distance from Star as predictors.

* **Model Summary**:
  + **Coefficients**:
    - **Intercept**: -0.1455750, representing the baseline log-odds of habitability when all features are zero.
    - **Solar Radiation**: -0.0003044, suggesting that increasing solar radiation slightly decreases the odds of habitability, but this result is not statistically significant (p = 0.569).
    - **Atmospheric Composition**: 0.0039738, indicating a positive but non-significant relationship (p = 0.417).
    - **Distance From Star**: 0.0005712, also not statistically significant (p = 0.918).
  + **Model Fit**:
    - Residual deviance = 275.90 on 196 degrees of freedom, indicating the model's goodness of fit.
    - AIC = 283.9 suggests a relatively moderate model fit.
    - **Accuracy**: The test set accuracy is approximately 51.67%, indicating the model's ability to predict planetary habitability is only slightly better than random guessing.
* **Decision Boundary and Probability Predictions**:
  + The logistic regression decision boundary is linear, separating habitable and non-habitable planets based on solar radiation and atmospheric composition. However, there is substantial overlap between the two classes, indicating a low discrimination capability of the logistic regression model.
  + The scattered nature of predicted probabilities across different levels of solar radiation suggests no clear trend, confirming the low predictive power.

**3.2 Linear Discriminant Analysis (LDA) for Planetary Classification**

LDA was applied to classify the planets' habitability based on the same features.

* **Group Means**:
  + The group means for each feature (Solar Radiation, Atmospheric Composition, and Distance from Star) were computed for the two groups (0 for non-habitable, 1 for habitable).
* **Coefficients of Linear Discriminants**:
  + **Solar Radiation**: 0.0014319968
  + **Atmospheric Composition**: 0.024367678
  + **Distance from Star**: 0.002635648
* **Accuracy**:
  + The LDA model achieved a test accuracy of 52.54%, slightly higher than the logistic regression model, indicating a marginally better classification performance.

**4. Results and Analysis**

The results of the logistic regression and LDA models suggest that the prediction of planetary habitability using these features is challenging due to low discriminatory power. Both models show moderate performance, with test accuracies around 51-55%.

* **Confusion Matrix Analysis**:
  + The confusion matrix for the logistic regression model shows 21 true negatives, 23 true positives, 6 false negatives, and 10 false positives, resulting in an overall accuracy of 51.67%.
  + A simpler model achieved a slightly higher accuracy of 55%, indicating potential overfitting in more complex models.

**5. Discussion of Findings**

The analysis reveals that while traditional statistical models like logistic regression and LDA can provide initial insights into planetary habitability, their predictive power is limited by the overlap between the habitable and non-habitable classes. The decision boundaries are not well-defined, and the models struggle to distinguish between the two classes accurately.

* **Solar Radiation**: Across all models, solar radiation shows a weak association with habitability, potentially indicating that the planets in this dataset have a diverse range of tolerances to solar exposure.
* **Atmospheric Composition**: The variability in atmospheric composition suggests that atmospheric elements play a critical role in determining habitability. However, the assumption of linear relationships may oversimplify these complex interactions.
* **Distance from Star**: The impact of the distance from the star is not consistently significant across models, implying that other unobserved factors may be influencing habitability.

**6. Conclusion and Future Directions**

This research explores the use of logistic regression, LDA, and simpler models to predict planetary habitability based on key attributes. The findings indicate that while these models provide some insights, their performance is limited by the low discriminative power and assumptions of linearity. Future studies could benefit from richer datasets, advanced feature engineering, and state-of-the-art machine learning algorithms to improve predictive performance and deepen our understanding of exoplanet habitability.

* **Proposed Actions**:
  + Data Augmentation: Collect and integrate more granular planetary data to enrich feature sets.
  + Model Experimentation: Test alternative models like Random Forests, Gradient Boosting Machines, or Deep Neural Networks for improved accuracy.
  + Validation Techniques: Implement cross-validation techniques to assess model generalizability and mitigate overfitting risks.