EXOPLANET HABITABILITY PREDICTION SYSTEM

An AI & Machine Learning Approach for Predicting the Habitability Potential of Exoplanets

Name: Rushitha Konangi
Branch: Electronics and Communication Engineering

ABSTRACT

The rapid discovery of exoplanets through astronomical missions such as NASA's Kepler and TESS has resulted in massive datasets that are difficult to analyze manually. This project focuses on building an AI and Machine Learning–based system to predict the habitability potential of exoplanets using planetary and stellar parameters. The system includes data cleaning, feature engineering, handling class imbalance, machine learning model development, visualization, and deployment. The project follows a structured learning-based approach and demonstrates practical application of machine learning concepts in a real-world scientific problem.

INTRODUCTION

Exoplanets are planets that orbit stars outside our solar system. Although thousands of exoplanets have been discovered, only a very small number are potentially habitable. Habitability depends on factors such as planet radius, mass, orbital distance, temperature, and host star properties.

Manual analysis of such large astronomical datasets is inefficient. Machine learning provides an automated approach to analyze complex patterns and predict habitability. This project applies supervised machine learning techniques to classify exoplanets as habitable or non-habitable and rank them based on predicted habitability scores.

PROBLEM STATEMENT

The goal of this project is to predict whether an exoplanet is habitable or not using machine learning. The main challenges include missing values, noisy data, lack of a direct habitability label, and severe class imbalance where habitable planets are extremely rare.

OBJECTIVES

- To understand and analyze real-world astronomical datasets
- To clean and preprocess exoplanet data
- To create habitability labels and scores
- To handle imbalanced datasets using SMOTE
- To train and evaluate machine learning models
- To visualize insights using plots and dashboards
- To deploy the system on a cloud platform

DATASET DESCRIPTION

The dataset used in this project was obtained from the NASA Exoplanet Archive via Kaggle. It contains planetary parameters such as planet radius, mass, orbital distance (semi-major axis), equilibrium temperature, and stellar properties like star temperature and luminosity. Since the dataset does not contain a habitability label, a custom habitability classification was created based on scientific thresholds.

SYSTEM ARCHITECTURE

The system follows a pipeline-based architecture:

Dataset → Data Cleaning → Feature Engineering → Class Balancing → Model Training → Evaluation → Visualization → Deployment

DATA PREPROCESSING

Data preprocessing involved:

- Handling missing values using mean and median imputation
- Removing unnecessary columns
- Detecting and treating outliers using boxplots and Z-score
- Fixing inconsistent values
- Validating data quality using descriptive statistics

EXPLORATORY DATA ANALYSIS (EDA)

EDA was performed using:

- Histograms to analyze feature distributions
- Boxplots to detect outliers
- Scatter plots to understand feature relationships
- Correlation heatmaps to identify influential features

EDA revealed that planet radius, orbital distance, and temperature strongly influence habitability.

FEATURE ENGINEERING

Feature engineering included:

- Creation of a **Habitability Score Index** using planetary conditions
- Creation of a **Stellar Compatibility Index** using star properties
- Feature scaling using StandardScaler
- Dimensionality reduction using PCA
- Visualization of clusters using t-SNE

These steps improved model efficiency and performance.

HANDLING CLASS IMBALANCE

The dataset was highly imbalanced with very few habitable planets. To address this:

- Stratified train-test split (80:20) was applied
- SMOTE was used only on the training dataset

This ensured unbiased learning and fair evaluation.

MACHINE LEARNING MODELS

The following models were implemented:

- Logistic Regression (baseline model)
- Support Vector Machine (SVM)
- Random Forest Classifier
- XGBoost (multi-class habitability prediction)

Random Forest performed best due to its ability to handle non-linear relationships.

MODEL TRAINING AND EVALUATION

Models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

- Confusion Matrix

Random Forest achieved the highest and most stable performance. Exoplanets were ranked based on predicted habitability probability.

VISUALIZATION AND DASHBOARD

Visualization included:

- Histogram of planet radius
- Boxplot of orbital distance
- Scatter plot of radius vs distance
- Correlation heatmap
- Feature importance plots

An interactive dashboard was developed using Chart.js to display rankings and insights.

REST API AND BACKEND

A REST API was developed using FastAPI to send planetary data to the trained model and receive predictions. FastAPI was chosen for its support of asynchronous requests, high performance, and scalability.

DEPLOYMENT

The application was deployed on the Render cloud platform. The deployment process included pushing code to GitHub, configuring the cloud environment, and running the API on a Linux-based server.

RESULTS

The system successfully predicts and ranks exoplanets based on habitability potential. Visualization outputs validate the importance of key planetary features. The deployed application demonstrates end-to-end functionality.

LIMITATIONS

- Dependence on indirect habitability indicators
- Limited availability of complete astronomical data
- Possible overfitting due to small habitable samples

FUTURE ENHANCEMENTS

- CNN-based light curve analysis
- SHAP-based explainable AI
- Real-time dataset integration
- Advanced interactive dashboards

CONCLUSION

This project demonstrates how machine learning can be effectively applied to astronomical data to predict exoplanet habitability. Through structured learning, preprocessing, modeling, visualization, and deployment, the system provides a scalable and practical solution for habitability analysis.

REFERENCES

- NASA Exoplanet Archive
- Kaggle Exoplanet Datasets
- Research papers on exoplanet detection and habitability
- GitHubRepository:

https://github.com/springboardmentor74280b-design/Habitability-of-Exoplanets/tree/rushitha-konangi

Deployed Application:

https://habitability-of-exoplanets-2.onrender.com/

Project Video Demo:

https://onedrive.live.com/?qt=allmyphotos&photosData=%2Fshare%2F7455FACDCC191830%21s7ba3f36466c141838e8686b421bd15d1%3Fithint%3Dvideo%26e%3DGx7JbM%26migratedtospo%3Dtrue&cid=7455FACDCC191830&id=7455FACDCC191830%21s7ba3f36466c141838e8686b421bd15d1&redeem=aHR0cHM6Ly8xZHJ2Lm1zL3YvYy83NDU1ZmFjZGNjMTkxODMwL0lRQms4Nk43d1dhRFFZNkdoclFodlJYUkFaUUhGSDJZb3ItSVAxand4RzQzQ1dnP2U9R3g3SmJN&v=photos