5ª SIEPE SEMANA INTEGRADA ENSINO | PESQUISA | EXTENSÃO
EICTI XII ENCONTRO ANUAL DE INICIAÇÃO CIENTÍFICA E VIII ENCONTRO DE INICIAÇÃO EM DESENVOLVIMENTO TECNOLÓGICO E INOVAÇÃO
UNILA

IDENTIFICAÇÃO DE EXOPLANETAS UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

Bruno Henrique Dourado Macedo[(1)*], Joylan Nunes Maciel[(2)], Willian Zalewski[(3)]

[(1)] Bolsista ITI - FA, Engenharia Física, Instituto Latino-Americano De Ciências Da Vida E Da Natureza (ILACVN), UNILA.

[(2)] Coorientador(a), Instituto Latino-Americano de Tecnologia, Infraestrutura e Território (ILATIT), Universidade Federal da Integração Latino Americana, UNILA.

[(3)] Orientador(a), Instituto Latino-Americano de Tecnologia, Infraestrutura e Território (ILATIT) Universidade Federal da Integração Latino Americana, UNILA.

*E-mail de contato: bhd.macedo.2017@aluno.unila.edu.br

1. RESUMO

Nas últimas décadas, o progresso tecnológico e a redução de custos em equipamentos astronômicos levaram a uma expansão significativa dos recursos de coleta e armazenamento de dados pelos cientistas. Missões espaciais como CoRoT, NuSTAR, NEOWISE, Gaia, Hubble, Kepler, TESS e o mais recente Telescópio Espacial James Web aprimoraram nossa compreensão do universo. Os dados coletados por essas missões, principalmente na forma de curvas de luz, foram essenciais para detectar exoplanetas usando métodos como a técnica de trânsito planetário Em especial, na missão Kepler, 76% dos exoplanetas foram encontrados por meio dessa técnica. No entanto, o acúmulo contínuo de dados temporais, especialmente na forma de curvas de luz, resultou em um grande volume de dados. Como exemplo, o projeto espacial Kepler da NASA, totalizou cerca de 678 GB de dados coletados ao final do projeto. Isso tornou os métodos analíticos tradicionais insuficientes para a eficaz exploração e interpretação dos dados. Nesse contexto, o objetivo desta pesquisa é enfrentar esse desafio, empregando algoritmos de *machine learning* e métodos de representação de séries temporais para a detecção automatizada de exoplanetas. Para atingir esse objetivo, neste estudo, foi conduzida uma avaliação experimental abrangente usando a plataforma de computação de alto desempenho do Google Cloud e as bibliotecas de programação *Python* (*numpy, Lightkurve, Scikit-Learn, pandas*). O conjunto de dados analisado foi construído usando as curvas de luz do catálogo online da *NASA Exoplanet Archive* com 9564 curvas de luz. Considerando os 17 trimestres, cada um dos objetos possui aproximadamente 60 mil pontos de leitura. Os dados foram pré-processados para remoção de ruídos e para redução de dimensionalidade utilizando a representação global de Shallue & Vanderburg (2017)[3]. Ao final da etapa de pré-processamento o conjunto de dados totalizou 5302 curvas de luz, sendo 3107 (58,60%) falsos positivos e 2195 (41,40%) confirmados cada uma com 2001 pontos. Para a construção dos modelos de predição foram utilizados os seguintes algoritmos de transformação da biblioteca *Sktime*: *MINImally RandOm Convolutional KErnel Transform (MiniRocket), Canonical Time-series Characteristics (Catch22)*. O *MiniRocket* é um método desenvolvido exclusivamente para lidar com séries temporais univariadas, esse método usa convoluções de tamanho 9, aplicadas com pesos limitados a dois valores distintos. Ele

emprega um conjunto fixo de 84 convoluções, consistindo em seis convoluções com um peso específico e três convoluções com o segundo peso. O método *Catch22* tem uma abordagem que se baseia em um conjunto de 22 características calculadas a partir das séries temporais. Essas características são projetadas para capturar informações temporais diversas. Para a indução dos modelos os seguintes algoritmos de *machine learning*: *Random Forest Classifier (RF), Multi-layer Perceptron Classifier (MLPClassifier), Naive Bayes (NB)*. Com o intuito de explorar uma melhor combinação dos parâmetros desses algoritmos foi aplicada a função *BayesSearchCV* da biblioteca *scikit-optimize*. A avaliação dos modelos induzidos foi realizada por meio da estratégia *Nested Cross Validation*. Nessa estratégia de avaliação os dados são divididos em *n_splits* partições treino/teste e os experimentos são repetidos *n_repeats* vezes. Assim, em cada repetição são selecionados dados diferentes para cada partição, minimizando assim um possível viés sobre os dados. Os parâmetros utilizados para esta avaliação foram: *cv_outer = RepeatedStratifiedKFold(n_splits=2, n_repeats=5, random_state=1)* totalizando 10 repetições treino/teste e *cv_inner = StratifiedKFold(n_splits=3, shuffle=True, random_state=1)* para a otimização dos parâmetros em cada repetição. Na **Tabela 1** são apresentados os resultados dos experimentos realizados no conjunto de treino e teste em cada repetição, considerando as métricas de desempenho em termos de acurácia média (Acc), desvio padrão (Dp). A partir dos resultados encontrados é possível observar que o modelo *RF* com a transformação *MINIROCKET* e *CATCH22* apresentaram os melhores desempenhos em termos de acurácia. Em trabalhos futuros, temos a intenção de avaliar algoritmos baseados em redes neurais convolucionais (*CNNs*) para séries temporais.

Tabela 1 - Resultado dos experimentos.

Modelo	Transformação	Teste (Acc %)	Teste (Dp %)	Treino (Acc %)	Treino (Dp %)
RF	**CATCH22**	83,05	0,48	82,65	0,58
	MINIROCKET	83,30	0,68	83,35	0,60
NB	**CATCH22**	60,70	0,69	60,70	0,73
	MINIROCKET	69,20	0,41	69,65	0,70
MLP	**CATCH22**	77,50	0,81	76,65	0,84
	MINIROCKET	80,05	0,32	79,05	0,35

Fonte: Autoria própria.

2. REFERÊNCIAS

1. MONTANGER, P. O.; ZALEWSKI, W. **Programa computacional para a identificação automática de exoplanetas.** Revista Brasileira de Iniciação Científica, p. 195-208, abr. 2020. ISSN 2359-232X.

2. ZALEWSKI, W. **Modelagem Simbólica de Padrões Morfológicos para a Classificação de Séries Temporais.** Dissertação (Doutorado) – Universidade Federal do Paraná- UFPR, 2015.

3. SHALLUE, C. J.; VANDERBURG, A. **Identifying exoplanets with deep learning: A five planet resonant chain around Kepler-80 and an eighth planet around Kepler-90**. The Astronomical Journal, 2017. DOI 10.3847/1538-3881/aa9e09

3. AGRADECIMENTOS

Gostaria de agradecer a UNILA por abrir as portas da universidade, à PRPPG/UNILA e à Fundação Araucária/PR pelo seu apoio através da bolsa ITI, ao professor Willian Zalewski pela orientação neste trabalho, aos professores da Engenharia Física e a minha companheira Renata Benedet pelo apoio. Também à PRPPG/UNILA pela promoção de recursos por meio das chamadas 104/2020 e 105/2020.