ORBITALIA: PADRONIZAÇÃO, ESTIMATIVA AUTOMATIZADA DE QUALIDADE E CONTEXTUALIZAÇÃO DE CURVAS DE LUZ PARA ASTRONOMIA COLABORATIVA

[TÍTULO EM INGLÊS]

FELIPE PEREIRA UMPIERRE
Graduando em Ciência da Computação
[e-mail]

NICHOLE DA SILVA NICOLINI
Graduanda em Ciência da Computação
[e-mail]

HOWARD CRUZ ROATTI
Prof.º Msc. Centro Universitário FAESA
howard.cruz@faesa.br

RESUMO

A diversidade de formatos, unidades e condições de aquisição das curvas de luz dificulta sua preparação e avaliação automatizadas em iniciativas de astronomia colaborativa. Este trabalho desenvolveu o núcleo computacional da plataforma Orbitalia, integrando padronização fotométrica e temporal, classificação automatizada da qualidade das curvas e consulta a catálogos astronômicos para complementar o contexto dos objetos observados. Das 449 curvas obtidas no VarAstro, foram selecionadas 353 observações elegíveis e não duplicadas, distribuídas em 297 curvas de treinamento e 56 de teste, sem compartilhamento de estrelas entre os conjuntos. Utilizou-se um modelo Random Forest, composto por 200 árvores de decisão e alimentado por 14 características extraídas das curvas, para estimar o índice de qualidade DQ, de 1 a 5, tomando as notas do VarAstro como referência; no teste, o modelo reproduziu exatamente a classe catalogada em 50,0% das observações, com F1 macro de 0,455 e erro absoluto médio de 0,518 ponto de DQ. O sistema reúne preparação dos dados, estimativa de qualidade e contextualização das observações em um fluxo reproduzível, oferecendo apoio à avaliação de curvas de luz e à organização de contribuições observacionais, embora a precisão da classificação ainda demande aprimoramento.

Palavras-chave: curvas de luz; qualidade de dados; Random Forest; ciência cidadã; astronomia.

ABSTRACT

The diversity of formats, units, and acquisition conditions of light curves complicates their automated preparation and assessment in collaborative astronomy. This study

developed the computational core of the Orbitalia platform, integrating photometric and temporal standardization, automated light-curve quality classification, and astronomical catalogue queries to provide context for the observed objects. Of the 449 light curves obtained from VarAstro, 353 eligible, non-duplicate observations were selected and divided into 297 training and 56 test curves, with no stars shared between the sets. A Random Forest model comprising 200 decision trees and using 14 features extracted from the light curves estimated the quality index DQ, ranging from 1 to 5, with VarAstro ratings as the reference; on the test set, the model exactly reproduced the catalogue class for 50.0% of observations, with a macro F1-score of 0.455 and a mean absolute error of 0.518 DQ points. The system combines data preparation, quality estimation, and observational contextualization in a reproducible workflow, supporting light-curve assessment and the organization of observational contributions, although classification accuracy still requires improvement.

Keywords: *light curves; data quality; random forest; citizen science; astronomy.*

INTRODUÇÃO

A investigação de planetas fora do Sistema Solar envolve técnicas como trânsito, velocidade radial e microlente gravitacional, cujas sensibilidades e limitações são distintas. O trânsito corresponde à diminuição do brilho observado quando um planeta passa diante de sua estrela, na perspectiva do observador; a velocidade radial investiga o movimento da estrela associado à interação gravitacional; e a microlente utiliza a amplificação gravitacional da luz de uma fonte mais distante. Informações complementares podem contribuir para caracterizar sistemas e avaliar hipóteses planetárias, sem que a combinação de todos esses métodos seja obrigatória em cada descoberta (PERRYMAN, 2018).

O acompanhamento fotométrico de trânsitos também admite contribuições de observadores com pequenos telescópios, dependendo do brilho do alvo, da profundidade do evento, da precisão instrumental e das condições de observação. Essas contribuições ampliam a cobertura de eventos conhecidos e auxiliam sua continuidade observacional; a descoberta inicial de um planeta não encerra a necessidade de observá-lo (ZELLEM et al., 2020). A identificação e a validação de determinados candidatos podem demandar campanhas prolongadas, sobretudo em sistemas de período longo. As medições de brilho registradas ao longo do tempo constituem curvas de luz, objeto central deste trabalho.

Entretanto, arquivos produzidos por diferentes observadores podem apresentar unidades, convenções de tempo, nomes de colunas e formatos distintos. Também

podem conter lacunas, medições inválidas e informações instrumentais incompletas. A dificuldade computacional não se restringe a armazenar esses arquivos: é necessário interpretar sua estrutura, preservar sua procedência e produzir representações comparáveis sem apagar características relevantes da observação.

O problema investigado foi como apoiar a preparação e a avaliação da qualidade de curvas de luz produzidas por diferentes observadores. Esses dados podem apresentar formatos, unidades e condições de aquisição distintos, dificultando sua comparação e utilização conjunta. Nesse contexto, a Orbitalia integra o recebimento das observações, a padronização dos dados, a estimativa automatizada de qualidade e a consulta a catálogos astronômicos para complementar as informações sobre o objeto observado.

O objetivo geral foi desenvolver e avaliar o núcleo computacional da plataforma Orbitalia para preparar curvas de luz, estimar sua classe de qualidade e complementar o contexto do objeto observado por consultas a catálogos. Foram definidos os seguintes objetivos específicos:

- organizar arquivos e metadados com rastreabilidade entre observação, curva e rótulo de qualidade;
- implementar interpretação e padronização fotométrica e temporal, registrando hipóteses e impedimentos;
- extrair características numéricas das curvas e treinar um classificador supervisionado de DQ;
- avaliar o desempenho em estrelas separadas do treinamento e examinar o efeito da limpeza adicional;
- disponibilizar o processamento por uma interface HTTP e consultar catálogos sem transferir suas classificações ao modelo;
- documentar limitações, resultados e requisitos de integração com a interface da plataforma.

A hipótese do trabalho foi que uma sequência documentada de preparação permitiria padronizar dados de diferentes origens e apoiar a classificação de qualidade. Também foi avaliado se a remoção adicional de medições inválidas melhoraria a classificação. Essa comparação utilizou curvas já normalizadas, e não arquivos brutos.

A pesquisa sobre exoplanetas depende da continuidade das observações, tanto para acompanhar eventos conhecidos quanto para investigar mudanças ao longo do tempo. Astrônomos amadores podem contribuir com esse acompanhamento, ampliando a cobertura observacional e fornecendo dados complementares aos obtidos por grandes projetos. Essa participação constitui uma contribuição importante da ciência cidadã à astronomia (ZELLEM et al., 2020).

Iniciativas como a Amateur Exoplanet Archive (AXA), conduzida por Bruce Gary, e o Exoplanet Transit Database (ETD), ligado à Sociedade Astronômica Tcheca, demonstram a importância de reunir e disponibilizar observações de diferentes colaboradores. Esses acervos também evidenciam a necessidade de considerar as diferenças de qualidade entre as curvas de luz (PODDANÝ; BRÁT; PEJCHA, 2010).

A Orbitalia propõe apoiar esse processo ao oferecer um ambiente colaborativo para centralizar observações, organizar seus metadados e padronizar os dados recebidos. A classificação automatizada busca ajudar o participante a interpretar a qualidade de sua curva e orientar sua revisão antes do compartilhamento ou de análises posteriores. Dessa forma, a contribuição da plataforma envolve tanto a avaliação das observações quanto a organização de um acervo que facilite seu aproveitamento científico.

A proposta também prevê a formação de uma comunidade ativa, cujas contribuições possam ampliar continuamente esse acervo. Após revisão e atribuição de rótulos confiáveis, novas observações poderão integrar futuros treinamentos do modelo, permitindo avaliar e buscar melhorias na classificação. Assim, a participação dos usuários poderá contribuir para a evolução da base de dados e do sistema.

METODOLOGIA DE DESENVOLVIMENTO

O trabalho envolveu o desenvolvimento de um protótipo e a avaliação inicial de seu desempenho por medidas numéricas. O procedimento foi organizado em seis etapas:

1. Definição do problema e levantamento de referências: adotou-se a classificação da qualidade das curvas, usando as notas DQ do catálogo como respostas para o treinamento.

2. Organização da base: os arquivos VarAstro foram preservados na camada bruta, convertidos para representação tabular e associados aos rótulos por identificador de observação.
3. Preparação das curvas: foram implementadas regras de interpretação de cabeçalhos, conversão de brilho e tempo, propagação de incertezas e identificação de valores inválidos, mantendo relatórios de pendências e hipóteses.
4. Construção do conjunto de aprendizado: extraíram-se 14 características por curva, eliminaram-se duplicações identificadas e dividiram-se as observações por estrela em treinamento e teste.
5. Treinamento e avaliação: treinou-se um modelo Random Forest, calculou-se seu desempenho nas cinco classes DQ e compararam-se os resultados antes e depois da limpeza adicional, mantendo os mesmos conjuntos de treinamento e teste.
6. Integração e demonstração: implementou-se o processamento HTTP, a consulta independente a catálogos e a devolução de resultados e metadados, com registro dos testes realizados.

O treinamento foi realizado localmente em CPU, e o Modal foi utilizado para disponibilização do serviço. A avaliação apresentada corresponde ao artefato versionado `dq_limpo_20261029T040341933863Z`. Os testes de integração abrangeram o processamento e a resposta da API.

FUNDAMENTAÇÃO TEÓRICA

Curvas de luz, fotometria e tempo

Uma curva de luz reúne medições do brilho de uma estrela ao longo do tempo. Cada linha da tabela representa uma medição, e cada curva é um exemplo utilizado pelo modelo. Fluxo e magnitude expressam o brilho de maneiras diferentes: o fluxo aumenta quando a estrela fica mais brilhante, enquanto a magnitude diminui em uma escala logarítmica. Na fotometria diferencial, o brilho da estrela observada é comparado ao de estrelas de referência. Isso ajuda a reduzir efeitos que atingem ambas, como mudanças na transparência da atmosfera.

A notação V-C identifica a diferença entre a magnitude da estrela variável ou alvo e a da comparação. O nome genérico dmag, isoladamente, não documenta qual quantidade foi subtraída da outra. Por isso, a interpretação depende do cabeçalho ou de uma hipótese registrada. Confundir a convenção pode inverter a evolução do brilho convertido. Um valor negativo de magnitude diferencial não é, por si só, inválido.

Datas julianas (JD) representam instantes por uma contagem em dias. HJD e BJD acrescentam correções pelo tempo de viagem da luz, tomando como referência o centro do Sol e o centro de massa do Sistema Solar, respectivamente. UTC e TDB indicam a escala usada para contar o tempo: UTC é a base do horário civil, enquanto TDB é usada em cálculos astronômicos referidos ao Sistema Solar. A conversão para BJD_TDB exige conhecer a escala original, a direção da estrela e a localização da observação. Subtrair uma data de referência apenas cria tempo relativo; não realiza essas correções (ASTROPY PROJECT, [s.d.]b).

Arquivos FITS podem conter imagens, tabelas e cabeçalhos com informações sobre a observação. Uma imagem de estrelas ainda precisa ser processada para medir o brilho e produzir uma curva de luz. A interface deste trabalho aceita FITS com tabelas de medições compatíveis, e não imagens brutas (ASTROPY PROJECT, [s.d.]a).

Qualidade observacional e índice DQ

O índice DQ é um rótulo de qualidade associado à curva inteira. Na convenção adotada, 1 corresponde à melhor classe e 5 à pior; a ordem não implica distâncias físicas iguais entre as classes. O ETD descreve um índice baseado na relação entre dispersão, profundidade do trânsito e densidade das medições. Assim, DQ não é apenas uma medida do ruído instrumental nem uma porcentagem do arquivo que contém informação útil (PODDANÝ; BRÁT; PEJCHA, 2010).

As notas de qualidade DQ foram obtidas do catálogo VarAstro e associadas às respectivas curvas de luz. Com esses exemplos, o modelo aprendeu a estimar a qualidade de novas observações a partir de suas características estatísticas, sem executar diretamente o cálculo utilizado pelo ETD. Os dados podem ter recebido tratamentos diferentes antes do envio ao catálogo, o que deve ser considerado na análise dos resultados.

Uma curva de uma estrela com planeta confirmado pode ter cobertura incompleta ou baixa precisão. Inversamente, uma binária eclipsante pode produzir medições precisas, embora não represente um trânsito planetário. Por esse motivo, os estados confirmado, candidato e falso positivo não foram usados como substitutos dos rótulos DQ. Uma probabilidade de classe DQ não representa uma fração de ruído, e baixa qualidade não comprova ausência de planeta.

Astronomia colaborativa e trabalhos relacionados

Como antecedente histórico, a Amateur Exoplanet Archive (AXA), conduzida por Bruce Gary, reuniu fotometria amadora. Poddaný, Brát e Pejcha (2010) a situam entre os arquivos anteriores ao ETD. Esse precedente contextualiza a motivação colaborativa, sem atribuir ao presente protótipo a substituição dessas iniciativas.

Bases como ETD e VarAstro reúnem observações de diferentes colaboradores e oferecem contexto para acompanhamento fotométrico. A identificação do observador, do objeto, da observação e do equipamento permite reconhecer autoria e investigar diferenças entre as condições de aquisição. A procedência de um arquivo de uma plataforma colaborativa, entretanto, não comprova que ele foi produzido por um observador amador; instrumentos e instituições devem ser considerados individualmente (PODDANÝ; BRÁT; PEJCHA, 2010; HYKEL et al., 2024).

Zellem et al. (2020) discutem o uso de pequenos telescópios operados por cientistas cidadãos no acompanhamento de exoplanetas. Essa contribuição não deve ser confundida com a necessidade de redescobrir objetos já conhecidos: monitoramento e continuidade temporal permanecem relevantes. A Orbitalia foi delimitada como apoio computacional à organização e à triagem dessas observações, sem alegar substituir bases existentes.

No aprendizado aplicado a dados de missões espaciais, Shallue e Vanderburg (2018) desenvolveram uma rede neural para classificação de sinais em dados Kepler, frequentemente associada ao AstroNet. Esse estudo trata da identificação de eventos planetários e não da estimativa do índice DQ, que é o problema abordado aqui. A diferença de objetivo, de representação e de domínio observacional impede comparar diretamente as respectivas acurácias.

Modelo Random Forest e aprendizado supervisionado

No aprendizado supervisionado, cada exemplo de treinamento possui uma resposta conhecida. Neste trabalho, cada curva é representada por 14 características numéricas, e a resposta é sua nota DQ. O modelo Random Forest reúne várias árvores de decisão, treinadas com seleções aleatórias de exemplos e características. A combinação de suas previsões reduz a dependência de uma única árvore (BREIMAN, 2001).

Cada árvore organiza os exemplos por regras, como verificar se a dispersão do brilho supera determinado valor. O critério Gini ajuda a escolher regras que separem melhor as classes: $G = 1 - \sum_k p_k^2$, em que p_k é a proporção de exemplos da classe k no grupo avaliado. Quanto menor o valor, menor a mistura de classes nesse grupo. O modelo calcula a média das probabilidades fornecidas pelas árvores para cada classe e escolhe a de maior valor (SCIKIT-LEARN, [s.d.]c).

A escolha do modelo Random Forest considerou o tamanho relativamente pequeno da base, sua capacidade de aprender relações complexas entre as características e a possibilidade de treinamento em CPU. Seu desempenho foi comparado à estratégia simples de prever sempre a classe mais frequente. Essa comparação não estabelece superioridade sobre outros algoritmos.

Métricas de avaliação

A matriz de confusão compara as notas do catálogo com as previsões do modelo. As linhas representam o DQ catalogado e as colunas, o DQ previsto; os acertos aparecem na diagonal. Nas fórmulas abaixo, C representa essa matriz e N, o total de curvas. Para cada classe k, TP_k conta os acertos, FP_k conta as curvas de outras classes classificadas como k e FN_k conta as curvas de k classificadas incorretamente em outras classes.

Tabela 1 - Métricas utilizadas

Métrica	Definição	Interpretação
Acurácia	$\sum_k C_{kk}/N$	Proporção de DQs exatamente iguais ao catálogo.
Precisão da classe k	$TP_k/(TP_k + FP_k)$	Entre as previsões k, quantas correspondem a k.
Sensibilidade (recall) da classe k	$TP_k/(TP_k + FN_k)$	Entre as curvas que realmente são k, proporção classificada corretamente.
F1 da classe k	$2 \times$ precisão $\times$ sensibilidade$/($precisão $+$ sensibilidade$)$	Combina precisão e sensibilidade em uma medida.
Média macro	média simples das métricas das cinco classes	Dá o mesmo peso a cada classe.
MAE	média de $\lvert$DQ estimado $-$ DQ catalogado$\rvert$	Distância média em pontos de DQ.
Previsão com diferença de até um ponto	número de previsões com diferença ≤ 1 / N	Inclui acertos exatos e previsões com erro de um ponto.

Fonte: elaboração dos autores; dados do experimento e da implementação.

A sensibilidade, também chamada de *recall*, indica quantas curvas de uma classe foram classificadas corretamente. Se o modelo acertar três de dez curvas DQ 2, a sensibilidade dessa classe será de 30%. A acurácia balanceada é a média dessas taxas nas cinco classes, dando a mesma importância a cada uma, mesmo quando há quantidades diferentes de exemplos. Já o erro absoluto médio (MAE) indica quantos pontos, em média, separam a nota prevista da nota do catálogo. Por exemplo, prever DQ 3 para uma curva DQ 2 representa um erro de um ponto. Essa diferença numérica não significa que a mudança física de qualidade entre DQ 1 e 2 seja igual à mudança entre DQ 4 e 5 (SCIKIT-LEARN, [s.d.]b).

IMPLEMENTAÇÃO E PROTOCOLO EXPERIMENTAL

Organização dos dados e rastreabilidade

A organização separou arquivos originais, dados processados e arquivos de treinamento. As tabelas foram convertidas para Parquet para facilitar sua leitura. O conversor também preservou cabeçalhos e textos de origem nos casos previstos pelas suas regras. Os arquivos originais foram mantidos para consulta e reprocessamento,

pois a tabela convertida não substitui necessariamente todo o conteúdo de um FITS ou TXT.

Um índice relaciona `observation_id`, `star_name`, `parquet_path` e `dq`. Cada linha desse índice aponta para uma curva, e não para uma medição. No Parquet da curva, cada linha contém uma medição de tempo, brilho e, quando disponível, incerteza. Separar o rótulo global da tabela de medições evita sua repetição e permite rastrear erros de pareamento.

Tabela 2 - Composição e seleção da base

Etapa	Quantidade	Significado
Arquivos VarAstro brutos	449	Universo organizado nesta versão do projeto.
Curvas com brilho e tempo normalizados	355	Disponíveis antes da exclusão de duplicações.
Pendências anteriores à elegibilidade	94	Não incorporadas ao treinamento; não tratadas como curvas ruins.
Duplicações excluídas	2	Identificadores 89704 e 107262.
Curvas efetivamente usadas	353	Uma amostra de aprendizado por curva.
Treinamento	297 curvas / 36 estrelas	Ajuste do imputador e do classificador.
Teste	56 curvas / 9 estrelas	Avaliação em estrelas ausentes do treinamento.

Fonte: elaboração dos autores; dados do experimento e da implementação.

Uma curva com formato difícil de interpretar não é necessariamente uma curva de baixa qualidade. Por isso, as exclusões podem afetar quais tipos de observação estão representados no treinamento. As 353 curvas não formam uma seleção aleatória de todo o VarAstro. A base inclui observações terrestres de diferentes origens, não apenas amadoras.

Padronização do brilho e das incertezas

A interpretação foi baseada em cabeçalhos e metadados, e não exclusivamente na posição ou na aparência numérica das colunas. Regras específicas foram adicionadas para nomes de tempo, magnitude, fluxo e erro. Quando existiam vários alvos ou convenções contraditórias, a escolha exigia identificação ou hipótese

explícita; o fato de uma coluna conter números próximos de 1 não prova que sua unidade é fluxo.

Para magnitude m ou magnitude diferencial na convenção alvo menos comparação, utilizou-se a transformação:

$$F_{\mathrm{rel}} = 10^{-0,4\,(m-m_{\mathrm{ref}})}. \tag{1}$$

A referência m_{ref} foi tomada como mediana das medições válidas da curva. Para fluxos já existentes, dividiu-se o fluxo pela referência de normalização e aplicou-se o mesmo fator ao erro. Essas operações tornam as escalas comparáveis, mas não removem automaticamente tendências atmosféricas nem isolam a contribuição de um trânsito.

A incerteza em magnitude foi propagada, em aproximação linear, por:

$$\sigma_F = \frac{\ln(10)}{2,5}\, F_{\mathrm{rel}}\, \sigma_m. \tag{2}$$

Essa expressão trata a referência como fixa e não inclui a incerteza de sua estimação. Uma incerteza ausente ou nula não foi interpretada como medição perfeita. Nos casos dmag sem convenção completa, hipóteses foram registradas; não se presume que o nome dmag resolva sozinho a direção da subtração. A conversão não produziu novas informações observacionais nem recuperou medições inexistentes.

Padronização temporal e limpeza

O processamento temporal buscou expressar as medições em BJD_TDB com recursos do Astropy. Foram considerados o padrão original, as coordenadas do alvo e a localização do observador. Para HJD, a correção heliocêntrica precisa ser desfeita antes da aplicação da correção baricêntrica correspondente; substituir a etiqueta HJD por BJD seria incorreto.

Parte dos arquivos históricos exigiu hipóteses sobre a escala UTC ou aproximações quanto à localização. Essas decisões foram registradas nos relatórios e constituem limitação de precisão temporal. Um arquivo de saída em BJD_TDB não prova, sozinho, que o relógio ou os metadados originais estavam corretos. Para as características do classificador, foram utilizados intervalos relativos em minutos, não o valor absoluto da data como identificador do objeto.

A limpeza adicional removeu linhas com sentinelas reconhecidas, como 99.99999 e 99999 nos campos pertinentes, tempos ou fluxos não finitos e fluxos não positivos. Sentinelas de incerteza e erros ausentes foram tratados especificamente, sem transformar ausência de informação em erro zero. A interpretação de uma sentinela deve respeitar o campo, pois o mesmo número pode ser legítimo em outro contexto.

Nas 355 curvas preparadas antes da exclusão das duplicações, havia 57.317 medições. Foram removidas 400 linhas, restando 56.917; seis curvas perderam mais de 10% das linhas. Esses totais descrevem a etapa de limpeza das 355 curvas e não devem ser confundidos com a contagem exata de linhas do conjunto final de 353 amostras. A fração inválida original foi preservada como informação de procedência para o cálculo de características.

Não foram aplicadas indiscriminadamente suavização, interpolação de lacunas, retirada estatística de pontos discrepantes ou correção de massa de ar. Tais operações poderiam esconder problemas observacionais relevantes à tarefa de qualidade. AIRMASS pode ser preservada para inspeção, mas não integra as 14 características usadas neste modelo.

Características usadas pelo classificador

O extrator calcula uma representação de tamanho fixo por curva, permitindo trabalhar com séries que possuem diferentes números de medições. A versão utilizada do conjunto de características foi a 2.

Tabela 3 - Características do modelo

Característica	Informação representada
n_pontos	Número de medições válidas.
fracao_invalida	Fração de medições inválidas registrada no processamento.
duracao_min	Intervalo entre primeiro e último instante, em minutos.
cadencia_mediana_min	Mediana dos intervalos entre medições.
maior_lacuna_min	Maior intervalo entre medições consecutivas.
fracao_tempos_repetidos	Proporção de instantes duplicados.
dispersao_fluxo	Desvio-padrão do fluxo normalizado.
mad_fluxo	Desvio absoluto mediano em relação à mediana do fluxo.
amplitude_p95_p05	Diferença entre percentis 95 e 5 do fluxo.
dispersao_diferencas	Dispersão de diferenças consecutivas, ajustada por raiz de 2.
correlacao_vizinhos	Correlação entre medições consecutivas.
inclinacao_por_min	Inclinação linear do fluxo em função do tempo relativo.
erro_mediano	Mediana das incertezas válidas do fluxo.
fracao_sem_erro	Fração de medições sem incerteza utilizável.

Fonte: elaboração dos autores; dados do experimento e da implementação.

Essas características resumem a quantidade e a distribuição das medições no tempo, as variações de brilho e os erros informados. Elas não identificam a causa de cada variação, que pode ser astronômica, instrumental ou atmosférica. Nome da estrela, observador, equipamento e informações sobre planetas não são entradas do modelo.

Separação dos dados e configuração do modelo

As curvas foram separadas por estrela com GroupShuffleSplit. Foram reservadas 20% das estrelas para teste e usado o valor fixo 42 para permitir a repetição do sorteio. Como cada estrela possui quantidades diferentes de curvas, a divisão resultou em 84,14% das curvas no treinamento e 15,86% no teste. A avaliação utilizou esses dois conjuntos e a configuração do modelo apresentada a seguir (SCIKIT-LEARN, [s.d.]a).

A separação por estrela reduz o risco de resultados excessivamente otimistas por presença do mesmo alvo nos dois conjuntos. Ela não garante separação por observador, instrumento ou campanha. Não se pode, portanto, interpretar o resultado principal como garantia de desempenho em qualquer novo observador ou telescópio.

Tabela 4 - Distribuição de classes

DQ	Treinamento	Teste
1	46	9
2	59	10
3	96	18
4	65	14
5	31	5
Total	297	56

Fonte: elaboração dos autores; dados do experimento e da implementação.

Valores ausentes nas características foram preenchidos com a mediana calculada no treinamento, usando SimpleImputer. Também foram criados indicadores para registrar quais valores estavam ausentes, mantendo as colunas sem dados. Em seguida, treinou-se RandomForestClassifier com a configuração abaixo. O modelo salvo foi treinado com 297 curvas; as outras 56 permaneceram separadas para teste.

Tabela 5 - Configuração do modelo Random Forest

Parâmetro	Valor
`n_estimators` / `criterion`	200 / gini
`max_depth`	None
`min_samples_split` / `min_samples_leaf`	2 / 3
`max_features` / `bootstrap`	sqrt / True
`class_weight`	balanced_subsample
`random_state` / `n_jobs`	42 / 2

Fonte: elaboração dos autores; dados do experimento e da implementação.

O balanceamento dá maior peso às classes menos frequentes, sem criar novas observações. Cada folha, ou grupo final de uma árvore, deve conter pelo menos três exemplos. Essa restrição ajuda a evitar regras baseadas em casos isolados, mas não elimina o risco de o modelo aprender detalhes específicos do treinamento e errar em curvas novas. A configuração foi avaliada neste experimento, sem busca pela melhor combinação possível de parâmetros.

Ferramentas e execução

Python foi utilizado na preparação, no aprendizado e na integração. Pandas e NumPy apoiaram operações tabulares e numéricas; PyArrow foi utilizado para Parquet; Astropy para estruturas astronômicas e tempo; scikit-learn para imputação, separação e classificação; e joblib para persistência do modelo. A avaliação salva registra Python 3.14.7, scikit-learn 1.9.1, pandas 3.0.6, NumPy 2.5.4 e Astropy 8.0.1. Esses números descrevem o ambiente registrado do experimento e não uma exigência universal da solução.

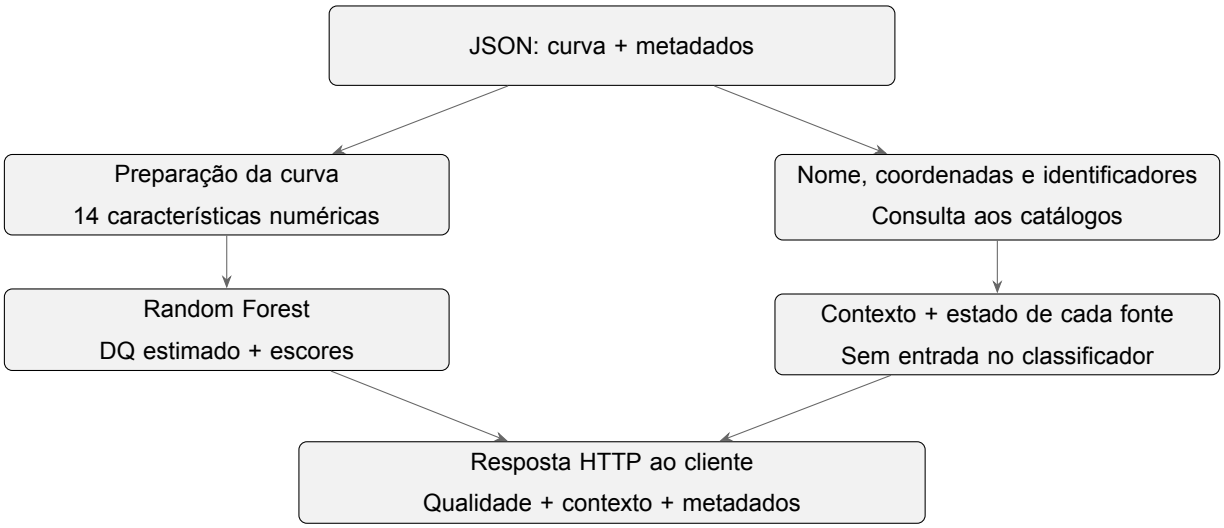
O serviço foi estruturado com funções Python disponibilizadas no Modal e interface HTTP compatível com FastAPI. O processamento online reutiliza a preparação e o extrator de características do treinamento; a execução remota não implica novo treinamento a cada envio. Os resultados foram mantidos em artefatos de avaliação para permitir comparação e auditoria.

INTEGRAÇÃO, CATÁLOGOS E DEMONSTRAÇÃO

Fluxo de recebimento e resposta

O sistema cliente envia um JSON com o identificador da observação, informações do objeto, local, equipamento e dados da curva. O CSV é enviado como texto. Para FITS, o arquivo é codificado em Base64, uma forma de representar seu conteúdo como texto no JSON, respeitando o limite de tamanho do leitor. O FITS precisa conter uma tabela de medições; imagens brutas exigem uma etapa anterior de extração do brilho.

Figura 1 - Separação entre avaliação da curva e contexto catalográfico



Fonte: elaboração dos autores a partir da arquitetura implementada.

Após validação, o serviço interpreta as colunas, prepara brilho e tempo, verifica medições válidas e extrai as características. O classificador devolve DQ estimado e escores por classe. A interface exige pelo menos dez medições utilizáveis e duração positiva como condições operacionais mínimas; esses critérios não certificam que um trânsito esteja bem coberto.

Metadados de telescópio, câmera e local são preservados como informação complementar. Quando presentes no formulário e no cabeçalho FITS, suas origens devem permanecer distinguíveis, sem que o serviço presuma equivalência ou veracidade automática. Esses campos não são resultados inferidos pelo classificador nem entradas das suas 14 características.

A saída é devolvida diretamente na resposta HTTP ao sistema cliente. Os testes com Postman verificaram o recebimento da observação, o processamento e a devolução dos resultados nesse fluxo.

Consulta a catálogos separada da classificação

O cruzamento de catálogos complementa a observação com informações sobre o objeto, mesmo quando não há planeta confirmado. O módulo consulta SIMBAD para identificação e informações da estrela, Gaia DR3 quando há um identificador correspondente e NASA Exoplanet Archive para informações sobre planetas e candidatos. Cada resultado mantém a identificação de sua fonte; divergências

entre catálogos não são resolvidas automaticamente (WENGER et al., 2000; GAIA COLLABORATION et al., 2023; NASA EXOPLANET ARCHIVE, [s.d.]).

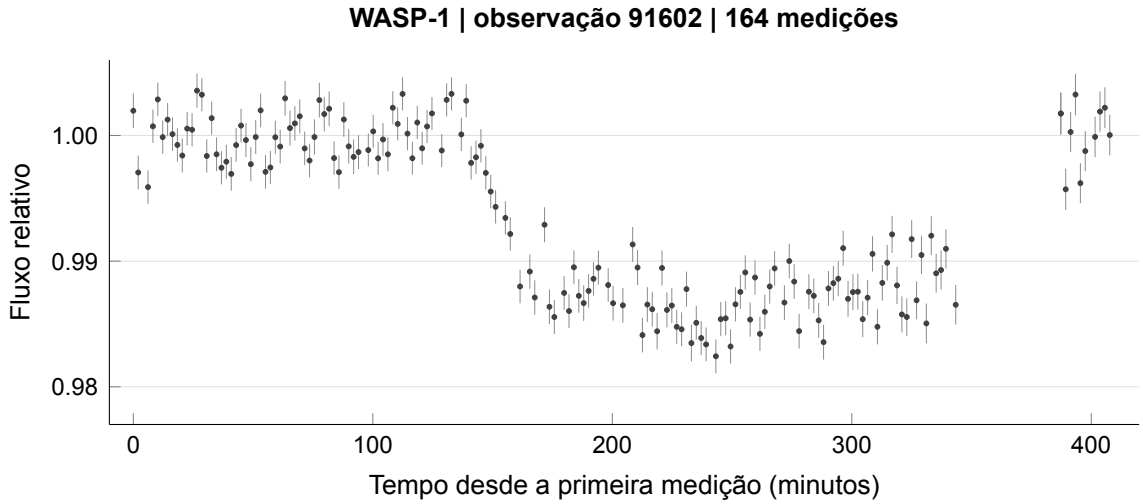
As consultas utilizam nome, identificadores e posição no céu, quando disponíveis. Dois registros próximos no céu não são necessariamente o mesmo objeto: é preciso considerar a data das coordenadas, o movimento da estrela e a presença de estrelas vizinhas. A consulta fornece contexto, mas não garante a identificação em todos os casos. Seus resultados não alteram a previsão de qualidade nem substituem a nota DQ do catálogo.

Cada serviço pode retornar sucesso, ausência de correspondência ou erro. Uma falha de rede, como as indisponibilidades registradas em consultas Gaia, não significa que o objeto não exista. Da mesma forma, ausência em uma tabela de planetas confirmados não demonstra ausência de planetas. A resposta preserva estados parciais em vez de interromper ou inventar contexto.

Demonstração com uma observação reservada

A observação VarAstro 91602, de WASP-1, foi reservada para demonstração de inferência e não integrou o treinamento nem o conjunto principal de teste. Ela contém 164 medições e foi atribuída a Manfred Raetz nos metadados de origem. O resultado registrado foi DQ estimado 2, correspondente ao DQ 2 do catálogo; o escore da classe foi aproximadamente 0,6412.

Figura 2 - Observação 91602 utilizada na demonstração



Pontos e barras representam fluxo e incerteza presentes no CSV de envio; o tempo é relativo ao primeiro registro. Não foi ajustado um modelo físico de trânsito nesta figura.

Fonte: fotometria de Manfred Raetz, VarAstro; representação dos autores.

Tabela 6 - Resultado da demonstração reservada

Campo da demonstração	Resultado registrado
Identificador de envio	teste-reservado-91602
Estado do processamento	concluido
Medições	164
DQ estimado / DQ catalogado	2 / 2
Escore da classe estimada	0,6412

Fonte: elaboração dos autores; dados do experimento e da implementação.

O valor de 64,12% é a probabilidade atribuída pelo modelo à classe DQ 2. Ele não representa a porcentagem de qualidade ou ruído na curva, nem a probabilidade de existir um planeta. Também não foi verificado se valores como esse correspondem à frequência real de acertos em novos dados. Como outras curvas da mesma estrela podem estar no treinamento, esta demonstração não substitui o teste com estrelas separadas.

A demonstração compreendeu o recebimento CSV e a resposta em nuvem, além da leitura local de uma tabela FITS compatível. O fluxo e os resultados registrados da API documentam esses testes; a validação do formato FITS apresentada neste estudo está restrita ao ambiente local.

RESULTADOS E DISCUSSÃO

Desempenho no conjunto de teste

No conjunto de 56 curvas de nove estrelas separadas do treinamento, o modelo acertou exatamente o DQ em 28 observações. O resultado principal foi, portanto, acurácia de 50,0%, e não uma medida da proporção de ruído em cada curva. As métricas a seguir foram extraídas da avaliação salva, sem novo ajuste para esta revisão textual.

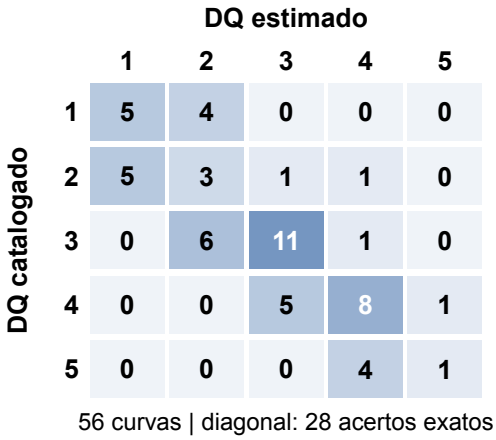
Tabela 7 - Métricas no conjunto de teste

Métrica	Resultado
Acurácia exata	50,00% (28/56)
Acurácia balanceada	44,76%
Precisão macro	48,99%
Sensibilidade macro (recall)	44,76%
F1 macro	45,46%
Erro absoluto médio	0,518 ponto de DQ
Previsão com diferença de até um ponto	98,21% (55/56)
Referência: prever sempre DQ 3	32,14% (18/56)

Fonte: elaboração dos autores; dados do experimento e da implementação.

A estratégia de comparação prevê sempre DQ 3, a classe mais frequente no treinamento. O modelo obteve 17,86 pontos percentuais a mais de acurácia no teste. Não foi realizado um teste estatístico para avaliar se essa diferença se mantém além da amostra estudada. Em 98,21% das curvas, a previsão ficou a até um ponto da nota do catálogo, mas isso não corresponde à acurácia exata: prever DQ 4 para uma curva DQ 5 ainda é um erro.

Figura 3 - Matriz de confusão



Linhas representam DQ catalogado e colunas, DQ estimado.
Fonte: avaliação do modelo `dq_limpo_20261029T040341933863Z`, elaborada pelos autores.

Tabela 8 - Desempenho por classe

DQ	Precisão	Sensibilidade (recall)	F1	Curvas no teste
1	0,500	0,556	0,526	9
2	0,231	0,300	0,261	10
3	0,647	0,611	0,629	18
4	0,571	0,571	0,571	14
5	0,500	0,200	0,286	5

Fonte: elaboração dos autores; dados do experimento e da implementação.

A classe DQ 2 apresentou baixa precisão e sensibilidade. Cinco curvas DQ 2 receberam previsão DQ 1, e quatro curvas DQ 5 receberam previsão DQ 4. Nesses casos, o modelo indicou qualidade melhor que a registrada no catálogo, embora tenha errado por apenas um ponto. Como havia somente cinco curvas DQ 5 no teste, cada acerto representa 20 pontos percentuais na sensibilidade dessa classe.

A matriz permite identificar padrões de erro, mas não atribuir automaticamente causas físicas a cada falha. Não se conclui, apenas a partir de uma classificação incorreta, que houve nuvens, uma binária eclipsante ou problema de guiagem. Essa atribuição exigiria inspeção específica e informação adicional.

Comparação antes e depois da limpeza adicional

O modelo foi avaliado antes e depois da remoção adicional de medições inválidas, mantendo os mesmos conjuntos de treinamento e teste. O histórico da limpeza foi usado para calcular a proporção de medições inválidas de cada curva. As notas DQ do catálogo permaneceram como referência. As duas versões já estavam normalizadas, e produziram as mesmas métricas e matrizes de confusão no teste.

Tabela 9 - Comparação da limpeza adicional

Métrica	Antes da limpeza adicional	Depois	Diferença
Acurácia	50,00%	50,00%	0,00 p.p.
Precisão macro	48,99%	48,99%	0,00 p.p.
Sensibilidade macro (recall)	44,76%	44,76%	0,00 p.p.
F1 macro	45,46%	45,46%	0,00 p.p.
MAE em DQ	0,518	0,518	0,000

Fonte: elaboração dos autores; dados do experimento e da implementação.

A limpeza retirou medições inválidas, mas não alterou as métricas de classificação neste teste. Isso não significa que a preparação dos dados seja desnecessária: as duas versões comparadas já estavam normalizadas. O resultado se refere apenas à limpeza adicional.

Como o mesmo conjunto de teste foi consultado durante a comparação, a análise deve ser considerada exploratória. Uma confirmação independente exigiria novos dados separados antes de qualquer decisão metodológica. A avaliação do modelo Random Forest foi baseada nas métricas de classificação e na matriz de confusão apresentadas.

Limitações e estado do protótipo

As limitações principais incluem o tamanho e a seleção da base, o desequilíbrio de classes, as curvas excluídas por dificuldade de interpretação, as hipóteses sobre convenções fotométricas e temporais e a ausência de validação independente em outros instrumentos ou observadores. A referência de qualidade é catalográfica; seu uso não garante uniformidade de todos os tratamentos anteriores realizados pelos colaboradores.

O classificador aprende a distinguir as cinco classes DQ, mas não foi configurado para considerar diretamente sua ordem, da melhor à pior qualidade. As características usadas também não incluem a profundidade e a cobertura obtidas pelo ajuste de um modelo físico de trânsito. Assim, o sistema estima as notas a partir das características das curvas, sem determinar a causa de cada ruído.

Tabela 10 - Estado dos componentes

Componente	Evidência e limite atual
Preparação e treinamento	Implementados e avaliados no conjunto descrito.
Recebimento CSV em nuvem	Teste de requisição e resposta registrado.
Leitura FITS tabular	Implementada e testada localmente; validação em nuvem a completar.
Cruzamento de catálogos	Implementado com estados parciais; dependente dos serviços externos.
Metadados de equipamento/local	Informação complementar, sem entrada no classificador.
Interface, persistência e callback externo	Integração final e verificação dependem do sistema cliente.
Segurança e operação de produção	Autenticação, limites e testes operacionais ainda devem ser consolidados.

Fonte: elaboração dos autores; dados do experimento e da implementação.

Novos envios ampliam o acervo, mas não entram automaticamente no treinamento. Antes disso, precisam ser revisados, receber notas confiáveis e passar pela verificação de duplicações. Cada nova versão do modelo deve ser registrada e avaliada em dados separados para verificar se houve melhora.

CONCLUSÃO

Foi desenvolvido o núcleo computacional da Orbitalia para preparar curvas de luz, estimar classes DQ e consultar contexto catalográfico de forma separada. A organização por observação e a preservação de procedência permitiram associar tabelas, metadados, transformações e rótulos. O escopo final é avaliação automatizada de qualidade, não classificação binária de exoplanetas nem quantificação percentual de sinal e ruído.

O modelo supervisionado utilizou 353 curvas VarAstro elegíveis, com 297 no treinamento e 56 no teste separado por estrela. A acurácia exata foi de 50,0%, com

F1 macro de 0,455. Embora tenha superado a referência majoritária e apresentado erros geralmente próximos na escala DQ, o desempenho por classe evidencia a necessidade de revisão humana e de avaliação mais ampla antes de decisões automáticas de aceitação ou rejeição.

A preparação viabilizou o fluxo computacional descrito, mas a hipótese de ganho após a limpeza adicional não foi confirmada: a diferença de acurácia foi de zero ponto percentual. A conclusão se limita ao experimento realizado, sem extrapolação para uma comparação integral entre dados brutos e tratados.

Como trabalhos futuros, propõem-se ampliar a base com curvas de qualidade verificada, testar o modelo com observadores e instrumentos separados e avaliar métodos que considerem a ordem das notas DQ. Também é importante verificar a confiabilidade das probabilidades estimadas e incluir características relacionadas à cobertura do trânsito. Na integração, devem ser concluídos o acordo sobre os dados trocados com o sistema cliente, os testes FITS em nuvem, o armazenamento e a autenticação. A finalidade permanece apoiar a avaliação e a organização colaborativa das observações.

REFERÊNCIAS

ASTROPY PROJECT. **FITS File Handling (astropy.io.fits)**. [s.d.]. Documentação online. Disponível em: https://docs.astropy.org/en/stable/io/fits/index.html. Acesso em: 1 nov. 2026.

ASTROPY PROJECT. **Time and Dates (astropy.time)**. [s.d.]. Documentação online. Disponível em: https://docs.astropy.org/en/stable/time/index.html. Acesso em: 1 nov. 2026.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, p. 5–32, 2001. DOI: 10.1023/A:1010933404324.

GAIA COLLABORATION et al. Gaia Data Release 3: Summary of the content and survey properties. **Astronomy & Astrophysics**, v. 674, art. A1, 2023. DOI: 10.1051/0004-6361/202243940.

HYKEL, J. et al. **VarAstro Database**. 2024. Base de dados. Forma de citação indicada pela administração da base; não se trata de artigo de periódico. Disponível em: https://var.astro.cz/en/. Acesso em: 1 nov. 2026.

NASA EXOPLANET ARCHIVE. **Retrieving Exoplanet Archive Data With Table Access Protocol**. [s.d.]. Documentação online. Disponível em: https://exoplanetarchive.ipac.caltech.edu/docs/TAP/usingTAP.html. Acesso em: 1 nov. 2026.

PERRYMAN, M. **The Exoplanet Handbook**. 2. ed. Cambridge: Cambridge University Press, 2018.

PODDANÝ, S.; BRÁT, L.; PEJCHA, O. Exoplanet Transit Database. Reduction and processing of the photometric data of exoplanet transits. **New Astronomy**, v. 15, p. 297–301, 2010. DOI: 10.1016/j.newast.2009.09.001.

SCIKIT-LEARN. **GroupShuffleSplit**. [s.d.]. Documentação online. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GroupShuffleSplit.html. Acesso em: 1 nov. 2026.

SCIKIT-LEARN. **Metrics and scoring: quantifying the quality of predictions**. [s.d.]. Documentação online. Disponível em: https://scikit-learn.org/stable/modules/model_evaluation.html. Acesso em: 1 nov. 2026.

SCIKIT-LEARN. **RandomForestClassifier**. [s.d.]. Documentação online. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Acesso em: 1 nov. 2026.

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**, v. 155, art. 94, 2018. DOI: 10.3847/1538-3881/aa9e09.

WENGER, M. et al. The SIMBAD astronomical database. **Astronomy and Astrophysics Supplement Series**, v. 143, p. 9–22, 2000. Disponível em: https://arxiv.org/abs/astro-ph/0002110. Acesso em: 1 nov. 2026.

ZELLEM, R. T. et al. Utilizing Small Telescopes Operated by Citizen Scientists for Transiting Exoplanet Follow-up. **Publications of the Astronomical Society of the Pacific**, v. 132, art. 054401, 2020. DOI: 10.1088/1538-3873/ab7ee7.

AGRADECIMENTOS E PROCEDÊNCIA

Reconhece-se a contribuição da Variable Stars and Exoplanet Section of the Czech Astronomical Society e dos observadores que disponibilizaram fotometria por meio do VarAstro e do ETD. Os identificadores de observação e os metadados de autoria foram preservados nos arquivos e manifestos do projeto. O caso de demonstração 91602, de WASP-1, é atribuído a Manfred Raetz. A utilização e redistribuição dos arquivos devem respeitar as permissões e a citação dos respectivos observadores; a centralização não transfere a autoria dos dados aos desenvolvedores da plataforma.