UNIVERSIDADE FEDERAL DE OURO PRETO

KASSIO RODRIGUES FERREIRA 19.1.8139
MATEUS COSTA 20.1.8025

MINERAÇÃO DE DADOS: CLASSIFICAÇÃO DE PULSARES

CSI605 - SISTEMAS DE APOIO À DECISÃO

João Monlevade
2025

1. Introdução

Este projeto tem como objetivo a aplicação de técnicas de mineração de dados para resolver um problema de classificação binária: distinguir entre pulsares reais e falsos positivos em sinais de rádio captados por telescópios.

A base de dados utilizada é a **HTRU2 (High Time Resolution Universe Survey)**, disponível no repositório da UCI Machine Learning. A tarefa consiste em prever, a partir de estatísticas extraídas dos sinais, se uma amostra representa ou não um pulsar.

- Link do dataset: https://archive.ics.uci.edu/dataset/372/htru2

O modelo de classificação adotado foi o ***Random Forest Classifier*** disponível via biblioteca python *Scikit-learn*. Na fase de treinamento foi adotada a validação cruzada em conjunto com a otimização de hiperparâmetros do modelo via ***GridSearchCV***.

1.1 Sobre o Dataset HTRU2

O conjunto de dados HTRU2 foi criado a partir de sinais de rádio captados por telescópios, com o objetivo de identificar pulsares – estrelas de nêutrons que emitem feixes de radiação de forma periódica, como um farol no espaço.

Detectar esses sinais é desafiador, pois a maioria dos registros captados é composta por ruídos ou interferências, vindos de transmissões humanas ou outros fenômenos cósmicos. O grande desafio é diferenciar os sinais legítimos (pulsares reais) dos falsos positivos.

Para isso, cada amostra do dataset é descrita por estatísticas extraídas de duas representações do sinal de rádio captado pelos telescópios:

- **Perfil Integrado:** mostra como o sinal varia ao longo do tempo.
- **Curva DM-SNR:** mostra como o sinal se comporta ao variar o valor da dispersão (um fenômeno causado pela presença de partículas entre a estrela e a Terra).

De cada uma dessas curvas, foram extraídas quatro medidas estatísticas que ajudam a caracterizar o sinal:

- Média: valor médio da curva; sinais de pulsares tendem a ter médias mais elevadas em função da presença de picos regulares.
- Desvio padrão: mostra o quanto o sinal oscila em torno da média; pulsares costumam apresentar maior variação por conta da repetição de pulsos intensos.
- Curtose: detecta picos acentuados; sinais de pulsares geralmente exibem curtose alta

devido aos pulsos bem definidos.

- Assimetria: avalia se o sinal é balanceado ou inclinado para um dos lados; em muitos casos, pulsares apresentam assimetria positiva ou negativa dependendo do formato do pulso captado.

1.2 Características gerais do Dataset

- Total de amostras: 17.898
- Classes:
    - 0: Não-pulsar (16.259 amostras)
    - 1: Pulsar (1639 amostras)

1.2.1 Atributos

Cada amostra do dataset tem 8 atributos que representam medidas estatísticas descritivas dos sinais baseadas nas duas curvas extraídas do sinal original: **curva integrada** e **curva DM-SNR**.

Para cada tipo de curva, foram extraídas 4 estatísticas (os atributos foram renomeados para português visando facilitar a análise):

- Perfil Integrado:
    - Média, Desvio Padrão, Curtose, Assimetria
- Curva DM-SNR:
    - Média, Desvio Padrão, Curtose, Assimetria

2. Preparação dos dados

Antes da aplicação dos algoritmos de classificação, foi realizada uma etapa de pré-processamento dos dados com o objetivo de garantir a qualidade e a consistência do conjunto de dados. Essa etapa incluiu análise exploratória das variáveis, remoção de atributos altamente correlacionados, verificação de valores ausentes, normalização das features e balanceamento das classes por meio de undersampling com a técnica *NearMiss*.

2.1 Correlação entre os atributos

As Figuras 1 e 2 mostram o levantamento da correlação entre os atributos, onde é possível observar algumas features altamente correlacionadas. Nesta implementação, optamos por remover duas destas features: assimetria_perfil_integrado e assimetria_curva_dm_snr por apresentarem correlação superior a 90% com outras variáveis.

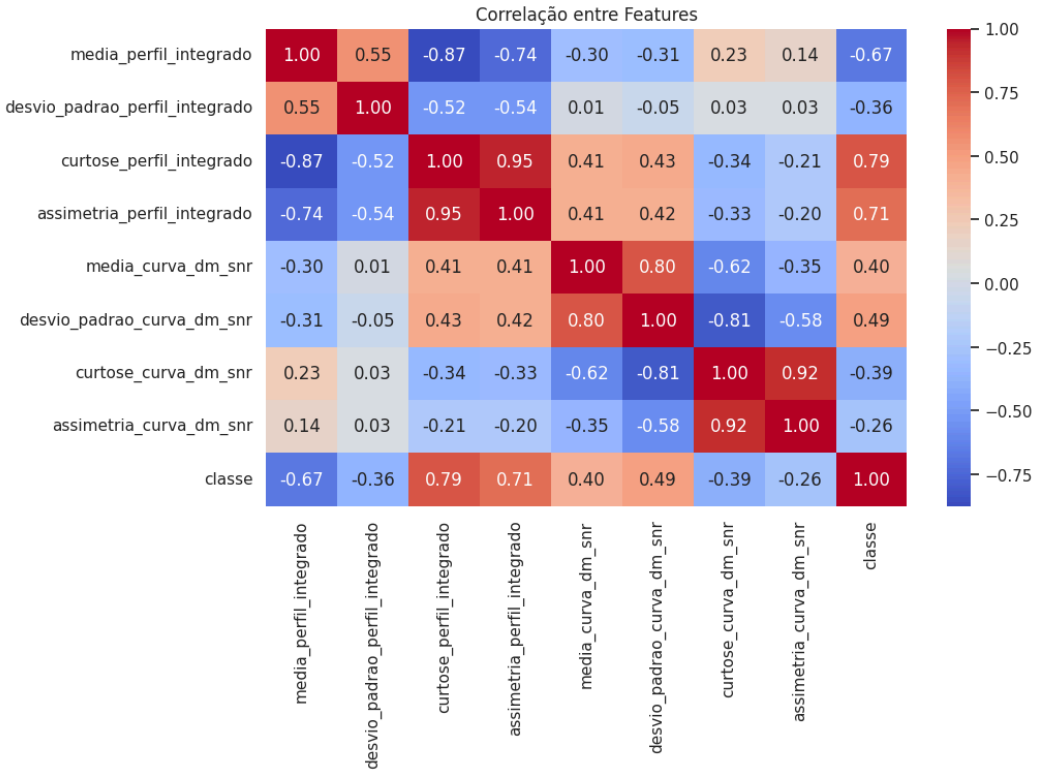

Figura 1: Matriz de correlação dos atributos

Correlações acima de 50% entre features:

Variável 1	Variável 2	Correlação (%)
curtose_perfil_integrado	assimetria_perfil_integrado	94.5729
curtose_curva_dm_snr	assimetria_curva_dm_snr	92.3743
media_perfil_integrado	curtose_perfil_integrado	-87.3898
desvio_padrao_curva_dm_snr	curtose_curva_dm_snr	-80.9786
media_curva_dm_snr	desvio_padrao_curva_dm_snr	79.6555
curtose_perfil_integrado	classe	79.1591
media_perfil_integrado	assimetria_perfil_integrado	-73.8775
assimetria_perfil_integrado	classe	70.9528
media_perfil_integrado	classe	-67.3181
media_curva_dm_snr	curtose_curva_dm_snr	-61.5971
desvio_padrao_curva_dm_snr	assimetria_curva_dm_snr	-57.58
media_perfil_integrado	desvio_padrao_perfil_integrado	54.7137
desvio_padrao_perfil_integrado	assimetria_perfil_integrado	-53.9793
desvio_padrao_perfil_integrado	curtose_perfil_integrado	-52.1435

Figura 2: Correlação dos atributos acima de 50%

2.2 Normalização dos atributos

Por se tratar de medidas estatísticas, os atributos são todos numéricos, mas contendo, em algumas colunas valores altos, e em outras colunas valores relativamente baixos. Por tanto, foi aplicada uma normalização com a ferramenta *StandarScaler* (da lib scikit-learn) que utiliza a abordagem z-score, garantindo média zero para os atributos com desvio padrão 1. A Figura 3 mostra os primeiros atributos do dataset antes da normalização, a Figura 4 mostra os atributos após a normalização.

	media_perfil_integrado	desvio_padrao_perfil_integrado	curtose_perfil_integrado
0	140.562500	55.683782	-0.234571
1	102.507812	58.882430	0.465318
2	103.015625	39.341649	0.323328
3	136.750000	57.178449	-0.068415
4	88.726562	40.672225	0.600866

Figura 3: Atributos antes da normalização (Mostrando apenas as primeiras 3 colunas)

	media_perfil_integrado	desvio_padrao_perfil_integrado	curtose_perfil_integrado
0	1.149317	1.334832	-0.669570
1	-0.334168	1.802265	-0.011785
2	-0.314372	-1.053322	-0.145233
3	1.000694	1.553254	-0.513409
4	-0.871402	-0.858879	0.115609

Figura 4: Atributos depois da normalização (Mostrando apenas as primeiras 3 colunas)

2.1 Balanceamento das classes

A distribuição das classes para cada amostra no dataset é muito desproporcional, tendo mais de 90% das amostras pertencentes à classe Não-Pulsar, fator este que pode prejudicar a eficiência do modelo de classificação. Como nosso objetivo é classificar os registros que são pulsares, a estratégia adotada foi aplicar um *undersampling* para assim equilibrar a distribuição das classes no dataset. Este balanceamento foi feito via estratégia *Near Miss*, que seleciona os elementos da classe alvo mais próximas – porque são as mais “difíceis de prever” para permanecer no dataset.

A Figura 5 mostra a distribuição das classes antes e depois do balanceamento. Onde após o balanceamento o dataset ficou com um total de 1.639 amostras de cada classe, totalizando 3.278 amostras.

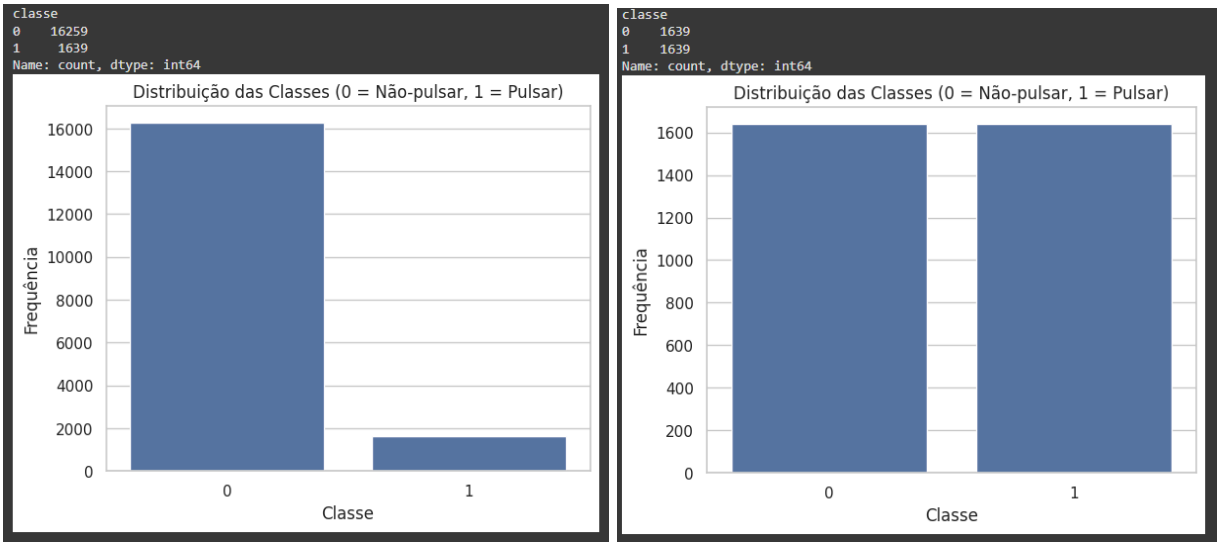

Figura 5: Distribuição das Classes antes e depois do balanceamento

3. Resultados

Para a tarefa de classificação, foi utilizado o algoritmo ***Random Forest Classifier***, disponível no módulo *sklearn.ensemble* da biblioteca *Scikit-learn*. Esse algoritmo, baseado em um conjunto de árvores de decisão, é reconhecido por sua robustez, capacidade de lidar com variáveis numéricas e resiliência ao *overfitting*, sendo ideal para problemas supervisionados com dados tabulares.

Visando uma avaliação confiável e generalizável, foi aplicada a técnica de validação cruzada estratificada com a classe ***StratifiedKFold***, do módulo *sklearn.model_selection*, garantindo que a proporção entre as classes fosse preservada em cada divisão (fold). O número de *folds* utilizado foi 5.

Além disso, foi utilizado o ***GridSearchCV*** (*sklearn.model_selection*) para realizar a otimização dos hiperparâmetros do modelo, explorando combinações de parâmetros como número de estimadores e profundidade máxima para encontrar a configuração com melhor desempenho com base na métrica *f1_macro*.

O modelo foi treinado e avaliado de forma independente em cada *fold*, permitindo uma análise mais robusta e representativa dos resultados obtidos.

3.1 Métricas

Após o treinamento e validação do modelo Random Forest com validação cruzada estratificada em 5 *folds*, foram calculadas as principais métricas de desempenho para cada partição do conjunto de dados.

3.1.1 Acurácia, Precisão, Recall e F1-Score

As métricas utilizadas para avaliação do modelo foram:

- Acurácia: proporção total de classificações corretas.
- Precisão: proporção de amostras classificadas como positivas que são realmente positivas.
- Recall (Sensibilidade): proporção de amostras positivas corretamente identificadas.
- F1-Score: média harmônica entre precisão e recall, útil especialmente em contextos com classes desbalanceadas.

A Figura 6 mostra as métricas para cada *fold* e média geral destas métricas

	Accuracy	Precision	Recall	F1-Score
0	0.940549	0.943123	0.940549	0.940462
1	0.955793	0.956306	0.955793	0.955780
2	0.955793	0.957020	0.955793	0.955763
3	0.955725	0.955841	0.955713	0.955721
4	0.954198	0.954259	0.954208	0.954198

Figura 6: Métricas de Acurácia, Precisão Recall (sensibilidade) e F1-Score

3.1.2 Matriz de Confusão

Para avaliar o desempenho do modelo em termos de acertos e erros entre as classes, foi gerada a soma das matrizes de confusão dos cinco folds, como mostra a Figura 7:

- A diagonal principal representa os acertos (verdadeiros positivos e verdadeiros negativos).
- Os valores fora da diagonal indicam erros de classificação:
    - 47 falsos positivos: não-pulsares classificados incorretamente como pulsars.
    - 109 falsos negativos: pulsares classificados incorretamente como não-pulsars.

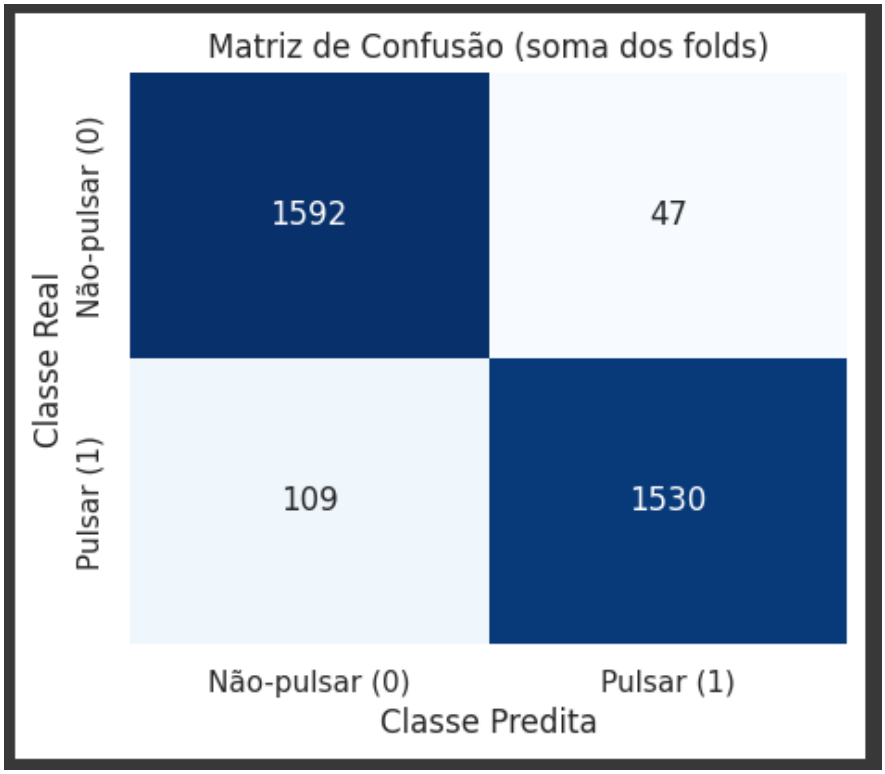

Figura 7: Matriz de confusão

3.1.3 Importância dos Atributos

Por fim, a importância relativa das variáveis no modelo foi calculada com base nos critérios internos do algoritmo Random Forest, que considera o ganho de informação em cada divisão das árvores. A Figura 8 mostra esta análise, onde observa-se que as variáveis associadas ao **perfil integrado** concentram a maior parte da importância atribuída pelo modelo, indicando maior contribuição dessas features na tomada de decisão do classificador.

	Feature	Importância
2	curtose_perfil_integrado	0.279018
0	media_perfil_integrado	0.233619
3	media_curva_dm_snr	0.204692
4	desvio_padrao_curva_dm_snr	0.140447
5	curtose_curva_dm_snr	0.098436
1	desvio_padrao_perfil_integrado	0.043788

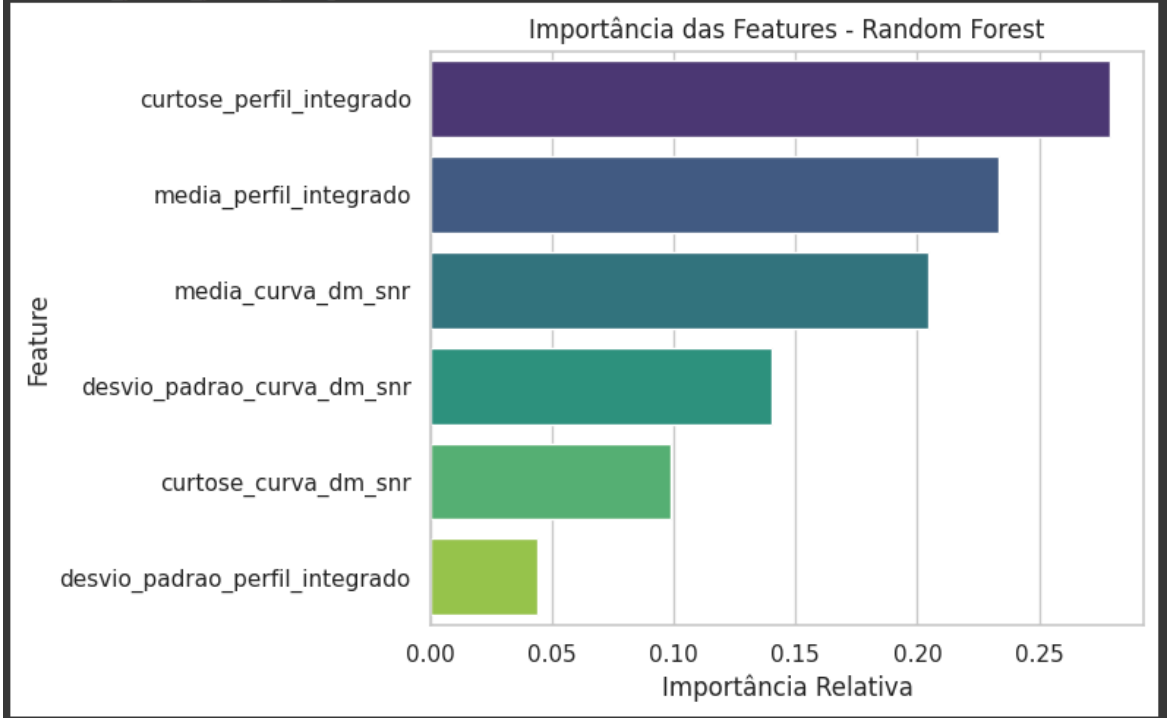

Figura 8: Importância dos atributos - Random Forest

4. Considerações Finais

A partir do conjunto de dados HTRU2, foi desenvolvido um fluxo completo de preparação, balanceamento e modelagem voltado à classificação binária de sinais de rádio em pulsares reais e falsos positivos. As etapas envolveram análise exploratória, remoção de variáveis redundantes, normalização dos dados e aplicação da técnica *NearMiss* para balancear as classes.

O algoritmo ***Random Forest***, aliado à validação cruzada estratificada e otimização de hiperparâmetros com *GridSearchCV*, apresentou desempenho consistente, com métricas médias próximas de 95% para acurácia, precisão, recall e F1-score. A análise da matriz de confusão mostrou um número moderado de erros, com maior incidência de falsos negativos.

A análise de importância dos atributos revelou que as variáveis extraídas do perfil integrado foram as mais relevantes para a tomada de decisão do modelo. Esses resultados indicam que o modelo é capaz de capturar padrões relevantes nos dados, oferecendo uma base sólida para futuras investigações ou aplicação em cenários reais de detecção de pulsares.

Referências

Materiais fornecidos durante o semestre letivo.

R. Lyon. "HTRU2," UCI Machine Learning Repository, 2015. [Online]. Available: https://doi.org/10.24432/C5DK6R.

Lyon, R.J., Stappers, B.W., Cooper, S., Brooke, J.M., & Knowles, J.D. (2016). Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. Monthly Notices of the Royal Astronomical Society, 459, 1104-1123.