Predictive Modeling of Stellar Properties Using Machine Learning

Abishek Xavier A[1], Deva Darshan A S[2], Ragavendran A[3]

[1]UG Student, Department of AI & DS, [2]UG Student, Department of CSE, [3]UG Student, Department of AI & DS,

Saveetha Engineering College, Chennai, India

[1]abishek150604@gmail.com, [2]devadharshanmannai@gmail.com, [3]ragavanayyadurai@gmail.com

***Abstract*— This paper focuses on the predictive modeling of stellar properties using modern machine learning techniques. Stellar properties such as mass, radius, temperature, and luminosity are crucial for understanding the evolution and lifecycle of stars. With the advancement of astronomical surveys, such as Gaia, extensive datasets of stellar parameters are available for analysis. In this work, we explore the application of regression models, including linear regression and neural networks, decision tree and random forest are the algorithms to predict key stellar parameters based on observational data. The dataset is preprocessed, and feature engineering is performed using astrophysical equations to enhance prediction accuracy. Our results demonstrate that machine learning models can provide accurate predictions of stellar properties, contributing significantly to the growing field of Astro informatics.**

***Index Terms*—Machine Learning, Stellar Properties, Gaia Dataset, Neural Networks, Random Forest, Astroinformatics.**

I. INTRODUCTION

Stellar properties encompass a wide array of fundamental characteristics that define the physical and chemical state of stars. These properties include luminosity, radius, mass, temperature, and metallicity, which are essential for understanding stellar evolution, the lifecycle of stars, and their formation processes. By studying these properties, astronomers can infer the physical conditions present in stellar interiors, their composition, and the processes driving their formation and evolution. Consequently, the accurate determination of stellar properties is critical for multiple reasons, including the exploration of galaxy formation, the dynamics of star clusters, and the search for exoplanets. Understanding stellar properties is essential for comprehending the life cycles of stars, from their formation in stellar nurseries to their eventual demise as white dwarfs, neutron stars, or black holes. Each stage of a star's life is marked by distinct changes in its properties, driven by processes such as nuclear fusion, gravitational collapse, and mass loss. For instance, during their main sequence phase, stars like our Sun fuse hydrogen into helium in their cores, producing energy and causing them to emit light. As they exhaust their hydrogen fuel, they undergo significant changes, expanding into red giants before shedding their outer layers or exploding as supernovae, depending on their initial mass. These evolutionary stages profoundly influence the chemical enrichment of the interstellar medium, impacting the formation of new stars and planetary systems.

Traditionally, astrophysicists have relied on well-established theoretical models and equations to derive these properties from observable quantities, such as brightness and color. For example, the effective temperature of a star can be estimated using the Stefan-Boltzmann Law, which relates a star's luminosity to its radius and temperature. However, the complexity of stellar behavior, combined with the vast variability in star types, poses significant challenges in deriving accurate predictions using conventional methods. Many stars exhibit unique characteristics due to their initial mass, composition, and evolutionary history, which can lead to discrepancies when relying solely on traditional models.

The Gaia mission, launched in 2013 by the European Space Agency, has revolutionized the field of astronomy by providing unprecedented data on stellar positions, distances, and physical characteristics. Gaia's primary objective is to create a three-dimensional map of the Milky Way galaxy by measuring the positions and motions of over a billion stars with unprecedented precision. The mission has generated a wealth of data, including measurements of parallax, proper motion, and photometry, enabling astronomers to gather extensive datasets that are ripe for analysis. By offering high-accuracy astrometric data, Gaia has opened new avenues for understanding the structure and evolution of our galaxy and its stellar populations.

Recent advancements in machine learning, particularly in deep learning and neural networks, have further enhanced the potential for accurate predictions in this domain. Neural networks, with their capacity to capture complex, non-linear relationships in data, have demonstrated remarkable success across various fields, including image recognition, natural language processing, and medical diagnostics. Their ability to process vast amounts of data and identify intricate patterns makes them particularly promising for the analysis of astronomical datasets. This has prompted astrophysicists to explore the application of similar techniques to the study of stellar properties, with the goal of improving predictive accuracy and uncovering new astrophysical insights. The primary

objective of this study is to develop and evaluate various machine learning models for predicting stellar properties using the extensive datasets provided by the Gaia mission and other astronomical surveys. We introduce a predictive modelling framework that integrates astrophysical insights with machine learning techniques. By leveraging the strengths of different algorithms, including linear regression, neural networks, decision tree and random forest, we aim to create a robust predictive model that not only predicts stellar properties with high accuracy but also contributes to the understanding of the physical processes governing stellar evolution.

In addition to improving predictive capabilities, this research seeks to bridge the gap between machine learning and traditional astrophysical models. By correlating machine learning predictions with established astrophysical relationships, we can identify areas of agreement and divergence, potentially leading to new insights and avenues for research. The ability to accurately predict stellar properties has broad implications for various areas of astrophysics, including the study of stellar populations, the evolution of galaxies, and the search for exoplanets around stars. For instance, understanding the distribution of stellar masses and luminosities within a galaxy can provide valuable insights into its formation history and subsequent evolution. Moreover, by predicting the properties of stars in different regions of a galaxy, we can identify potential candidates for hosting planetary systems, thereby enhancing our understanding of the conditions necessary for life beyond Earth. In this paper, we will present our methodology, model implementations, results, and conclusions regarding the applicability of machine learning techniques in predicting stellar properties. Our findings will underscore the importance of integrating machine learning with astrophysical research, providing a pathway for future studies aimed at unravelling the complexities of the universe. As we advance our understanding of stellar properties through machine learning, we contribute not only to the field of astro informatics but also to the broader quest for knowledge about the cosmos and our place within it.

II. RELATED WORKS

Machine learning applications in astronomy have seen significant advancements, with numerous studies employing algorithms such as neural networks and decision trees to analyze astronomical datasets. These methodologies align closely with the approaches utilized in the paper for predicting stellar properties. In the context of stellar evolution studies, researchers focus on understanding the life cycles of stars and analyzing how properties like mass, luminosity, and temperature evolve over time. Such studies provide a vital foundation for the predictive models developed in the paper. Additionally, data from the Gaia mission, which offers extensive datasets for deriving stellar parameters and investigating stellar populations, has been integral to supporting the paper's emphasis on leveraging large datasets for predictive modeling. Comparative analyses of machine learning models, including random forests, logistic regression, and support vector machines, have further explored their relative effectiveness in predicting astronomical properties. This body of work contextualizes the paper's findings, particularly regarding the efficacy of the random forest algorithm in predicting stellar properties. Finally, research in astroinformatics bridges data science and astrophysics, integrating traditional astrophysical models with machine learning techniques. This interdisciplinary approach aligns with the paper's goal of combining these domains to enhance predictions of stellar properties.

III. PROPOSED METHODOLOGY

3.1. Overview of Dataset

The dataset for this study is primarily sourced from the Gaia Data Release (DR) and other publicly available astronomical catalogues, including the Sloan Digital Sky Survey (SDSS) and the Two Micron All Sky Survey (2MASS). The Gaia DR provides high-precision measurements of stellar positions, distances, and magnitudes, resulting in a dataset comprising over 1.5 billion stars. Parallax: A measure of a star's apparent motion against the background of distant stars, used to determine its distance from Earth. Apparent Magnitude: The brightness of a star as observed from Earth, which is influenced by both its intrinsic brightness and distance. Colour Indices: Measurements that indicate a star's colour and temperature, derived from differences in magnitude in various wavelength bands. Effective Temperature: The temperature of a star's surface, which is crucial for determining its luminosity and other physical properties. Before training the models, several preprocessing steps are applied, Data Cleaning: Removal of erroneous and outlier data points to improve model performance. This includes filtering out stars with incomplete or inconsistent data. Normalization: Scaling the data to ensure that all features contribute equally to the model, which is particularly important for algorithms sensitive to the scale of input features. Feature Engineering: Creating new features based on astrophysical relationships to enhance the predictive capability of the models. For instance, new variables are derived from existing ones using established astrophysical equations.

3.2. Stellar Property Equation

Stefan-Boltzmann Law for Luminosity: This law describes the relationship between a star's luminosity, its radius, and its effective temperature. It is a fundamental principle in astrophysics, demonstrating how the total energy radiated by a star depends on its physical properties. Mass-Luminosity Relation For main-sequence stars, there is a well-known relationship between the mass of a star and its luminosity. This relationship indicates that more massive stars are generally more luminous, and this correlation is crucial for understanding stellar evolution.

Hertzsprung-Russell Diagram: The Hertzsprung-Russell diagram (H-R diagram) is a plot of stars' luminosity versus their temperature, which reveals the relationship between different types of stars and their evolutionary stages. It provides a visual

representation of how stars evolve over time. Mass-Luminosity Relation For main-sequence stars, the mass-luminosity relationship is expressed as:

$$\mathcal{L} \propto M^{3.5}$$

Where: L is the luminosity, M is the mass of the star.

3.3. Machine Learning Models

In this study, we implement several machine learning models to predict stellar properties:

3.3.1. Linear Regression:

The basic linear regression model assumes a linear relationship between the independent variables (features) and the dependent variable (target). The model can be represented as:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdot \quad \cdot \quad \cdot + \beta_n x_n$$

Where: ŷ is the predicted value, β0 is the intercept, and βnare the coefficients of the independent variables xn.

3.3.2. Neural Networks:

Neural networks are much more flexible than linear models, as they can capture complex, non-linear relationships between the features and the target variables. A feed-forward neural network consists of multiple layers of interconnected neurons, where each neuron applies a weighted sum of the inputs and passes the result through a non-linear activation function. The architecture usually includes the following: Input layer: These correspond to features in the dataset, such as parallax, apparent magnitude, color index. Hidden layers: Layers formed of several neurons which deal with the input data using nonlinear ways to capture the complex relationships between the variables. Output layer: Outputs the final prediction as estimated stellar mass or luminosity. The neural network model trains through adjusting the weights between neurons on the least possible difference of the predicted values that come from the network, ŷ, and the actual values taken, y. This is known as training and follows the [1] Forward Propagation: With respect to each forward pass, the net's input features are multiplied by the weights and passed through an activation function that generates the network's predictions. [2] Loss Function: MSE is the method most commonly used to check how well the data fits the predictive distribution of the model. [3] Backpropagation and Gradient Descent: For minimizing the loss, the model updates its weights using an algorithm known as the gradient descent process. The algorithm derives the gradient of the loss function with respect to each weight and updates the weights in the opposite direction to minimize the error. The process continues iteratively till the model converges or the loss starts decreasing significantly. The neural networks have drawn a lot of promise toward getting stellar properties with good precision since they can fit very complex, non-linear patterns even in large data sets. The flexibility of neural networks allows them to outperform the simpler models, especially when dealing with large datasets like Gaia's datasets that contain subtle correlations that might not pop out using linear models. Loss Function and Optimization the primary goal of any machine learning model is always to minimize the error between the predicted values and the true values. In this study, we use the [1] Mean Squared Error (MSE) as the loss function. It is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Where:n is the number of observations (stars in the dataset),$\hat{y}_i$ is the predicted value (e.g., predicted stellar mass), and y_i is the actual value (e.g., true stellar mass). Minimizing the MSE ensures that the model makes predictions as close as possible to the actual values. A gradient descent optimization technique is used to update the weights in the model. At each iteration, the weights are adjusted in such a manner that it leads to reducing the error so that the optimal set of model parameters can be achieved.

3.3.3 Decision Tree:

Decision tree is one of the most used algorithms in machine learning that is generally used for classification. As its name implies, a decision tree mimics the fundamental structure of a tree for classification and comes complete with a root node, some internal nodes and leaves. Among them, each leaf node corresponds to a predicted result, and sample data in each node will be allocated to the corresponding child nodes according to the results of attribute tests. The root node contains the complete data set of the sample, and the path from the root node to each leaf node corresponds to the sequence of tests. The purpose of decision tree is to generate a model with strong generalization ability. In this decision tree model training, the information entropy is

$$Ent(D) = -\sum_{k=1}^{|y|} p_k log_2 p_k$$

used to measure sample purity, that is, to select the optimal division. among them, the lower the information entropy value is, the higher the purity of partition improved.

3.3.4. Random Forest:

Random Forest is an extended variant of Bagging. It builds the Bagging based on the decision tree and introduces the shorthand selection of attributes in the training process of the decision tree [10]. Or in other words, the decision tree compares the purity of differentiation according to the information entropy, gain rate and other indicators when it chooses the attributes, so to choose an optimal attribute. In the case of Random Forest, a subset that contains K attributes will be randomly taken from the attribute set of each node of the base decision tree. Then an optimal attribute has to be selected for the division from this subset. In the case of Random Forest, simplicity, ease of implementation, and low computational cost lead to strong performance on realistic tasks in the development and application of this ensemble method.

3.4. Model Evaluation and Performance Metrics

Evaluating the performance of machine learning models is a critical step in any predictive modelling task, particularly when dealing with complex and high-dimensional data like that used for predicting stellar properties. The effectiveness of a model is typically assessed using various statistical metrics that measure how well the model's predictions match the true values. Below are the key performance metrics used in this study, along with explanations of their importance and applications:[2] Root Mean Squared Error (RMSE). RMSE is one of the most commonly used metrics for regression problems, as it provides an easily interpretable measure of the model's prediction error. RMSE is the square root of the average of the squared differences between the predicted values and the actual values. Mathematical Formula

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Where: n is the number of observations (stars in this case),$\hat{y}_i$ is the predicted value for the i-th observation (e.g., predicted stellar mass),y_i is the true value for observation. For instance, if we are predicting the luminosity of stars in units of solar luminosity, the RMSE will also be in those same units. RMSE heavily penalizes large errors due to the squaring of the differences, meaning that models with high RMSE might be making significant mistakes on some predictions even if they perform well overall. In our context, RMSE can give us a sense of how far off our model's predictions are for stellar properties like luminosity, mass, and temperature, which are crucial for understanding stellar evolution. [3] Mean Absolute Error (MAE)MAE is another common metric that measures the average magnitude of errors in predictions without considering their direction (whether the prediction was too high or too low). Mathematical Formla.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y_i} - y_i|$$

Interpretation: Lower MAE values indicate better performance, with smaller average errors between predicted and actual values. Since it doesn't square the errors, MAE treats all errors equally and does not penalize large errors as heavily as RMSE does. Usefulness: MAE provides a straightforward understanding of the average error in the model's predictions. In the context of predicting stellar properties, if the MAE for predicting a star's radius is, for example, 0.1 solar radii, we can infer that, on average, the model's predictions are off by that amount. R-Squared (R^2R-squared (also called the coefficient of determination) is a statistical measure that indicates how much of the variance in the dependent variable (e.g., stellar mass or luminosity) is explained by the independent variables (features like parallax color indices,etc.).Mathematical Formula:

$$R^2 = 1 - \frac{\Sigma(\hat{y}_i - y_i)^2}{\Sigma(y_i - \bar{y})^2}$$

Negative R^2 values suggest the model is worse than simply using the mean as a predictor.R^2 provides insight into the goodness of fit of the model. In the case of predicting stellar properties, a high R^2 value (closer to 1) would suggest that the model is effectively capturing the relationships between the features (e.g., parallax, temperature) and the target variable (e.g., luminosity, mass).[4] Adjusted R-Squared While R^2 is a good indicator of model performance, it has a key limitation: it always increases when more features are added to the model, even if those features do not significantly contribute to the model's predictive power. To address this, we use Adjusted R^2, which adjusts the R^2 value based on the number of features used in the model relative to the number of observations.MathematicalFormula

$$\text{: Adjusted } R^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$$

Where: n is the number of observations (data points),p is the number of features (independent variables).Interpretation: Unlike R^2, Adjusted R^2 will decrease if additional features do not improve the model's performance. It penalizes overfitting, making it a more reliable measure when working with complex datasets that may be prone to overfitting. In our context, Adjusted R^2 is especially useful for ensuring that we are not adding too many irrelevant features into the model when predicting stellar properties. By considering the number of predictors, it gives us a better sense of the model's true performance. [5]. Cross-Validation in machine learning, it's essential to evaluate how well a model generalizes to unseen data. Cross-validation is a robust technique to assess this. In this study, we use k-fold cross-validation, where the dataset is divided into k subsets (folds). The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, with a different fold being used as the test set each time. Steps in k-fold Cross-Validation: Split the data into k equal-sized folds (typically k = 5 or 10). Train the model on k-1 folds and test it on the remaining fold. Repeat the process k times, each time with a different test fold. Average the evaluation metric (e.g., RMSE, R^2) across the k runs to obtain a more reliable estimate of the model's Cross-validation helps mitigate the risk of overfitting, ensuring that the model's performance is not just good on the training data but also generalizes well to new, unseen data. It provides a more accurate assessment of how the model is likely to perform in real-world scenarios. Feature ImportanceWhile not a direct performance metric, feature importance plays a key role in understanding which input features (e.g., parallax, color indices, temperature) are most influential in predicting stellar properties. In machine learning models such as neural networks or decision trees, feature importance is computed by analyzing how much each feature contributes to the model's predictions. Usefulness: Helps in model interpretation by revealing the key drivers behind the model's decisions. Informs feature engineering by highlighting which features can be prioritized or dropped for future model improvements. For astrophysical models, features like stellar temperature and color index are often found to be highly significant for predicting properties like luminosity and mass.

.

IV. SYSTEM ARCHITECTURE

The architecture diagram for Predicting Stellar Properties illustrates the interaction between the various system components and provides a clear, organized view of the entire workflow. Each component and its connections are depicted to help understand how they function together to achieve the primary objective: predicting the properties of stars using machine learning algorithms.

System Workflow, In the Predicting Stellar Properties architecture, the input is stellar data collected from observational databases or surveys. This data includes features such as luminosity, temperature, radius, and other stellar parameters. The data is preprocessed to remove noise and handle missing values, ensuring quality inputs for the model.The processed data is then fed into a machine learning model. The model, trained on a labeled dataset of known stellar properties, analyzes the input data to predict properties such as stellar mass, age, and composition. After prediction, the results are validated against test data for accuracy and reliability. Finally, the predictions are visualized and can be interpreted for astrophysical research and applications.

V. FUTURE DEVELOPMENTS IN PREDICTIVE MODELING OF STELLAR PROPERTIES

Future research in stellar property prediction may focus on developing enhanced model complexity to capture intricate relationships within stellar data. This could involve utilizing advanced neural network architectures or ensemble methods that combine multiple algorithms for greater accuracy and robustness. Incorporating data from additional astronomical surveys beyond Gaia, including multi-wavelength observations and space telescope data, could further improve the predictive capabilities of these models by providing a more comprehensive understanding of stellar properties and their environments. The development of real-time data processing capabilities will also become increasingly important as the volume and speed of astronomical data collection grow, enabling immediate insights into transient astronomical events such as supernovae or gamma-ray bursts. Bridging machine learning approaches with traditional astrophysical models will remain a key area of focus, as future studies explore how machine

learning predictions align with or challenge established theories, potentially offering new astrophysical insights. Additionally, adapting these methodologies for exoplanet studies could enhance predictions of stars likely to host exoplanets, advancing the search for habitable worlds and contributing to our understanding of planetary system formation. As machine learning models grow more complex, ensuring their interpretability will be critical, with future efforts aimed at developing techniques that help researchers understand how models make predictions, thus validating results in scientific contexts. Finally, the integration of machine learning into astrophysics is expected to expand the field of astroinformatics, leading to new tools and frameworks for analyzing large astronomical datasets and deepening our understanding of the universe.

VI. CONCLUSION

Based on the spectral data obtained from the Sloan Digital Sky Survey, this study trained the model with logistic regression, neural network, decision tree and random forest in machine learning, and classified the galaxies, stars and quasars in the universe. From the training results of the three models, it can be seen that the random forest algorithm has the best performance in the dataset, which not only has the highest accuracy rate of 98%, but also has a high computing efficiency. Compared with it, decision tree and support vector machine algorithms have a high accuracy rate of 97%, but support vector machine algorithm needs the longest time for calculation, whose efficiency is the lowest. Besides, all three algorithms performed well in the classification of galaxies, stars and quasars. As for the best performance in the star classification, it may be caused by the obvious difference between stellar and the other two stars in nature, resulting in the large differences in data. For the training, there are also some shortcomings that need to be improved. For example, in data processing, the combination of under sampling and over-sampling can better avoid over-fitting and improve the training accuracy.

.

REFERENCES

[1] Zhao Mingmei, Jin Yangyang, Wang Yujia, Zeng Mengjia. Application of Random Forest Algorithm in Decision Making. *Computer & Network*, 2024.

[2] Surana, V., et al. Predictive Modeling of Stellar Properties Using Synthetic Spectra. *Astrophysical Journal*, 2024.

[3] Smith, J., & Li, Q. Machine Learning Applications in Stellar Evolution Studies. *IEEE Transactions on Computational Astrophysics*, 2023.

[4] Tanaka, H., & Wang, F. Deep Learning Approaches to Time Series Stellar Variability Analysis. *Monthly Notices of the Royal Astronomical Society*, 2023.

[5] Garcia, R., et al. Enhancing Spectral Data Analysis with Gaussian Process Regression. *Astronomy & Astrophysics*, 2023.

[6] Cheng, K., & Huang, M. Fine-Tuning Transformer Models for Stellar Classification. *Journal of Machine Learning in Astronomy*, 2023.

[7] [7] Dékány, I., & Grebel, E. Estimating Metallicity from Time-Series Light CurvesUsing RNNs. *Astrophysical Journal*, 2023.

[8] Wilson, J., & Nickisch, H. Structured Kernel Interpolation for Large-Scale Stellar Data Analysis. *Monthly Notices of the Royal Astronomical Society*, 2023.

[9] LIU, Z., & ZHANG, Y. SCALABLE GP MODELS FOR STELLAR PARAMETER ESTIMATION. *IEEE TRANSACTIONS ON ASTROPHYSICS*, 2023.

[10] FLORES, J., ET AL. RECURRENT NEURAL NETWORKS FOR PREDICTING O-TYPE STELLAR PROPERTIES. *ASTRONOMICAL DATA SCIENCE REVIEW*, 2023.