Star Classification: EDA, Preprocessing, Modeling with scikit-learn, Feature Importance & Optimization

1st Anaya Jain  
*Oberoi International School JVLR*Mumbai,India  
jainanaya2020@gmail.com

2nd Dr.Reetu Jain  
*Mentor, On My Own Technology*, Mumbai, India  
reetu.jain@onmyowntechnology.com

***Abstract*—This electronic document is a “live” template and already defines the components of your paper [title, text, heads, etc.] in its style sheet. *\*CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract*. (***Abstract***)**

***Keywords—component, formatting, style, styling, insert (****key words****)***

# Introduction (*Heading 1*)

This template, modified in MS Word 2007 and saved as a “Word 97-2003 Document” for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

# Literature Review

The study by Brian R. Kent[1] Astronomical data in the form of catalogs, data cubes, images, and simulations can be effectively visualized using advanced 3D graphics software like Blender. This software allows astronomers to leverage high-quality, interactive 3D rendering capabilities, enhancing their ability to explore complex parameter spaces in astronomy. Blender's user-friendly interface provides essential features for creating detailed models, textures, animations, and renders, making it accessible for astronomers to produce insightful visualizations. The ability to animate and showcase various astronomical phenomena through video and sample files facilitates better understanding and communication of complex data, thus benefiting the astronomical community. While Blender offers powerful tools for visualizing astronomical data, it has a steep learning curve for those unfamiliar with 3D graphics software. The process of generating models, textures, animations, and camera work requires significant time and expertise, which may be challenging for astronomers with limited experience in these areas. Additionally, the complexity of rendering high-quality graphics can demand substantial computational resources, potentially leading to performance issues. Despite the availability of sample visualizations and scripts, there may still be a gap in fully integrating Blender's capabilities with specific astronomical data needs and workflows.

The study by Jacob VanderPlas, Andrew J. Connolly, Zeljko I vezic, Alex Gray [2] The increase in astronomical data volume due to advanced detectors and telescopes necessitates robust data handling tools. AstroML, built on Python and scikit-learn, addresses this need by providing a comprehensive set of machine learning tools tailored for astronomical data analysis. This initiative supports the next generation of students and researchers by offering practical solutions for handling large, complex datasets. AstroML’s integration with established libraries ensures that users can leverage powerful statistical techniques without needing extensive training in data science. Example applications demonstrate the package’s capabilities, making it a valuable resource for analyzing vast amounts of astronomical data. Despite its advantages, AstroML faces challenges due to the rapid growth in data volume, potentially overwhelming users with large-scale datasets. Students and researchers accustomed to traditional methods may struggle with the learning curve associated with advanced machine learning tools. The effectiveness of AstroML depends on users' familiarity with Python and scikit-learn, which may not be universally applicable. Additionally, the package may not fully address all specific needs of every astronomical survey, potentially requiring supplementary tools or custom solutions. The sheer scale of data entering the petabyte domain may still pose difficulties in efficient processing and analysis.

The study by Nicholas M. Ball, Robert J. Brunner, and Adam D. Myers [3] Classifying 143 million nonrepeat photometric objects from the SDSS using decision trees offers significant insights into star and galaxy populations. By training on 477,068 objects with spectroscopic data, the approach achieves reliable classifications for approximately 22 million objects. The integration of Data-to-Knowledge and supercomputing resources allows for thorough exploration of decision tree parameters, enhancing the accuracy of classifications. This public release provides a comprehensive dataset, categorized into galaxy, star, or neither, with associated probabilities, facilitating targeted analyses. The blind tests on SDSS, 2dFGRS, and 2QZ surveys validate the robustness of classifications, demonstrating reliability even at photometric limits. The reliance on decision trees trained with spectroscopic data poses limitations in extrapolating beyond the training set’s photometric range. Despite robust classifications up to r≈20r \approx 20r≈20, the efficiency of extrapolation for objects at higher magnitudes remains uncertain. The method's effectiveness is constrained by the photometric limits of the spectroscopic training data, potentially affecting accuracy for fainter objects. Additionally, while the classification is comprehensive, it relies heavily on the quality and completeness of the training set, which may not capture all nuances of the vast SDSS data, leading to potential gaps in classification accuracy.

The study by Christopher Kraussa, Xuan Anh Dob, Nicolas Huckc [4] The integration of deep neural networks (DNN), gradient-boosted trees (GBT), and random forests (RF) in analyzing statistical arbitrage in finance showcases the power of advanced machine learning techniques. These models, trained on lagged returns from the S&P 500 stocks, demonstrate strong empirical performance, with an equal-weighted ensemble (ENS1) achieving significant out-of-sample returns exceeding 0.45 percent per day. This approach effectively leverages various methods' strengths, providing robust trading signals by converting high and low probability forecasts into actionable long and short positions. Despite declining profits, the results challenge the semi-strong form of market efficiency, suggesting potential for profitable trading strategies. The application of deep learning and ensemble methods to financial data, while promising, faces several challenges. The reliance on historical data and the elimination of survivor bias may not fully account for market changes or unforeseen events, potentially limiting model robustness. The observed decline in profits over recent years suggests that the effectiveness of these models may diminish, possibly due to increased competition and market adaptation. Additionally, the strategies may incur significant transaction costs, which can erode returns. The models' complexity also introduces a risk of overfitting, where performance on historical data may not translate into future success.

The study by Christopher Kraussa, Xuan Anh Dob, Nicolas Huckc [5] The research on applying deep learning and ensemble methods like deep neural networks (DNN), gradient-boosted trees (GBT), and random forests (RF) to statistical arbitrage demonstrates significant promise. By leveraging these advanced techniques, the study successfully generates effective trading signals based on historical stock returns from the S&P 500. The equal-weighted ensemble (ENS1) of DNN, GBT, and RF shows impressive out-of-sample returns exceeding 0.45 percent per day, highlighting the potential for these models to outperform traditional methods. This approach challenges the semi-strong form of market efficiency, suggesting that sophisticated machine learning techniques can identify profitable trading opportunities. Despite its promising results, the study faces several limitations. The performance of the models, particularly the equal-weighted ensemble, may be affected by transaction costs, which could diminish the reported returns. Additionally, the decline in profitability over recent years suggests potential issues with model robustness and adaptability in evolving market conditions. The reliance on historical data and the elimination of survivor bias might not fully capture future market dynamics or sudden changes, potentially impacting the models' long-term effectiveness. Moreover, the complexity of deep learning models increases the risk of overfitting, which could limit their predictive power in real-world scenarios.

The study by Edward J. Kim, and Robert J. Brunner [6] The use of deep convolutional neural networks (ConvNets) for star-galaxy classification offers several advantages. Unlike traditional classifiers that depend on manually extracted and selected features, ConvNets automatically learn relevant features directly from calibrated pixel values. This approach significantly reduces the need for human expertise in feature engineering, streamlining the classification process. Demonstrated with data from the Sloan Digital Sky Survey and the Canada–France–Hawaii Telescope Lensing Survey, ConvNets provide accurate and well-calibrated probabilistic classifications, rivaling conventional techniques. As deep learning technology advances, it holds promise for enhancing the analysis of current and future photometric surveys with minimal manual intervention. Despite its benefits, using deep convolutional neural networks (ConvNets) for star-galaxy classification has some drawbacks. The primary challenge is the need for extensive computational resources and large datasets to effectively train these models. Deep learning methods can be computationally intensive, requiring significant processing power and memory. Additionally, while ConvNets reduce manual feature engineering, they are not entirely free from human oversight. Model performance still depends on the quality and quantity of the training data, and ConvNets may struggle with noisy or incomplete data. Furthermore, the complexity of deep learning models can make them less interpretable, posing challenges for understanding and validating the results.

The study by Edward J. Kim, Robert J. Brunner and Matias Carrasco Kind [7] The proposed meta-classification framework offers significant advantages by integrating various star-galaxy classification techniques. This hybrid approach combines morphological classifiers, random forests, self-organizing maps, and hierarchical Bayesian methods to enhance classification robustness. The framework demonstrates improved performance over individual methods, particularly when trained with high-quality spectroscopic data from sources like DEEP2, SDSS, VIPERS, and VVDS. By addressing different scenarios, including mismatches between training and test data demographics, the framework proves versatile and effective. This ensemble strategy can be highly beneficial for current and future photometric surveys, such as the Dark Energy Survey and the Large Synoptic Survey Telescope, by providing more accurate and reliable classifications. While the meta-classification framework offers enhanced robustness, it also has notable limitations. The complexity of combining multiple techniques may introduce computational challenges and require significant resources for model training and evaluation. Managing and integrating diverse methods could complicate the workflow and make the system harder to maintain. Additionally, the framework's performance heavily relies on the quality and representativeness of the training data; discrepancies between training and test datasets can still impact accuracy. The need for high-quality spectroscopic labels and diverse data sources may limit applicability in scenarios where such data is unavailable or difficult to obtain.

The study by Ce Yu1,2 Kun Li1,2 Yanxia Zhang3 Jian Xiao1,2 Chenzhou Cui2,3 Yihan Tao3 Shanjiang Tang1,2 Chao Sun1,2 Chongke Bi [8] The integration of artificial intelligence (AI) into light curve analysis represents a significant advancement in handling the expanding scale of observational data in time domain astronomy. Traditional methods struggle to fully utilize the vast amount of data generated, whereas AI techniques, including machine learning and deep learning, offer enhanced capabilities for extracting valuable insights. By applying these advanced methods, researchers can identify a greater number of candidates with scientific research potential. This survey effectively reviews recent developments, summarizes key machine learning concepts, and outlines their applications in light curve analysis. Embracing AI enables a more comprehensive exploration of variable celestial objects, leveraging the full potential of big data in astronomy. Despite the advantages, the application of AI in light curve analysis comes with challenges. The effectiveness of AI techniques depends heavily on the quality and representativeness of the training data, which may not always be available or may introduce biases. The complexity of deep learning models requires substantial computational resources and expertise, potentially limiting accessibility for some researchers. Additionally, AI models can be prone to overfitting, where they perform well on training data but fail to generalize to new, unseen data. The evolving nature of AI technology means that continuous updates and adjustments are necessary, which can be resource-intensive. Furthermore, there is a risk of relying too heavily on automated methods without sufficient human oversight and validation.

The study by Ting-Yun Cheng, Christopher J. Conselice, Alfonso Aragon-Salamanca, Nan Li, Asa F. L. Bluck, Will G. Hartley, James Annis, David Brooks, Peter Doel, Juan Garc´ıa-Bellido, David J. James, Kyler Kuehn, Nikolay Kuropatkin, Mathew Smith , Flavia Sobreira and Gregory Tarle [9] The study provides a comprehensive comparison of various supervised machine learning methods for automated morphological classification of galaxies, including Convolutional Neural Networks (CNN), K-nearest neighbors, logistic regression, Support Vector Machines, Random Forests, and Neural Networks. By utilizing Dark Energy Survey (DES) data and visual classifications from the Galaxy Zoo 1 (GZ1) project, the research identifies CNN as the most effective method, achieving an accuracy of approximately 0.99 for classifying ellipticals and spirals. The study highlights the ability of CNNs to detect misclassifications in GZ1 data and reveal distinct galaxy types, such as lenticulars (S0), enhancing overall classification accuracy and improving dataset reliability. While the study demonstrates CNN's superior performance, it also has limitations. The reliance on visual classifications from GZ1, which are themselves prone to errors, may impact the accuracy of the comparison. Despite achieving high accuracy, the study identifies that approximately 2.5% of galaxies are misclassified by GZ1, suggesting that even high-performing models are not immune to dataset issues. Additionally, the correction of labels and the focus on CNNs might overshadow the potential strengths of other methods. The study's findings are based on a relatively small sample size of ∼2800 galaxies, which may not fully represent the diversity of galaxy types in larger datasets or other surveys.

The study by Christopher J. Fluke and Colin Jacobs [10] Machine learning (ML) and artificial intelligence (AI) have become integral to modern astronomy, offering transformative capabilities across various applications. By learning from examples, ML algorithms such as random forests, support vector machines, and neural networks enhance our ability to classify, predict, and discover new astronomical phenomena. These techniques are significantly impacting diverse fields, including the detection of extrasolar planets, analysis of transient objects, identification of quasars, and observation of gravitationally lensed systems. Additionally, ML and AI contribute to forecasting solar activity and distinguishing between genuine signals and instrumental noise in gravitational wave astronomy. The comprehensive review of contemporary literature highlights the versatility and growing maturity of these technologies in advancing astronomical research. Despite their advancements, the use of machine learning (ML) and artificial intelligence (AI) in astronomy faces several challenges. The reliance on ML algorithms can lead to issues with interpretability, making it difficult to understand how decisions are made or to validate results. Additionally, the performance of ML models is heavily dependent on the quality and representativeness of the training data, which may introduce biases or limit generalizability. The rapid pace of technological development also means that methods can quickly become outdated, requiring continuous adaptation and validation. Furthermore, the review of literature may not fully address the limitations of specific ML applications or the challenges associated with integrating AI into traditional astronomical workflows.

The study by Christopher J. Fluke and Colin Jacobs [11] Machine learning (ML) and artificial intelligence (AI) have become vital tools in astronomy, revolutionizing how we classify, predict, discover, and analyze astronomical data. With their ability to learn from examples, ML algorithms such as random forests, support vector machines, and neural networks are making significant impacts across a range of applications. These technologies are instrumental in discovering extrasolar planets, analyzing transient objects, identifying quasars, and studying gravitationally lensed systems. They also improve forecasting of solar activity and help distinguish genuine signals from instrumental noise in gravitational wave astronomy. The ongoing addition of new applications reflects the growing influence and maturity of ML and AI in advancing astronomical research. Despite their significant contributions, the application of machine learning (ML) and artificial intelligence (AI) in astronomy faces several challenges. The effectiveness of these technologies is highly dependent on the quality and representativeness of the training data, which can introduce biases and affect model performance. The complexity and opacity of ML models can also make them difficult to interpret and validate, potentially hindering trust and understanding. Furthermore, the rapid advancement of ML and AI techniques means that methods can quickly become obsolete, requiring continuous updates and adaptations. The review of literature may not fully address the practical challenges of integrating AI into existing astronomical workflows or the limitations of specific applications.

The study by Hongwen Zheng and Yanxia Zhang [12] As astronomical data grows exponentially in volume and complexity, feature selection and extraction become crucial for effective data analysis. This paper distinguishes between feature selection and extraction methods, providing a detailed taxonomy and characteristics of each approach. By comparing different feature selection methods through a case study, it highlights the practical benefits of filter methods, such as ReliefF and Fisher filter, over wrapper methods, which include CHAID, LDA, Naive Bayes, and C4.5. The study demonstrates that filter methods generally offer lower computational costs, making them more efficient. Furthermore, it shows that combining suitable learning algorithms with effective feature selection methods can enhance overall performance, addressing the challenge of extracting meaningful information from complex astronomical datasets. The study also reveals limitations and challenges associated with feature selection and extraction methods. While filter methods are computationally more efficient, they may not always capture the most relevant features compared to wrapper methods, which, despite their higher computational cost, can offer more precise feature selection. The case study's focus on specific methods may not generalize across all types of astronomical data or applications. Additionally, the performance of feature selection techniques is heavily dependent on the quality and nature of the data, which can vary widely. The paper may also overlook the potential benefits of hybrid approaches that combine elements of both filter and wrapper methods, which could offer a balance between computational efficiency and selection accuracy.

The study by Pablo Huijse, Pablo A. Estévez, Pavlos Protopapas,José C. Príncipe & Pablo Zegers [13] The paradigm shift in time-domain astronomy (TDA) due to the explosive growth in data size, complexity, and generation rates represents a significant advancement in the field. The Large Synoptic Survey Telescope (LSST), with its anticipated 150 Petabyte dataset and 2 Terabytes per hour streaming capability, will offer an unprecedented view of the sky, capturing both known and unknown phenomena. This transformation necessitates the development of new data-driven paradigms like astroinformatics and astrostatistics, which integrate statistics, data mining, machine learning, and computational intelligence. These approaches will enable automated, robust methods for detecting and classifying both known astrophysical objects and novel phenomena, pushing the boundaries of our understanding of the universe. The article provides a comprehensive overview of these applications and future research directions, highlighting the need for interdisciplinary collaboration to manage the vast data influx from the LSST. The rapid growth in data size and complexity also presents substantial challenges. The LSST's enormous data volume may overwhelm existing computational and analytical infrastructure, making it difficult to process and analyze in real-time. The integration of diverse techniques from astroinformatics, data mining, and machine learning introduces complexity in terms of methodology and implementation. The reliance on automated methods may also lead to challenges in ensuring accuracy and handling unexpected or anomalous data. Additionally, the need for interdisciplinary collaboration, while crucial, can be difficult to coordinate and may introduce logistical and communication hurdles. The sheer scale of data generated by the LSST could also exacerbate issues related to data storage, management, and sharing, requiring innovative solutions to address these pressing concerns.

The study by Zafiirah Hosenie, Robert J. Lyon, Benjamin W. Stappers and Arrykrishna Mootoovaloo [14] The upcoming synoptic surveys' immense data volumes necessitate advanced automatic frameworks for efficient classification of astronomical objects. This study uses the Catalina Real-Time Transient Survey (CRTS) data, focusing on 11 variable star types, to showcase robust methods for feature selection and evaluation using information theory. By applying and optimizing machine learning algorithms, such as random forests, the study achieves high classification accuracy. Notably, converting the multiclass classification into binary tasks results in a balanced accuracy rate of approximately 99% for specific star types. Additionally, transforming the classification problem into a hierarchical taxonomy enhances performance by accurately identifying subtypes of Cepheids, RR Lyrae, and eclipsing binary stars, demonstrating the effectiveness of advanced classification techniques in handling complex astronomical data. The reliance on automatic frameworks and machine learning algorithms presents challenges, including potential limitations in the generalizability of models across different datasets or survey conditions. While converting multiclass problems into binary tasks improves classification accuracy, it may oversimplify the complexity of the data and lead to potential loss of nuanced information. The hierarchical taxonomy approach, while beneficial, adds complexity to the classification process and may require extensive refinement and validation to ensure its effectiveness. Additionally, optimizing classifiers through cross-validation techniques can be computationally intensive and may not fully address issues related to data imbalance or inherent variability in astronomical observations. The study’s focus on specific datasets and features may also limit its applicability to other types of variable stars or future surveys with different characteristics.

The study by Amr HassanA,B and Christopher J. FlukeA[15] As astronomy progresses into an era of petabyte-scale data generation, the role of scientific visualization becomes increasingly crucial. This study highlights the importance of visualization in interpreting complex astronomical data, focusing on large-N particle data and spectral data cubes. By reviewing literature from the past two decades, the study identifies key areas of advancement and ongoing research challenges. It emphasizes the contributions of high-performance computing, such as distributed processing and GPUs, in enhancing visualization capabilities. Additionally, the study explores the benefits of collaborative visualization, metadata management via workflow systems, and the use of advanced interaction devices. These advancements collectively improve our ability to discover and understand astronomical phenomena amidst the growing data deluge.Despite its importance, scientific visualization in astronomy faces several challenges. The complexity of visualizing petabyte-scale data can lead to performance bottlenecks, requiring substantial computational resources and sophisticated technology. The study acknowledges limitations in current visualization methods and tools, which may not fully address the diverse needs of astronomical data analysis. Issues such as data integration, interoperability between different visualization tools, and the steep learning curve for advanced interaction devices can hinder widespread adoption. Additionally, the focus on specific types of data (large-N particle data and spectral data cubes) may overlook other significant areas of astronomical research. The identified grand challenges highlight the need for continued innovation, but also reflect the significant obstacles that must be overcome to advance scientific visualization in the Petascale Astronomy Era.

The study by Yudong Zhang, Siyuan Lu, Xingxing Zhou, Ming Yang, Lenan Wu, Bin Liu, Preetha Phillips and Shuihua Wang [16] The developed machine learning-based system for detecting multiple sclerosis (MS) from magnetic resonance imaging (MRI) represents a significant advancement in medical imaging analysis. By utilizing inter-scan normalization, the system effectively addresses gray-level differences, which is crucial for ensuring consistency across diverse datasets. Adjusting misclassification costs helps manage the imbalance between MS subjects and healthy controls (HCs), improving classification accuracy. The use of two-level stationary wavelet entropy (SWE) for feature extraction provides a robust method for capturing essential information from brain images. Among the evaluated classifiers—decision tree, k-nearest neighbors (kNN), and support vector machine—kNN demonstrated superior performance. Additionally, the SWE+kNN approach outperforms four existing methods, highlighting its effectiveness in MS detection. Despite the promising results, there are several limitations and challenges associated with the proposed system. The reliance on inter-scan normalization, while effective, may not fully address all sources of variability in MRI data, potentially affecting the system's robustness. Adjusting misclassification costs for class imbalance can introduce complexity in model tuning and might not fully resolve the challenges associated with small sample sizes. While kNN performed best among the classifiers tested, its performance can be sensitive to the choice of distance metrics and parameter settings. Additionally, the comparison with state-of-the-art methods, while positive, does not guarantee that the SWE+kNN approach will generalize well across different datasets or imaging protocols. The system's effectiveness is also contingent on the quality and representativeness of the data collected from the local hospital and the eHealth laboratory.

The study by Jia Wu, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, Si-Hao Deng [17] The use of Gaussian processes and Bayesian optimization for hyperparameter tuning represents a significant advancement in machine learning efficiency. This approach leverages Bayesian theorem to model the relationship between model performance and hyperparameters, effectively transforming hyperparameter tuning into an optimization problem. By setting a prior over the optimization function and updating it with information from previous samples, Bayesian optimization systematically explores the hyperparameter space. This method not only enhances the efficiency of finding optimal hyperparameters but also reduces the reliance on brute-force search and extensive expert knowledge. Experimental results on standard datasets demonstrate that this technique successfully identifies optimal hyperparameters for various machine learning models, including random forests, neural networks, and multi-grained cascade forests, while considering time cost. Despite its advantages, the Bayesian optimization approach has some limitations. The method requires careful selection of the utility function, which can impact the effectiveness of the optimization process. Additionally, while Bayesian optimization reduces the need for brute-force search, it still involves computational overhead associated with maintaining and updating the Gaussian process model. This may lead to increased computational time and complexity for large or high-dimensional hyperparameter spaces. Furthermore, the effectiveness of this approach depends on the quality of the prior assumptions and the chosen kernel function, which may not always align perfectly with all types of machine learning models. Therefore, while the method is promising, its practical implementation might require further refinement and adaptation to specific applications and datasets.

# Methodology

1. Identify the Problem

The mission is to predict star types based on their properties, which are represented by a dataset that includes attributes such as temperature, brightness, radius, and color. Our objective is to create a model capable of accurately classifying stars into already established groups. This categorization challenge is critical for understanding star properties and behaviors, which can be applied in astrophysics and other sciences. We hope to create a solid forecasting model for celestial examinations through investigating the dataset while employing machine learning techniques.

1. Evaluate the Literature

A thorough literature review is conducted to understand existing approaches in stellar classification. This includes studying previous works that use machine learning for star classification, examining feature selection techniques, and evaluating various algorithms like Random Forests and Support Vector Machines. The review helps in identifying the strengths and weaknesses of different methods, ensuring that our approach leverages the latest advancements and best practices in the field. This process also highlights gaps and opportunities for improvement.

1. Research Design

The research design involves creating a structured approach to solving the classification problem. This includes defining the problem, selecting appropriate features, and choosing suitable machine learning algorithms. The design encompasses the entire workflow from data collection and preprocessing to model training and evaluation. It also involves determining evaluation metrics to assess the model’s performance and ensure that it meets the desired accuracy and reliability.

1. Data Collection

Data collection involves gathering relevant data on star characteristics from available datasets. This includes attributes such as temperature, luminosity, radius, and color. The dataset should be comprehensive and representative of different star types. Data is often sourced from astronomical databases or research repositories. Ensuring data quality and relevance is crucial for building a robust predictive model.

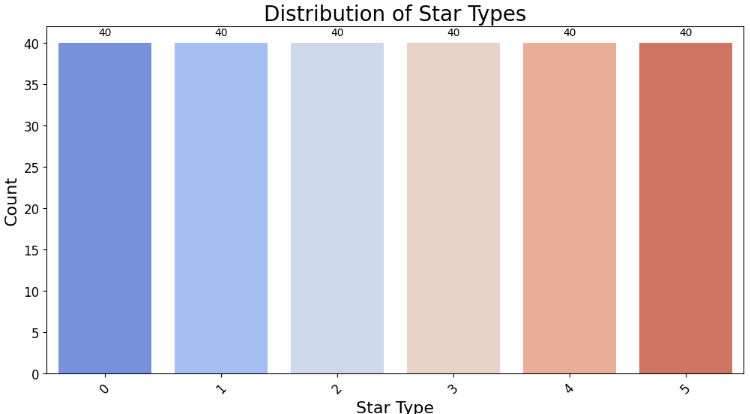


Figure 1: Distribution of Star Types

1. Data Cleaning

Data cleaning involves preprocessing the collected data to handle missing values, inconsistencies, and outliers. This step is essential to ensure the dataset is ready for analysis. Techniques like imputation for missing values and normalization for feature scaling are employed. Data cleaning also involves removing irrelevant or redundant features to improve the model's performance and reduce computational complexity.

1. Data Import and Initial Preview

Data import involves loading the dataset into a suitable environment for analysis. Initial preview includes examining the dataset's structure, such as columns and data types, to understand its content. This step helps in identifying any immediate issues with the data and provides a basis for further preprocessing. Tools like pandas are used to load and preview the data, ensuring it is ready for detailed analysis.

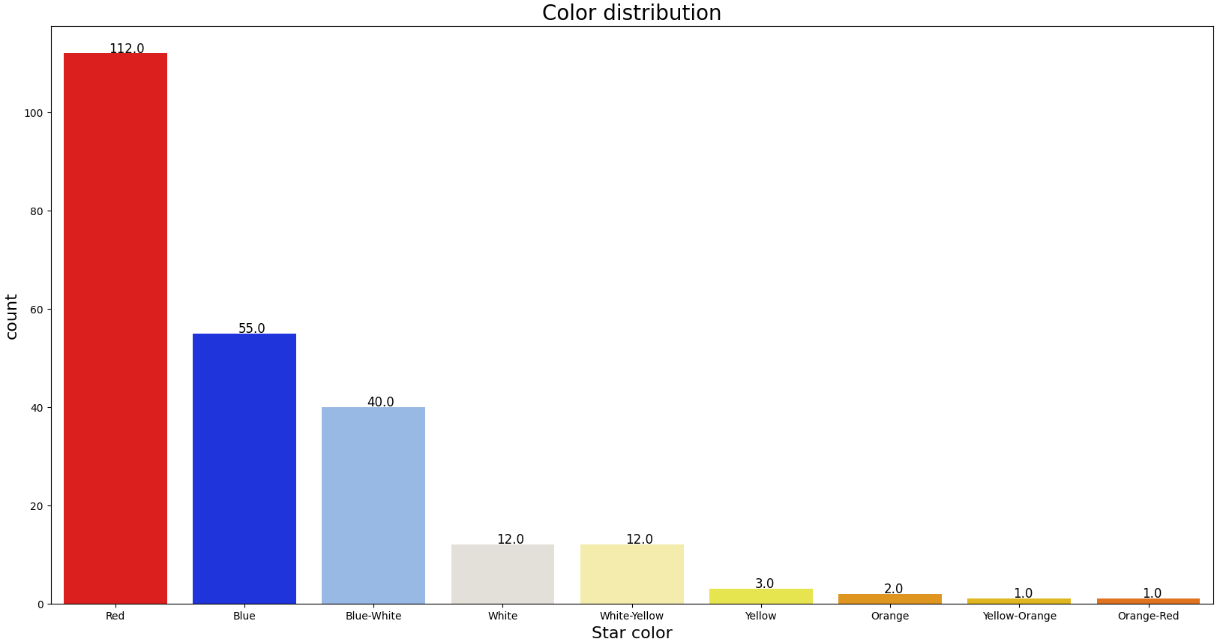


Figure 2: Color Distribution

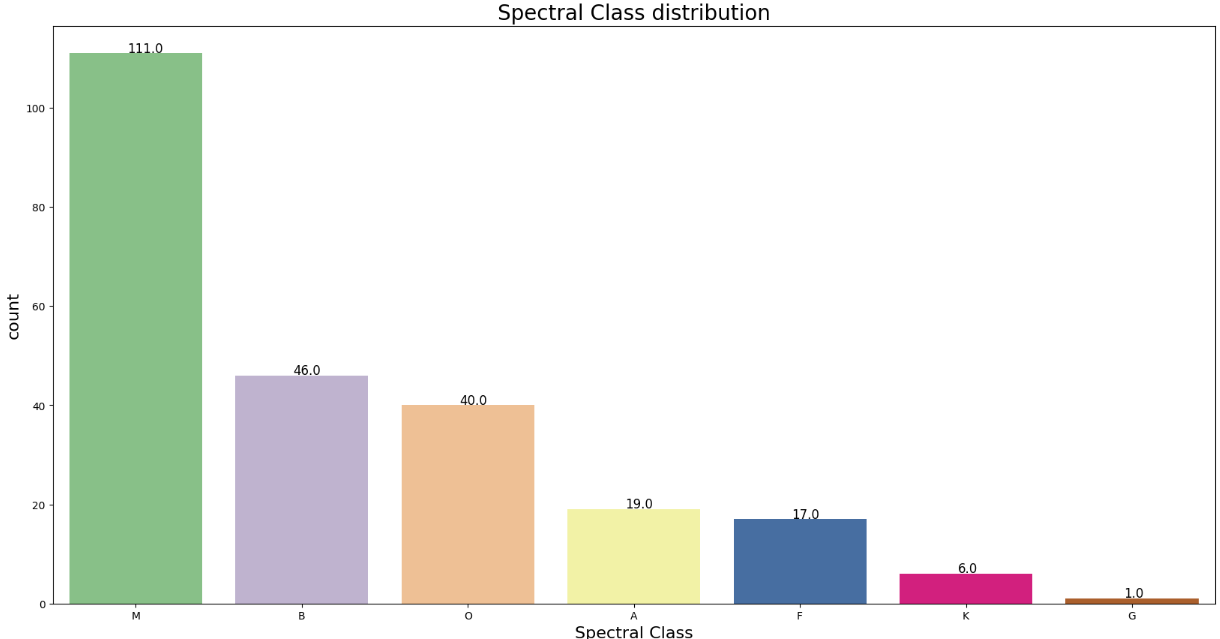


Figure 3:Spectral Class Distribution

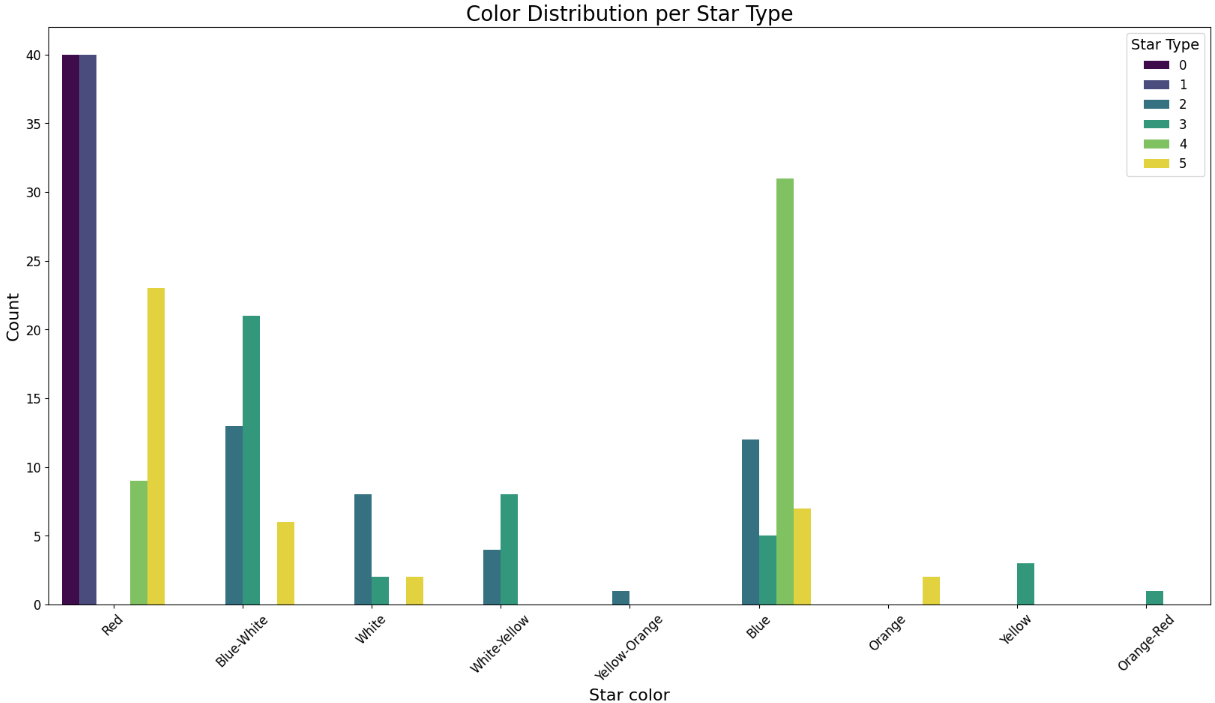


Figure 4: Color distribution per star Type

1. Choosing a Model

Choosing a model involves selecting an appropriate machine learning algorithm for the classification task. This decision is based on the nature of the data and the problem requirements. Common models for classification include Random Forests, Support Vector Machines, and Neural Networks. The choice of model is influenced by factors such as accuracy, interpretability, and computational efficiency.

1. Training the Model

Model training involves using the preprocessed data to teach the selected algorithm how to classify stars. This process includes splitting the data into training and test sets, fitting the model to the training data, and optimizing its parameters. The goal is to create a model that generalizes well to unseen data, effectively learning patterns that differentiate between star types.

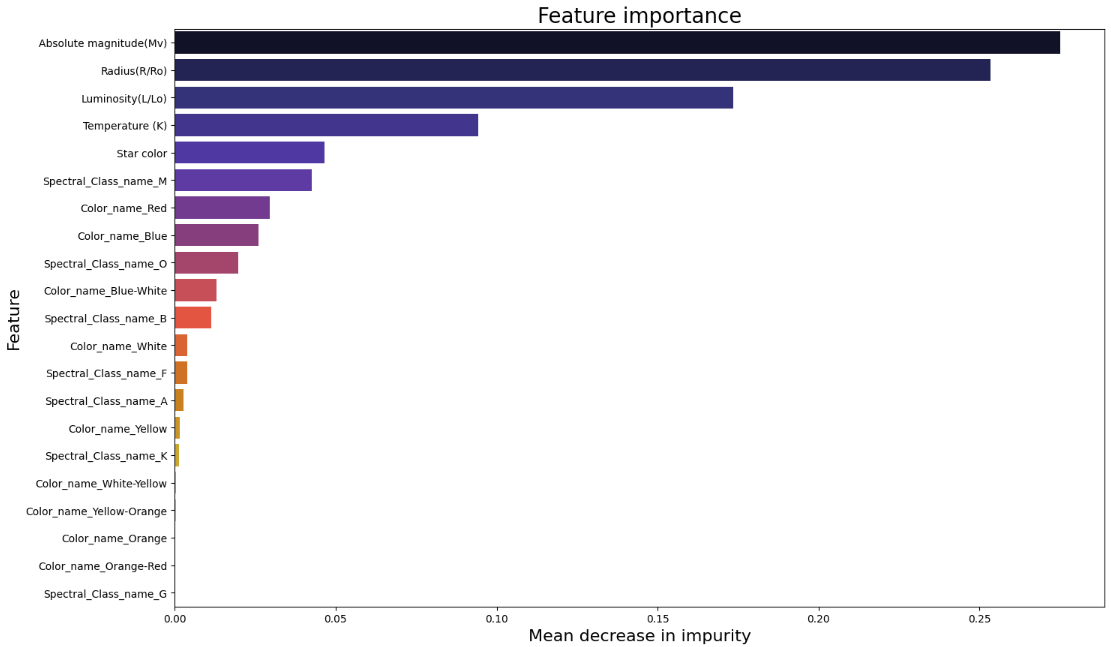


Figure 5: Feature Importance

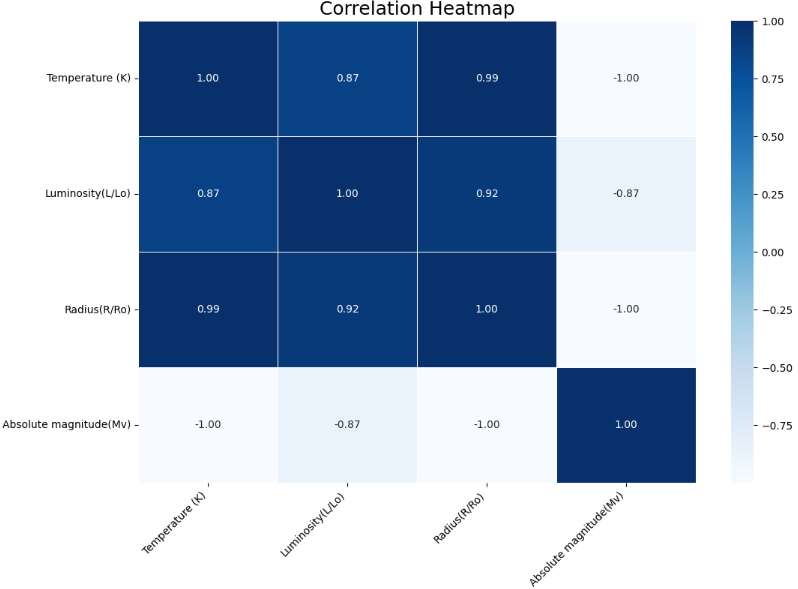


Figure 6:

1. Evaluating the Model

Model evaluation involves assessing the trained model's performance using various metrics such as accuracy, precision, recall, and F1 score. This step is crucial for understanding how well the model performs on unseen data. Evaluation helps in identifying any weaknesses or areas for improvement, ensuring that the model meets the desired performance standards.

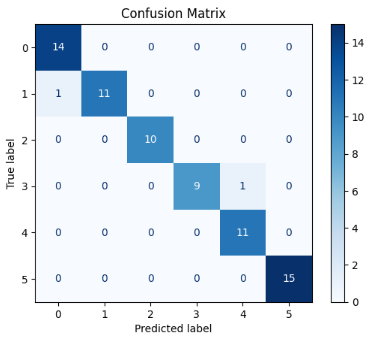


Figure 7:

1. Parameter Tuning

Parameter tuning involves optimizing the model's hyperparameters to enhance its performance. Techniques like grid search or random search are used to find the best combination of hyperparameters. This step is essential for improving the model's accuracy and ensuring it performs optimally on the given data.

1. Making Predictions

Making predictions involves using the trained model to classify new, unseen data. This step applies the model to real-world data to generate predictions about star types. Accurate predictions are crucial for practical applications and validating the model’s effectiveness.

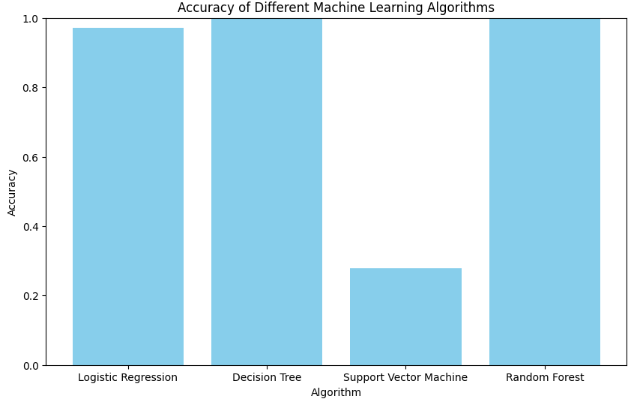


Figure 8: Accuracy of different models

1. Data Visualization

Data visualization involves creating visual representations of the data and model results to facilitate understanding and analysis. This includes plots, charts, and graphs that illustrate distributions, correlations, and model performance. Visualization aids in interpreting the data and communicating findings effectively to stakeholders.

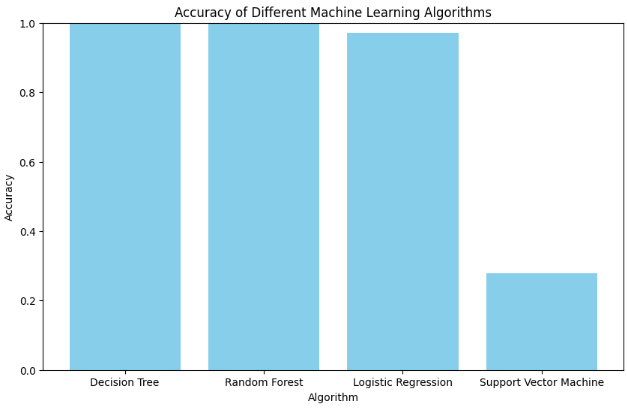


Figure 9: Accuracy of different models in ascending order

# Insights & Observations

The study

# Conclusion

The study

# Ease of Use

## *Selecting a Template (Heading 2)*

First, confirm that you have the correct template for your paper size. This template has been tailored for output on the A4 paper size. If you are using US letter-sized paper, please close this file and download the Microsoft Word, Letter file.

## *Maintaining the Integrity of the Specifications*

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

# Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

## *Abbreviations and Acronyms*

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## *Units*

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.

Identify applicable funding agency here. If none, delete this text box.

* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## *Equations*

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## *Some Common Mistakes*

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum *μ*0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## *Authors and Affiliations*

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### *For papers with more than six authors:* Add author names horizontally, moving to a third row if needed for more than 8 authors.

### *For papers with less than six authors:* To change the default, adjust the template as follows.

#### *Selection:* Highlight all author and affiliation lines.

#### *Change number of columns:* Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### *Deletion:* Delete the author and affiliation lines for the extra authors.

## *Identify the Headings*

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## *Figures and Tables*

#### *Positioning Figures and Tables:* Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| **Table Head** | **Table Column Head** | | |
| --- | --- | --- | --- |
| ***Table column subhead*** | ***Subhead*** | ***Subhead*** |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

1. Kent, B. R. (2013). Visualizing Astronomical Data with Blender. Publications of the Astronomical Society of the Pacific, 125(928), 731–748. doi:10.1086/671412
2. Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2507–2517. doi:10.1093/bioinformatics/btm344
3. Ball, N. M., Brunner, R. J., Myers, A. D., & Tcheng, D. (2006). Robust Machine Learning Applied to Astronomical Data Sets. I. Star‐Galaxy Classification of the Sloan Digital Sky Survey DR3 Using Decision Trees. The Astrophysical Journal, 650(1), 497–509. doi:10.1086/507440
4. Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research, 259(2), 689–702. doi:10.1016/j.ejor.2016.10.031
5. Kim, E. J., & Brunner, R. J. (2016). Star–galaxy classification using deep convolutional neural networks. Monthly Notices of the Royal Astronomical Society, 464(4), 4463–4475. doi:10.1093/mnras/stw2672
6. Kim, E. J., Brunner, R. J., & Carrasco Kind, M. (2015). A hybrid ensemble learning approach to star–galaxy classification. Monthly Notices of the Royal Astronomical Society, 453(1), 507–521. doi:10.1093/mnras/stv1608
7. Yu, C., Li, K., Zhang, Y., Xiao, J., Cui, C., Tao, Y., … Bi, C. (2021). A survey on machine learning based light curve analysis for variable astronomical sources. WIREs Data Mining and Knowledge Discovery, 11(5). doi:10.1002/widm.1425
8. Cheng, T.-Y., Conselice, C. J., Aragón-Salamanca, A., Li, N., Bluck, A. F. L., Hartley, W. G., … Tarle, G. (2020). Optimizing automatic morphological classification of galaxies with machine learning and deep learning using Dark Energy Survey imaging. Monthly Notices of the Royal Astronomical Society, 493(3), 4209–4228. doi:10.1093/mnras/staa501
9. Fluke, C. J., & Jacobs, C. (2019). Surveying the reach and maturity of machine learning and artificial intelligence in astronomy. WIREs Data Mining and Knowledge Discovery. doi:10.1002/widm.1349
10. Zheng, H., & Zhang, Y. (2008). Feature selection for high-dimensional data in astronomy. Advances in Space Research, 41(12), 1960–1964. doi:10.1016/j.asr.2007.08.033
11. Huijse, P., Estevez, P. A., Protopapas, P., Principe, J. C., & Zegers, P. (2014). Computational Intelligence Challenges and Applications on Large-Scale Astronomical Time Series Databases. IEEE Computational Intelligence Magazine, 9(3), 27–39. doi:10.1109/mci.2014.2326100
12. Hosenie, Z., Lyon, R. J., Stappers, B. W., & Mootoovaloo, A. (2019). Comparing Multiclass, Binary, and Hierarchical Machine Learning Classification schemes for variable stars. Monthly Notices of the Royal Astronomical Society, 488(4), 4858–4872. doi:10.1093/mnras/stz1999
13. Hassan, A., & Fluke, C. J. (2011). Scientific Visualization in Astronomy: Towards the Petascale Astronomy Era. Publications of the Astronomical Society of Australia, 28(02), 150–170. doi:10.1071/as10031
14. *Zhang, Y., Lu, S., Zhou, X., Yang, M., Wu, L., Liu, B., … Wang, S. (2016). Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine. SIMULATION, 92(9), 861–871.*
15. *10.11989/jest.1674-862x.80904120*