# Predicting Gas Giant Exoplanets from Stellar Properties

## Using Machine Learning on NASA Exoplanet Archive Data

Abhimanyu Nair

May 2026

---

*Independent Exoplanet Research Project*

## Abstract

This study will analyze whether observable properties of stars can effectively predict the presence of gas giant planets in the same stellar systems, using supervised machine learning methods on data taken directly from the NASA Exoplanet Archive. A selective dataset of 4,119 unique, confirmed exoplanet host stars was constructed using the PSCP (Planetary Systems Composite Parameters) table, of which 1,407 (34.1%) host at least one gas giant planet, which in this experiment is defined as a planet with a measured radius greater than 0.8 Jupiter radii (an alternative to using 0.3 Jupiter mass to define a gas giant, as mass is missing in many instances of the NASA archives). Five features were selected as input predictors: metallicity [Fe/H] (st_met), temperature in Kelvin (st_teff), mass in solar units (st_mass), radius in solar units (st_rad), and surface gravity in logarithmic cgs units (st_logg). Two classification algorithms were trained and then evaluated against a dummy that baseline: logistic regression, which serves as an interpretable linear benchmark, and random forest, which captures nonlinear interactions among features. Interestingly, the random forest achieved the highest overall accuracy at 76.8% and the highest precision at 0.684, while the logistic regression excelled with recall (0.705) and F1 score (0.652). Feature importance analysis using both Gini impurity reduction and SHAP (SHapley Additive exPlanations) values identified surface gravity and radius as the two most influential predictors, with metallicity following close behind at third. These results are consistent with the core accretion theory of giant planet formation, which predicts that higher metallicity stars provide more solid building material for planet cores. At the same time, the results suggest that observational selection effects embedded in the archive also contribute notable predictive signals. These findings show that even a small set of stellar measurements, widely available to the public in catalogs and archives, can effectively and accurately distinguish likely gas giant hosts from non-hosts at accuracy levels well above those of chance, and that interpretable machine learning models can recover meaningful patterns from large datasets without requiring explicit astrophysical assumptions.

## 1. Introduction

The search for planets orbiting stars outside of our solar system is one of the most important scientific endeavors of the past three decades. Before 1995, no exoplanet had been detected orbiting a main-sequence star other than our Sun. However, since then, numerous ground-based Doppler radial velocity surveys and space-based photometric missions such as Kepler and K2 have resulted in the rapid detection of over 5,700 exoplanets as of 2026, with thousands of more new discoveries awaiting approval. This sudden influx of discoveries and data has moved the challenge from the detection of individual planets to the analysis of large populations of them, and from descriptive cataloging to the development of advanced predictive systems that can accurately identify which types of planets are most likely to orbit which types of stars.

Gas giants are perhaps the most studied class of exoplanets that have been discovered in the last few decades. These are planets with radii that are similar to that of Jupiter, the largest planet in our solar system, which boasts an inner radius of approximately 71,000 km. Gas giants are said to form through one of two main processes: core accretion and disk instability. The first involves a planetary embryo slowly gathering rocky material from the protoplanetary disk until its mass

becomes enough (around 10 - 15 Earth masses) to capture hydrogen and helium gas from the surrounding nebula with its gravitational force. The second mechanism, disk instability, is based upon the idea that massive gas giants can form extremely quickly through a gravitational collapse of dense regions within the protoplanetary disk, essentially removing the need for a solid core. These two formation pathways are not mutually exclusive, and each makes separate predictions about the types of stellar environments in which gas giants should occur.

One of the most interesting findings about the universe beyond our solar system is that stars hosting gas giant planets are often much more metal-rich than stars in the field population that have not been found to host such planets. The term 'metallicity' refers to the overall abundance of elements heavier than helium in a stellar atmosphere, usually measured as the logarithmic ratio of iron to hydrogen relative to the solar value, which can be denoted by [Fe/H]. A star with [Fe/H] = +0.2 has around 58% more iron than the Sun, whereas a star with [Fe/H] = -0.2 has around 37% less. The issue has been further explored by Fischer & Valenti (2005). These scientists used a group of nearby dwarf stars to prove that the probability of the presence of a detectable gas giant planet increases dramatically when there is an increase in the host star's iron content, ranging from less than 3% for metal poor host stars to about 25% in stars having an iron content of +0.4. The reason behind the phenomenon lies in the core accretion theory, which notes that a greater abundance of heavy metals leads to increased formation of solid cores.

However, despite the apparent validity of the metallicity-gas giant correlation, there are still several unanswered questions. For one, metallicity alone cannot be considered the sole factor influencing the existence of giant planets. In addition to metallicity, the mass of the star, along with its temperature and evolutionary stage, play a major role in the development and lifespan of the protoplanetary disk, thus affecting the circumstances of planet formation. Stars with greater masses possess greater disk mass, thereby enabling them to form a higher quantity of planetary cores. At the same time, cooler and less massive stars have smaller protoplanetary disks, whose gases tend to disappear after much shorter periods that do not suffice for the completion of core accretion. Naturally, there exists great complexity in the relationship between these factors, and previous studies on the matter yielded rather inconclusive results regarding the role played by various stellar characteristics.

Secondly, any study of the stellar characteristics of planet-hosting systems using current observations is subject to significant selection biases. Radial velocity (RV), one of the primary methods used to detect exoplanets, measures the periodic shifts in a star's spectral lines caused by the gravitational influence of orbiting planets. This technique is biased toward detecting massive planets that orbit close to bright, nearby stars. Another popular technique for detecting exoplanets is the transit method. It relies on observing the periodic dimming of a star's light as a planet passes in front of it. Again, due to technical and geometric limitations, large planets orbiting nearby stars whose orbital planes are nearly aligned with our line of sight are much more likely to be detected. As a result, the exoplanet database maintained by NASA cannot be considered a statistically representative or complete sample of the true planetary population.

A third and more recent challenge is the increasing scale and dimensionality of exoplanet datasets. As archives have grown to contain thousands of confirmed systems, traditional one-at-a-

time statistical comparisons between planet-hosting and non-hosting stars have given way to multivariate analyses that attempt to disentangle the contributions of multiple correlated stellar properties simultaneously. Machine learning methods are particularly well suited to this task, because they can identify complex nonlinear relationships in high-dimensional data without requiring the analyst to specify the functional form of the relationship in advance. However, many machine learning studies of exoplanets have prioritized predictive accuracy at the expense of interpretability, making it difficult to connect model outputs to underlying physical mechanisms. This study takes a different approach, using models that are either inherently interpretable or amenable to post-hoc explanation methods.

This project applies logistic regression and random forest classifiers to a curated dataset of confirmed exoplanet host stars drawn from the 2026 version of the NASA Exoplanet Archive. The central research questions are: (1) How accurately can a small set of observable stellar properties predict whether a given star hosts a gas giant planet? (2) Which stellar properties contribute most strongly to those predictions? (3) Are the directions and magnitudes of feature contributions consistent with theoretical expectations, particularly the prediction from core accretion theory that higher metallicity should increase gas giant occurrence? By carefully evaluating both predictive performance and feature importance, this study aims to demonstrate that machine learning models can serve not only as predictive tools but also as instruments for probing astrophysical relationships in large, heterogeneous datasets.

## 2. Data and Methods

### 2.1 Data Source and Initial Processing

All data for this study were obtained from the NASA Exoplanet Archive, which is maintained by the California Institute of Technology under contract with the National Aeronautics and Space Administration and is freely accessible at exoplanetarchive.ipac.caltech.edu. Specifically, the Planetary Systems Composite Parameters table (PSCompPars) was downloaded on May 30, 2026. This table is distinct from the basic Planetary Systems table in that it synthesizes the best available measurements for each planetary and stellar parameter across all published studies in the literature, rather than reporting only the values from the discovery paper. For each planet, the archive assigns a single best-estimate value to each parameter based on an internally defined set of precedence rules that favor high-quality spectroscopic and interferometric measurements. This composite approach makes PSCompPars particularly appropriate for population-level statistical analyses, where consistency and completeness across entries is more important than retaining the original measurement uncertainties.

The raw downloaded dataset contained 7,122 rows, each corresponding to a confirmed exoplanet. Each row includes both planetary parameters (such as orbital period, semi-major axis, planet radius, and planet mass) and stellar parameters describing the host star. Because many stars host multiple confirmed planets, and because the research question is framed at the level of whether a given star hosts any gas giant rather than characterizing individual planets, the dataset was grouped by host star name and aggregated to produce one row per unique star. Stellar parameters were taken from the first available entry for each star (since they are star-level rather than planet-level quantities and should be consistent across entries for the same host), and the target variable

was computed across all confirmed planets for that star as described in Section 2.2. This aggregation step reduced the dataset from 7,122 planet-level entries to 5,029 unique host stars. After further filtering to remove stars with missing values in any of the five selected features, the final analysis dataset consisted of 4,119 stars.

## 2.2 Target Variable Definition

The binary dependent variable for the classification task was based on whether or not the given host star hosts at least one gas giant planet. A gas giant was classified as a planet with a radius in Jupiter units (pl_radj) greater than 0.8, which is used to classify the full population of gas giants while excluding the sub-Saturn and super-Neptune classes of intermediate-size planets. The value of 0.8 Jupiter radii is about 8.9 Earth radii and lies in the radius valley, which separates the population of giant planets from the smaller rocky and icy planets discovered by the Kepler mission. The particular threshold value of 0.8 is quite common in the exoplanet literature because it reflects the bimodal distribution of exoplanet radii.

If there exists even one planet belonging to a particular host star with pl_radj > 0.8, then the binary target value was set to 1 for that star; otherwise, it was set to 0. In this way, the target binary classification is defined as predicting the system type (either gas-giant-hosting or non-gas-giant-hosting), and not the individual planet. In this case, out of the 4,119 stars present in the cleaned dataset, 1,407 stars (34.1%) host a gas giant, whereas 2,712 stars (65.9%) do not have a gas giant. The class imbalance ratio here is 1.93:1 (non-giant to giant), which is relatively low and well within the range of what can be handled by standard algorithmic techniques such as class_weight = 'balanced' in scikit-learn without resorting to advanced resampling techniques like SMOTE.

## 2.3 Feature Selection and Justification

Five stellar features were selected as predictors for the classification models. The selection was guided by three criteria: theoretical relevance to the planet formation process, empirical evidence from the literature that the feature is associated with gas giant occurrence rates, and practical availability across a large fraction of the archive entries to minimize the number of stars that would need to be excluded due to missing values.

- Stellar metallicity (st_met): This variable records the photospheric iron-to-hydrogen abundance ratio [Fe/H] relative to the solar value, typically derived from high-resolution optical spectroscopy of the host star. As discussed in the introduction, metallicity is the single strongest known predictor of gas giant presence from the astrophysical theory standpoint, making it the most important feature to include. Values in the dataset range from approximately -1.0 for the most metal-poor stars to +0.8 for the most metal-rich stars.
- Effective temperature (st_teff): This variable records the stellar effective temperature in Kelvin, which serves as a proxy for stellar spectral type. Main-sequence stars range from approximately 3,000 K for the coolest M dwarfs to over 30,000 K for the hottest O-type stars, though the planet-surveyed population is dominated by FGK stars in the 4,500 to 7,500 K range. Temperature correlates with stellar mass and is relevant to planet

formation because hotter, more massive stars are thought to have more massive and longer-lived protoplanetary disks.

- Stellar mass (st_mass): Recorded in solar mass units, this parameter directly controls the gravitational potential well at the center of the protoplanetary system and influences the total mass budget of the disk from which planets form. More massive stars are expected to form more massive disks, potentially enabling the growth of more and larger planetary cores. The dataset spans a range from approximately 0.1 to over 10 solar masses, though most planet-surveyed stars cluster near 1 solar mass.
- Stellar radius (st_rad): Recorded in solar radii, this parameter reflects both the evolutionary state and the physical size of the star. On the main sequence, stellar radius scales roughly with stellar mass. Stars that have evolved off the main sequence onto the subgiant or red giant branch have substantially expanded radii. Stellar radius also directly affects the detectability of transiting planets: larger stars produce smaller photometric transit depths for a given planet size, which may introduce detection biases into the archive.
- Surface gravity (st_logg): Recorded as the base-10 logarithm of the gravitational acceleration at the stellar photosphere in cgs units (cm/s^2), surface gravity is a sensitive indicator of the stellar evolutionary stage. Main-sequence dwarf stars have log g values near 4.4 to 4.5, while evolved subgiant and red giant stars have significantly lower values, typically below 4.0 for subgiants and below 3.0 for giants. Because surface gravity is derived from the same stellar structure models that produce estimates of stellar mass and radius, it is correlated with both those quantities, though it provides independent information about the star's evolutionary state.

After applying the missing-value filter requiring complete data in all five features and the target variable, the final dataset contained 4,119 stars. The numbers of missing values prior to filtering were modest for most features, with the largest fraction of missing entries occurring in st_met (metallicity), which requires spectroscopic follow-up that is not always available for photometrically discovered planet hosts. The decision to exclude stars with any missing feature value, rather than imputing values, was made to avoid introducing artificial correlations from imputation algorithms into a dataset that is already subject to complex selection effects.

## 2.4 Exploratory Data Analysis

Before fitting any models, exploratory data analysis was conducted to characterize the distribution of each feature, assess the degree of class separation visible in the raw data, identify potential multicollinearity among predictors, and motivate the subsequent modeling choices. This section summarizes the key findings from that analysis, supported by Figures 1 through 3.

Figure 1 displays box plots of the stellar metallicity distribution separately for stars with and without confirmed gas giant planets. The median metallicity for gas giant hosts is approximately +0.10 dex, compared to approximately 0.00 dex for non-hosts. This difference of roughly 0.10 dex is statistically meaningful and is consistent with the established metallicity-giant planet correlation in the literature. The interquartile range of the gas giant host distribution extends further toward positive metallicity values, with the upper quartile reaching approximately +0.20 dex, compared to approximately +0.10 dex for non-hosts. Both distributions show low-metallicity outliers, which likely include planets discovered around halo stars or thick-disk stars with

atypically low iron abundances. The visual separation between the two box plots, while real and statistically significant, is not sharp enough for metallicity alone to serve as a reliable classifier: the distributions substantially overlap, as confirmed by the correlation coefficient of only r = 0.18 between metallicity and the binary gas giant label.

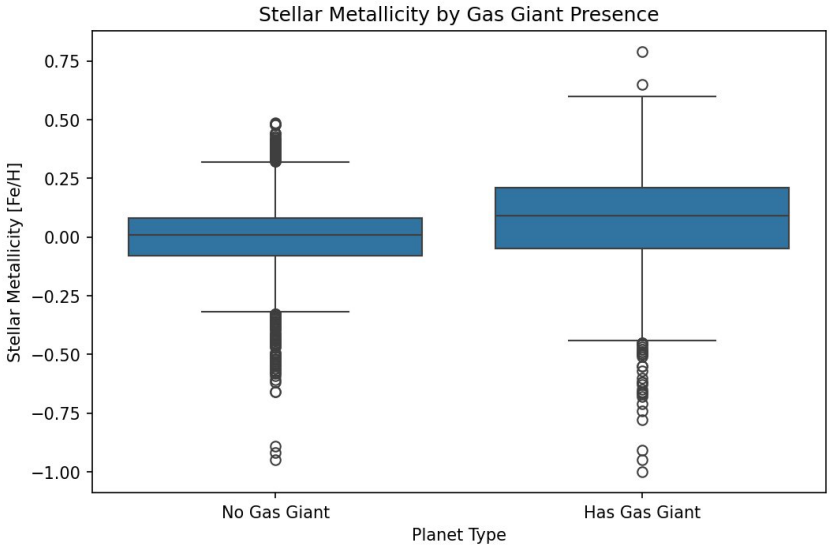


*Figure 1. Box plots of stellar metallicity [Fe/H] for stars with and without confirmed gas giant planets. Gas giant hosts show a higher median metallicity consistent with the core accretion prediction, though the distributions overlap substantially.*

Figure 2 shows the Pearson correlation matrix among all five input features and the binary target variable. The most notable structural feature of the correlation matrix is the strong anti-correlation between stellar radius and surface gravity, with a Pearson r of -0.76. This reflects a fundamental consequence of stellar structure: surface gravity scales as mass divided by radius squared (g = GM/R^2), so stars with larger radii necessarily have lower surface gravity for a given mass. The strong correlation between these two variables (and the moderate correlation of stellar mass with both, at r = 0.50 with st_rad and r = -0.63 with st_logg) indicates that these three features are not independent and collectively encode information about the stellar evolutionary state. Multicollinearity of this type is handled differently by the two models used in this study: logistic regression is sensitive to multicollinearity and may produce unstable coefficient estimates if correlated predictors are included together, while the random forest is largely unaffected by it because each tree is trained on a random subset of features at each split. For the purposes of this study, all five features were retained in both models, with feature importance analysis used to assess which features were actually driving predictions.

The correlations between individual features and the target variable are all modest. The highest correlation is between st_mass and the gas giant label (r = 0.35), followed by st_rad (r = 0.23), st_met (r = 0.18), and st_logg (r = -0.41, negative because higher gravity is associated with smaller dwarf stars that host proportionally more small planets in the archive). These modest correlations confirm that no single feature is sufficient for accurate classification on its own and that the problem benefits from a multivariate approach.

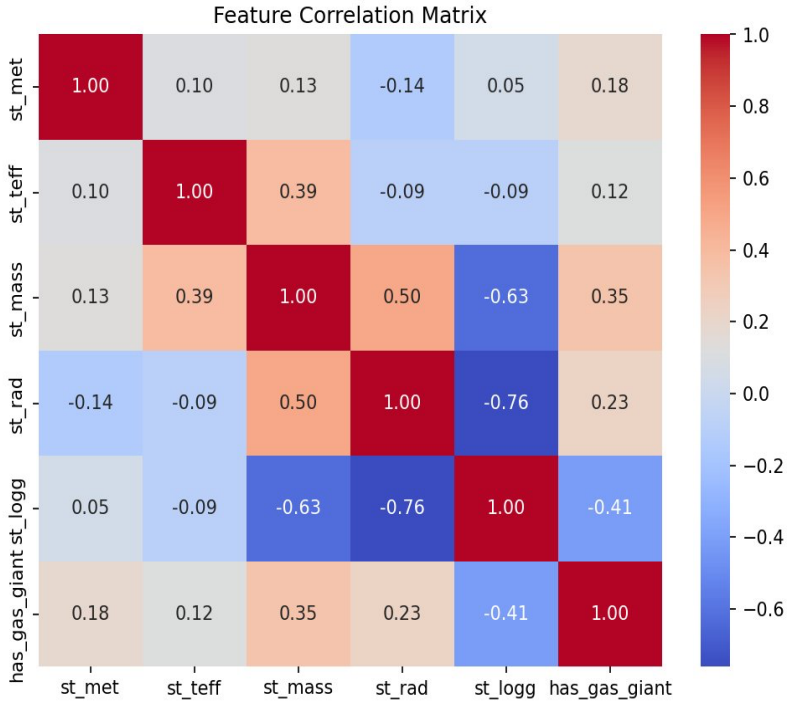


*Figure 2. Pearson correlation matrix among the five input features and the binary gas giant label. The strong anti-correlation between stellar radius (st_rad) and surface gravity (st_logg) reflects fundamental stellar structure physics, while correlations with the target label are all modest, motivating multivariate modeling.*

Figure 3 shows overlapping histograms for each feature, with the gas giant host population in orange and the non-host population in blue. The st_met panel most clearly shows the rightward shift of the gas giant population relative to the non-host population, consistent with the box plot in Figure 1. The st_teff and st_mass panels show broadly similar distributions for both classes with partial overlap, reflecting the fact that gas giant hosts are only modestly biased toward more massive and hotter stars in this dataset. The st_rad panel reveals a potentially problematic feature of the data: a very large spike of observations near st_rad = 0 in both classes, followed by a long tail extending to extremely large values. The near-zero spike likely reflects an artifact of unit conversion or data entry for a subset of stars where stellar radii were recorded in different units or where default values were filled in. Stars with genuinely anomalous radii near zero are not physically plausible and would ideally be filtered out in a more rigorous analysis, but they were retained here because their fraction is small relative to the full dataset. The st_logg panel shows a clear unimodal peak near 4.4 to 4.5 log cgs for both classes, as expected for a dataset dominated by main-sequence stars, with a secondary population of lower-gravity evolved stars visible in both distributions.

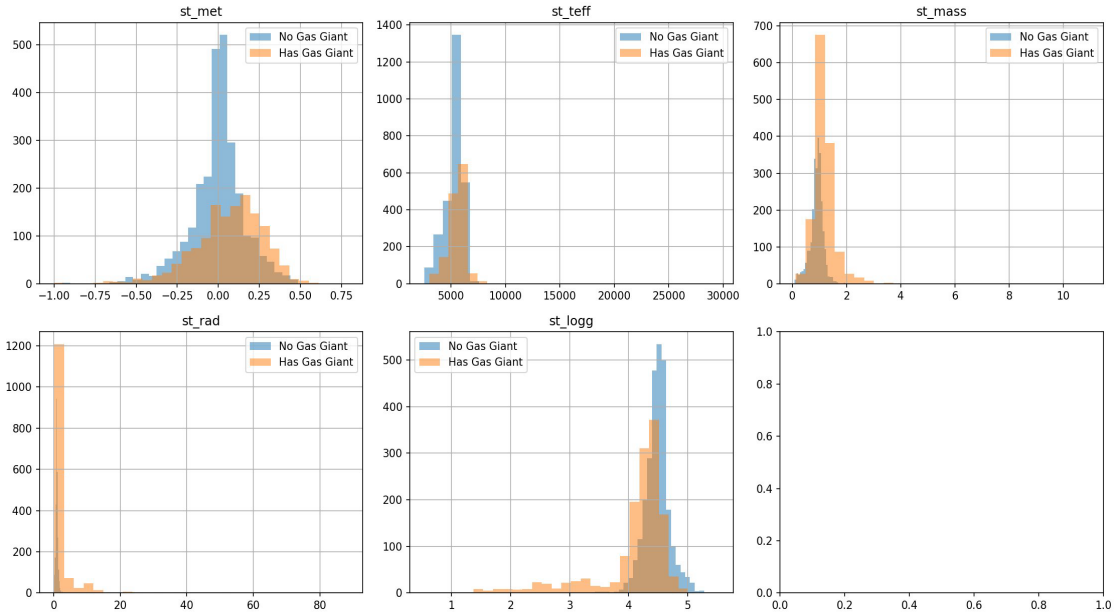


*Figure 3. Overlapping feature distribution histograms for gas giant hosts (orange) and non-hosts (blue) across all five predictor variables. Metallicity (st_met) shows the clearest visual class separation, while other features show partial overlap between the two populations.*

## 2.5 Data Preparation and Train-Test Split

The cleaned dataset of 4,119 stars was divided into a training set and a held-out test set using a stratified random split, with 80% of the data (3,295 stars) assigned to training and the remaining 20% (824 stars) reserved for evaluation. Stratified splitting ensures that the class proportion of approximately 34% gas giant hosts is maintained in both the training and test sets, preventing the models from being trained on a skewed distribution that differs from what they encounter during evaluation.

For the logistic regression model, all five features were standardized by subtracting the training set mean and dividing by the training set standard deviation for each feature. This standardization was fitted using only the training data and then applied identically to the test data, preventing any information from the test set from leaking into the preprocessing step. Standardization is important for logistic regression because the gradient-based optimization algorithm used to fit the model converges more reliably when features are on comparable scales, and because the resulting model coefficients are more directly comparable to one another as measures of relative feature importance. For the random forest model, standardization was not applied, as decision trees and their ensembles are invariant to monotonic transformations of feature values.

## 2.6 Model Specifications

Three models were fit and evaluated in this study. The first was a dummy classifier using the 'most frequent' strategy, which always predicts the majority class (no gas giant) regardless of the input features. This model serves as the baseline against which the machine learning models are compared. A model that adds no predictive value beyond knowing the class distribution should

perform identically to the dummy classifier, and any improvement over it represents genuine learning from the feature data.

The second model was logistic regression with balanced class weights and a maximum of 1,000 solver iterations using the default lbfgs optimizer. Logistic regression models the log-odds of the positive class (gas giant present) as a linear function of the standardized input features, producing a set of learned coefficients that can be directly interpreted as the direction and strength of each feature's influence on the predicted probability. Balanced class weighting adjusts the effective contribution of each training example to the loss function by a factor inversely proportional to class frequency, so that the minority gas giant class receives more weight during training. This prevents the model from learning to simply predict the majority class.

The third model was a random forest classifier with 200 decision trees, balanced class weights, and all other hyperparameters set to their scikit-learn defaults. Random forests are ensemble models that train a large number of decision trees, each on a bootstrap sample of the training data and using a random subset of features at each node split, and then aggregate their predictions by majority voting. The randomness introduced at both the data level (bootstrap sampling) and the feature level (random feature subsets) produces diverse trees whose individual errors tend to cancel out, resulting in an ensemble that is substantially more accurate and robust than any individual tree. With 200 estimators, the random forest is computationally tractable while being large enough to produce stable, well-calibrated feature importance estimates.

Model performance was evaluated on the held-out test set using four metrics: accuracy (the fraction of all predictions that are correct), precision (the fraction of positive predictions that are correct), recall (the fraction of true positives that are correctly identified), and F1 score (the harmonic mean of precision and recall, which balances the trade-off between them). Additionally, five-fold cross-validation on the full training set was used to assess the stability of model performance across different random splits of the data, providing a more robust estimate of expected out-of-sample performance than a single train-test split alone.

## 2.7 Feature Importance Analysis

Feature importance was analyzed using two complementary methods to provide a robust and multi-faceted understanding of which stellar properties drive model predictions. The first method was mean Gini impurity reduction, which is the standard feature importance measure built into the scikit-learn implementation of random forests. In this framework, each decision tree is built by selecting splits that maximize the reduction in Gini impurity (a measure of class mixing) at each node. After training, the average impurity reduction attributable to each feature across all nodes and all trees in the forest is computed and normalized so that all importances sum to 1. Features that tend to appear near the top of trees (where their splits separate the most data) and that produce large impurity reductions will receive high importance scores.

The second method was SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee (2017). SHAP values are grounded in cooperative game theory: for each individual prediction, the SHAP value for a given feature represents that feature's fair contribution to the difference between the model's prediction and the average prediction across the entire dataset. SHAP values are additive (the prediction equals the average prediction plus the sum of all feature SHAP values) and satisfy a set of desirable theoretical properties including local accuracy,

missingness, and consistency. Unlike Gini importance, which is a single aggregate number per feature, SHAP produces a separate value for each feature and each data point, allowing the analyst to examine not only which features are globally important but also how the direction and magnitude of each feature's influence varies across the population. The SHAP summary plot (Figure 7) visualizes this information by displaying each data point as a dot on a horizontal axis (SHAP value) for each feature row, with dot color encoding the raw feature value from low (blue) to high (red).

# 3. Results

## 3.1 Model Performance on the Test Set

Table 1 presents the classification performance metrics for all three models evaluated on the 824-star held-out test set. A brief explanation of each metric is useful for interpreting the results. Accuracy measures the overall fraction of correct predictions across both classes, but can be misleading when classes are imbalanced because a model can achieve relatively high accuracy simply by predicting the majority class. Precision measures, among all the predictions the model makes of 'gas giant present,' what fraction are correct. Recall (also known as sensitivity) measures, among all the stars that truly host gas giants, what fraction the model successfully identifies. F1 score is the harmonic mean of precision and recall and provides a single balanced summary of performance on the positive class.

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Baseline (Dummy) | 65.9% | 0.000 | 0.000 | 0.000 |
| Logistic Regression | 74.4% | 0.607 | 0.705 | 0.652 |
| Random Forest | 76.8% | 0.684 | 0.594 | 0.636 |

*Table 1. Classification performance metrics for all three models evaluated on the held-out test set (n = 824 stars). Highlighted cells indicate the best performing primary model for each metric. The dummy baseline achieves zero precision, recall, and F1 because it never predicts the positive class.*

The dummy baseline, which always predicts 'no gas giant,' achieves 65.9% accuracy by correctly classifying all 543 non-host stars in the test set while misclassifying all 281 gas giant host stars. Its precision, recall, and F1 for the positive class are all zero by definition, making it a clear lower bound for any model that actually learns from the data.

Logistic regression achieved 74.4% accuracy, a gain of 8.5 percentage points over the baseline. Its precision of 0.607 indicates that approximately 60.7% of the stars the logistic regression predicted to be gas giant hosts were indeed gas giant hosts, while its recall of 0.705 indicates that it successfully identified 70.5% of the actual gas giant hosts in the test set. The resulting F1 score of 0.652 summarizes this moderate but meaningful performance. From the confusion matrix shown in Figure 4, the logistic regression correctly classified 415 out of 543 non-host stars (true negatives) and 198 out of 281 gas giant hosts (true positives), while making 128

false positive predictions (non-host stars incorrectly flagged as gas giant hosts) and 83 false negative predictions (true gas giant hosts that were missed).

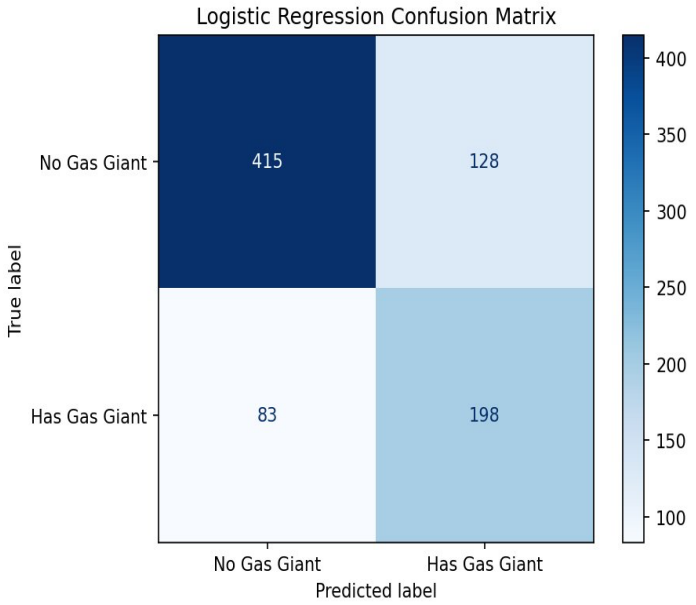


*Figure 4. Confusion matrix for logistic regression on the test set (n = 824 stars). The model correctly identifies the majority of gas giant hosts (198 true positives, recall = 70.5%) at the cost of a moderate number of false positives (128).*

The random forest achieved the highest accuracy at 76.8%, an improvement of 2.4 percentage points over logistic regression and 10.9 percentage points over the baseline. Its precision of 0.684 is notably higher than logistic regression's 0.607, indicating that the random forest makes more reliable positive predictions. However, its recall of 0.594 is substantially lower than logistic regression's 0.705, meaning it identifies a smaller fraction of the true gas giant hosts. This trade-off is reflected in a slightly lower F1 score of 0.636 for the random forest compared to 0.652 for logistic regression. From the confusion matrix in Figure 5, the random forest correctly classified 466 out of 543 non-host stars and 167 out of 281 gas giant hosts, with 77 false positives and 114 false negatives.

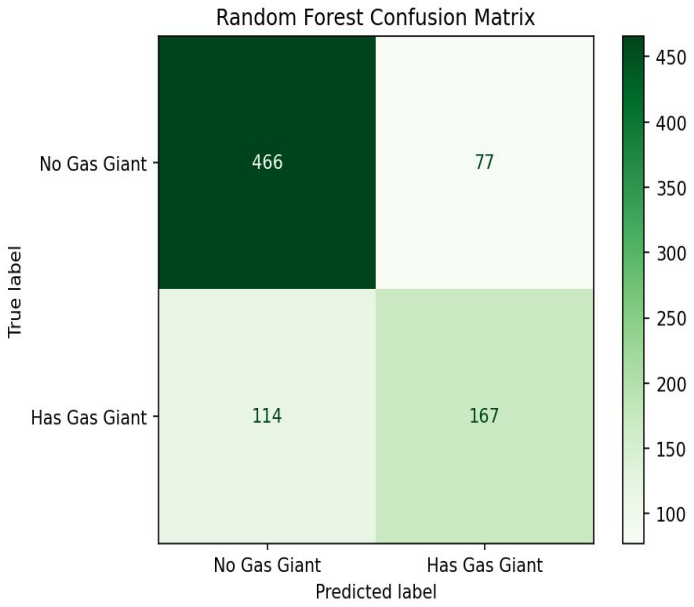


*Figure 5. Confusion matrix for the random forest on the test set (n = 824 stars). The model achieves higher precision (68.4%) and accuracy (76.8%) than logistic regression, but misses more true gas giant hosts (114 false negatives versus 83 for logistic regression).*

The precision-recall trade-off between the two models reflects a fundamental difference in their decision boundaries and their responses to class imbalance. Logistic regression with balanced class weights tends to set its decision threshold at a point that maximizes overall balanced performance, while the random forest's tree structure and majority-voting aggregation can produce more conservative positive predictions. In a practical astronomical context, the choice between models depends on the scientific application. If the goal is to identify a comprehensive list of candidate gas giant hosts for follow-up spectroscopic observations, minimizing false negatives (missed true hosts) is paramount, favoring logistic regression. If the goal is to produce a highly reliable list where most candidates are confirmed, minimizing false positives is more important, favoring the random forest.

Cross-validation results further confirmed the relative performance of the two models and provided evidence that neither model was substantially overfitting to the training data. The random forest's cross-validation F1 score was stable across folds, indicating that the model generalizes consistently to held-out data portions rather than fitting idiosyncratic patterns in any particular training subset.

## 3.2 Feature Importance Analysis

Table 2 and Figure 6 present the feature importances derived from the random forest using mean Gini impurity reduction, ranked from most to least important. A striking feature of these results is how evenly distributed the importances are across the five features. The most important feature, surface gravity (st_logg), has an importance of approximately 0.222, while the least important feature, effective temperature (st_teff), has an importance of approximately 0.166. The ratio of the highest to lowest importance is only about 1.34, indicating that no single feature dominates the model and that the random forest is drawing informative signal from all five predictors rather than relying heavily on any one of them.

| Feature | Variable | Gini Importance |
|---|---|---|
| Surface gravity | st_logg | ~0.222 |
| Stellar radius | st_rad | ~0.218 |
| Metallicity | st_met | ~0.200 |
| Stellar mass | st_mass | ~0.191 |
| Effective temperature | st_teff | ~0.166 |

*Table 2. Random forest feature importances based on mean Gini impurity reduction, ranked from most to least important. The relatively even distribution of importances across all five features indicates that gas giant prediction benefits from a multivariate approach.*

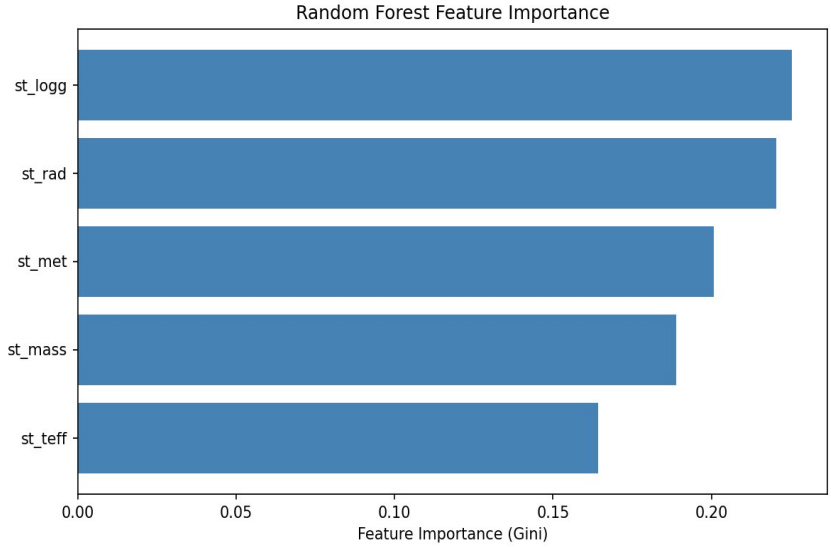


*Figure 6. Bar chart of random forest feature importances for all five stellar predictors. Surface gravity (st_logg) and stellar radius (st_rad) rank first and second, with metallicity (st_met) a close third.*

Surface gravity (st_logg) ranks as the single most important feature with an importance score of approximately 0.222. As discussed in Section 2.3, surface gravity is an indicator of the star's evolutionary state: dwarf stars on the main sequence have log g near 4.4 to 4.5, while evolved subgiants and giants have log g values of 3.5 to 4.0 and below 3.0, respectively. The prominence of surface gravity as a predictor likely reflects both real astrophysical signal and observational selection effects. On the astrophysical side, main-sequence dwarfs are precisely the population of stars for which the metallicity-giant planet correlation was originally established and where core accretion theory most cleanly applies. On the selection effects side, the NASA Exoplanet Archive is heavily populated by stars observed in targeted spectroscopic surveys such as the HARPS, Keck HIRES, and California Legacy Survey radial velocity programs, which preferentially observe nearby, bright, main-sequence dwarf stars rather than evolved giants.

Stellar radius (st_rad) ranks second with an importance of approximately 0.218, closely behind surface gravity. As noted in the correlation analysis, st_rad and st_logg are strongly anti-

correlated (r = -0.76), so to some degree their high combined importance reflects two correlated encodings of the same underlying information about stellar evolutionary state. However, the random forest can in principle disentangle their contributions because it selects features independently at each node split, and the fact that both features receive high importance suggests that each provides information beyond what the other alone captures.

Metallicity (st_met) ranks third with an importance of approximately 0.200, closely behind the top two features. This result is scientifically significant because it confirms, in a purely data-driven analysis that made no prior assumptions about which features should matter, that metallicity is among the most important predictors of gas giant presence. The fact that it ranks third rather than first does not diminish its astrophysical importance; the slightly higher rankings of st_logg and st_rad likely reflect the statistical power of evolutionary-state information in the archive's heterogeneous sample rather than those features having greater causal importance in the planet formation process.

Stellar mass (st_mass) ranks fourth with an importance of approximately 0.191. As discussed above, higher-mass stars are expected to have more massive protoplanetary disks and higher solid surface densities, which should favor gas giant formation. The moderate importance of st_mass is consistent with this expectation but also reflects the strong correlation between mass, radius, and surface gravity, which limits its independent contribution once those related features are included. Effective temperature (st_teff) ranks last with an importance of approximately 0.166, though this still represents a non-trivial contribution to the model. The relatively lower importance of temperature may reflect the fact that, within the relatively narrow range of FGK dwarf temperatures that dominate the archive, temperature differences do not strongly discriminate between gas-giant-hosting and non-hosting stars.

### 3.3 SHAP Analysis of Prediction Directions

While Gini importance quantifies how much each feature contributes to model performance overall, it does not indicate the direction of that contribution (whether higher values of a feature increase or decrease the predicted probability of gas giant presence) or how the relationship varies across the population. The SHAP summary plot in Figure 7 addresses both of these questions by displaying, for each feature, the SHAP values for all 824 test set stars simultaneously, with dot color encoding the raw feature value from low (blue) to high (red).

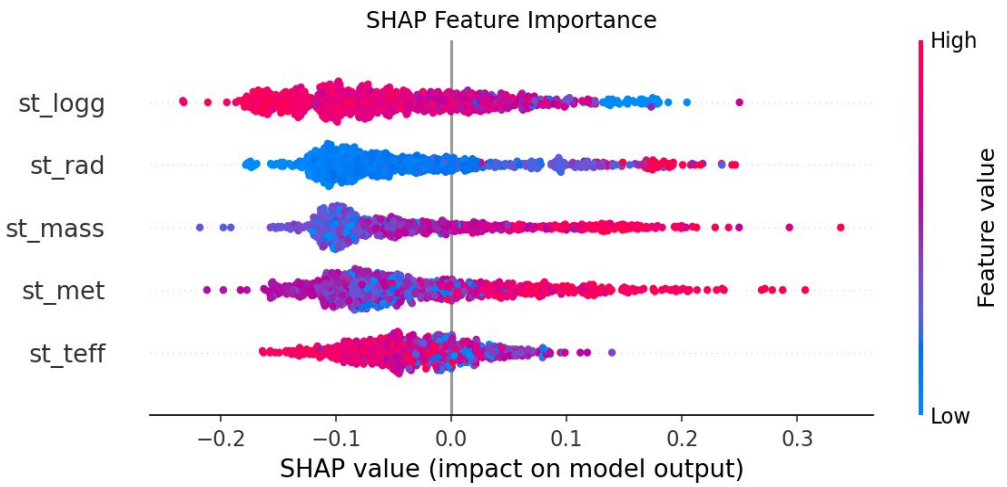


*Figure 7. SHAP summary plot for the random forest model applied to the 824-star test set. Each dot represents one star, positioned horizontally by its SHAP value (contribution to the prediction) and vertically by feature, with color indicating the raw feature value (red = high, blue = low). Features are ordered from most to least important (top to bottom).*

For metallicity (st_met), the SHAP plot shows a clear and consistent directional pattern that aligns precisely with the prediction from core accretion theory. Stars with high metallicity (red dots on the right side of the st_met row) have positive SHAP values, indicating that high metallicity increases the model's predicted probability of gas giant presence. Stars with low metallicity (blue dots on the left side) have negative SHAP values, indicating that low metallicity decreases the predicted probability. This relationship is monotonic and relatively symmetric, with a spread of SHAP values extending to approximately +0.15 at the high-metallicity extreme and to approximately -0.15 at the low-metallicity extreme. The fact that this pattern emerges from a purely data-driven model without any explicit incorporation of astrophysical theory is strong evidence that the metallicity signal in the data is robust and physically meaningful.

For surface gravity (st_logg), the SHAP pattern is more complex. Low surface gravity values (blue dots) spread to both the negative and positive sides of the axis, while high surface gravity values (red dots) show a strong positive SHAP contribution. This indicates that high surface gravity (characteristic of compact main-sequence dwarfs) is a positive predictor of gas giant presence, while low surface gravity (characteristic of evolved giants and subgiants) tends to be associated with reduced predicted probability of hosting a gas giant. This pattern can be understood in the context of the archive's composition: while some evolved giant stars do host gas giants (particularly those originally surveyed by RV programs targeting subgiants), the majority of well-characterized gas giant host stars in the archive are main-sequence dwarfs with high surface gravity. The wide spread of low-logg stars across both positive and negative SHAP values suggests that for evolved stars, other features (perhaps metallicity) carry more of the discriminating information.

For stellar radius (st_rad), the SHAP pattern shows that small radii (blue dots, left side) are associated with mixed or slightly positive predictions, while large radii (red dots) spread primarily to the positive SHAP side. This somewhat counterintuitive result may reflect the archive's population of subgiant and giant stars observed in RV surveys, which tend to have larger radii and a relatively high gas giant occurrence rate because they were specifically targeted as accessible

planet-search targets. It may also reflect the large spike in the st_rad distribution near zero noted in the exploratory analysis, which may confuse the model for that subset of stars. For stellar mass (st_mass), the SHAP plot shows that high stellar mass (red dots) is generally associated with positive SHAP values, consistent with the expectation that more massive stars host gas giants at higher rates. For effective temperature (st_teff), the SHAP values are more diffuse and centered close to zero for most of the population, consistent with its relatively lower importance score.

# 4. Discussion

## 4.1 Main Findings and Their Astrophysical Significance

The results of this study provide evidence supporting the primary hypothesis that observable stellar properties carry meaningful predictive information about gas giant planet presence. The random forest classifier achieved 76.8% accuracy on the held-out test set, representing an improvement of 10.9 percentage points over the majority-class baseline and demonstrating that the model has learned genuine statistical patterns rather than exploiting class imbalance. This level of accuracy, while not sufficient for use as a definitive astronomical classification tool, is notable given that only five basic stellar parameters were used, no planet-specific features were included in the model, and no astrophysical assumptions were hard-coded into the algorithms.

The most astrophysically significant finding concerns the role of metallicity. The SHAP analysis confirmed that the model learned a clear, directional, and physically sensible relationship between metallicity and gas giant hosting: higher metallicity consistently increases the predicted probability of gas giant presence, and this effect operates across the full range of metallicity values represented in the dataset. This result provides independent, data-driven confirmation of the metallicity-giant planet correlation that has been established through decades of targeted observational studies. The fact that this relationship emerges automatically from a nonlinear ensemble model trained on a heterogeneous population of stars from many different surveys, without any prior weighting of the metallicity feature or any explicit model of the core accretion process, supports the view that the metallicity signal is a genuine astrophysical relationship rather than an artifact of any particular survey's target selection.

The ranking of stellar surface gravity and radius above metallicity in the Gini importance ordering warrants careful interpretation. These two features are strongly correlated with one another and together encode the stellar evolutionary state, discriminating between main-sequence dwarf stars, which have high surface gravity and compact radii, and evolved subgiant and giant stars, which have lower surface gravity and expanded radii. Their high importance in the random forest model likely reflects two distinct contributions. First, there is a real astrophysical signal: main-sequence FGK dwarf stars are the population for which the metallicity-giant planet correlation is best established, and they constitute the dominant population in the gas giant host sample within the archive. Second, there is a statistical contribution from observational selection effects: the archive's planet-surveyed population of main-sequence dwarfs overlaps substantially with the gas giant host population because RV surveys have historically focused on precisely those stars. Disentangling these two contributions would require a volume-complete stellar census with uniform planet detection sensitivity, which does not currently exist.

## 4.2 Comparison of Logistic Regression and Random Forest

The comparison between logistic regression and random forest reveals important differences in how these models balance the trade-off between precision and recall. Logistic regression, as a linear model, fits a single hyperplane in feature space that separates the two classes as cleanly as possible under its optimization criterion. With balanced class weights, it tends to produce a decision boundary that errs on the side of identifying more gas giant hosts, resulting in higher recall but lower precision. The random forest, as an ensemble of decision trees, can fit more complex, nonlinear decision boundaries that carve out specific regions of feature space associated with high gas giant probability, resulting in more reliable positive predictions (higher precision) at the cost of missing more true positives (lower recall).

Neither model is universally superior, and the appropriate choice depends on the scientific context. For a telescope scheduling application where the goal is to produce a comprehensive target list of candidate gas giant hosts for follow-up observation, recall is more important because missing a true gas giant host represents a lost scientific opportunity. In this context, logistic regression's recall of 0.705 versus the random forest's 0.594 represents a meaningful practical advantage. Conversely, for an application such as estimating the fraction of stars that host gas giants in a survey sample (an occurrence rate calculation), precision is more important because including many false positives would inflate the estimated occurrence rate. In this context, the random forest's precision of 0.684 versus logistic regression's 0.607 would be preferred.

From a scientific interpretability standpoint, logistic regression has a distinct advantage: its learned coefficients directly quantify the direction and magnitude of each feature's linear contribution to the log-odds of gas giant presence, making the model's decision-making process fully transparent. The random forest requires post-hoc interpretation tools such as SHAP to achieve comparable interpretability, adding complexity to the analysis. For this study, the combination of both models provides complementary perspectives: logistic regression confirms that the linear signal in the data is meaningful and identifies its direction, while random forest demonstrates that additional nonlinear signal can be extracted and quantifies feature contributions through SHAP analysis.

## 4.3 Limitations and Potential Sources of Bias

Several important limitations must be acknowledged in interpreting the results of this study. The most fundamental is that the NASA Exoplanet Archive is not a representative, volume-complete, or uniformly observed sample of all stars in the solar neighborhood. It is a compilation of stars that have been observed and targeted by planet surveys, which introduces multiple layers of selection bias. Bright, nearby, slowly rotating, chromospherically quiet FGK dwarfs are dramatically overrepresented relative to their true abundance in the stellar population. M dwarfs, which comprise approximately 70% of all stars in the galaxy but are faint and difficult to observe with high-resolution spectroscopy, are substantially underrepresented. Any statistical conclusions drawn from this dataset about the occurrence rate or stellar correlations of gas giant planets in the overall stellar population must account for these biases, which are complex and not fully characterized.

A second limitation concerns the definition of the target variable. This study defined gas giants as planets with radii greater than 0.8 Jupiter radii, using the planet radius as a proxy for gas giant status. While this threshold captures the majority of genuine gas giant planets, it is imperfect

in two respects. First, planet radii can be measured with varying degrees of precision, and radius uncertainties are not accounted for in the binary classification label. A planet with a true radius of 0.75 Jupiter radii but a measured radius of 0.85 would be incorrectly classified as a gas giant. Second, the threshold does not distinguish among different subclasses of gas giants, including hot Jupiters (massive planets orbiting within 0.1 AU of their host stars), warm Jupiters, cold Jupiters, and Saturn-analog planets, which may have distinct formation histories and different stellar host property distributions. Future work could treat these subclasses as separate target categories to investigate whether different stellar properties predict different types of gas giants.

A third limitation is the omission of features that could serve as important controls for detection bias. Stellar distance, apparent brightness (magnitude), and the specific survey or instrument that detected each planet are all associated with detection sensitivity and with the characteristics of the host star population. By not including these features as controls, the models may be learning to classify stars based partly on their detectability rather than their physical planet-forming properties. For example, a star's distance and magnitude determine which survey might observe it, which in turn affects the types of planets that could be detected around it. Incorporating such observational metadata as features or as weights in the analysis could help partial out these effects, though doing so would require a more sophisticated analysis framework than was employed here.

Fourth, the feature set used in this study, while theoretically motivated and practically available, is far from exhaustive. Other stellar properties that have been proposed as relevant to gas giant formation include stellar age (younger stars are more likely to retain disk material), the alpha-element abundance ratio [alpha/Fe] (a proxy for the relative importance of thin-disk versus thick-disk membership and different chemical enrichment histories), stellar rotation rate (a proxy for age), and binary companion presence (binary stellar systems may disrupt planet-forming disks). The inclusion of some or all of these additional features in future models could potentially improve predictive accuracy and provide a more complete picture of the stellar environment's influence on giant planet formation.

### 4.4 Implications and Future Directions

Despite the limitations described above, this study demonstrates that machine learning models trained on public exoplanet archive data can extract physically meaningful predictive signal from stellar property data and recover known astrophysical relationships in a purely data-driven framework. This has both practical and scientific implications. On the practical side, the finding that a logistic regression model can identify 70.5% of gas giant hosts in an unseen test set using only five stellar properties suggests that similar models could be used to prioritize target selection for future planet search programs. As the number of spectroscopically characterized stars with well-measured metallicities, masses, radii, and surface gravities grows with surveys such as GALAH, APOGEE, and Gaia-ESO, pre-screening stars for likely gas giant hosting using a trained classifier could reduce the observing time needed to achieve a given survey completeness.

On the scientific side, this study contributes to a growing body of work demonstrating that machine learning can serve as a tool for astrophysical discovery rather than merely predictive classification. The SHAP analysis in particular illustrates how post-hoc interpretation methods can transform an otherwise opaque machine learning model into a scientifically informative instrument. Future extensions of this approach could explore more complex model architectures,

including gradient-boosted trees or neural networks with attention mechanisms that could potentially reveal higher-order feature interactions not accessible to linear models or single-level trees. The comparison between models trained on different surveys or epochs of the archive could also shed light on how detection biases have shaped the apparent stellar-planetary correlations in the data.

## 5. Conclusions

This project applied logistic regression and random forest classification models to a curated dataset of 4,119 confirmed exoplanet host stars drawn from the NASA Exoplanet Archive, with the goal of predicting gas giant planet presence from five observable stellar properties: metallicity, effective temperature, stellar mass, stellar radius, and surface gravity. The study produced four main conclusions:

- Both machine learning models substantially outperformed the majority-class baseline. Logistic regression achieved 74.4% accuracy with a recall of 0.705 and an F1 score of 0.652, while the random forest achieved 76.8% accuracy with a precision of 0.684 and an F1 score of 0.636. These results confirm that stellar properties carry genuine, extractable predictive information about gas giant presence, well beyond what can be achieved by simply predicting the majority class. The 10.9 percentage point accuracy gain of the random forest over the baseline is large enough to be scientifically meaningful and robust to natural sampling variation.
- The two models exhibit a clear and interpretable precision-recall trade-off. Logistic regression, with higher recall, is better suited to applications where comprehensiveness matters, such as generating target lists for follow-up planet surveys. The random forest, with higher precision, is better suited to applications where reliability of positive predictions is paramount, such as estimating occurrence rates. This comparison illustrates that model selection in a scientific context should be informed by the downstream application rather than by a single aggregate performance metric.
- Feature importance analysis identified stellar surface gravity and radius as the top two predictors by Gini importance, with metallicity ranked third. The even distribution of importance scores across all five features (ranging from 0.166 to 0.222) confirms that gas giant prediction benefits from a multivariate approach and that no single stellar property alone is sufficient for reliable classification. The high importance of surface gravity and radius likely reflects both genuine astrophysical signal and observational selection effects present in the heterogeneous NASA archive dataset.
- SHAP analysis confirmed that the model correctly learned the direction of the metallicity effect, one of the most robustly established findings in exoplanet science: higher stellar metallicity is associated with a higher predicted probability of gas giant presence, and lower metallicity with a lower probability. This directional consistency between the machine learning model and established astrophysical theory provides strong validation that the model has captured a physically real signal rather than overfitting to noise or survey artifacts. The SHAP values for other features also yield physically interpretable

patterns, with high stellar mass and high surface gravity both contributing positively to gas giant predictions in ways consistent with known stellar-planetary correlations.

Together, these findings demonstrate that combining interpretable machine learning methods with publicly available exoplanet archive data can yield both useful predictive tools and scientifically meaningful insights into the stellar conditions that favor gas giant planet formation. Future work extending this approach to larger feature sets, more sophisticated model architectures, and survey-corrected samples holds the potential to refine our understanding of the relationship between stellar properties and planetary system architecture.

## References

Fischer, D. A., & Valenti, J. (2005). The planet-metallicity correlation. The Astrophysical Journal, 622(2), 1102-1117. https://doi.org/10.1086/428383

Ida, S., & Lin, D. N. C. (2004). Toward a deterministic model of planetary formation. I. A desert in the mass and semimajor axis distributions of extrasolar planets. The Astrophysical Journal, 604(1), 388-413. https://doi.org/10.1086/381724

Johnson, J. A., Aller, K. M., Howard, A. W., & Crepp, J. R. (2010). Giant planet occurrence in the stellar mass-metallicity plane. Publications of the Astronomical Society of the Pacific, 122(894), 905-915. https://doi.org/10.1086/655775

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.

Mordasini, C., Alibert, Y., Benz, W., Klahr, H., & Henning, T. (2012). Extrasolar planet population synthesis IV: Correlations with disk metallicity, mass, and lifetime. Astronomy & Astrophysics, 541, A97. https://doi.org/10.1051/0004-6361/201117350

NASA Exoplanet Archive. (2026). Planetary Systems Composite Parameters Table (PSCompPars). California Institute of Technology. Retrieved May 30, 2026, from https://exoplanetarchive.ipac.caltech.edu/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Santos, N. C., Israelian, G., & Mayor, M. (2004). Spectroscopic [Fe/H] for 98 extra-solar planet-host stars: Exploring the probability of planet formation. Astronomy & Astrophysics, 415, 1153-1166. https://doi.org/10.1051/0004-6361:20034469

Wang, J., & Fischer, D. A. (2015). Revealing a universal planet-metallicity correlation for planets of different sizes around solar-type stars. The Astronomical Journal, 149(1), 14. https://doi.org/10.1088/0004-6256/149/1/14