Project Title: Machine Learning for Enhanced Lightcurve Classification

Introduction

Lightcurve classification is a crucial task in astronomy for identifying variable stars and detecting exoplanet transits. Traditional approaches rely on human expertise and classical algorithms (e.g. periodicity searches) to classify lightcurves, but these methods can be time-consuming and less effective with large datasets. Modern **machine learning (ML)** techniques have the potential to improve accuracy and automate the classification of lightcurves from large-scale surveys. This project will focus on the **KELT-South** telescope's commissioning dataset (a 46-day run in January–February 2010) to explore how ML can enhance classification accuracy and validate existing non-ML methods. By applying ML-based techniques and comparing them to traditional methods, the project aims to identify various types of variable stars (such as pulsating stars and eclipsing binaries) and potential exoplanet transit events more effectively.

Project Goals

- **Improve Classification Accuracy:** Apply modern ML algorithms to lightcurve data to increase the accuracy of classifying stellar variability types compared to traditional methods.
- **Validate Existing Methods:** Cross-check and validate the results of ML classifications against results from established (non-ML) classification techniques, ensuring that new methods are reliable.
- **Develop a Processing Pipeline:** Create a Python-based pipeline capable of ingesting lightcurve data from multiple telescopes (starting with KELT-South) and processing them for ML classification. This pipeline will handle data preprocessing, feature extraction, model prediction, and output generation.
- **Catalogue Variable Stars:** Produce a scientifically valuable catalogue of variable stars identified in the KELT-South 2010 commissioning dataset, including classification of known variables and highlighting candidates for exoplanet transits.

Methodology

1. Literature Review: Conduct a thorough review of current literature on machine learning techniques used in astronomy, specifically for time-series and lightcurve analysis. This includes examining previous studies on variable star classification and exoplanet detection using ML (e.g., random forests, neural networks) and understanding the traditional methods (like periodogram analysis or template fitting) used for lightcurve classification. The literature review will guide the choice of algorithms and features for this project.

2. Data Preprocessing: Prepare the KELT-South lightcurve data for analysis. Key steps include:

- Cleaning the dataset (removing bad data points, handling missing values).
- Normalising or scaling the lightcurves for consistent analysis.

- Extracting relevant features from each lightcurve (e.g. period, amplitude, variability indices) that can feed into ML models.
- If needed, label a subset of lightcurves using known classifications (from catalogues or manual inspection) to create a training set for supervised learning.

3. Model Training and Evaluation: Train several machine learning models to classify lightcurves:

- Start with algorithms like **random forests, XGBoost**, or other tree-based classifiers using features extracted from the lightcurves. Evaluate their performance.
- Experiment with advanced approaches (such as neural networks or LSTM recurrent neural networks) that can potentially learn directly from the time-series data.
- Split data into training and test sets (and use cross-validation) to ensure models generalise well. Use appropriate metrics (accuracy, precision, recall, F1-score) to evaluate classification performance for each class of variable star and for detecting transit events.

4. Comparison with Traditional Methods: Implement one or more traditional lightcurve classification methods to serve as a baseline. For example, use a period-search algorithm to identify periodic variables and classify them by known period ranges or lightcurve shape, or use statistical thresholds to detect variability. Compare the results of these classical techniques with the ML model results in terms of:

- Accuracy of correctly identified variable stars and exoplanet candidates.
- The types of objects each method might miss or classify incorrectly.
- Computational efficiency and scalability to larger datasets.

5. Pipeline Development: Integrate the data preprocessing, model prediction, and result validation steps into a cohesive **Python pipeline**. This pipeline will be modular, allowing easy input of lightcurve data from KELT-South and eventually other telescopes. It will output classified lightcurves with labels (e.g., “delta Scuti star”, “eclipsing binary”, “transiting exoplanet candidate”, etc.) and probabilities or confidence measures from the ML model. The pipeline will be documented and user-friendly so it can be extended or applied to future datasets.

6. Validation and Optimisation: Validate the classification results thoroughly:

- Cross-match the newly created variable star catalogue with existing catalogues (if available for the KELT-South field) to verify known objects are correctly identified.
- Inspect a subset of lightcurves manually to ensure the ML model is making reasonable classifications, especially for high-value candidates like potential exoplanet transits.
- Based on validation, iterate on the ML model and preprocessing: adjust the feature set, try different model hyperparameters, or incorporate additional data (e.g., colours or external data) to improve performance.
- Ensure the final pipeline is optimised for both accuracy and speed, making it feasible to run on larger surveys or real-time data in the future.

Expected Outcomes

- **Functional ML Pipeline:** A working Python-based pipeline that can process raw lightcurve data and output classified variable stars. This tool will be adaptable to multiple telescope datasets and include documentation for future users.

- **Enhanced Classification Accuracy:** Demonstrated improvement in classification performance (accuracy and other metrics) using ML techniques compared to traditional methods. A report on the performance of different ML models and what types of variables each method handles best will be included.
- **Variable Star Catalogue:** A new catalogue of variable stars identified from the KELT-South 46-day commissioning run. This catalogue will list the stars, their variability class (as determined by the ML pipeline, with validation), and any notable objects such as candidate exoplanet transits that merit follow-up.
- **Comparison Study:** An analysis summarising how ML-based classification compares with non-ML methods. The project will detail cases where ML excels and cases where traditional methods might still be advantageous, providing insights for astronomers considering ML for similar problems.
- **Thesis and Publications:** The work will form the basis of the student's MSc thesis. If results are significant, they could be developed into a research paper, for example, describing the new variable stars found or the efficacy of the ML approach in an astronomical journal.

Student Skills Required

- **Python Programming:** Strong proficiency in Python for scientific computing. Experience with libraries such as NumPy and pandas for data manipulation, and matplotlib or other libraries for basic visualisation of lightcurves.
- **Data Science & ML Libraries:** Familiarity with scikit-learn for implementing machine learning algorithms and XGBoost (or similar libraries) for advanced gradient boosting techniques. Some exposure to neural network frameworks (TensorFlow or PyTorch) would be beneficial for experimenting with deep learning models.
- **Astronomy Background:** Basic understanding of stellar variability and exoplanet transits to make informed decisions on feature extraction and to interpret model results in a physically meaningful way. Prior exposure to time-series data analysis (e.g., Fourier transforms, period finding) would be helpful for comparing ML methods with classical approaches.
- **Analytical Skills:** Ability to perform rigorous data analysis, evaluate model performance critically, and optimise algorithms. This includes skills in cross-validation, hyperparameter tuning, and error analysis.
- **Problem-Solving and Adaptability:** Since research can yield unexpected challenges (e.g., data quality issues or models not performing as expected), the student should be resourceful, willing to troubleshoot problems, and adapt the approach as needed.

Conclusion

By leveraging modern machine learning techniques, this project aims to push the boundaries of how efficiently and accurately we can classify stellar lightcurves. The anticipated outcome is a robust classification pipeline and a new catalogue of variable stars from KELT-South's data, demonstrating the value of ML in astrophysical data analysis. This project will not only enhance the student's skills in data science and astronomy but also contribute tools and knowledge that can benefit the broader astrophysical community in the era of ever-growing datasets.

Contact details:

Dr Rudi Kuhn

SALT Astronomer

Southern African Large Telescope / South African Astronomical Observatory

1 Observatory Road, Cape Town, South Africa

P.O. Box 9, Observatory, 7935, Cape Town, South Africa

Tel: +27 21 460 9304

Email: rudi@saao.ac.za or r.kuhn@saao.nrf.ac.za