Extending Frontiers: A Statistical Analysis of Characteristics Influencing Pulsar Classification for Optimized Search by Applying Ensemble Machine Learning Techniques

Lima, Lucas (School: Lee Virtual Instruction Program)

When analyzing the search pipeline for pulsars, many sub-optimal and inefficient procedures are implemented due to the need for manual data analysis and the presence of physical interference. This project's goal is to develop a low-cost method for optimizing the way pulsars are detected by identifying their most important characteristics and creating an automated model capable of classifying possible candidates. First, an exploratory statistical analysis on the publicly available pulsar datasets was performed to better understand the overall structure of the data and the features. Then, ensemble machine learning algorithms were selected and applied for training models on various sets of input data and would be evaluated with appropriate performance metrics. The models would then be optimized by analyzing the significance of characteristics on the model's performance and then removing redundant or negative features from the data. Finally, the models that proved to be the most robust in terms of performance and generalization were used to evaluate each of the statistical characteristics based on their relative importance. In terms of consistency and reliability, the most important characteristics for distinguishing between real pulsars and radio frequency interference proved to be the skewness of the Integrated Pulse Profile and the chi-squared value from double Gaussian fit to Pulse Profile. These results have the potential to enable considerably more efficient observations and have highly extensive implications in related fields where the frontiers of further understanding on phenomena are dependent on such discoveries.