Machine Learning Analysis Report

S. Palis

Udacity Master's Degree in AI Capstone Applied Machine Learning: Model Design, Training, And Performance Evaluation

February 9, 2026

GitHub repository:

https://github.com/MinervaRose/applied-machine-learning

Overview

This project addresses a supervised multi-class classification problem aimed at predicting whether an observed astronomical signal corresponds to a confirmed exoplanet, a candidate object, or a false positive. The dataset used is the *Exoplanet Classification Dataset* (Silva, n.d.), a publicly available tabular dataset combining multiple astronomical catalogs. A series of machine learning classifiers were implemented using scikit-learn to model signal classification and evaluate predictive performance. The goal is to compare models and assess how well structured astrophysical features support automated classification.

Dataset Description

The dataset represents processed astronomical measurements derived from exoplanet observation catalogs. It contains 19,761 rows and 17 numerical columns, including astrophysical features such as orbital period, stellar temperature, planetary radius, and principal component representations of signal structure. The target variable, labeled `label`, encodes four signal

classes corresponding to different exoplanet classifications. All features are numerical and contain no missing values, making the dataset well suited for supervised machine learning. The dataset structure allows models to learn relationships between observed signal characteristics and classification outcomes..

Modeling Approach

The dataset was prepared by separating input features from the categorical target variable and performing a stratified train–test split to preserve class distribution. Feature scaling was applied using standard normalization, which improves convergence and prevents features with large numeric ranges from dominating model behavior (Géron, 2019). Three classification models were trained: Logistic Regression as an interpretable linear baseline, Random Forest as a nonlinear ensemble model, and Gradient Boosting as a boosted decision-tree method.

Logistic regression is a probabilistic linear classifier that models class membership using decision boundaries in feature space (Bishop, 2006). Ensemble tree methods such as Random Forest and Gradient Boosting combine multiple weak learners to improve predictive performance and robustness (Géron, 2019). Accuracy, precision, recall, and F1-score were used to evaluate performance because these metrics provide complementary views of classification quality and class balance (Pedregosa et al., 2011). Accuracy measures the proportion of correct predictions, while precision and recall quantify class-specific prediction reliability and completeness (Pedregosa et al., 2011). These models assume that historical labeled examples are representative of future observations and that feature relationships remain stable over time.

Results

Model comparison shows meaningful performance differences. Logistic Regression achieved approximately 61% test accuracy, serving as a baseline linear classifier. Gradient Boosting improved performance to roughly 72%, while Random Forest achieved the highest test accuracy at approximately 77%. Class-wise metrics reveal that most prediction errors occur among the primary signal categories, while the rarest class shows unstable performance due to very limited sample size. Confusion matrix analysis indicates overlapping feature characteristics between classes, which limits perfect separability..

Interpretation for a Non-Technical Audience

The system attempts to automatically categorize astronomical signals into different types of exoplanet detections. The best-performing model correctly classifies about three out of four signals it has never seen before. While this is strong performance, some categories are harder to distinguish because their characteristics look similar in the data. The rarest type of signal appears so infrequently that the model has limited opportunity to learn its patterns. In practical terms, the model can support scientific analysis but should be used alongside expert review rather than as a fully autonomous decision system.

Limitations and Potential Bias

One major limitation is class imbalance: some signal categories appear far less frequently than others. When rare classes are underrepresented, models tend to favor common classes, which can reduce reliability for unusual events. This imbalance introduces evaluation uncertainty and may bias predictions toward majority categories. Additionally, machine learning systems trained on historical astronomical data assume that future signals follow similar statistical patterns, which may not hold if observation technology changes. Responsible machine learning

practice requires acknowledging such uncertainty and monitoring performance over time (Sculley et al., 2015).

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Géron, A. (2019). *Hands-on machine learning with scikit-learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

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, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in neural information processing systems* (pp. 2503–2511).

DataTales by Agos. (n.d.). *Exoplanet Classification Dataset* [Dataset]. Kaggle. Retrieved February 9, 2026, from https://www.kaggle.com/datasets/datatalesbyagos/exoplanet-classification-dataset