Predicting the Habitability of Exoplanets Using Machine Learning

A Machine Learning–Based Classification of Exoplanet Habitability

By Mohite Swaraj Sanjay

swarajmohite16@gmail.com

Infosys Springboard Internship Project

Comprehensive Technology Stack

Frontend

HTML5, CSS3, JavaScript for an interactive and responsive user interface.

Backend

Python and Flask for robust REST API development.

Machine Learning

XGBoost, Random Forest, SVM, Logistic Regression, KNN, Naive Bayes for diverse model training.

Data & Visualisation

Pandas, NumPy, Scikit-learn, imbalanced-learn, Matplotlib, Seaborn, PCA, t-SNE for data handling and analysis.

Deployment

Netlify for frontend and Render for backend, with Git & GitHub for version control.

Core Features of Our System

Habitability Prediction

Accurately predicts exoplanet habitability classes.

Class Imbalance Handling

Effectively manages extreme data imbalance for rare habitable planets.

Multi-Model Comparison

Trains and evaluates various machine learning models.

Data Visualisation

Utilises PCA, t-SNE, and confusion matrices for insightful data representation.

Full-Stack Web System

A complete web-based prediction system, accessible and interactive.

Live Deployment

The system is fully deployed and available for live access.

Navigating Key Challenges

Severe Class Imbalance

Habitable planets constitute less than 1% of the dataset, posing a significant challenge for model training.

Missing & Noisy Data

Astronomical data often contains high levels of missing values and inherent noise, impacting data quality.

High-Dimensional Feature Space

Dealing with numerous features requires careful handling to avoid the curse of dimensionality.

Oversampling Overfitting Risk

Techniques to address imbalance, like oversampling, risk introducing overfitting into the models.

Integration & Deployment Hurdles

Backend–frontend integration and resolving CORS configurations presented complex deployment challenges.

System Architecture Flow

- User Interface
- Flask Backend
- Preprocessing
- XGBoost Model

This diagram illustrates the sequential data flow, from user interaction through machine learning processing, to the final display of results.

Addressing Challenges with Strategic Solutions

Challenge: Class Imbalance

Habitable planets are extremely rare in the dataset.

Solution: Resampling & Weighting

Implemented SMOTE, SMOTE-Tomek, and class weighting.

Challenge: Overfitting

Risk of models performing poorly on unseen data.

Solution: Feature Engineering

Applied feature selection and regularisation techniques.

Challenge: Missing Data

Incomplete and sparse astronomical observations.

Solution: Imputation Strategies

Utilised median and mode imputation for missing values.

Challenge: High Dimensionality

Too many features complicating model training.

Solution: Dimensionality Reduction

Employed PCA and correlation analysis.

Challenge: Deployment Issues

Integrating frontend and backend for live accessibility.

Solution: Separate Hosting

Hosted frontend and backend independently to resolve issues.

Measurable Outcomes and Positive Impact

Exceptional Performance

XGBoost achieved a Macro F1-score of approximately 0.96, ensuring high performance.

High Minority-Class Recall

The model demonstrates strong recall for the crucial, rare habitable planet class.

Stable & Consistent Predictions

Delivers reliable and consistent predictions for exoplanet habitability.

Astronomical Assistance

Aids astronomers in rapidly identifying potentially habitable exoplanets.

Real-World ML Application

Showcases a practical application of machine learning in scientific discovery.

Educational & Reusable Tool

Serves as an educational resource for learners and a fully deployed, reusable ML system.

Future Enhancements: A Roadmap to Discovery

1. **Deep Learning Integration**

   Explore deep learning models using light curve data for enhanced accuracy.

2. **New Data Integration**

   Incorporate the latest datasets from NASA and JWST missions.

3. **Explainability Improvement**

   Develop methods for improved explainability of model predictions.

4. **Advanced Analytics**

   Implement sophisticated dashboards and analytical tools for deeper insights.

5. **Continuous Learning**

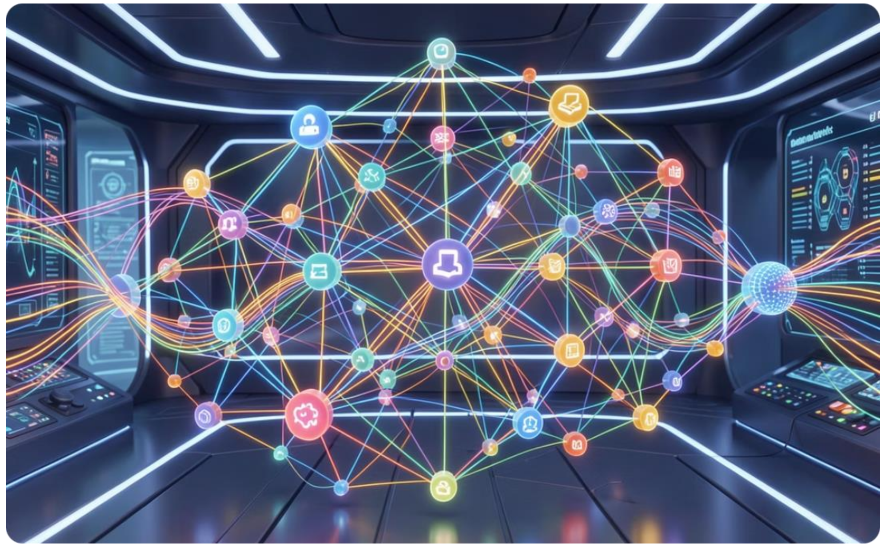
   Establish a framework for continuous model improvement with new astronomical discoveries.

Project Summary

Our project successfully developed a machine learning-based system to predict exoplanet habitability, addressing significant data challenges.

Exoplanet Discovery

Leveraging ML for new insights into distant worlds.

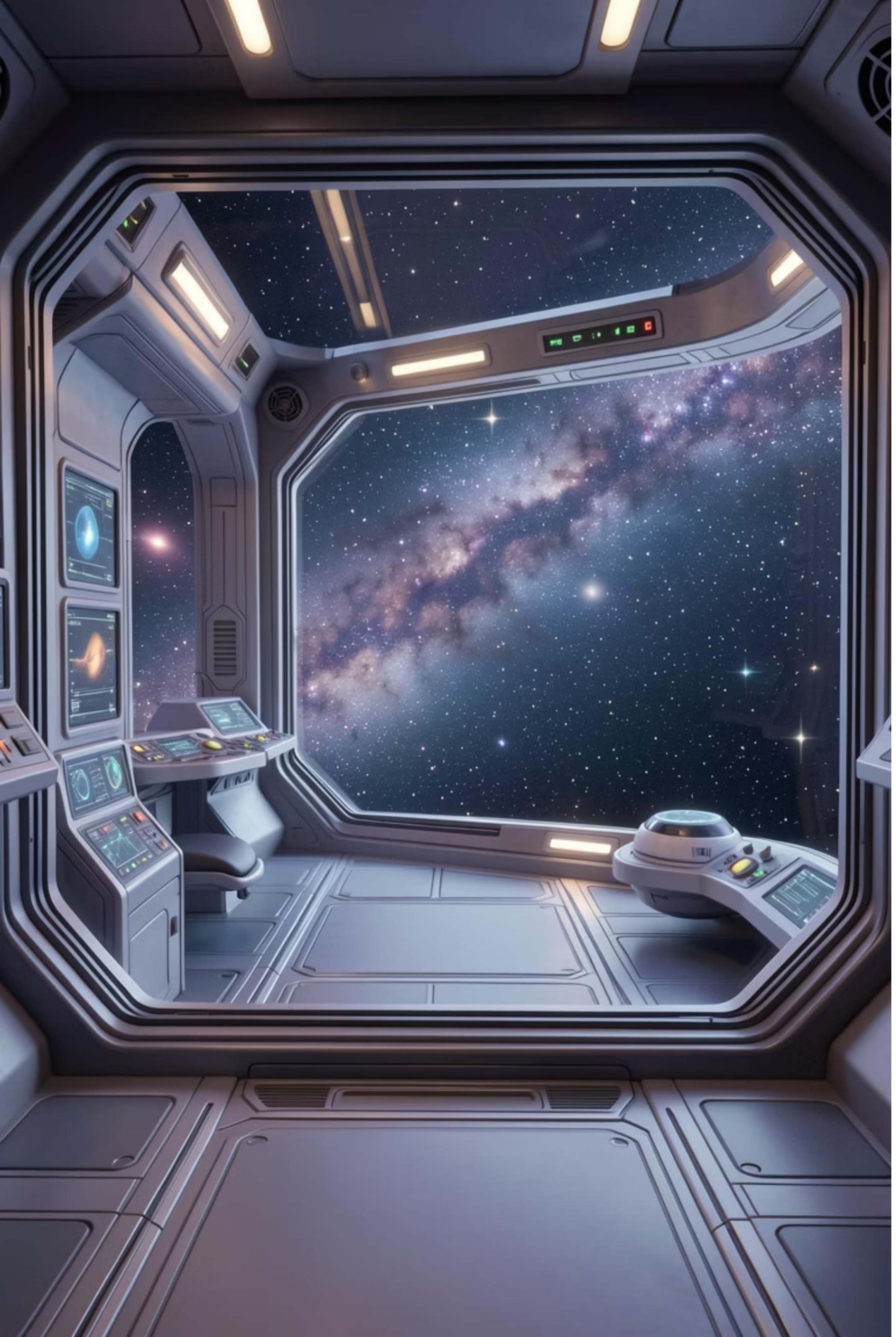


Advanced ML Techniques

Robust models handling complex astronomical data.

Interactive Web Platform

User-friendly access to powerful predictive analytics.



Thank You

Mohite Swaraj Sanjay