Tarefas de Mineração de Dados: Classificação de Pulsares

- Kassio Ferreira
- Mateus Costa

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.

Sobre a Base de Dados

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

Sobre a Base de Dados

Detectar esses sinais não é simples, 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.

Sobre a Base de Dados

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).

Sobre a Base de Dados

Para cada medida - Perfil integrado e Curva DM-SNR, são coletados valores estatísticos que caracterizam o objeto observado, totalizando 8 atributos para cada amostra:

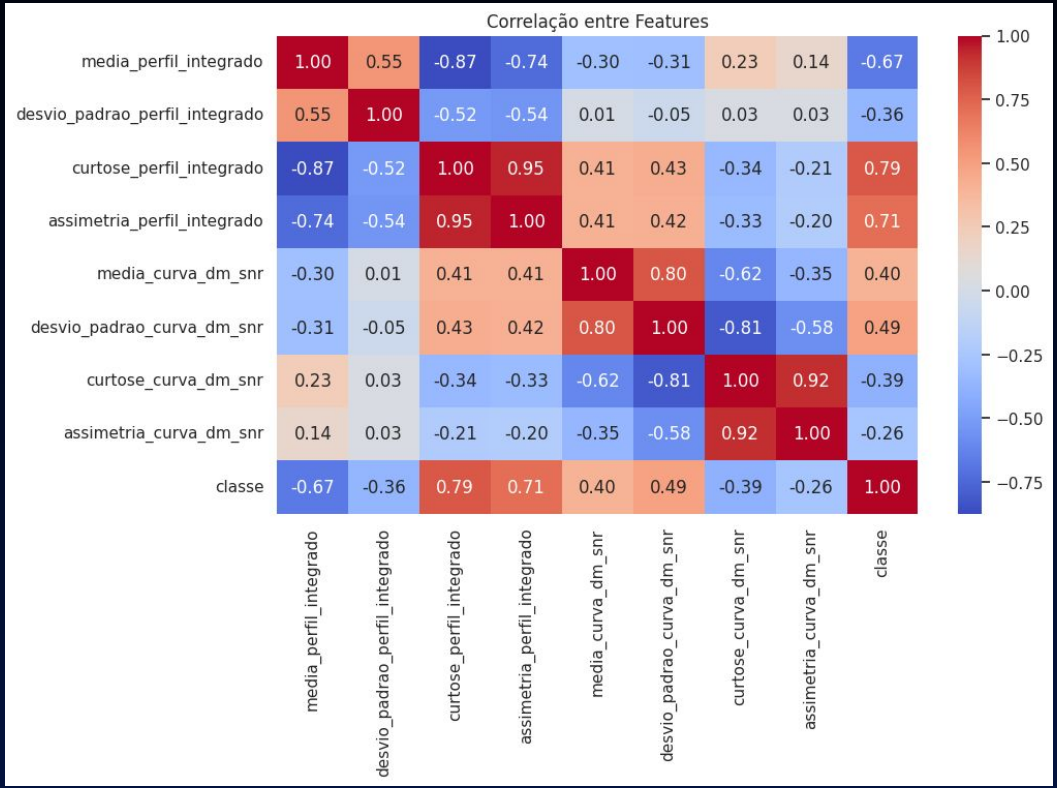
Média: valor médio da curva; sinais de pulsares tendem a ter médias mais elevadas em função da presença de picos regulares.

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.

Análise e Preparação: Correlação entre os atributos



Análise e Preparação: Normalização dos valores

Foi aplicada uma normalização com a ferramenta StandardScaler (da lib Scikit-learn) que utiliza a abordagem z-score, garantindo média zero para os atributos

	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)

Análise e Preparação: Distribuição das Classes

como a classe 0 representava mais de 90% das amostras, foi utilizado um *undersampling* com *NearMiss*. Resultando em:

- 1.639 amostras de cada classe | Total: 3.278 amostras

Figura 5: Distribuição das Classes antes e depois do balanceamento

Resultados: Treinamento

Algoritmo de Classificação:

- **Random Forest Classifier (sklearn.ensemble)**
  - Classificação supervisionada com dados tabulares
  - Lida bem com dados numéricos e reduz risco de overfitting

Validação e ajuste:

- **Validação Estratificada com StratifiedKFold (sklearn.model_selection)**
  - Número de Folds: 5
  - Garante a mesma proporção entre classes em cada divisão

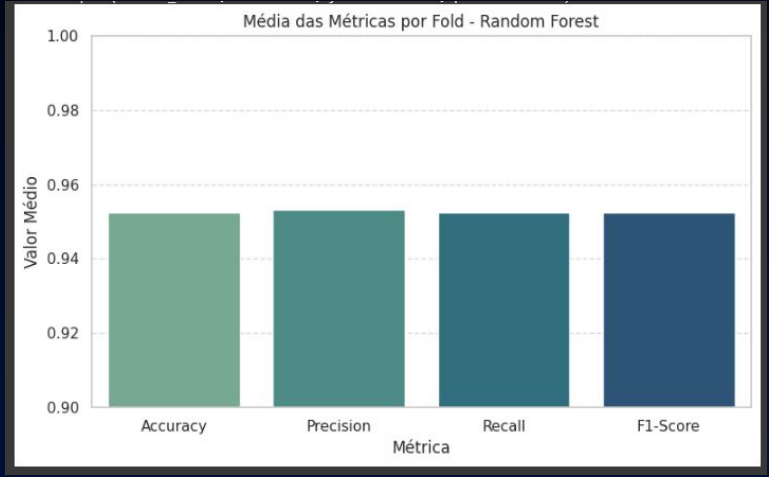
Otimização de Hiperparâmetros:

- **GridSearchCV (sklearn.model_selection)**

Resultados: Métricas

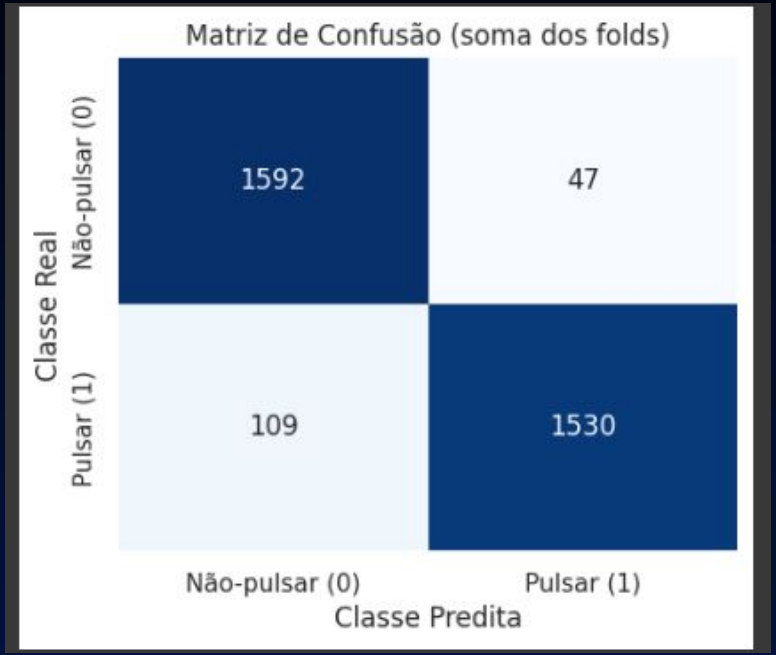
Acurácia, Precisão, Recall e F1-Score

	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



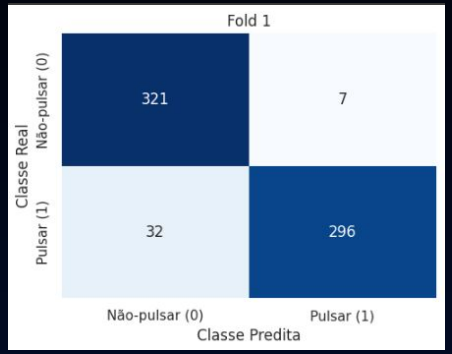
Resultados: Métricas

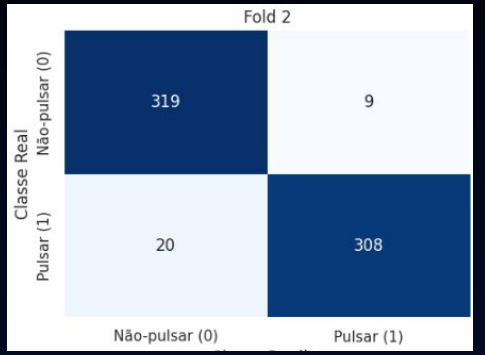
Matriz de confusão

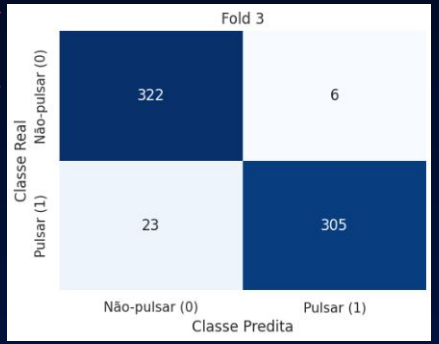


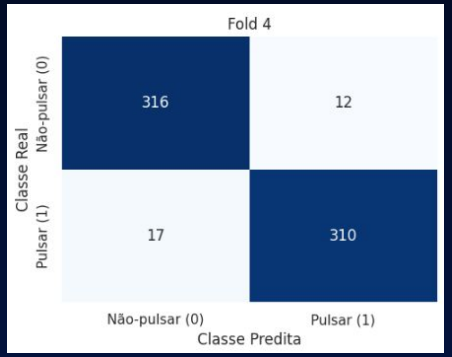
Resultados: Métricas

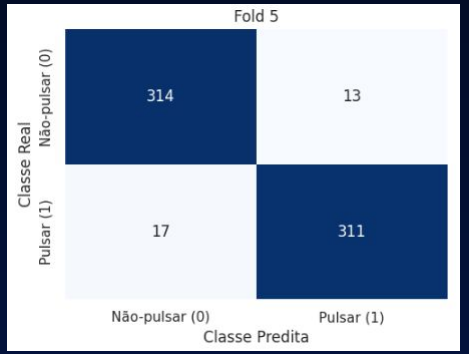
Matriz de confusão
em cada Fold





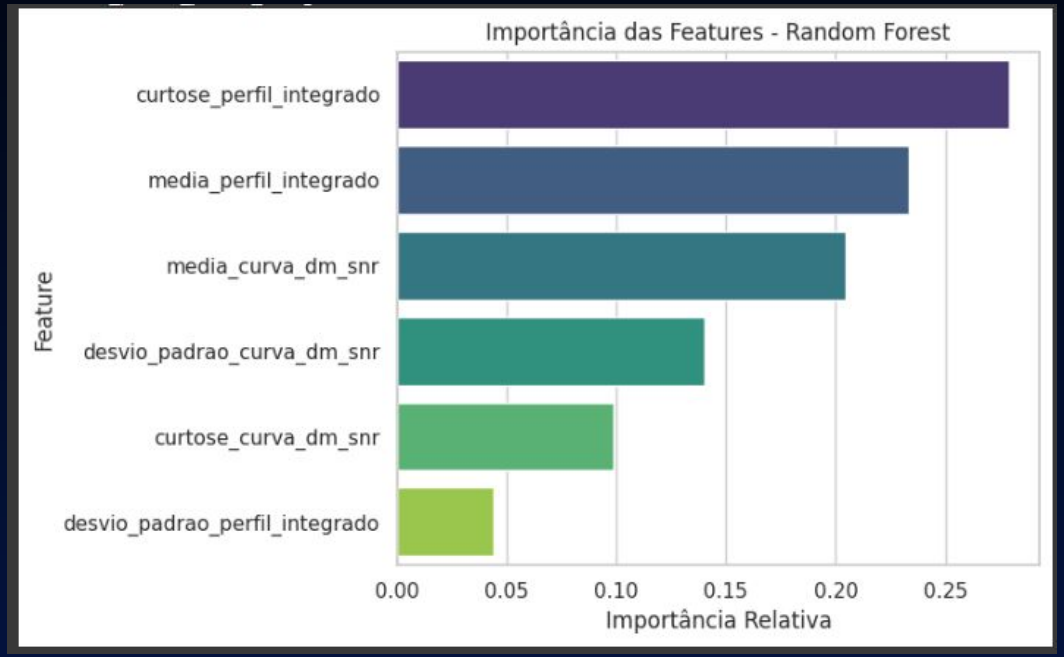






Resultados: Métricas

Nível de “Importância” dos atributos durante a classificação



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.