Identification of Exoplanet Orbitals

Name:	**ASHISH KUMAR**
Registration No./Roll No.:	21058
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE, BS-Engineering
Problem Release date:	August 17, 2023
Date of Submission:	November 19, 2023

1 Introduction

Identification of Exoplanet Orbitals data analysis is the study of identifying known periodic drops in light intensity of the star, known as "Transit Drops".A regular and prominent occurrence of this "Transit drop" in data can indicate the presence of an exoplanet or exomoon. An exomoon or extra-solar moon is a natural satellite that orbits an exomoon or other non-stellar extrasolar body. Hence successful classification of these transit drops is an important part of Exoplanet detection in the Kepler Mission of NASA.Traget variable values has in range between (-5095,7506), average is 1.606558 and median is 0.0832 and mode is 0, number of instances in training data is 17971 and in test data is 1998, number of features in training data is 253 and in test data is 253.Various machine learning algorithms were applied to achieve the desired Regression outcome. Among these,K-Nearest Neighbors Algorithm emerged as the most effective, yielding optimal evaluation results.

2 Methods

The comprehensive methodology for this exoplanet identification project is designed to navigate through various stages, ensuring a systematic and effective approach to address the complexities inherent in the dataset. It commences with the initialization of essential libraries, including pandas, numpy, and scikit-learn, setting the stage for subsequent data exploration. The initial data loading involves reading the 'exoplanet_trn_data.csv' dataset, a pivotal step in understanding its inherent structure and characteristics.Moving into the data preprocessing phase, the project strategically addresses missing values. Columns exceeding a 50 percent threshold of missing data are systematically removed to maintain the integrity and reliability of the dataset. Further enhancing the dataset's suitability for regression tasks, categorical columns ('soltype,' 'discoverymethod,' 'pl_letter,' and 'disc_locale') undergo the transformation of One-Hot Encoding, facilitating the inclusion of categorical information in subsequent modeling.A discerning step involves the removal of irrelevant columns ('pl_name,' 'hostname,' etc.) that do not contribute significantly to the regression task. This streamlining ensures a focused consideration of pertinent features during model training and evaluation.The process then shifts to handling the target variable, which is loaded from The dataset is judiciously split into training and validation sets, a critical step in gauging model performance accurately.The heart of the methodology lies in the selection and training of regression models. The project explores a diverse set of models, starting with the Linear Regression approach. This involves incorporating feature scaling, selection, and hyperparameter tuning to optimize model performance. Following this, Decision Tree Regression and Random Forest Regression are employed, each subject to meticulous hyperparameter tuning processes.The methodology extends its purview to encompass K-Nearest Neighbors (KNN) Regression, Ridge Regression, and Support Vector Regression (SVR). For SVR, additional considerations include feature scaling and selection, tailoring the approach to the unique characteristics of the dataset. The ensemble methods are rounded off with the implementation of AdaBoost Regression, incorporating hyperparameter tuning for fine-tuning.The evaluation phase is marked by the meticulous assessment of each regression model's performance using robust metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R2-Score. These metrics provide a comprehensive understanding of the models' efficacy and their suitability for the specific exoplanet identification task. Github-link

3 Experimental Setup

The experimental setup for the exoplanet identification project is a methodical process designed to rigorously evaluate and compare various regression models. In evaluating the performance of the proposed regression models for the exoplanet identification task, key metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R2-Score were employed. The Mean Squared Error, a measure of the average squared difference between the predicted and true values, provided insight into the overall accuracy of the models. Additionally, the Root Mean Squared Error, representing the square root of MSE, offered a normalized measure of the average magnitude of prediction errors. Complementing these, the R2-Score, indicating the proportion of variance in the target variable explained by the models, served as a valuable metric for assessing predictive performance.Upon analysis of the results, a lower MSE and RMSE, along with a higher R2-Score, collectively indicated superior predictive capabilities of the regression models in the exoplanet identification context. These metrics not only quantified the accuracy of predictions but also provided a measure of the models' ability to capture underlying patterns in the dataset.Furthermore, for a more comprehensive understanding and meaningful comparison, it is imperative to consider state-of-the-art results in the field of exoplanet identification. Benchmarking against existing literature or established models allows for a nuanced assessment of the proposed methodology's efficacy. State-of-the-art results, encompassing metrics achieved by cutting-edge models or methods on analogous datasets, offer a valuable benchmark for gauging the relative performance of the proposed regression models. This comparative analysis aids in discerning the strengths and potential areas for improvement in the proposed approach, paving the way for advancements in the field of exoplanet identification.

Table 1: Performance Of Different Regression Using preprocessed data

Models	MSE	RMSE	R2-Score
Adaptive Boosting	580.34	24.09	-3.045
Decision Tree	225.04	15.00	-0.56
K-Nearest Neighbor	126.57	11.25	0.11
Linear Regression	349.51	18.69	-1.43
Random Forest Regression	142.07	11.91	0.0096
SVM Regressor	143.04	11.96	0.0029
Ridge Regression	347.26	18.63	-1.42

4 Results and Discussion

In the evaluation of various regression models for exoplanet identification Table 1, the Adaptive Boosting model exhibited challenges with a high MSE of 580.34 and a negative R2-Score of -3.045, indicating a significant disparity between predicted and actual values. The Decision Tree model performed relatively better with an MSE of 225.04 and a negative R2-Score of -0.56. The K-Nearest Neighbor model demonstrated improved predictive capabilities, yielding an MSE of 126.57 and a positive R2-Score of 0.11. However, Linear Regression, Random Forest Regression, SVM Regressor, and Ridge Regression faced challenges, displaying negative R2-Scores and limited accuracy in capturing the underlying patterns in the dataset. Notably, the K-Nearest Neighbor model emerged as the most promising, though further refinement may be necessary for optimal performance. These findings highlight the intricacies of the exoplanet identification task and suggest avenues for improvement, such as feature engineering and hyperparameter tuning.

5 Conclusion

The evaluation of regression models for exoplanet identification revealed that the K-Nearest Neighbor model demonstrated the most promising predictive capabilities, with relatively lower MSE and a positive R2-Score. This suggests that the model captured a notable portion of the target variable's variance. However, the overall negative R2-Scores and challenges observed in other models indicate the complexity of the task. Future work could focus on further refining the K-Nearest Neighbor model, exploring advanced ensemble methods, and delving into feature engineering techniques to enhance the models' ability to discern underlying patterns in the dataset. Additionally, investigating alternative regression approaches and leveraging more sophisticated algorithms may contribute to improved performance in exoplanet identification.

References

[1]The dataset was retrieved from the NASA Archive https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTblsconfig=PS

[2]https://towardsdatascience.com/detecting-habitability-of-exoplanets-with-machine-learning-b28c2d82576

[3]https://github.com/ashishkumar0803/DSE317–Semester-Project-Machine-Learninng.git