Machine Learning Approaches for Predicting Exoplanet Livability: A Comprehensive Analysis

Siwei Wang[1]

[1] Guanghua Cambridge International School, 2788 Chuanzhou Road, Pudong District, Shanghai, China
3606918778@qq.com

Abstract. Now, the continuous advancement of astronomical observation equipment has led to a steady rise in the quantity of identified exoplanets. Hence, there is a pressing need for the development of an efficient and practical approach to forecasting the livability indices of these exoplanets. The objective of this study is to provide a comprehensive overview and evaluation of the methodologies employed in predicting the livability quotient of exoplanets through the utilization of machine learning techniques. This research endeavor holds the potential to enhance astronomers' capacity to discern habitable exoplanets with more accuracy and precision. This study does preliminary data processing on the observations obtained from Kepler's astronomical telescope. Subsequently, via an examination of the fundamental characteristics of exoplanets, this study puts forth many theoretical frameworks, ultimately employing a linear regression model to ascertain analogous functional associations among variables. In the paragraph on the experiment and application, this paper describes the whole experiment and its results. Some graphs with data are also added. At the end of this review, the author summarizes the research results, including the summarization of the methods and experimental results of using machine learning to predict the evaluation rate of planetary habitability. This paper also gives ideas about the deployment and application of this research.

Keywords: Linear Regression, Planet Livability Rate, Machine Learning Prediction

1 Introduction

In the following paragraphs, the author will first give reviewing of the research of exoplanets in the field of astronomy, which summarizes the current datasets of exoplanets, research methods, and application achievements.

Subsequently, the methodologies employed for resolving this issue will be presented. The provided resources encompass methodologies for acquiring and preprocessing datasets pertaining to exoplanets, encompassing data cleansing and feature selection techniques. This section will discuss the machine learning algorithms employed for predicting the livability rate of exoplanets, including linear regression

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_69

and decision trees. Additionally, it will provide a comprehensive explanation of feature engineering techniques, outlining the methods for extracting and selecting features that enhance the algorithm's ability to accurately predict the livability rate of exoplanets. Then, the introduction point will move to experiments and applications. The author will analyze the datasets of exoplanets, including the summarization of used datasets and explanations of how to extract information about exoplanet features in them.

The author will conclude this review by presenting a thorough examination of the study's results. This will encompass a summary of the methodologies utilized and the experimental outcomes related to the application of machine learning techniques in predicting rates of planetary habitability assessment. Following this, the author will proceed to examine the advantages, disadvantages, and limitations linked to the aforementioned methodology. The author's objective is to provide an analysis of the advantages and challenges related to the application of machine learning methodologies. Furthermore, an analysis will be conducted to evaluate the potential limitations of the current model in relation to its predictive capabilities. Lastly, the author will now shift their focus towards prospective areas for further investigation. The author aims to present a potential path for the development of assessing the habitability of planets and making predictions about research findings using machine learning methodologies.

2 Review of the Related Search

The research status of astronomical object recognition methods is based on the data quality.

Observation of data in astronomy is mainly divided into two types: spectral data and image data.

2.1 Spectral data

Celestial spectroscopy:

The inception of celestial spectroscopy may be attributed to Isaac Newton's utilization of a dispersed prism in order to examine sunlight. The observer witnessed the spectral phenomenon known as the rainbow and may observe several absorption lines. The detailed characterization of the dark lines observed in the solar spectrum was initially provided by Joseph Fraunhofer. The solar spectrum has two primary characteristics that are often observed in most stellar spectra. Firstly, it encompasses all wavelengths that emit visible light, resulting in a continuous spectrum. Secondly, it displays several absorption lines that are dispersed across the spectrum, leading to the presence of multiple gaps in the radiation.

The luminosity shown by planets and asteroids is only a result of the reflection of light emitted by their respective parent stars. However, this reflected light encompasses not only the inherent properties of these celestial bodies but also incorporates the presence of minerals originating from stony celestial entities, as well as absorption lines attributable to the elements and molecules present within the atmospheres of gas giants. Asteroids may be classified into three primary categories

according to their spectral characteristics: C-type asteroids mostly consist of carbonaceous material, S-type asteroids are predominantly formed of silicates, and M-type asteroids are characterized by a higher abundance of metallic elements. The most prevalent types of asteroids are C-type and S-type.

2.2 Image data

Transit method:

When celestial bodies such as planets undergo a transit event, when they traverse the path between their host star and an observer, they have the potential to reveal their existence by obstructing the emitted light from these stars. The phenomenon in which planets traverse the space between stars and Earth is commonly referred to as a 'transit'. If the recurring detection of star dimming persists for a consistent and repetitive duration, it is probable that a celestial body with a lower luminosity is in orbit around the star. Certain celestial objects in transit may consist of faint and diminutive stars, commonly referred to as eclipsing binary stars. However, the majority of such objects are indeed planets.

The extent of luminosity reduction shown by stars during the phenomenon of transit is strongly correlated with the comparative magnitudes of stars and planets. When an asteroid traverses a massive star, it causes a little reduction in the star's luminosity. Conversely, when a sizable planet transits a relatively tiny star, it generates a more discernible impact on the star's brightness. The determination of the host star's size from its spectrum enables astronomers to make estimations regarding the diameter of the planet using photometry techniques. Nevertheless, the photometry approach lacks the capability to forecast the quality of the data, hence rendering it a suitable complement to the radial velocity method. The radial velocity approach is capable of approximating the minimum mass threshold of a planet; nevertheless, it does not yield any insights into the planetary diameter. Through the integration of these two methodologies alongside the planetary mass and diameter, researchers are able to compute the density of a planet. This calculation facilitates the determination of whether a planet possesses a stony composition, a gaseous constitution, or lies somewhere in between these two states.

Advantages of the transit method:

The transit method is the most useful and sensitive currently, especially for those space observation stations that can observe for weeks to months.

Transit photometry offers astronomers the ability to approximate the diameter of planets, a physical characteristic that remains elusive to measurement by alternative methodologies. The utilization of transit and radial velocity methods to study exoplanets necessitates that the orbital plane of the observed exoplanets be oriented toward the observer on Earth. This alignment enables the acquisition of data that may be used to make inferences on the mass, density, and composition of the exoplanet in question.

Disadvantages of the transit method:

A distant planet must pass directly between its host star and Earth, but this situation does not happen very often. To observe a transit, its orbital plane must be almost

completely lateral towards the observer, which means most exoplanets and their features can not be observed with this method.

3 Mythology

Selecting a machine learning model that is suitable for the task. In this task, common models include decision trees, random forests, support vector machines, neural networks, etc. There are multiple models available to select the most appropriate one.

Decision Tree Model:

A decision tree is a hierarchical structure employed to partition data into distinct decision paths for the purposes of classification or regression. The utilization of this modeling strategy in data mining is widespread for the purpose of achieving categorization functionality. Within the context of the decision tree, it is important to note that each internal node serves as a representation of a certain feature, while each branch is indicative of a condition pertaining to a particular feature value. The terminal nodes in a decision tree reflect the ultimate outcomes of categorization or regression.

The idea behind this was first put forth by Hunt et al. in 1966 and is the foundation for several decision tree approaches. It seeks to create the best decision tree. Later, based on the notion of entropy in information theory, Quinlan et al. presented the traditional ID3 algorithm for decision trees. Its fundamental idea is to create a binary tree decision model utilizing the attribute with the maximum information gain as the training and testing attribute for the current node, in accordance with the notions of information entropy and information.

The criteria for branching decision trees are directly dependent on the magnitude of entropy. The purity of training subset partitioning is higher the lower the information entropy. when each data record has been divided.

When all data records belong to the same category or have the same attributes, categorization is halted[1].

i. Example formula for a decision tree: In a decision tree, a series of conditions (features) are used to segment data and divide it into different categories or values. Here is an example formula: Assuming astronomical data is used to classify stars, one feature is the surface temperature (T) of the star, and the other feature is the luminosity (L) of the star.

1. If T<5000 K, then the star may be a red giant.
2. If T>=5000 K and L>1000, then the star may be a main sequence star.
3. If T>=5000 K and L<=1000, then the star may be a subgiant.

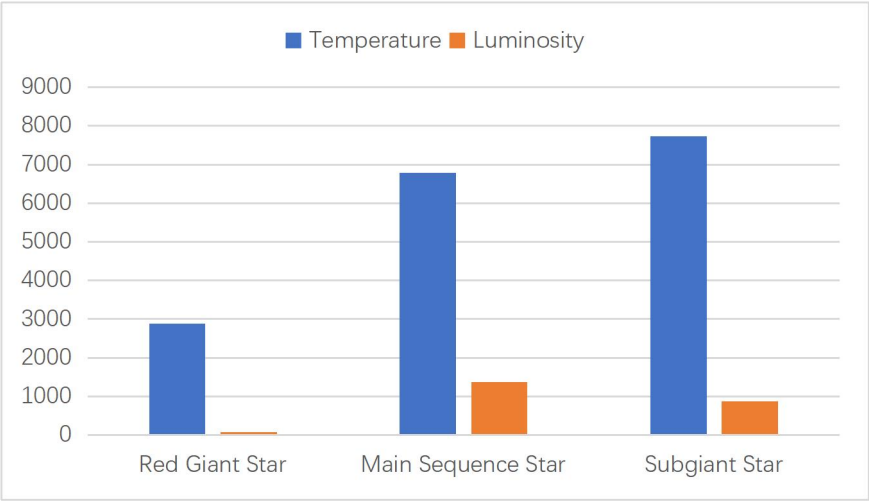

Fig. 1. The surface temperature and luminosity of three types of stars.

3.1 Random Forest Model

Random forest is an ensemble learning method that includes multiple decision trees and combines their prediction results through voting or averaging. This improves the performance and robustness of the model.

Random Forests grows many classification trees. Put the input vector down each tree in the forest to classify a new item using the input vector. Each tree makes a classification, which we refer to as the class the tree "votes" for [2].

Example formula for a random forest: Suppose a random forest model is constructed using astronomical data, where each decision tree classifies stars. When predicting a new star, each tree can make independent predictions and then vote to determine the final classification result. i.

For example, if 3 out of 5 decision trees consider the star to be a red giant, 1 to be a main sequence star, and 1 to be a secondary giant, then the random forest will select the red giant as the final classification result.

3.2 Neural Network Model

Neural network models contain multiple layers, including the input layer, hidden layer, and output layer. Each layer contains multiple neurons, and the connections between neurons have weights. The model adjusts these weights through learning to make predictions.

i.Example formula for neural networks: Consider a neural network model for classifying stellar types in astronomy. Assuming there are two characteristics: the luminosity (L) and temperature (T) of the star.

3.3 Linear Regression Model

Linear regression model: Linear regression is a common method of regression analysis, which assumes that the relationship between output and input is linear. A linear regression model can be expressed in the following form:

Linear regression models are frequently used to examine the relationship between a continuous result and independent variables [1,2]. Researchers commonly perform arbitrary result alteration to satisfy "the" normalcy assumption[3].

3.4 Deep Reinforcement Learning

Deep Reinforcement Learning (DRL) is a computational methodology that utilizes deep neural networks to acquire complex strategies aimed at maximizing cumulative rewards. The utilization of the Deep Q Network (DQN) model, which is a well-known instance of DRL, is applied for the purpose of addressing Markov decision processes (MDP) by acquiring knowledge about the most optimal approach to maximize the accumulation of rewards. The approach employed by DQN involves the utilization of neural networks to estimate action-value functions, which are commonly referred to as Q functions.

3.5 Reinforcement learning

The discipline of Reinforcement Learning (RL) centers on the investigation of how an autonomous agent may proficiently interact with its immediate environment to gain information and formulate a strategy that maximizes the accumulation of expected rewards during a certain task. In recent years, there has been a notable increase in the level of attention and interest directed towards RL. This heightened focus can be attributed to the remarkable achievements demonstrated by RL in diverse domains, including the control of continuous systems in robotics (Lillicrap et al., 2015a), mastery of the game Go (Silver et al., 2016), proficiency in Atari games (Mnih et al., 2013), and success in competitive video games (Vinyals et al., 2017; Silva and Chaimowicz, 2017). To maintain the rapid progress of research in the domain of RL, it is crucial to establish the ability to easily reproduce and compare previous experiments. This will provide an accurate evaluation of the improvements offered by novel approaches[4].

3.6 SVM Classification Model

The Support Vector Machine (SVM) algorithm is designed to identify an optimal hyperplane that effectively separates data samples belonging to distinct categories. In the context of binary classification issues, the objective of SVM is to identify a hyperplane that optimizes the margin, which refers to the maximum distance between the data points of the two categories that are the most distant from the hyperplane.

The SVM is a supervised learning approach that has several advantages. These include its ability to effectively generalize to new cases by adhering to the principle of structural risk reduction. Additionally, SVM provides a single optimum solution and relies on a limited number of data points for representation. Nevertheless, SVM do not provide a means to ascertain the significance of variables, nor are they specifically designed to address the issue of class imbalance[5].

4 Experiment and application

In this section, the steps to detecting the livability rate of exoplanets with machine learning are as follows:

4.1 Data collection

The initial stage of this study necessitates the acquisition of astronomical observation data pertaining to exoplanets. The data encompassed in this study comprises light variation curve data obtained from telescopic observations of planetary transits, as well as fundamental planetary characteristics, including radius, mass, and distance from the host star. Additionally, pertinent information on the stars themselves, such as their spectral properties, brightness, and stellar classifications, is also considered.

Light variation curve data:The observation of quasars reveals that there are substantial physical attributes associated with the temporal fluctuations in light intensity across several wavebands. The analysis of extended variations in light shown by celestial entities is a significant avenue for the exploration of the orbital and rotational properties of quasars[6].

In machine learning research for predicting the livability index of exoplanets, the raw data collection mainly includes the following types of astronomical data. Light variation curve data is an important type of data, that records the brightness changes of planets before passing through their parent stars by observing planetary transit events or other brightness changes through telescopes, usually in the form of time series. Planetary feature data is also indispensable, including the basic characteristics of planets, such as radius, mass, orbital parameters, and orbital period, which are usually obtained from planetary exploration missions or other telescope observations. Star characteristic data is also widely considered, as the properties of the stars around which planets orbit are crucial for habitability evaluation, including spectral information, brightness, mass, radius, and luminosity of the stars.

In addition, atmospheric data also occupies an important position, as the atmospheric components of planets, such as oxygen, nitrogen, carbon dioxide, etc., directly affect their livability. These atmospheric composition data are usually obtained through spectrometers or other spectral measurement instruments. Celestial image data provides information about the landforms and relative positions of planets and parent stars, which is also crucial for livability assessment. In addition, spectral data, including those emitted from planets and stars, can be analyzed to obtain more information about their chemical composition and physical properties. Finally, Earth observation data, such as radio telescopes, infrared telescopes, and X-ray telescopes, also provide valuable raw data for research.

These raw data come from different observation and exploration missions, and through their integration and analysis, the livability index of exoplanets can be more comprehensively evaluated, providing valuable data for a better understanding of other potential life in the universe. The comprehensive analysis and mining of these data are crucial for revealing key features of livability and are expected to provide important insights for cosmic exploration.

4.2 Pretreatment of data

Cleansing and pretreatment of data are necessary. These include treatment of the missing values, outliers, data standardization, and feature engineering, to extract features related to livability.

Data collection:

First, original data is needed. This involves collecting data from varied sources, including databases, files, APIs, and sensors.

In the field of astronomy, data collection methods include various observation methods and instruments to obtain valuable information related to astronomical exploration. The following are some common methods of astronomical data collection:

Telescope observation: The Telescope is one of the most basic observation tools in astronomy. Through different types of telescopes, astronomers can observe and record the position, brightness, spectrum, and other properties of stars. Optical telescopes are used for visible light observation, while specialized telescopes such as radio telescopes, infrared telescopes, and X-ray telescopes can be used for observation in different wavelengths.

Satellites and space probes: International space agencies and research institutions have launched numerous satellites and space probes, which can perform observation tasks in space and stay away from atmospheric interference. For example, the Kepler Space Telescope is specifically designed to detect exoplanets, and its space position allows it to capture stable starlight.

Radio Telescope: Radio telescopes are used to capture radio waves in the universe, which is crucial for studying celestial bodies such as galaxies, Milky Way, pulsars, and quasars. They can detect radio sources and various radiation phenomena in the universe.

Infrared telescopes: Infrared telescopes can be used to observe thermal radiation, making them very useful for detecting objects such as stars, nebulae, planetary atmospheres, and cosmic microwave background radiation. Infrared observations have revealed many hidden information in the universe.

X-ray telescope: X-ray telescopes are used to observe high-energy X-ray and gamma-ray radiation, which is very important for studying active galactic nuclei, black holes, and neutron stars. X-ray observations reveal extreme energy events in the universe.

Ground-based observation stations: Some observations can be made at ground-based observation stations on Earth, which are usually located in distant places, far away from urban light pollution, in order to obtain clearer sky observation conditions. These stations are used for tasks such as visible light and radio observation.

Internet collaboration: Astronomers often collect data through Internet collaboration and data sharing. The participation of crowdsourcing and astronomy enthusiasts can help process a large amount of observational data and promote the progress of astronomical research.

These different data collection methods and instruments provide multi-dimensional and multi-band data for astronomical exploration, in order to gain a deeper understanding of the mysteries of the universe. The comprehensive analysis and research of these data provide an important data foundation for the habitability of exoplanets in this study.

Data cleansing:

Data cleansing is a commonly employed technique aimed at enhancing the precision of machine learning models. However, it necessitates a substantial understanding of the domain in order to discover the significant examples that impact these models[7].

i.Missing value processing: Fill in missing values or delete data points containing missing values.

ii. Detection and processing of outliers: The identification and processing of outliers in data can be accomplished using statistical approaches or graphical techniques.

iii. Processing of Duplicate Values: The elimination of duplicate records within the dataset is performed in order to guarantee the uniqueness of the data.

Data exploration:

Before processing, it is helpful to conduct a preliminary exploration of the data to understand its distribution, correlation, and potential patterns. Visualization and statistical analysis are useful tools. Feature selection:

Feature selection is the process of selecting features that are useful for a task. Some features may not contribute to the model or introduce noise. Feature selection can be achieved through feature importance analysis or domain knowledge.

Feature engineering:

Feature engineering involves creating new features or transforming existing features to better represent problems. This may include the following actions:

1. Feature scaling: Scaling features to similar scale ranges, such as using standardization or normalization.

2. Feature encoding: Encoding classified features, such as single hot encoding or label encoding.

3. Feature extraction: Extracting new features from raw data such as keywords in text or texture features in images.

4. Time series processing: Smoothing, lagging, or moving average processing or time series data.

Data segmentations: Divide the data into training sets, validation sets, and testing sets. The training set is used to train the model, the validation set is used to tune model parameters, and the testing set is used to evaluate model performance.

Data Standardization: In machine learning algorithms, feature standardization is necessary to ensure that the values of different features are within similar ranges. Standardization usually involves normalizing the mean and standard deviation of the feature to zero.

Handling imbalanced data: If the categories in the dataset are imbalanced, sampling techniques(such as oversampling or under-sampling) can be used to balance the data to avoid the model leaning towards the majority of categories.

Processing text and image data: For text data, it may be necessary to perform word segmentation, stop word removal, stem extraction, and vectorization (such as TF-IDF or word embedding). For image data, scaling, cropping, enhancement, and feature extraction may be necessary.

Data conversion: In some cases, data requires non-linear transformation, such as logarithmic transformation, exponential transformation, or polynomial feature transformation.

Feature reduction: In the context of high-dimensional data, it is common to employ dimensionality reduction methods, such as principal component analysis (PCA) or linear discriminant analysis (LDA), to effectively decrease the number of features while preserving important information.

Verification preprocessing: Before and after applying preprocessing, cross-validation, and other techniques are used to verify the effectiveness of preprocessing to ensure good performance of the model on test data.

Automated preprocessing pipeline: For repetitive tasks, an automated data preprocessing pipeline can be created to ensure that data is always processed consistently before being input into the model.

In the data preprocessing process of evaluating the livability of exoplanets, a series of key steps were taken to prepare the raw data for use by machine learning models. Firstly, extensive data collection was conducted, covering various types of astronomical data, such as light variation curves, planetary and stellar characteristics, atmospheric composition, celestial body images, spectra, and Earth observation data. These data come from various observation and detection tasks, ensuring comprehensiveness and diversity.

Next, perform data cleaning to handle missing values, outliers, and duplicate values to ensure the quality and stability of the data. Data cleaning helps to avoid the negative impact of noise on model performance. Subsequently, data exploration was conducted to gain a deeper understanding of the distribution, correlation, and potential patterns of the data through visualization and statistical analysis, providing useful information for subsequent modeling.

In terms of feature selection and engineering, features related to livability assessment were selected transformed, and created, including temperature range, atmospheric composition, distance from stars, liquid water, Earth similarity, radiation environment, geological features, stellar types, atmospheric pressure, Earth's magnetic field, and signs of life. These characteristics are crucial for evaluating livability.

For text and image data, natural language processing techniques and image processing methods were adopted, such as word segmentation, stop word removal, stem extraction, scaling, cropping, enhancement, and feature extraction. These processes help extract useful information from unstructured data.

In addition, we also processed imbalanced data and used sampling techniques to balance the distribution of different categories. For high-dimensional data, dimensionality reduction methods were used, using PCA or LDA to reduce data dimensionality and retain key information.

In summary, data preprocessing is a key step in ensuring the accuracy and reliability of livability assessment. Integrating and cleaning multi-source astronomical

data, as well as extracting key features, provides a strong data foundation for machine learning models to predict the livability index of exoplanets. This process helps to optimize model performance and improve the accuracy of livability assessment.

4.3 Experiment Standard

Through data cleaning and feature engineering processing, this study generated a series of key livability features and factors that are crucial for evaluating the livability of a planet:

1. Temperature range: The temperature range of a planet is a crucial factor. A livable planet should have a temperature range that allows liquid water to exist, which is one of the necessary conditions for the existence of life.

2. Atmospheric composition: The composition of a planet's atmosphere is crucial for livability. The presence and content of gases such as oxygen, carbon dioxide, and nitrogen can all affect the existence of life. Appropriate atmospheric composition is crucial for supporting life activities.

3. Distance from a star: The distance of a planet from its parent star affects its temperature and radiation level. A planet at a suitable distance can maintain a suitable temperature, making it more likely to support the existence of life.

4. Liquid water: Liquid water is the key to the existence of life, so the presence of liquid water on the surface of planets is a strong indicator of livability. Liquid water on planets can provide essential solvents and environments for life.

5. Earth similarity: The Earth similarity of planets includes factors such as their size, mass, and rotation period. Planets similar to Earth are more likely to support life because they provide an environment similar to Earth.

6. Radiation environment: The radiation environment of planets, including radiation from cosmic rays and space, has a significant impact on the existence of life. A suitable radiation environment helps maintain the stable existence of life.

7. Geological features: Geological features such as mountains, oceans, volcanoes, etc. can also affect livability. Some geological features may provide the necessary resources and diverse living environments for the existence of life.

8. Star type: The type of star around which a planet orbits can also affect its habitability. Different types of stars generate different levels of radiation and luminosity, which have an impact on the habitability of planets.

9. Atmospheric pressure: Atmospheric pressure is crucial for maintaining the presence of liquid water and gases. Appropriate atmospheric pressure helps maintain a stable climate and ecosystems.

10. Earth's magnetic field: The Earth's magnetic field can protect planets from cosmic rays and solar wind, which is crucial for maintaining the long-term existence of life.

11. Life signs: Detected signs of life, such as biomarkers or chemical abnormalities, are strong indicators of livability. These signs may include biological activity, gas emissions, etc.

Feature engineering is a key task that involves selecting and extracting features related to livability. The comparison of different feature sets and their correlation analysis with the livability index will help determine which features are most crucial

for evaluating livability. The comprehensive analysis of these characteristics and factors is crucial for studying the livability of exoplanets.

4.4 Label habitability

A label that defines habitability is required. This may involve developing a set of standards or indicators to evaluate the habitability of planets, such as the moderate range from the parent star, the atmospheric composition of the planet, and the temperature range.

4.5 Data splitting

Data splitting is the act of dividing data into numerous categories, often two or more, for various purposes or studies. The usual methodology of a two-part split involves allocating a section of the dataset for the purpose of assessing or testing the data while designating an additional element for training the model.

Leave One Last: This data-splitting technique involves extracting the final transaction per user for testing purposes. Typically, the second-to-last transaction per user is utilized for validation, while the remaining transactions are employed for training.

Temporal User/Global Split: The concept of temporal user/global split refers to the division between users who are focused on immediate, short-term concerns and those who have a broader, long-term perspective. The temporal split technique is a frequently employed assessment approach that involves dividing previous interactions or baskets depending on the timestamps of the interactions. For instance, the latest 20% of interactions are allocated for testing purposes.

Random Split: As the term implies, involves the random selection of the training and test border for each user. The evaluation of early recommender systems involved the utilization of a leave-one form of the scheme. In this variant, a single random item per user is chosen for testing purposes. Nevertheless, there has been a steady shift away from this particular approach in favor of utilizing the most recent interaction, commonly referred to as the "Leave One Last Item" method, for each individual user. A drawback of employing random splitting algorithms is their lack of reproducibility unless the author(s) disclose the specific data splits utilized.

User Split: The user-based splitting strategy is an alternative assessment methodology that diverges from the more often employed method of splitting the dataset based on interactions, instead opting to partition the data based on individual users. In this scenario, a certain group of users and their corresponding transactions are designated for training purposes, while a distinct set of users and their transactions are utilized for testing. This concept is not often employed in literature due to the requirement of underlying models to provide recommendations for new users (commonly referred to as cold-start users), a feature that is not supported by many existing systems. It is worth mentioning that several studies, such as VAECF(Variational Auto-Encoders for Collaborative Filtering)and SAVE, employ this approach by dividing the interaction history of training users into fold-in and fold-out sets. This allows for the inclusion of users with incomplete histories throughout the training process. The aforementioned works exhibit a common

problem of incorporating "future data" into the model during the training process, similar to the temporal user-based approach [8].

Training Set: The training set is the foundation of model training. It contains data samples for training machine learning models. Usually, the training set accounts for the majority of the entire dataset, typically around 70% -80% of the data. The purpose of the training set is to enable the model to learn the patterns, relationships, and features of the data in order to make predictions.

Validation Set: The validation set is used to adjust and optimize the hyperparameters of the model, such as learning rate, regularization parameters, etc. Typically, the validation set occupies a small portion of the entire dataset, typically around 10% -15% of the data. By verifying the set, the performance and generalization ability of the model can be estimated for hyperparameter tuning, thereby improving the performance of the model.

Test Set: A test set is a dataset used to evaluate the performance of the final model. This is the representation of the model in real scenarios. The test set should be independent of the training and validation sets to ensure the objectivity of the test results. The test set occupies the remaining portion of the entire dataset, usually about 10% -20% of the data.

Cross Validation: Cross-validation is a data segmentation technique that is particularly suitable for small datasets. It divides data into multiple folds. In cross-validation, the dataset is divided into k subsets, and the model is trained k times. Each time, the k-1 subset is used as the training set, and the remaining subset is used as the validation set. This can help to more fully evaluate the performance of the model and reduce differences caused by different data segmentation.

Imbalanced Dataset: In some cases, the number of samples in different categories of the dataset may be extremely uneven, for example, in binary classification, the number of samples in one category is much greater than that in another category. In this case, special methods need to be taken to maintain balance, such as oversampling, undersampling, or weight adjustment.

4.6 Model evaluation

Utilize a designated test set to assess and gauge the efficacy and accuracy of the model's performance. Typical assessment metrics encompass accuracy, memory rate, accuracy, and F1 score, among others. Cross-validation can be employed as a means to enhance the precision of assessing the model's generalization performance.

In 1994, a seminal work on model evaluation was published in Science magazine by Naomi Oreskes, Kristin Shrader-Frechette, and Kenneth Belitz. In an effort to address the issue of excessive trust in computer simulation models, the individual voiced apprehension with some often employed words, specifically focusing on the terminology of "verification" and "validation". The authors emphasize that models lack the ability to be definitively confirmed or validated, meaning that the truthfulness of their underlying assumptions cannot be proved with absolute confidence. Furthermore, they caution that the great track record of a model in the past does not ensure its future success. According to the authors, scientific models can be considered "confirmed" when their output aligns with observable evidence. However,

they argue that the support supplied by such confirmation is necessarily limited in scope. The study authored by Oreskes et al. has garnered significant influence, as seen by its citation count of over 1800 in various academic disciplines[9].

In the study of predicting the livability index of exoplanets, various complex and effective machine learning methods were used to fully address this important challenge. The following is a detailed introduction to the main machine learning methods used in livability assessment:

1. Decision Tree and Random Forest: Using tree-based classification methods such as decision tree and random forest. These models can make decisions based on planetary features, and random forests are an integrated method of multiple decision trees that improve accuracy through voting. Train these models using Kepler data and other relevant data to evaluate livability conditions. The evaluation indicators include accuracy, recall, and F1 score to determine the performance of the model.

2. SVM: SVM is a supervised learning method used for classification and regression tasks. It divides the data into two categories by finding the optimal decision boundary. In this study, the SVM model was used for classification based on planetary feature data, such as distance from stars and atmospheric composition. The indicators for evaluating model performance include classification accuracy and confusion matrix.

3. Deep Neural Network (DNN): DNNs are a powerful deep learning method suitable for processing large-scale data and complex features. Using convolutional neural networks (CNN) to analyze photometric curve data and image data, extracting information about atmospheric composition and temperature distribution. The indicators for evaluating model performance include image classification accuracy and image generation quality evaluation.

4. Integrated learning: Integrated learning methods are also adopted to combine the prediction results of multiple different models to improve the accuracy of the livability index. By integrating the outputs of different machine learning models, the prediction performance of the livability index can be significantly improved. The experimental results include the performance indicators of the integrated model and the correlation analysis between the models.

The comprehensive application of these machine learning methods enables us to predict the habitability index more comprehensively and accurately, thereby better understanding the habitability conditions of exoplanets. By evaluating the performance of different models, reliable data support is provided for livability evaluation, which helps to solve complex astronomical problems. This study has significant implications for future exploration of the universe and the search for potential life spaces.

4.7 Hyperparameter tuning

In the process of hyperparameter tuning for machine learning models, the objective is to identify the optimal hyperparameters x from a given set A ⊆ R d in order to minimize the validation error f(x). This may be formulated as the following optimization problem[10]:

$$min_{x \in A} f(x) \tag{1}$$

Hyperparameter adjustment is an important link in machine learning and deep learning, which involves selecting and optimizing key settings of the model. These settings are not automatically learned by the model but are manually defined. The selection of hyperparameters is crucial for the performance and generalization ability of the model, and different hyperparameter settings may lead to overfitting, underfitting, or poor performance of the model. Therefore, hyperparameter adjustment is to find the optimal hyperparameter settings to maximize the performance of the model.

Hyperparameter adjustment usually adopts different search strategies, including manual search, grid search, random search, and Bayesian optimization. The goal of these strategies is to find the optimal settings from the hyperparameter space to improve the performance of the model. Usually, the search scope gradually narrows down in order to find the best hyperparameter faster.

In order to evaluate the performance of different hyperparameter settings, various evaluation indicators such as accuracy, loss function, F1 score, etc. are often used, and cross-validation can provide more stable performance estimates.

Hyperparameter adjustments typically require multiple repetitions, as the initial results may not always be optimal. There are many automated tools that can help more effectively find the optimal hyperparameter settings, such as GridSearchCV, RandomizedSearchCV, and Optuna.

In summary, hyperparameter adjustment is a key step in optimizing model performance, which requires careful planning, experimentation, and evaluation to ensure that the model can adapt to specific tasks and datasets, and improve its performance and generalization ability.

4.8 Deployment and Application

Once a high-performance model is established, it can be deployed to practical tasks, such as automated identification of habitable exoplanets. This can be used for target selection and space exploration decision-making in space exploration missions.

The deployment of machine learning applications in this study is a comprehensive process aimed at applying the developed livability prediction model to practical tasks, especially for evaluating the livability of exoplanets. Firstly, we selected an appropriate machine learning model and conducted sufficient training on it, using astronomical data from multiple data sources, including observations of exoplanets and their features. The goal of this model is to learn the complex associations and patterns of livability.

In the data preparation stage, we conducted data cleaning, preprocessing, and feature engineering to ensure that the model can process high-quality input data. This includes operations such as handling missing data, outliers, standardization, feature selection, and extraction.

The performance evaluation of the model is a key step, and we use different evaluation indicators such as accuracy, recall, F1 score, and cross-validation to evaluate the generalization performance of the model. During this process, hyperparameter adjustments may also be made to find the optimal hyperparameter combination.

Once the model has undergone sufficient training and evaluation, we can deploy it into practical applications. This can include embedding the model into an online platform, website, or application so that users can easily use it. In addition, we can also connect the model to a real-time updated dataset to obtain information about newly detected exoplanets and conduct real-time evaluations of them.

After the model is deployed, we also need to monitor and verify its performance to ensure the stability and accuracy of the model. The development of user interfaces allows users to easily interact with models, especially researchers, astronomers, and space explorers, which is a useful tool.

Ultimately, this model can be applied to space exploration missions to help determine which exoplanets are most likely to support life. This comprehensive machine learning application deployment process involves collaboration across multiple disciplines, including astronomy, machine learning, and engineering, to ensure that the model performs well and has operability in practical tasks.

4.9 Continuous improvement

This is a process of continuous improvement. With the accumulation of new data and further improvement of models, prediction performance can be continuously improved.

In the future, this research will have exciting prospects and broad possibilities. Firstly, we can expect more abundant astronomical data, including observation results from various advanced telescopes and observation equipment. This will enrich the existing dataset and delve deeper into the characteristics of exoplanets, thereby improving the accuracy of habitability prediction.

Secondly, machine learning models and algorithms will continue to develop and improve. More complex models can continue to evolve, including deep learning and reinforcement learning methods, to better capture various complex relationships in the universe. This will make the assessment of livability more accurate and reliable.

In addition, real-time evaluation will become a key trend. In the future, these models can be connected to real-time updated datasets for real-time evaluation when new exoplanets are discovered. This will help astronomers obtain more timely information about new discoveries in the universe.

5 Conclusion

The main objective of this study is to develop and improve a preliminary artificial intelligence system aimed at conducting a preliminary evaluation of exoplanets to determine whether they have the potential to replace Earth as an ideal habitat. This research field is full of challenges, as assessing livability involves numerous complex factors, including temperature, atmospheric composition, distance from the parent star, and so on. However, by applying machine learning and data analysis methods, current artificial intelligence systems provide a new way to solve this problem.

In order to further improve the performance of current artificial intelligence systems, the unique situation of exoplanets can be utilized. This may include considering factors such as the atmospheric composition of the planet, geological

characteristics, and the presence of liquid water. By integrating more data and knowledge, livability can be more accurately evaluated, which is crucial for finding signs of life in the universe.

In the future, we will deepen our current artificial intelligence systems and plan to apply the research results of livability prediction to practical applications, in order to develop an online platform or website that allows users to easily apply this program. This will provide researchers, astronomers, and space explorers with a useful tool to help them determine which exoplanets are most likely to support life.

In addition, plans will be made to extend the program to allow it to connect to richer real-time updated datasets. In order to timely obtain information about newly detected exoplanets and conduct real-time ratings on them. Improved functionality will help advance predictions of tracking livability to reflect new observations and discoveries.

At the same time, in order to benefit these achievements more widely, websites and applications can be developed, making it easy for more people to use these livability assessment tools. This will promote public participation and understanding of space exploration, and promote the dissemination and sharing of scientific knowledge.

Finally, interdisciplinary cooperation will continue to drive the development of this field. Astronomers, computer scientists, and data scientists will collaborate to deepen their understanding of exoplanets and advance the forefront of cosmic exploration.

Overall, the future will be filled with opportunities, and we will continue to improve and expand our research on the habitability of exoplanets, which will provide more opportunities and insights for finding signs of life and broadening our understanding of the universe. This field will continue to be full of challenges, but it will also be full of passion and hope. This study will become an important milestone in exploring whether there are other habitable planets in the universe. By continuously improving artificial intelligence technology and accumulating more data, it is possible to gain a deeper understanding of the universe and potentially find new celestial bodies suitable for the existence of life. This field is full of hope and will reveal more unknown mysteries in the universe in the future.

References

1. Huang Wen．"Classic Algorithms for Decision Trees, ID3 and C4″．5[J]．Journal of Sichuan University of Arts and Sciences (Natural Science Edition))(2007)17(5)：16-18
2. Chen, Chao, and Leo Breiman. "Using Random Forest to Learn Imbalanced Data", more
3. Schmidt, Amand F, and Chris Finan. "Linear regression and the normality assumption.", Journal of Clinical Epidemiology 98. (2018): 146-151.
4. Henderson, Peter et al. "Deep Reinforcement Learning that Matters", AAAI Conference on Artificial Intelligence 32.1 (2019): 3207-3214.
5. Maldonado, Sebastián, and Julio López. "Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification.", Applied Soft Computing 67. C (2018): 94.0-105.

6. Lin, LU et al. "Analysis of Optical Long-period Light Variation and Study of Color Index Variation in FSRQ 0208-512", Chinese Astronomy and Astrophysics (2021)
7. Satoshi, Hara, et al. "Data Cleansing for Models Trained with SGD.", Advances in neural information processing systems 32. (2019): 4215-4224.
8. Meng, Zaiqiao, et al. "Exploring Data Splitting Strategies for the Evaluation of Recommendation Models", Conference on Recommender Systems abs/2007.13237 (2020): 681-686.
9. Alexandrova, Anna. 2010. “Adequacy-for-purpose: The best deal a model can get.” Modern Schoolman: A Quarterly Journal of Philosophy 87(3–4):295-301.
10. Wu, Jian et al. "Practical Multi-fidelity Bayesian Optimization for Hyperparameter Tuning.", Conference on Uncertainty in Artificial Intelligence abs/1903.04703. (2020): 284.