Predicting Gas Giant Exoplanets from Stellar Properties

Using Machine Learning on NASA Exoplanet Archive Data

Abhimanyu Nair

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 used for this analysis came from the NASA Exoplanet Archive, which is supported by the California Institute of Technology under a contract with the National Aeronautics and Space Administration, and is free and publicly available at exoplanetarchive.ipac.caltech.edu. More specifically, the Planetary Systems Composite Parameters table (PSCompPars) was acquired on May 30, 2026. Unlike the regular Planetary Systems table, PSCompPars provides not only the data extracted from the discovery paper, but the best estimate of the planetary and stellar parameters for every planet using the information published in all of the available literature sources. Namely, the NASA Exoplanet Archive provides a single estimate of each of the parameters for each of the planets using the internal precedence rules that give priority to the high-quality spectroscopic and interferometric measurements.

The initial dataset, which had been downloaded as is, contained 7,122 rows, each of which corresponded to a confirmed exoplanet. The rows contained not only parameters for the planets themselves (like orbital period, semi-major axis, planet radius, and planet mass), but also parameters for their host stars. However, since there are several planets for many stars and the research question relates to whether or not a certain star has any gas giants, rather than describing individual planets, the data was aggregated on host stars and one row per unique star was created. The parameters for the stars were selected as the ones from the first entry for the particular star (as they are star-specific, not planet-specific parameters). The target variable was calculated for each host star across all of its confirmed planets using the approach described in Section 2.2. As a result of aggregation, the number of rows decreased from 7,122 to 5,029 (unique host stars). After excluding the stars which had missing values in any of the selected features, the number of stars for analysis became 4,119.

2.2 Target Variable Definition

The target binary variable for classification is defined based on whether the host star has any gas giants orbiting it. To qualify as gas giants, the planets had to have a radius more than 0.8 times that of Jupiter (pl_radj). This limit was set such that all giant

planets were considered and the intermediate size planets, which fall under the category of sub-Saturn and super-Neptune, were excluded. A radius of 0.8 times that of Jupiter translates to about 8.9 Earth radii and lies in the valley that exists between the population of gaseous giant planets and small rocky and icy planets found by the Kepler space telescope.

A binary target label value of 1 was assigned to each unique host star if any of its confirmed planets satisfied the criterion pl_radj > 0.8, and 0 otherwise. Such an approach enables us to consider the prediction task as a classification of systems (gas-giant hosts or not), rather than individual planets, thus circumventing the statistical challenges brought about by considering the planets of a single host as separate instances. According to this criterion, 1,407 out of the total 4,119 stars in the filtered dataset (34.1%) host one or more gas giants, whereas 2,712 stars (65.9%) do not have any gas giants among their planets. The achieved class imbalance ratio of 1.93:1 (not gas giant to gas giant) is moderate enough to be handled by most machine learning algorithms, including the built-in class_weight='balanced' setting in scikit-learn.

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 measures the [Fe/H] ratio, i.e., the iron-to-hydrogen abundance ratio in the photosphere compared to the sun's value, which is usually measured through high-resolution optical spectra of the star. In the introduction, we mentioned that metallicity is the best predictor of gas giants from an astrophysical theory perspective, which means it should be the first one on our list. The metallicity of stars in the dataset varies between -1.0 (least metallic stars) and +0.8 (most metallic stars).
- Effective Temperature (st_teff): This is the measure of the effective temperature of the star in units of Kelvin, and acts as a stand-in for stellar spectral class. Main sequence stars can be as cool as 3,000K for M dwarfs to 30,000K for hot O-type stars, although the survey planets are mostly in the spectral classes of FGK and have temperatures between 4,500 and 7,500 Kelvin. Temperature is related to mass and plays an important role in planet formation due to higher mass disks around higher mass stars.

- 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): In units of solar radii, this quantity describes the state of evolution and physical dimensions of the star. As it is on the main sequence, the stellar radius is approximately proportional to stellar mass. Once the star has moved from the main sequence to the subgiant or red giant branch, its radius has greatly increased. The stellar radius also directly influences the detectability of transiting planets; larger stars will give transit depths for a given planet radius that are smaller.
- 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 filtering out any rows with any missing values in all five features and the target label, the resulting dataset had 4,119 stars. The amount of missing values before the filtering procedure was relatively low for most of the features, although the highest number of missing values belonged to st_met (metallicity), as it requires a spectroscopic observation that might not be available for some photometrically observed planet host candidates. The choice to filter out the rows instead of filling in the values was done to prevent any false correlations from appearing due to the imputation algorithm.

2.4 Exploratory Data Analysis

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

Figure 1 shows box plots of stellar metallicity for stars with and without confirmed gas giants. The median metallicity of gas giant hosts is about +0.10 dex, compared to

about 0.00 dex for non-hosts. The upper quartile of the host distribution reaches roughly +0.20 dex, versus about +0.10 dex for non-hosts. Both distributions include low-metallicity outliers that may correspond to planets orbiting halo and thick-disk stars with unusually

low iron abundance. Although the two distributions are visibly separated, they overlap considerably, consistent with a modest correlation of r = 0.18 between metallicity and gas giant status.

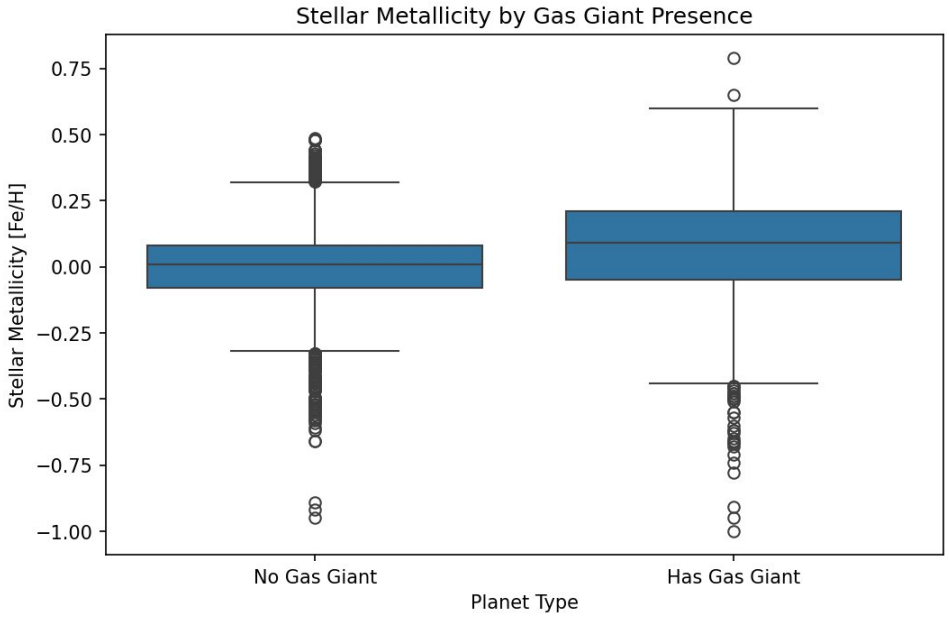

Figure 1. Box plots of stellar metallicity [Fe/H] for stars with and without confirmed gas giant planets. Gas giant hosts show a higher median metallicity consistent with the core accretion prediction, though the distributions overlap substantially.

Figure 2 shows the Pearson correlation matrix among the five input features and the binary target. The most notable structural feature is the strong anti-correlation between stellar radius and surface gravity (r = -0.76), a direct consequence of stellar structure, since surface gravity scales as mass divided by radius squared. Stellar mass is moderately correlated with both radius (r = 0.50) and surface gravity (r = -0.63), indicating that these three features jointly encode information about the star's evolutionary state rather than being fully independent.

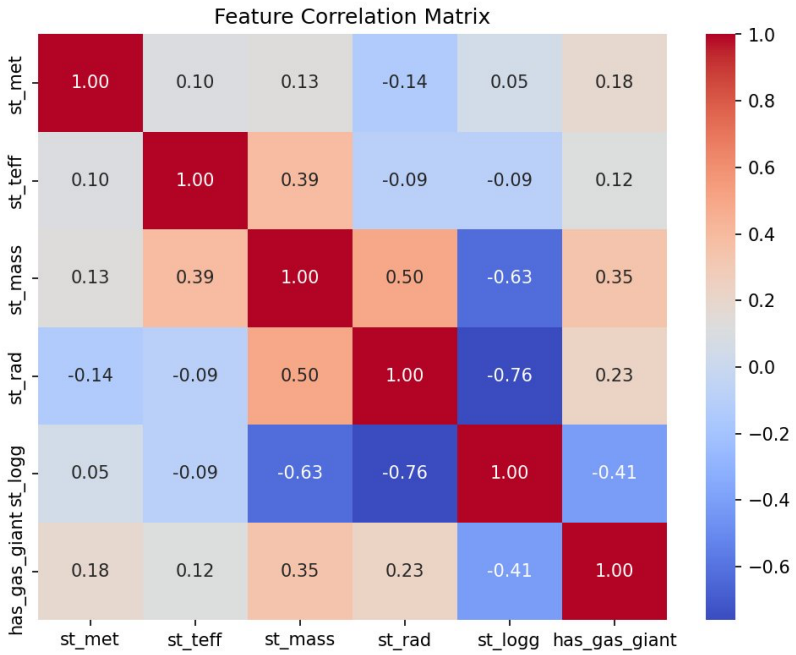

Figure 2. Pearson correlation matrix among the five input features and the binary gas giant label. The strong anti-correlation between stellar radius (st_rad) and surface gravity (st_logg) reflects fundamental stellar structure physics, while correlations with the target label are all modest, motivating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

Correlations between individual features and the target variable are all modest: st_mass shows the strongest association (r = 0.35), followed by st_rad (r = 0.23), st_met (r = 0.18), and st_logg (r = -0.41, negative because higher surface 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 on its own and that a multivariate approach is warranted. Multicollinearity affects the two models differently: logistic regression can produce unstable coefficients when correlated predictors are included together, whereas random forests are largely unaffected because each tree considers only a random subset of features at each split. All five features were retained in both models, with feature importance analysis used afterward to assess which features actually drove predictions.

Figure 3 shows overlapping histograms of each feature for gas giant hosts (orange) and non-hosts (blue). The st_met panel shows the clearest rightward shift for

hosts, consistent with Figure 1. The st_teff and st_mass panels show broadly similar distributions for both classes with partial overlap. The st_rad panel reveals a large spike of observations near zero in both classes, likely an artifact of unit conversion or default-value entry for a subset of stars; these values are not physically plausible but were retained since they represent a small fraction of the dataset. The st_logg panel shows a clear unimodal peak near 4.4-4.5, as expected for a dataset dominated by main-sequence stars, with a secondary population of lower-gravity evolved stars visible in both classes.

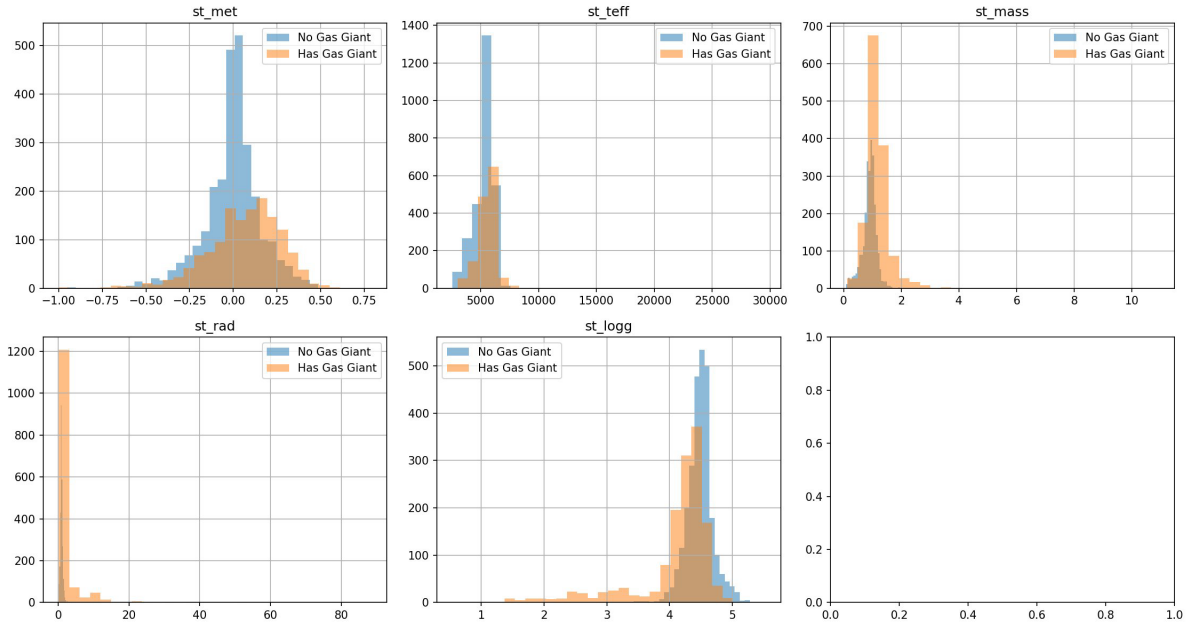

Figure 3. Overlapping feature distribution histograms for gas giant hosts (orange) and non-hosts (blue) across all five predictor variables. Metallicity (st_met) shows the clearest visual class separation, while other features show partial overlap between the two populations.

2.5 Data Preparation and Train-Test Split

The cleaned dataset of 4,119 stars was split into training (80%, 3,295 stars) and held-out test (20%, 824 stars) sets using a stratified random split, preserving the roughly 34% gas giant host rate in both sets. For logistic regression, all five features were standardized by subtracting the training-set mean and dividing by the training-set standard deviation; this transformation was fit on the training data only and then applied identically to the test data, preventing test-set leakage into preprocessing. Standardization improves convergence of the gradient-based solver and makes the resulting coefficients more directly comparable as measures of relative feature

importance. For the random forest, standardization was not applied, since tree-based ensembles are invariant to monotonic transformations of feature values.

2.6 Model Specifications

Three models were fit and evaluated. The first was a dummy classifier using the 'most frequent' strategy, which always predicts the majority class (no gas giant) regardless of input features; this serves as the baseline against which the machine learning models are compared. The second was logistic regression with balanced class weights and up to 1,000 solver iterations using the default lbfgs optimizer, which models the log-odds of gas giant presence as a linear function of the standardized features and yields directly interpretable coefficients. Balanced class weighting increases the effective training weight of the minority (gas giant) class so the model does not simply default to predicting the majority class. The third was a random forest classifier with 200 decision trees, balanced class weights, and scikit-learn's default hyperparameters otherwise. Random forests train many trees on bootstrap samples of the data using random feature subsets at each split, then aggregate predictions by majority vote; the resulting diversity across trees typically yields greater accuracy and robustness than any individual tree.

Model performance was evaluated on the held-out test set using accuracy, precision, recall, and F1 score. Five-fold cross-validation on the training set was additionally used to assess the stability of performance across different data splits, 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 assessed using two complementary methods. The first, mean Gini impurity reduction, is the standard importance measure built into scikit-learn's random forest implementation: each tree's splits are chosen to maximize the reduction in class-mixing (Gini impurity) at each node, and the average reduction attributable to each feature across all nodes and trees is computed and normalized to sum to 1. The second method, SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee (2017), is grounded in cooperative game theory: for each individual prediction, a feature's SHAP value represents its fair contribution to the difference between that prediction and the average prediction across the dataset. Unlike Gini importance, which yields a single aggregate score per feature, SHAP produces a value for every feature and every data point, allowing examination of both global importance and how a feature's direction and magnitude of influence vary across the population.

3. Results

3.1 Model Performance on the Test Set

Table 1 presents classification metrics for all three models on the 824-star held-out test set. Accuracy measures the overall fraction of correct predictions, but can be misleading under class imbalance, since a model can score well simply by predicting the majority class. Precision measures, among stars predicted to host a gas giant, what fraction actually do; recall measures, among stars that truly host a gas giant, what fraction the model correctly identifies; and F1 score is the harmonic mean of precision and recall.

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). The dummy baseline achieves zero precision, recall, and F1 because it never predicts the positive class.

The dummy baseline achieves 65.9% accuracy by correctly classifying all 543 non-host stars while misclassifying all 281 gas giant hosts; its precision, recall, and F1 are zero by definition, making it a clear lower bound for any model that genuinely learns from the data.

Logistic regression achieved 74.4% accuracy, an 8.5 percentage point gain over the baseline. Its precision of 0.607 indicates that about 60.7% of its positive predictions were correct, while its recall of 0.705 indicates that it identified 70.5% of true gas giant hosts, for an F1 score of 0.652. It correctly classified 415 of 543 non-hosts and 198 of 281 hosts, with 128 false positives and 83 false negatives.

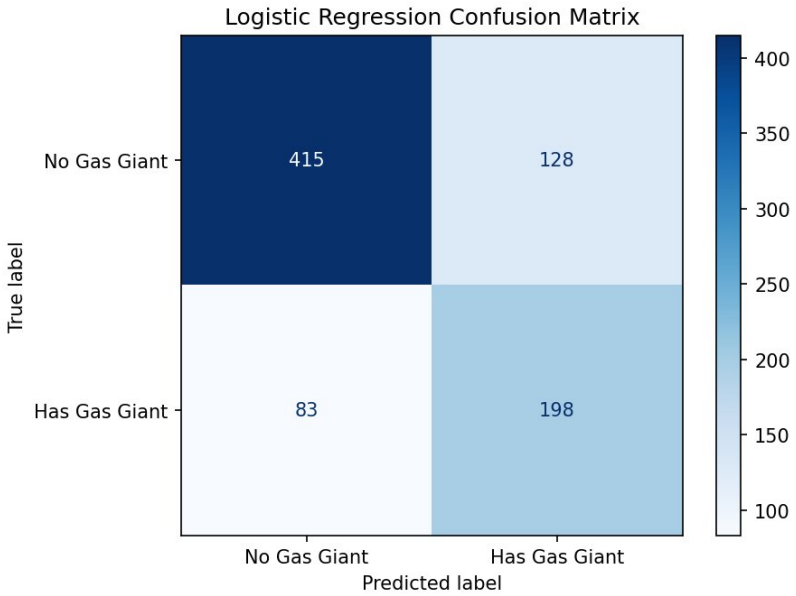

Figure 4. Confusion matrix for logistic regression on the test set (n = 824 stars). The model correctly identifies the majority of gas giant hosts (198 true positives, recall = 70.5%) at the cost of a moderate number of false positives (128).

The random forest achieved the highest accuracy at 76.8%, a 2.4 percentage point improvement over logistic regression and 10.9 points over the baseline. Its precision of

0.684 is notably higher than logistic regression's 0.607, indicating more reliable positive predictions, but its recall of 0.594 is substantially lower, meaning it misses more true hosts. This trade-off is reflected in a slightly lower F1 score (0.636 versus 0.652). It correctly classified 466 of 543 non-hosts and 167 of 281 hosts, with 77 false positives and 114 false negatives.

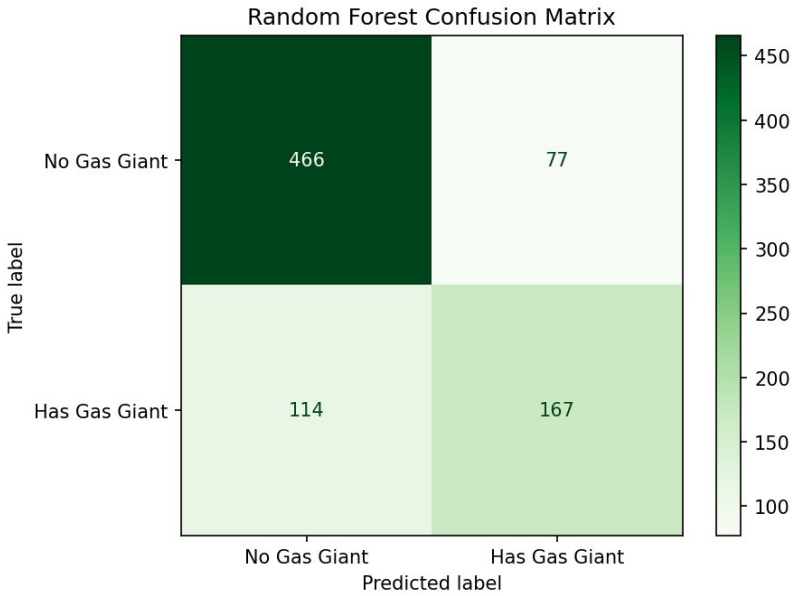

Figure 5. Confusion matrix for the random forest on the test set (n = 824 stars). The model achieves higher precision (68.4%) and accuracy (76.8%) than logistic regression, but misses more true gas giant hosts (114 false negatives versus 83 for logistic regression).

This precision-recall trade-off reflects a fundamental difference in decision boundaries: logistic regression with balanced class weights tends to maximize overall balanced performance, while the random forest's majority-voting structure produces more conservative positive predictions. The appropriate model therefore depends on the application. If the goal is a comprehensive candidate list for follow-up spectroscopy, minimizing false negatives favors logistic regression; if the goal is a highly reliable list of confirmed-likely candidates, minimizing false positives favors the random forest. Cross-validation results confirmed the relative ranking of the two models and showed no evidence of substantial overfitting, with the random forest's cross-validation F1 score remaining stable across folds.

3.2 Feature Importance Analysis

Table 2 and Figure 6 present random forest feature importances based on mean Gini impurity reduction, ranked from most to least important. Importances are relatively evenly distributed: the most important feature, surface gravity, scores about 0.222, while

the least important, effective temperature, scores about 0.166, a ratio of only 1.34. This indicates that no single feature dominates and that the model draws informative signal from all five predictors.

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.

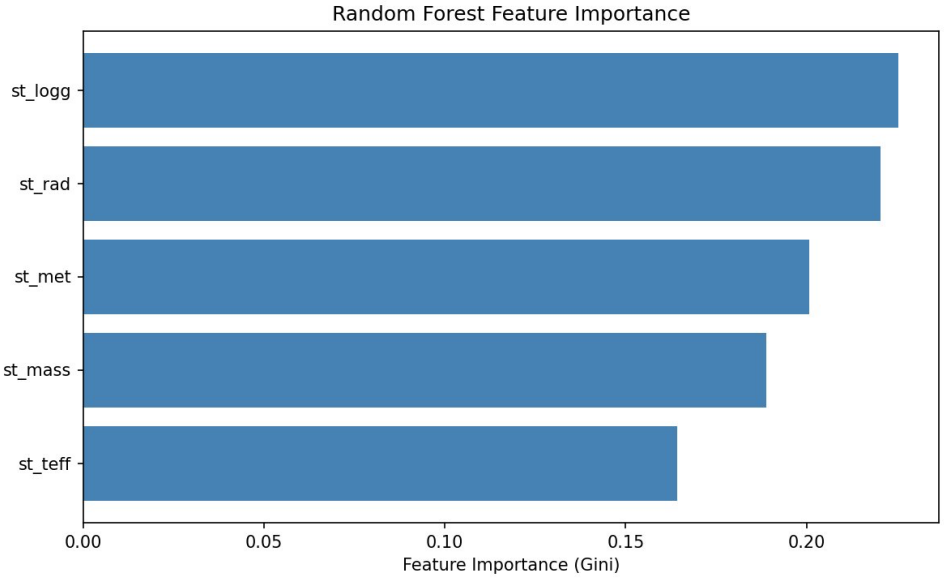

Figure 6. Bar chart of random forest feature importances for all five stellar predictors. Surface gravity (st_logg) and stellar radius (st_rad) rank first and second, with metallicity (st_met) a close third.

Surface gravity ranks first (about 0.222). As an indicator of evolutionary state, its prominence likely reflects both real astrophysical signal, since main-sequence dwarfs are the population for which the metallicity-giant planet correlation was originally established,

and observational selection effects, since the archive is dominated by targeted radial velocity surveys (HARPS, Keck HIRES, the California Legacy Survey) that preferentially observe nearby, bright, main-sequence dwarfs rather than evolved giants.

Stellar radius ranks second (about 0.218), closely behind surface gravity. Because radius and surface gravity are strongly anti-correlated ($r = -0.76$), their combined high importance partly reflects two correlated encodings of the same evolutionary-state information; nonetheless, because the random forest selects features independently at each split, both receiving high importance suggests each contributes information beyond the other.

Metallicity ranks third (about 0.200), close behind the top two. This is scientifically significant: a purely data-driven analysis, with no prior weighting toward metallicity, confirms it as one of the most important predictors of gas giant presence. Its third-place rank likely reflects the statistical power of evolutionary-state information within this heterogeneous archive sample rather than metallicity having a smaller causal role in planet formation.

Stellar mass ranks fourth (about 0.191), consistent with the expectation that more massive stars host more massive, higher-surface-density disks favorable to gas giant formation, though its independent contribution is limited once radius and surface gravity are included. Effective temperature ranks last (about 0.166) but still contributes non-trivially; its comparatively low importance may reflect the narrow range of FGK dwarf temperatures that dominate the archive, within which temperature does not strongly discriminate between hosts and non-hosts.

3.3 SHAP Analysis of Prediction Directions

While Gini importance quantifies each feature's overall contribution to performance, it does not reveal the direction of that contribution or how the relationship varies across the population. Figure 7 addresses this by displaying SHAP values for all 824 test-set stars, with dot color encoding the raw feature value from low (blue) to high (red).

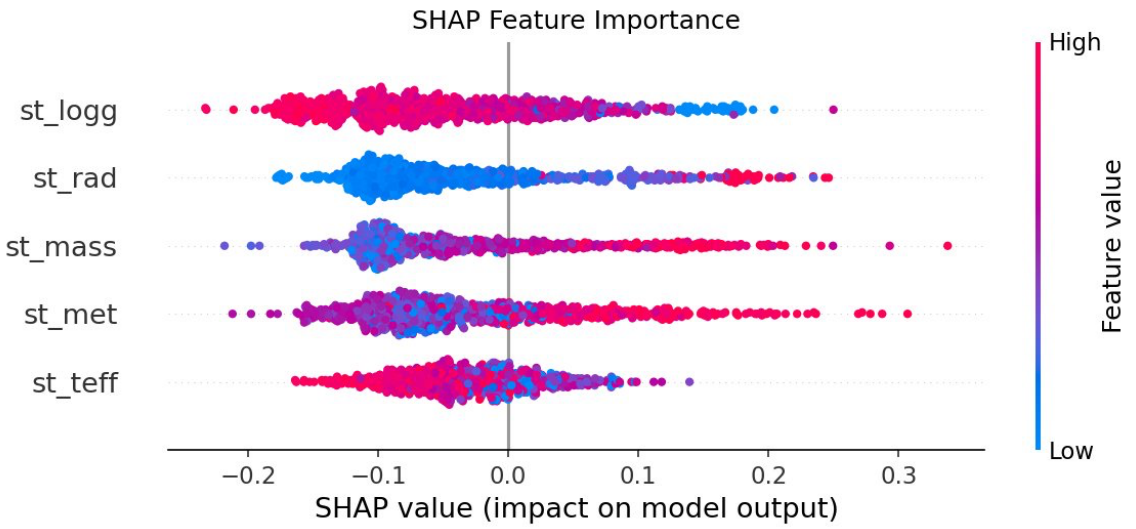

Figure 7. SHAP summary plot for the random forest model applied to the 824-star test set. Each dot represents one star, positioned horizontally by its SHAP value (contribution to the prediction) and vertically by feature, with color indicating the raw feature value (red = high, blue = low). Features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

For metallicity, the SHAP plot shows a clear, monotonic pattern aligned with core accretion theory: high-metallicity stars (red dots) show positive SHAP values, increasing the predicted probability of gas giant presence, while low-metallicity stars (blue dots) show negative values. This relationship is roughly symmetric, spanning about +0.15 to -0.15 across the metallicity range. Because this pattern emerges from a purely data-driven model with no explicit astrophysical assumptions built in, it is strong evidence that the metallicity signal in the data is robust and physically meaningful.

For surface gravity, the pattern is more complex: high-surface-gravity stars (compact main-sequence dwarfs) show a strong positive SHAP contribution, while low-surface-gravity stars (evolved subgiants and giants) spread across both positive and negative values. This likely reflects that most well-characterized gas giant hosts in the archive are main-sequence dwarfs, even though some evolved giant hosts exist, particularly among subgiants targeted by radial velocity programs; for evolved stars, other features such as metallicity appear to carry more of the discriminating signal.

For stellar radius, small radii are associated with mixed or slightly positive predictions, while large radii spread mainly toward positive SHAP values, a somewhat counterintuitive pattern that may reflect the archive's population of subgiant and giant stars specifically targeted by radial velocity surveys, as well as the near-zero radius spike noted in the exploratory analysis. For stellar mass, high mass is generally associated with positive SHAP values, consistent with the expectation that more massive stars host gas

giants at higher rates. For effective temperature, SHAP values are diffuse and centered near zero for most of the population, consistent with its lower overall importance.

4. Discussion

4.1 Main Findings and Their Astrophysical Significance

These results support the hypothesis that observable stellar properties carry meaningful predictive information about gas giant presence. The random forest's 76.8% test accuracy, a 10.9 percentage point improvement over the majority-class baseline, indicates genuine learned structure rather than exploitation of class imbalance. This performance is notable given that only five basic stellar parameters were used, no planet-specific features were included, and no astrophysical assumptions were hard-coded into either algorithm.

The most astrophysically significant finding concerns metallicity: SHAP analysis confirmed a clear, directional, physically sensible relationship in which higher metallicity consistently increases the predicted probability of gas giant presence across the full range of observed values. This provides independent, data-driven confirmation of the metallicity-giant planet correlation established over decades of targeted observational studies, emerging automatically from a nonlinear ensemble model with no explicit weighting toward metallicity and no built-in model of core accretion.

The higher Gini-importance ranking of surface gravity and radius relative to metallicity warrants careful interpretation. These two features are strongly correlated and jointly encode stellar evolutionary state, distinguishing compact, high-gravity main-sequence dwarfs from expanded, low-gravity subgiants and giants. Their prominence likely reflects two distinct contributions: a real astrophysical signal, since the metallicity-giant planet correlation is best established for main-sequence FGK dwarfs, which dominate the gas giant host sample in the archive, and a statistical contribution from selection effects, since radial velocity surveys have historically focused on precisely this population. 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 the two models highlights how they balance precision against recall. Logistic regression, as a linear model with balanced class weights, tends to favor identifying more gas giant hosts, yielding higher recall but lower precision. The random forest, as an ensemble of nonlinear decision trees, carves out more specific

regions of feature space associated with high gas giant probability, yielding higher precision at the cost of lower recall.

Neither model is universally superior; the appropriate choice depends on scientific context. For telescope scheduling aimed at a comprehensive target list, where missing a true host represents a lost opportunity, logistic regression's higher recall (0.705 versus 0.594) is preferable. For estimating occurrence rates, where false positives would inflate the estimate, the random forest's higher precision (0.684 versus 0.607) is preferable. Logistic regression also offers a distinct interpretability advantage: its coefficients directly quantify each feature's linear contribution to the log-odds of gas giant presence, whereas the random forest requires post-hoc tools such as SHAP to achieve comparable transparency. Used together, the two models provide complementary perspectives: logistic regression confirms and signs the linear signal in the data, while the random forest extracts additional nonlinear signal and quantifies feature contributions through SHAP.

4.3 Limitations and Potential Sources of Bias

Several limitations should be acknowledged. Most fundamentally, the NASA Exoplanet Archive is not a representative, volume-complete, or uniformly observed sample of stars in the solar neighborhood; bright, nearby, slowly rotating FGK dwarfs are dramatically overrepresented, while M dwarfs, which make up roughly 70% of all stars in the galaxy but are faint and difficult to characterize spectroscopically, are substantially underrepresented. Any conclusions about occurrence rates or stellar correlations drawn from this dataset must account for these complex, not fully characterized biases.

A second limitation concerns the target variable definition. Using a 0.8 Jupiter radius threshold as a proxy for gas giant status captures most genuine gas giants but is imperfect: radius measurement uncertainty is not reflected in the binary label, and the threshold does not distinguish among gas giant subclasses (hot, warm, and cold Jupiters, and Saturn analogs), which may have distinct formation histories and host-star property distributions. Future work could treat these subclasses as separate target categories.

A third limitation is the omission of features that could control for detection bias, such as stellar distance, apparent magnitude, and the specific survey or instrument used, all of which relate to detection sensitivity and to the characteristics of the observed host-star population. Without these controls, the models may partly be learning detectability rather than purely physical planet-forming properties. A fourth limitation is that the five-feature set, while theoretically motivated and practically available, is not exhaustive; other candidate predictors include stellar age, the alpha-element abundance ratio [alpha/Fe], stellar rotation rate, and the presence of binary companions, any of which could improve future models.

4.4 Implications and Future Directions

Despite these limitations, this study demonstrates that machine learning models trained on public exoplanet archive data can extract physically meaningful predictive signal and recover known astrophysical relationships in a purely data-driven framework. Practically, the finding that logistic regression can identify 70.5% of gas giant hosts in an unseen test set using only five stellar properties suggests that similar models could help prioritize target selection for future planet-search programs; as the number of spectroscopically characterized stars grows through surveys such as GALAH, APOGEE, and Gaia-ESO, pre-screening with a trained classifier could reduce the observing time needed to reach a given survey completeness.

Scientifically, this study contributes to a growing body of work showing that machine learning can serve as a tool for astrophysical discovery, not merely predictive classification, with SHAP analysis in particular illustrating how post-hoc interpretation can turn an otherwise opaque model into a scientifically informative instrument. Future work could explore more complex architectures, such as gradient-boosted trees or attention-based neural networks capable of revealing higher-order feature interactions, and could compare models trained on different surveys or archive epochs to better characterize how detection biases have shaped the apparent stellar-planetary correlations in the data.

5. Conclusions

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

First, 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, confirming that stellar properties carry genuine, extractable predictive information well beyond simply predicting the majority class.

Second, the two models show a clear, interpretable precision-recall trade-off. Logistic regression's higher recall suits applications where comprehensiveness matters, such as generating target lists for follow-up surveys, while the random forest's higher precision suits applications where reliability of positive predictions is paramount, such as estimating occurrence rates. Model selection in a scientific context should therefore be guided by the downstream application rather than by a single aggregate metric.

Third, feature importance analysis identified surface gravity and radius as the top two predictors by Gini importance, with metallicity ranked third; the relatively even distribution of importances (0.166 to 0.222) confirms that gas giant prediction benefits from a multivariate approach, and that the prominence of surface gravity and radius likely reflects both genuine astrophysical signal and observational selection effects present in the heterogeneous archive.

Fourth, SHAP analysis confirmed that the model correctly learned the direction of the metallicity effect, one of the most robustly established results 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 model and established theory validates that the model captured a physically real signal rather than overfitting to noise or survey artifacts; SHAP values for the remaining features likewise yielded physically interpretable patterns, with high stellar mass and high surface gravity both contributing positively in ways consistent with known stellar-planetary correlations.

Together, these findings show that combining interpretable machine learning methods with publicly available exoplanet archive data can yield both useful predictive tools and scientifically meaningful insight into the stellar conditions that favor gas giant planet formation. Extending this approach to larger feature sets, more sophisticated model architectures, and survey-corrected samples holds promise for refining 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