Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Intelligent Identification of Astronomical Objects: a physics-motivated benchmark of supervised classification

Mohit Trivedi
Indian Institute of Technology Roorkee
Mayank Goswami
mayankfph@iitr.ac.in

Indian Institute of Technology Roorkee

Method Article

Keywords: astronomical source classification; SDSS; photometry; spectroscopic redshift; random forest; gradient boosting.

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11648697/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Intelligent Identification of Astronomical Objects: a physics-motivated benchmark of supervised classification

Mohit Trivedi[1] and Mayank Goswami,[*1,2]

[1]Department of Physics, IIT Roorkee
[2]MF School of AI and DS, IIT Roorkee
mayank.goswami@ph.iitr.ac.in

November 6, 2026

Abstract

We examine supervised classification of stars, galaxies, and quasars using a 100,000-object catalogue distributed as the Stellar Classification Dataset (SDSS17). The final input vector contains five optical magnitudes and spectroscopic redshift. We develop the mathematical connection between broadband measurements, thermal radiation, colour indices, and wavelength shifts, then formulate the classification problem through posterior probabilities, empirical risk, decision-tree impurity, and ensemble prediction. The saved project notebook compares logistic regression, a decision tree, random forest, and XGBoost on one stratified 80/20 train–test split. Random forest has the highest recorded accuracy, 0.97970, corresponding to 406 errors among 20,000 test objects, whereas XGBoost has the highest macro one-versus-rest ROC-AUC, 0.995771. An independent exploratory forest assigns 0.638909 of its normalised impurity importance to redshift. We derive the identity between support-weighted recall and accuracy, establish the information-preserving role of a colour-plus-anchor transformation, and explain why an additional measurement can improve Bayes-optimal prediction without guaranteeing improvement for a finite fitted model. These are standard analytical results used to interpret the experiment, rather than claims of a new learning algorithm. The evidence supports an internal catalogue benchmark, not demonstrated photometry-only inference or transfer to another survey. Results come from recorded notebook outputs; the original CSV, fitted estimators, and object-level predictions were unavailable for independent rerunning and further empirical validation.

Keywords: astronomical source classification; SDSS; photometry; spectroscopic redshift; random forest; gradient boosting.

1 Introduction

The identification of an astronomical source as a star, galaxy, or quasar is a basic step in constructing scientific samples. A stellar photosphere, an unresolved galaxy, and a luminous active galactic nucleus can produce similar angular appearances even though their emission mechanisms and physical scales differ. Classification must therefore use the measured distribution of light and the observing context, rather than appearance alone.

Large surveys make numerical classification useful because a reproducible mapping from measurements to labels can screen many sources. SDSS provides imaging, spectroscopy, and catalogue parameters [1]. The present project uses a tabular derivative of these observations distributed on Kaggle [2]. Its three labels are operational catalogue categories. In particular, a quasar is an active nucleus in a host galaxy, although the classification target separates QSO from GALAXY.

Previous work already demonstrates source classification at much larger scale. Clarke et al. trained a random forest with SDSS and WISE photometry from 3.1 million spectroscopically labelled sources and classified 111 million SDSS sources without existing spectra [3]. Our task differs because spectroscopic redshift remains an input. High performance in this setting does not establish the ability to classify an unobserved source using imaging alone.

The objectives are to compare four established classifiers under a common split, explain the physical relevance of their inputs, and identify the statistical limits of the recorded evidence. The term *physics-motivated* refers to feature interpretation. The training procedure does not incorporate a radiative-transfer equation, conservation constraint, or physics-based penalty into its loss.

The paper combines three types of material: the experiment documented by the notebook, arithmetic derived from its reported quantities, and analytical explanations of standard physical and statistical models. Proposed evaluations are identified explicitly and are not reported as completed results.

2 Problem formulation and dataset

2.1 Inputs, labels, and the target risk

Write the labelled sample as

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}, \qquad N = 100000, \tag{1}$$

where $y_i \in \mathcal{Y} = \{0, 1, 2\}$ denotes GALAXY, QSO, and STAR, respectively. The six-dimensional feature vector is

$$\boldsymbol{x} = (m_u, m_g, m_r, m_i, m_z, z_{\text{spec}})^T \in \mathbb{R}^6. \tag{2}$$

Here m_z is a band magnitude, whereas z_{spec} is dimensionless spectroscopic redshift. The band label i identifies the SDSS i filter; the sample index on $\boldsymbol{x}_i$ identifies an individual object.

A probabilistic classifier produces

$$\widehat{\boldsymbol{p}}(\boldsymbol{x}) \in \Delta^2, \quad \Delta^2 = \{\boldsymbol{p} \in \mathbb{R}^3 : p_k \geq 0,\ \textstyle\sum_k p_k = 1\}, \tag{3}$$

and predicts $\widehat{y}(\boldsymbol{x}) = \arg\max_k \widehat{p}_k(\boldsymbol{x})$. For a target population distribution P, the classification risk is

$$R_P(f) = \mathbb{E}_{(X,Y)\sim P}\big[\mathbf{1}\{f(X) \neq Y\}\big]. \tag{4}$$

The test error estimates this risk only to the extent that the test sample represents P and the evaluation remains independent of model development.

Bayes' rule gives the ideal posterior

$$p_k(\boldsymbol{x}) = \frac{\pi_k\, p(\boldsymbol{x} \mid Y = k)}{\sum_{\ell=0}^{2} \pi_\ell\, p(\boldsymbol{x} \mid Y = \ell)}, \tag{5}$$

where $\pi_k = P(Y = k)$. Consequently, class prevalence affects posterior probabilities even when the class-conditional measurement distributions remain fixed. Neither the density functions nor the true population priors are estimated explicitly by the notebook.

2.2 Recorded composition and provenance

The notebook reports 100,000 rows and 18 columns, including one target column. Table 1 gives the saved class counts. Every column has 100,000 non-null entries in the recorded data summary. This is not an audit of measurement validity, duplicates, sentinel values, or selection effects. The available evidence does not establish the exact SDSS magnitude estimator, extinction treatment, or quality-flag selection used by the redistributed CSV. Those definitions should be verified before interpreting small photometric differences physically.

For the constant predictor $f_{\text{maj}}(\boldsymbol{x}) = \text{GALAXY}$,

$$\widehat{A}_{\text{maj}} = \frac{59445}{100000} = 0.59445. \tag{6}$$

This is a full-sample reference baseline. Stratification makes the held-out class fractions approximately the same, but their realised integer counts are not listed in the saved output.

Table 1. Class counts recorded in notebook cell 16. Fractions use the full sample, not a separately reported test-set class table.

Class	Count	Fraction
GALAXY	59,445	0.59445
STAR	21,594	0.21594
QSO	18,961	0.18961

2.3 Column reduction

The initial cleaning removes `obj_ID`, `spec_obj_ID`, `run_ID`, `rerun_ID`, `cam_col`, `field_ID`, and `fiber_ID`. Eleven columns then remain: ten numerical predictors and the target. An exploratory forest uses all ten predictors, including sky coordinates, plate, and modified Julian date (MJD). Final classification retains only Eq. (2); it excludes those four coordinate and observing variables. Identifier removal reduces arbitrary catalogue associations but does not by itself remove all survey dependence.

3 Physical basis of the measurements

3.1 Thermal radiation and the observed spectrum

For an ideal blackbody, spectral radiance per unit wavelength is

$$B_\lambda(T) = \frac{2hc^2}{\lambda^5}\left[\exp\!\left(\frac{hc}{\lambda k_{\text{B}} T}\right) - 1\right]^{-1}. \tag{7}$$

Here h is Planck's constant, c the speed of light, and k_{B} Boltzmann's constant. Differentiating with respect to λ and writing $x = hc/(\lambda k_{\text{B}} T)$ gives the peak condition

$$5(1 - e^{-x}) = x, \qquad x \simeq 4.9651. \tag{8}$$

Thus

$$\lambda_{\text{peak}} T = \frac{hc}{4.9651 k_{\text{B}}} \simeq 2.898 \times 10^{-3}\ \text{m K}. \tag{9}$$

A 5800 K blackbody therefore peaks near 500 nm. Figure 1 illustrates this relationship, with each curve normalised to its own maximum. The curves are theoretical examples, not measured spectra from this dataset.

For an unresolved thermal sphere without intervening attenuation, a schematic relation is $f_\lambda \simeq \pi B_\lambda(T)(R/d)^2$, with radius R and distance d. A galaxy instead combines many stellar populations and gas emission. An active nucleus includes accretion-powered radiation. Dust, spectral lines, and redshift further modify all these idealised pictures. The classifier need not recover a unique physical model to exploit the resulting statistical differences in the measurements.

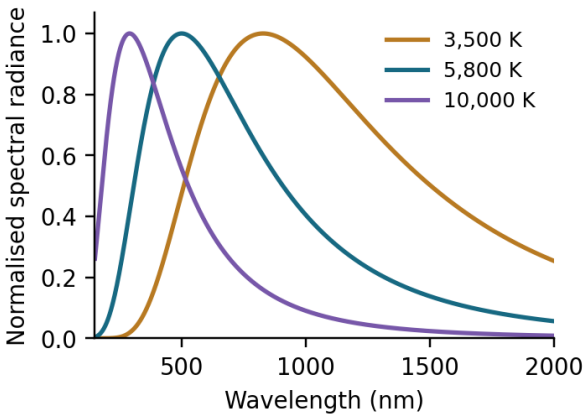


Figure 1. Ideal blackbody spectra computed from Eq. (7). Independent normalisation compares spectral shape rather than absolute luminosity. A stellar atmosphere, a galaxy, and a quasar generally require more detailed emission models.

3.2 Band integration and the magnitude transformation

Let $S_b(\lambda)$ denote a dimensionless photon-detection response in band b. For effective collecting area $A_{\rm tel}$ and exposure time t, an ideal expected source photon count is

$$C_b = \frac{A_{\rm tel} t}{hc} \int_0^\infty \lambda f_\lambda(\lambda) S_b(\lambda)\, d\lambda. \tag{10}$$

This forward model explains why different filters retain different information about the same spectrum. It is not a reconstruction of the SDSS reduction pipeline or an operation performed by the notebook.

For positive calibrated band flux F_b, the conventional magnitude is

$$m_b = -2.5 \log_{10}\left(\frac{F_b}{F_{0,b}}\right). \tag{11}$$

Within one band, $m_1 - m_2 = -2.5 \log_{10}(F_1/F_2)$; hence a five-magnitude difference corresponds to a factor of 100 in flux. The negative sign makes brighter objects numerically smaller in magnitude.

SDSS documents an inverse-hyperbolic-sine magnitude convention [4, 5]. Defining dimensionless flux $f_b = F_b/F_{0,b}$ and softening parameter $\beta_b > 0$, its form is

$$m_b^{\rm asinh} = -\frac{2.5}{\ln 10}\left[\operatorname{asinh}\left(\frac{f_b}{2\beta_b}\right) + \ln \beta_b\right]. \tag{12}$$

For $f_b \gg \beta_b$, $\operatorname{asinh}(f_b/2\beta_b) \simeq \ln(f_b/\beta_b)$, recovering Eq. (11). At weak or negative measured flux, Eq. (12) remains finite. The exact magnitude definitions in the redistributed data require a provenance check; a column name alone does not identify an aperture estimator.

3.3 Colours, distance dependence, and uncertainty

For two bands, the logarithmic colour is

$$c_{ab} = m_a - m_b = -2.5 \log_{10}(F_a/F_b) + 2.5 \log_{10}(F_{0,a}/F_{0,b}). \tag{13}$$

When both fluxes share the same multiplicative distance factor, that factor cancels in their ratio. Colour can therefore describe spectral shape while an anchor magnitude retains apparent-brightness information. Extinction, redshift, population mixtures, and measurement noise still affect colour. If A_b is extinction in magnitudes, then

$$m_b^{\rm obs} = m_b^{\rm intrinsic} + A_b, \quad c_{ab}^{\rm obs} = c_{ab}^{\rm intrinsic} + A_a - A_b. \tag{14}$$

No extinction correction is documented in the final notebook workflow.

Write $\boldsymbol{m} = (m_u, m_g, m_r, m_i, m_z)^T$ and define the proposed representation

$$\boldsymbol{q} = (m_u - m_g, m_g - m_r, m_r - m_i, m_i - m_z, m_r)^T. \tag{15}$$

This representation is not used in the reported models. Its information content is established in Appendix A. For a linear transformation $\boldsymbol{q} = A\boldsymbol{m}$, covariance propagates as

$$\Sigma_q = A \Sigma_m A^T, \quad \operatorname{Var}(c_{ab}) = \sigma_a^2 + \sigma_b^2 - 2\operatorname{Cov}(m_a, m_b). \tag{16}$$

Thus colours generally have correlated errors, even when a simplified model treats the original magnitudes as independent. The dataset summary supplies no per-object covariance matrices, so Eq. (16) is explanatory rather than an uncertainty calculation carried out here.

3.4 Redshift and shared spectroscopic information

Spectroscopic redshift is

$$z_{\rm spec} = \frac{\lambda_{\rm obs} - \lambda_{\rm rest}}{\lambda_{\rm rest}}, \quad \lambda_{\rm obs} = (1 + z_{\rm spec})\lambda_{\rm rest}. \tag{17}$$

For cosmological sources this shifts spectral features through the filter responses. In the standard luminosity-distance convention, neglecting attenuation and lensing, spectral luminosity per unit wavelength obeys [6]

$$f_{\lambda,\rm obs}(\lambda) = \frac{L_{\lambda,\rm rest}(\lambda/(1+z))}{4\pi D_L^2 (1+z)}. \tag{18}$$

Integrating over observed wavelength recovers the bolometric relation $F = L/(4\pi D_L^2)$. Equation (18) describes an ideal cosmological source and is not a distance estimator used by the classifier. The low-redshift approximation $v \simeq cz$ must not be extended to large cosmological redshift.

Many sampled stars have redshifts near zero; sampled extragalactic objects extend to larger positive

values. These trends reflect both source populations and survey selection. SDSS estimates class and redshift through spectral template comparisons [7]. Their common origin introduces shared information: a predictor that already contains spectroscopic redshift addresses a different task from predicting labels using photometry alone. This is not evidence of a coding leak, but it changes the interpretation and the intended deployment population.

4 Statistical models

4.1 Empirical risk and multinomial logistic regression

For model parameters θ, a standard regularised learning objective is

$$\widehat{\theta} = \arg\min_{\theta} \left[\frac{1}{N_{\mathrm{tr}}} \sum_{i \in \mathcal{D}_{\mathrm{tr}}} \ell(y_i, \boldsymbol{p}_\theta(\boldsymbol{x}_i)) + \lambda \Omega(\theta) \right]. \quad (19)$$

This is a mathematical framework for the compared methods, not a claim that they optimise identical objectives. Trees use recursive impurity reduction, whereas probabilistic linear and boosting methods use differentiable losses.

In the multinomial linear model, let

$$a_k(\boldsymbol{x}) = \boldsymbol{w}_k^T \boldsymbol{x} + b_k, \qquad p_k(\boldsymbol{x}) = \frac{e^{a_k(\boldsymbol{x})}}{\sum_{\ell=0}^{2} e^{a_\ell(\boldsymbol{x})}}. \quad (20)$$

With $t_{ik} = \mathbf{1}\{y_i = k\}$, cross-entropy is

$$\ell_i = -\sum_{k=0}^{2} t_{ik} \log p_k(\boldsymbol{x}_i). \quad (21)$$

Its score gradient and Hessian are

$$\frac{\partial \ell_i}{\partial a_{ik}} = p_{ik} - t_{ik}, \qquad H_i = \mathrm{diag}(\boldsymbol{p}_i) - \boldsymbol{p}_i \boldsymbol{p}_i^T. \quad (22)$$

For any $\boldsymbol{v}$, $\boldsymbol{v}^T H_i \boldsymbol{v}$ is the variance of a discrete variable taking value v_k with probability p_{ik}, and is nonnegative. This explains the convex score-space loss. Adding an identical constant to all scores leaves the probabilities unchanged, accounting for an identifiable parameter redundancy unless a reference convention or penalty resolves it.

The notebook calls `LogisticRegression(max_iter=5000, random_state=42)`. Equation (20) gives the standard multiclass formulation [8, 9]; the saved environment does not pin solver, library version, or regularisation settings beyond that call. The final feature matrix is not standardised, although standardisation appears in exploratory PCA. Because feature scale interacts with regularisation and optimisation, this limits conclusions about the strongest attainable linear baseline.

4.2 Decision-tree partitions

A classification tree partitions feature space into leaves $\{R_\ell\}_{\ell=1}^{L}$. At node t, let n_t be its training count and p_{tk} its class fraction. Gini impurity is

$$G(t) = 1 - \sum_{k=0}^{2} p_{tk}^2. \quad (23)$$

A split using coordinate j and threshold s creates children $t_L = \{\boldsymbol{x} : x_j \le s\}$ and $t_R = \{\boldsymbol{x} : x_j > s\}$ within the parent region. Its local reduction is

$$\Delta G(t; j, s) = G(t) - \frac{n_L}{n_t} G(t_L) - \frac{n_R}{n_t} G(t_R). \quad (24)$$

Recursively choosing such splits gives piecewise-constant probabilities

$$\widehat{p}_k(\boldsymbol{x}) = \sum_{\ell=1}^{L} \mathbf{1}\{\boldsymbol{x} \in R_\ell\} \frac{n_{\ell k}}{n_\ell}. \quad (25)$$

This flexibility captures nonlinear boundaries but can produce high variance when leaves contain few observations. The recorded decision-tree invocation sets a random seed but no explicit depth or minimum-leaf constraint.

4.3 Random-forest averaging

A random forest combines trees fitted with bootstrap variation and random feature selection [10, 11]. With B trees,

$$\widehat{p}_k^{\mathrm{RF}}(\boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \widehat{p}_{bk}(\boldsymbol{x}), \quad \widehat{y}^{\mathrm{RF}} = \arg\max_k \widehat{p}_k^{\mathrm{RF}}. \quad (26)$$

For scalar tree outputs with common variance σ^2 and common pairwise correlation ρ, averaging gives

$$\mathrm{Var}\left(\frac{1}{B} \sum_{b=1}^{B} T_b \right) = \sigma^2 \left[\rho + \frac{1 - \rho}{B} \right]. \quad (27)$$

This exchangeable-output calculation, proved in Appendix B, explains why less-correlated trees can improve stability. It is not an estimated variance formula for the saved test score. The final forest uses $B = 300$, `random_state=42`, and `n_jobs=-1`.

4.4 Regularised gradient boosting

XGBoost constructs an additive ensemble [12]. For class scores $\boldsymbol{F}^{(t)}$, write

$$\boldsymbol{F}^{(t)}(\boldsymbol{x}) = \boldsymbol{F}^{(t-1)}(\boldsymbol{x}) + \eta \boldsymbol{f}_t(\boldsymbol{x}), \quad (28)$$

where a boosting round can contain a class-specific tree for each class. The softmax of the accumulated scores yields probabilities. A regularised objective has the form

$$\mathcal{L} = \sum_i \ell(y_i, \boldsymbol{p}(\boldsymbol{x}_i)) + \sum_{t,k} \left(\gamma L_{tk} + \frac{\lambda}{2} \sum_{\ell=1}^{L_{tk}} w_{tk\ell}^2 \right). \quad (29)$$

Table 2. Explicit settings in notebook cell 30. Other parameters depend on the unspecified installed library versions.

Model	Recorded settings
Logistic regression	Maximum iterations 5000; seed 42
Decision tree	Seed 42
Random forest	300 trees; seed 42; all CPU cores
XGBoost	300 estimators; depth 6; learning rate 0.1; seed 42; evaluation metric `mlogloss`

For a scalar score coordinate and a diagonal second-order approximation, the terms associated with a leaf reduce to

$$\widetilde{\mathcal{L}}_\ell(w) = G_\ell w + \tfrac{1}{2}(H_\ell + \lambda)w^2 + \gamma, \tag{30}$$

where $G_\ell = \sum_{i \in R_\ell} g_i$ and $H_\ell = \sum_{i \in R_\ell} h_i$. Differentiation gives

$$w_\ell^* = -\frac{G_\ell}{H_\ell + \lambda}, \quad \widetilde{\mathcal{L}}_\ell(w_\ell^*) = -\frac{G_\ell^2}{2(H_\ell + \lambda)} + \gamma. \tag{31}$$

Thus loss curvature and regularisation determine candidate leaf updates. These equations explain a standard scalar building block; the full softmax Hessian in Eq. (22) includes cross-class terms, so this expression is not a complete reconstruction of the library's multiclass solver. The notebook uses 300 estimators, maximum depth 6, learning rate $\eta = 0.1$, `eval_metric='mlogloss'`, and seed 42. Setting an evaluation metric does not itself implement early stopping.

5 Experimental protocol and metrics

5.1 Recorded workflow

The notebook encodes the target labels, selects Eq. (2), and uses a stratified split with

$$N_{\text{tr}} = 80000, \quad N_{\text{te}} = 20000, \quad \texttt{test_size} = 0.2, \quad \texttt{random_state} = 42. \tag{32}$$

All four classifiers use that same split. Stratification approximately preserves the sample class fractions; it does not create an independent spatial or survey-level evaluation. Table 2 lists only explicit settings and does not infer undocumented tuning.

Exploratory standardisation and two-component PCA precede final modelling. With column means μ_j and standard deviations s_j, the exploratory matrix has entries $Z_{ij} = (X_{ij} - \mu_j)/s_j$. Its covariance matrix is

$$S = \frac{Z^T Z}{N - 1}, \quad S\boldsymbol{v}_r = \lambda_r \boldsymbol{v}_r, \quad U = Z[\boldsymbol{v}_1\ \boldsymbol{v}_2]. \tag{33}$$

PCA selects directions of variance rather than directly optimising class separation [13]. The resulting coordinates U are not inputs to the reported classifiers. Standardisation, PCA, and the exploratory importance ranking use the full sample. Their descriptive value should be distinguished from an evaluation with all development choices isolated inside training folds.

5.2 Confusion matrix and support-weighted scores

For test predictions, define

$$C_{ab} = \sum_{i \in \mathcal{D}_{\text{te}}} \mathbf{1}\{y_i = a, \widehat{y}_i = b\}. \tag{34}$$

Rows correspond to true labels and columns to predictions. Let $n_k = \sum_b C_{kb}$ and $\widehat{n}_k = \sum_a C_{ak}$. Then

$$A = \frac{\sum_k C_{kk}}{N_{\text{te}}}, \tag{35}$$

$$P_k = \frac{C_{kk}}{\widehat{n}_k}, \qquad R_k = \frac{C_{kk}}{n_k}, \tag{36}$$

$$F_{1,k} = \frac{2P_k R_k}{P_k + R_k}. \tag{37}$$

For a metric M_k, support weighting gives

$$M_w = \sum_{k=0}^{2} \frac{n_k}{N_{\text{te}}} M_k. \tag{38}$$

The harmonic mean is applied before weighting in $F_{1,w}$; generally it is not equal to $2P_w R_w/(P_w + R_w)$. An aggregate weighted score can conceal poor performance on a smaller class.

Proposition 1 (Weighted-recall identity)**.** *For single-label multiclass classification, support-weighted recall equals accuracy whenever the nonempty classes use their observed supports.*

Proof. Substitute Eq. (36) into Eq. (38):

$$R_w = \sum_k \frac{n_k}{N_{\text{te}}} \frac{C_{kk}}{n_k} = \frac{\sum_k C_{kk}}{N_{\text{te}}} = A. \tag{39}$$

□

Thus the duplicate accuracy and recall values in the results table are expected, rather than independent confirmation of performance.

5.3 One-versus-rest ROC-AUC

For class k, let X_k^+ and X_k^- be independent draws from its positive and rest-of-classes populations, and let s_k be the class score. Its population AUC can be written

$$\begin{aligned} \text{AUC}_k = {} & \mathbb{P}\{s_k(X_k^+) > s_k(X_k^-)\} \\ & + \tfrac{1}{2}\mathbb{P}\{s_k(X_k^+) = s_k(X_k^-)\}. \end{aligned} \tag{40}$$

The notebook calls `roc_auc_score` with `multi_class='ovr'`, without setting `average`; the documented default is macro averaging [14]:

$$\text{AUC}_{\text{macro,OvR}} = \frac{1}{3}\sum_{k=0}^{2} \text{AUC}_k. \tag{41}$$

Each class has equal weight in this average, although the negative population inside each one-versus-rest task still reflects the class mixture. This averaging convention differs from the weighted precision, recall, and F1.

5.4 Calibration is a separate property

Ideal classwise calibration requires

$$\mathbb{P}(Y = k \mid \widehat{p}_k(X) = q) = q, \tag{42}$$

interpreted through suitable conditional distributions or score bins. Ranking performance does not imply this property [15]. For example, in a binary task, replacing p with

$$q(p) = \frac{p^a}{p^a + (1-p)^a}, \qquad a > 1, \tag{43}$$

preserves score ordering but increases confidence away from 1/2. The binary AUC is unchanged even when the new probabilities lose calibration. This is a mathematical illustration, not a transformation applied in the project.

Two proposed probability evaluations are log loss and the multiclass Brier score:

$$\mathrm{NLL} = -\frac{1}{N_{\mathrm{te}}} \sum_i \log \widehat{p}_{i,y_i}, \tag{44}$$

$$\mathrm{BS} = \frac{1}{N_{\mathrm{te}}} \sum_i \sum_{k=0}^{2} (\widehat{p}_{ik} - t_{ik})^2. \tag{45}$$

The Brier convention here sums over classes without an additional factor of 1/3. Neither score nor a reliability diagram is available in the saved results, so no calibration ranking is claimed.

6 Recorded results and derived quantities

6.1 Accuracy, F1, and error counts

Table 3 reports all available model scores, with the accuracy comparison visualised in Fig. 2. Random forest has the highest recorded accuracy and weighted F1. The linear baseline reaches 95.475% accuracy, and a single decision tree reaches 96.695%. The two ensembles reach 97.545% and 97.970%, respectively. The results show an ordering for these configurations and this split; they do not establish an ordering between optimally tuned model families.

For any displayed accuracy $\widehat{A}$, the implied error count is

$$e = N_{\mathrm{te}}(1 - \widehat{A}). \tag{46}$$

Random forest therefore has $20000(1-0.97970) = 406$ errors. Its difference from XGBoost is

$$\Delta A = 0.97970 - 0.97545 = 0.00425, \tag{47}$$

or 0.425 percentage points and 85 fewer incorrect predictions. Relative error reduction, using XGBoost as the reference, is

$$\frac{491 - 406}{491} = 0.173116\ldots \simeq 17.31\%. \tag{48}$$

Against logistic regression, the analogous reduction is $(905 - 406)/905 \simeq 55.14\%$. These are arithmetic comparisons of the saved point estimates, not estimates of repeatability or deployment benefit.

6.2 A conditional sampling calculation

For scale, suppose a fixed classifier is evaluated through n independent Bernoulli correctness indicators with common success probability p. Inverting the approximate score inequality

$$\frac{n(\widehat{A} - p)^2}{p(1-p)} \le q_{.975}^2 \tag{49}$$

gives the Wilson interval [16], with centre c_W and half-width h_W:

$$c_W = \frac{\widehat{A} + q^2/(2n)}{1 + q^2/n}, \tag{50}$$

$$h_W = \frac{q}{1 + q^2/n} \sqrt{\frac{\widehat{A}(1-\widehat{A})}{n} + \frac{q^2}{4n^2}}. \tag{51}$$

For $q = 1.959964$, $n = 20000$, and $\widehat{A} = 0.97970$, this produces approximately $[0.97765, 0.98156]$, or $[97.765\%, 98.156\%]$. This newly calculated interval is only an illustrative conditional binomial reference. The actual split is stratified, survey observations may be dependent, feature exploration used the full sample, and the best model was identified using the same test comparison. The calculation does not account for these effects or uncertainty from retraining, and is not presented as a validated deployment confidence interval. Comparing separate intervals would also ignore the pairing of the model predictions.

6.3 Ranking comparison

XGBoost has the largest recorded macro OvR ROC-AUC. The difference from random forest is

$$0.995771 - 0.995412 = 0.000359. \tag{52}$$

This is compatible with XGBoost providing better aggregate ranking on this split while random forest makes fewer errors at the argmax decision rule. The notebook's informal statement that XGBoost produces better probability estimates is stronger than the reported evidence supports: calibration and proper probability scores were not measured.

6.4 Exploratory impurity importance

For a tree fitted on N_{fit} observations, the unnormalised feature importance can be expressed as

$$I_j = \sum_{t:\, v(t)=j} \frac{n_t}{N_{\mathrm{fit}}} \Delta G(t), \qquad \widetilde{I}_j = \frac{I_j}{\sum_r I_r}, \tag{53}$$

Table 3. Recorded test metrics from notebook cell 30. Precision, recall, and F1 are support-weighted. ROC-AUC is macro one-versus-rest. Error counts are derived from the displayed accuracy and $N_{\text{te}} = 20000$; all other entries reproduce the saved output.

Model	Accuracy	Precision	Recall	F1	ROC-AUC	Errors
Random forest	0.97970	0.979611	0.97970	0.979563	0.995412	406
XGBoost	0.97545	0.975342	0.97545	0.975335	0.995771	491
Decision tree	0.96695	0.966944	0.96695	0.966947	0.970646	661
Logistic regression	0.95475	0.954913	0.95475	0.954369	0.987663	905

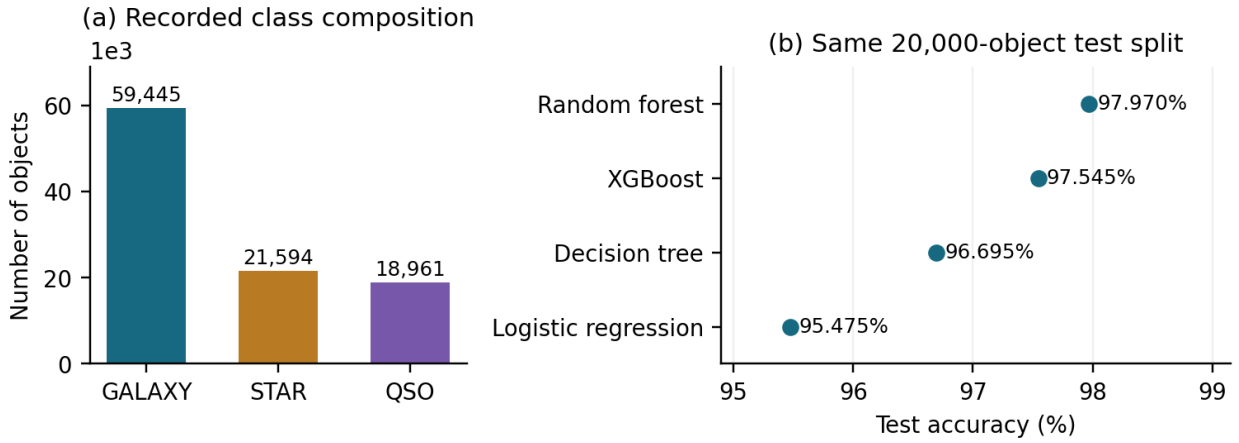


Figure 2. Recorded dataset composition and model accuracy. Panel (a) uses the full-sample class counts. Panel (b) shows point estimates from one stratified test split. Its restricted horizontal range is a dot-plot display, not a bar-length comparison or confidence interval.

Table 4. Exploratory feature importances reproduced from cell 26.

Feature	Importance	Feature	Importance
Redshift	0.638909	Plate	0.030996
m_z	0.084699	m_r	0.029321
m_u	0.059731	MJD	0.028284
m_g	0.056027	Declination	0.009467
m_i	0.053781	Right ascension	0.008786

where $v(t)$ is the splitting feature. A forest aggregates contributions across trees with appropriate normalisation. The notebook's exploratory 200-tree forest, fitted before the split to all ten remaining numerical predictors, gives the values in Fig. 3 and Table 4. The values sum to 1.000001 because of displayed rounding.

The large redshift contribution is consistent with the class-dependent distributions described by the notebook. Nevertheless, impurity importance depends on available splits and correlated predictors. It is neither a causal effect nor the accuracy loss that would result from removing a feature. The final six-feature model's own importance ranking is not reported.

7 Interpretation and limitations

7.1 Why extra physical information can help

Let M denote the magnitude vector and Z spectroscopic redshift. A useful population-level result separates the value of information from finite-sample algorithm behaviour.

Proposition 2 (Bayes risk with an additional measurement)**.** *If arbitrary measurable decision rules are available under a fixed joint distribution, the optimal zero–one risk using (M, Z) cannot exceed that using M alone:*

$$R^*(M, Z) \le R^*(M). \tag{54}$$

Proof. Any rule depending only on M is also a permissible rule on (M, Z) that ignores Z. Minimisation over the larger function set cannot yield a greater minimum. Equivalently, $R^*(M) = 1 - \mathbb{E}_M \max_k P(Y = k \mid M)$ and conditional Jensen's inequality for the convex maximum gives the result. □

Under optimal log loss, the improvement is the conditional mutual information

$$H(Y \mid M) - H(Y \mid M, Z) = I(Y; Z \mid M) \ge 0. \tag{55}$$

No mutual-information value is estimated here. These standard results describe ideal prediction when Z is

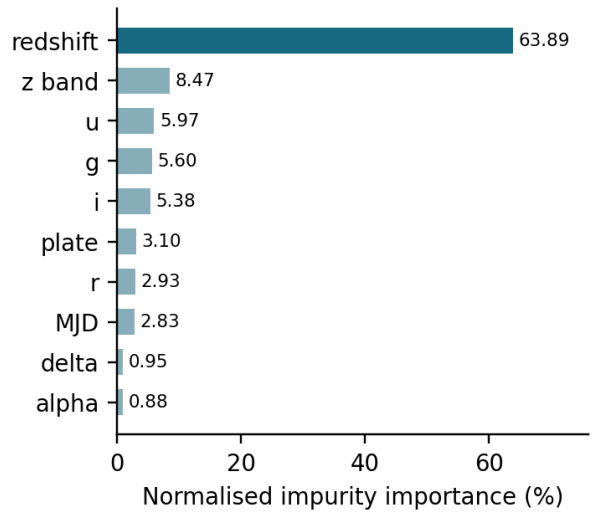


Figure 3. Saved impurity-importance ranking from the exploratory forest in cell 26. This model uses 200 trees and ten features on the full sample, whereas the final classifier uses 300 trees and six features. The plot is not a held-out permutation-importance result.

available; they do not guarantee that a particular finite forest improves, explain an impurity score causally, or justify use where Z is missing. An empirical ablation is still necessary.

7.2 Survey selection and distribution shift

Let $S = 1$ indicate selection into the available labelled catalogue. The data represent $P(X, Y \mid S = 1)$, which can differ from a target distribution $P_{\text{target}}(X, Y)$. A random internal split assesses that selected population more directly than it assesses a new survey. For any bounded loss $0 \le \ell \le 1$ and total variation defined as $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$,

$$|R_P(f) - R_Q(f)| \le \mathrm{TV}(P, Q). \tag{56}$$

The inequality follows from the layer-cake representation of a bounded non-negative function. No total-variation value is known for these data, so the expression establishes a dependence, not a numerical guarantee.

Even under the restrictive covariate-shift assumption $P_{\text{target}}(Y \mid X) = P_{\text{train}}(Y \mid X)$, a target risk would involve the density ratio

$$R_{\text{target}}(f) = \mathbb{E}_{\text{train}}\left[\frac{p_{\text{target}}(X)}{p_{\text{train}}(X)}\ell(Y, f(X))\right], \tag{57}$$

provided the required support overlap holds. Neither this assumption nor the ratio is verified in the project. Selection can also alter the conditional label distribution, making the simple weighting formula inapplicable.

7.3 What the available outputs cannot identify

Aggregate metrics do not determine the confusion matrix, object-level probabilities, calibration, runtime, or per-class failure modes. The result table supplies no held-out photometry-only or redshift-only score. It also does not report cross-validation, a separate validation set, a tuning search, or external testing.

Two classifiers evaluated on the same objects produce paired outcomes. Let b count objects misclassified only by model 1 and c objects misclassified only by model 2. A commonly used continuity-corrected paired comparison of the McNemar type [17] has statistic

$$\chi^2_{\text{paired}} = \frac{(|b - c| - 1)^2}{b + c}, \quad b + c > 0. \tag{58}$$

The marginal error difference supplies $b - c$, but not $b + c$. Object-level predictions are therefore needed to evaluate the observed paired statistic directly. We report no paired-test p-value or empirical significance claim.

Finally, the input CSV and fitted estimators are absent from the supplied project folder. The numerical evidence is the notebook's saved outputs. An exact independent rerun requires the dataset version, software environment, and training artifacts. The mathematics and newly generated plots do not substitute for those missing empirical checks.

8 Proposed validation programme

A next experiment should compare input sets

$$\begin{aligned} \mathcal{X}_{\text{phot}} &= (m_u, m_g, m_r, m_i, m_z), \\ \mathcal{X}_z &= (z_{\text{spec}}), \\ \mathcal{X}_{\text{both}} &= (m_u, m_g, m_r, m_i, m_z, z_{\text{spec}}). \end{aligned} \tag{59}$$

All branches should share evaluation folds so that differences reflect features rather than different test populations. A colour-plus-anchor version of the photometry should be assessed alongside the original magnitudes. Its invertibility means that any improvement would reflect model geometry, regularisation, or finite-sample optimisation, rather than newly acquired information.

Preprocessing, feature selection, and hyperparameter search should remain inside training folds. If $\theta^{(-v)}$ is selected without fold v, an outer-fold risk estimate is

$$\widehat{R}_{\text{outer}} = \frac{1}{N}\sum_v \sum_{i \in \mathcal{D}_v} \mathbf{1}\{f_{\theta^{(-v)}}(\boldsymbol{x}_i) \ne y_i\}. \tag{60}$$

Spatial, plate, or survey grouping should be considered where observations share acquisition conditions; duplicates and repeated sources must remain in the same group. A final independent survey sample should be reserved for the claimed deployment setting.

Per-class precision and recall, macro F1, balanced accuracy, and the confusion matrix should accompany weighted metrics. Calibration curves, NLL, and Brier score should assess probability quality. Repeated or grouped resampling should quantify variability at the relevant sampling unit. A scaled logistic regression pipeline would strengthen the linear baseline.

Where uncertain sources require review, a proposed selective classifier is

$$\widehat{y}_\tau(\boldsymbol{x}) = \begin{cases} \arg\max_k \widehat{p}_k(\boldsymbol{x}), & \max_k \widehat{p}_k(\boldsymbol{x}) \geq \tau, \\ \text{review}, & \text{otherwise.} \end{cases} \tag{61}$$

Its coverage is $P(\max_k \widehat{p}_k(X) \geq \tau)$, and its selective risk conditions Eq. (4) on that accepted subset. A confidence threshold requires calibration and a measured coverage–risk trade-off. No thresholded deployment system is implemented in the recorded study.

9 Conclusion

The saved experiment demonstrates three-class classification of a selected SDSS-derived catalogue using five optical magnitudes and spectroscopic redshift. On one stratified 20,000-object test set, random forest has the highest recorded accuracy (97.970%) and weighted F1 (0.979563), while XGBoost has the highest macro OvR ROC-AUC (0.995771). The exploratory importance ranking identifies redshift as the dominant splitting variable in a separate full-sample model.

The analytical development explains how spectra produce photometric inputs, how linear and tree models use them, and why the evaluation metrics need different interpretations. The study supports an internal benchmark with physically meaningful measurements. It does not establish photometry-only classification, calibrated probability estimates, causal feature effects, or transfer to an independent survey. Those questions require the validation programme described above and access to the original data and predictions.

Data, code, and reproducibility statement

The empirical source is the supplied `Stellar_Classification.ipynb`. Dataset dimensions and counts appear in cells 6, 7, and 16; column removal in cell 10; exploratory importance in cell 26; and final features, split, and model outputs in cells 28–30. The dataset URL is listed in Ref. [2]. This manuscript does not represent a new training run. The accompanying `make_figures.py` generates figures from the displayed counts, metric values, importance values, and Planck's law. A machine-readable arithmetic summary is included as `derived_values.json`. Software versions of the original model run, dataset checksum, fitted estimators, and test predictions remain unrecorded in the available materials.

A Invertibility of colours with an anchor

For Eq. (15), the transformation matrix is

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}. \tag{62}$$

Writing $\boldsymbol{q} = (c_{ug}, c_{gr}, c_{ri}, c_{iz}, a)^T$ with anchor $a = m_r$, the inverse is explicit:

$$\begin{aligned} m_r &= a, & m_g &= a + c_{gr}, \\ m_u &= a + c_{gr} + c_{ug}, & m_i &= a - c_{ri}, \\ m_z &= a - c_{ri} - c_{iz}. \end{aligned} \tag{63}$$

Thus A is nonsingular. A measurable classifier on the original magnitudes can be written as $f(A^{-1}\boldsymbol{q})$ without information loss, and conversely. The four colours alone remove the common-brightness direction: $\boldsymbol{m}$ and $\boldsymbol{m} + \delta\mathbf{1}$ have identical colours. The anchor restores that missing degree of freedom.

Information preservation does not imply equal performance within a restricted model family. In particular, an axis-aligned split on $m_g - m_r$ is oblique in the original (m_g, m_r) coordinates. Explicit colours can therefore change how many tree splits are needed to represent a useful boundary, even though the colour-plus-anchor transformation is invertible.

B Variance of an exchangeable ensemble

For $\overline{T} = B^{-1}\sum_b T_b$, suppose $\mathrm{Var}(T_b) = \sigma^2$ and $\mathrm{Cov}(T_b, T_{b'}) = \rho\sigma^2$ for $b \neq b'$. Then

$$\begin{aligned} \mathrm{Var}(\overline{T}) &= \frac{1}{B^2}\left[\sum_b \sigma^2 + \sum_{b\neq b'} \rho\sigma^2\right] \\ &= \frac{\sigma^2}{B} + \frac{B-1}{B}\rho\sigma^2, \end{aligned} \tag{64}$$

which is Eq. (27). As $B \to \infty$, the expression approaches $\rho\sigma^2$. Under these assumptions, increasing the number of trees removes only the uncorrelated part of the variance. The calculation does not include bias, varying tree variances, or the sampling distribution of the final accuracy score.

C Why cross-entropy targets the posterior

For a fixed measurement x and a candidate probability vector q, the conditional expected cross-entropy is

$$\begin{aligned} \mathbb{E}[-\log q_Y \mid X = x] &= -\sum_k p_k(x)\log q_k \\ &= H(p(x)) + \mathrm{KL}(p(x)\|q). \end{aligned} \tag{65}$$

Non-negativity of KL divergence makes $q = p(x)$ an optimum, with the usual support conditions. The minimum expected loss is therefore $H(Y \mid X)$. Applying this identity to $X = M$ and $X = (M, Z)$ yields Eq. (55). This derivation explains the predictive value of additional information at the population optimum without asserting that the notebook estimates entropy, mutual information, or a calibrated posterior.

D video

https://youtu.be/BlD-pKUphZ0

References

[1] Abdurro'uf et al., The Seventeenth Data Release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar, and APOGEE-2 Data, *Astrophys. J. Suppl. Ser.* **259**, 35 (2022), arXiv:2112.02026.

[2] fedesoriano, *Stellar Classification Dataset—SDSS17*, Kaggle, https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17 (dataset source recorded in the project notebook).

[3] A. O. Clarke, A. M. M. Scaife, R. Greenhalgh, and V. Griguta, Identifying galaxies, quasars, and stars with machine learning: A new catalogue of classifications for 111 million SDSS sources without spectra, *Astron. Astrophys.* **639**, A84 (2020), doi:10.1051/0004-6361/201936770.

[4] R. H. Lupton, J. E. Gunn, and A. S. Szalay, A modified magnitude system that produces well-behaved magnitudes, colors, and errors even for low signal-to-noise ratio measurements, *Astron. J.* **118**, 1406–1410 (1999), arXiv:astro-ph/9903081.

[5] SDSS Collaboration, *Measures of Flux and Magnitude, DR17*, https://www.sdss4.org/dr17/algorithms/magnitudes/ (accessed October 29, 2026).

[6] D. W. Hogg, *Distance measures in cosmology*, arXiv:astro-ph/9905116 (1999; revised 2000).

[7] SDSS Collaboration, *Redshifts, Classifications and Velocity Dispersions, DR17*, https://www.sdss4.org/dr17/algorithms/redshifts/ (accessed October 29, 2026).

[8] F. Pedregosa et al., Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.* **12**, 2825–2830 (2011), https://jmlr.org/papers/v12/pedregosa11a.html.

[9] scikit-learn developers, *LogisticRegression API reference*, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (accessed October 29, 2026).

[10] L. Breiman, Random forests, *Machine Learning* **45**, 5–32 (2001), doi:10.1023/A:1010933404324.

[11] scikit-learn developers, *Ensembles: gradient boosting, random forests, bagging, voting, stacking*, https://scikit-learn.org/stable/modules/ensemble.html (accessed October 29, 2026).

[12] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in *Proceedings of KDD 2016*, pp. 785–794 (2016), doi:10.1145/2939672.2939785.

[13] scikit-learn developers, *PCA API reference*, https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html (accessed October 29, 2026).

[14] scikit-learn developers, *roc_auc_score API reference*, https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html (accessed October 29, 2026).

[15] scikit-learn developers, *Probability calibration*, https://scikit-learn.org/stable/modules/calibration.html (accessed October 29, 2026).

[16] E. B. Wilson, Probable inference, the law of succession, and statistical inference, *J. Am. Stat. Assoc.* **22**, 209–212 (1927), doi:10.1080/01621459.1927.10502953.

[17] Q. McNemar, Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika* **12**, 153–157 (1947), doi:10.1007/BF02295996.