47.**Modeling Planetary Habitability A Comparative Study of Logistic Regression, LDA, and Other Classifiers**

**Abstract**

This research paper investigates the application of various statistical models, including Logistic Regression and Linear Discriminant Analysis (LDA), to predict the habitability of planets based on their attributes such as Solar Radiation, Atmospheric Composition, and Distance from Star. The dataset comprises 200 observations of hypothetical planets with diverse characteristics. The study aims to identify the strengths and weaknesses of each modeling approach and evaluate their effectiveness in predicting planetary habitability. The results suggest that while Logistic Regression provides a basic linear decision boundary, LDA offers a more flexible approach with varying predictive accuracy. The findings highlight the need for more sophisticated models to capture the complexities of planetary classification.

**1. Introduction**

Predicting the habitability of exoplanets is a crucial challenge in planetary science. This study explores different statistical models to classify planets based on their likelihood of being habitable, using various features such as Solar Radiation, Atmospheric Composition, and Distance from Star. We apply Logistic Regression and Linear Discriminant Analysis (LDA) to evaluate their performance and understand the underlying relationships between planetary characteristics and habitability status.

**2. Data Overview**

The dataset comprises 200 observations of planets, each described by four key features:

* **Solar Radiation:** The solar energy received by the planet, which impacts its surface temperature and potential to support life.
* **Atmospheric Composition:** The makeup of the planet's atmosphere, which may include gases necessary for sustaining life.
* **Distance from Star:** The planet's distance from its star, influencing its thermal environment.
* **Habitability:** A binary classification indicating whether the planet is habitable (1) or not (0).

**3. Methodology and Model Interpretations**

**A. Logistic Regression Model for Planetary Habitability**

The Logistic Regression model was applied to predict the habitability of planets based on Solar Radiation, Atmospheric Composition, and Distance from Star. The model provides a linear decision boundary for classifying planets into habitable and non-habitable categories.

* **Model Coefficients:**
  + **Intercept:** -0.1455750, which represents the baseline log-odds of habitability when all other features are zero.
  + **Solar Radiation:** -0.0003044, suggesting a slight decrease in the odds of habitability with increasing solar radiation (p = 0.569).
  + **Atmospheric Composition:** 0.0039738, indicating a positive but statistically insignificant relationship (p = 0.417).
  + **Distance from Star:** 0.0005712, also not statistically significant (p = 0.918).
* **Model Fit:**
  + The model's residual deviance is 275.90 on 196 degrees of freedom, with an AIC of 283.9, suggesting a moderate fit.
  + Test set accuracy is approximately 51.67%, indicating the model's predictive ability is only marginally better than random guessing.

**B. Decision Boundary and Probability Predictions**

* **Logistic Regression Decision Boundary:**
  + The decision boundary plot shows a linear separation between habitable and non-habitable planets based on Solar Radiation and Atmospheric Composition. However, there is significant overlap between the two classes, demonstrating a low discrimination capability.
* **Predicted Probability Plot:**
  + The scattered predicted probabilities across different levels of Solar Radiation indicate no clear trend, confirming the low predictive power of the Logistic Regression model.

**C. Linear Discriminant Analysis (LDA) for Planetary Classification**

The LDA model aims to maximize the separation between the classes by modeling the differences in the means of the classes, assuming equal variance-covariance structures.

* **Group Means and Prior Probabilities:**
  + Prior probabilities of habitability classes are approximately 0.5177 (non-habitable) and 0.4827 (habitable).
  + Mean values for Solar Radiation, Atmospheric Composition, and Distance from Star differ between the two groups, suggesting some degree of class separation.
* **Coefficients of Linear Discriminants:**
  + **Solar Radiation:** 0.00143961, indicating a small contribution to the discriminant function.
  + **Atmospheric Composition:** 0.02436768, suggesting a moderate impact on the classification.
  + **Distance from Star:** 0.002635648, showing a minimal effect.
* **LDA Model Accuracy:**
  + The LDA model achieves an accuracy of approximately 52.54%, similar to the Logistic Regression model, indicating that it does not significantly improve classification performance.

**D. Visualization of Results**

* **Decision Boundary Plots:**
  + The Logistic Regression decision boundary is linear and fails to provide a clear separation between the habitable and non-habitable planets.
  + The LDA decision boundary suggests a better but still limited ability to differentiate between classes.
* **Predicted Habitability Plots:**
  + The scatter plots of predicted habitability by Logistic Regression and LDA reveal overlapping regions, confirming the models' limited discriminative power.

**4. Discussion of Findings**

The analysis demonstrates that both Logistic Regression and LDA have limited success in accurately classifying planetary habitability. The low accuracy rates (approximately 52.54%) indicate that these linear models struggle to capture the complexities inherent in planetary data.

* **Solar Radiation Impact:**
  + Both models show a weak association between Solar Radiation and habitability, suggesting that this feature alone is not a strong predictor.
* **Atmospheric Composition and Distance from Star:**
  + These features also exhibit low significance in the models, indicating that additional factors may be critical in determining habitability.

**5. Conclusion and Future Directions**

The study highlights the limitations of simple linear models like Logistic Regression and LDA in predicting planetary habitability. The low accuracy rates suggest the need for more advanced models that can account for non-linear relationships and complex interactions among features.

**Proposed Actions:**

* **Data Augmentation:** Incorporate more diverse planetary features, such as surface temperature and gravitational force, to enhance model inputs.
* **Advanced Modeling Techniques:** Experiment with more sophisticated models like Random Forests, Gradient Boosting Machines, or Neural Networks to improve predictive accuracy.
* **Cross-Validation:** Implement robust cross-validation techniques to evaluate model performance and prevent overfitting.