



Geração de dados sintéticos tabulares para detecção de malware Android: um estudo de caso



UFAM



Angelo Gaspar, Diego Kreutz,
Hendrio Bagança,
Rodrigo Mansilha,
Kayuã Oleques Paim

Motivação



Motivação

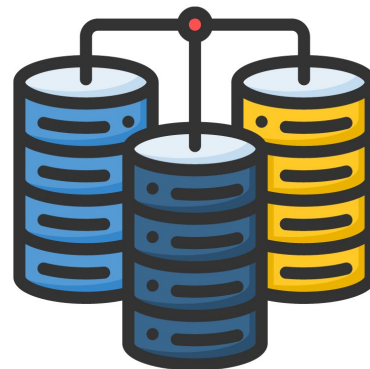


modelos de
aprendizado de
máquina

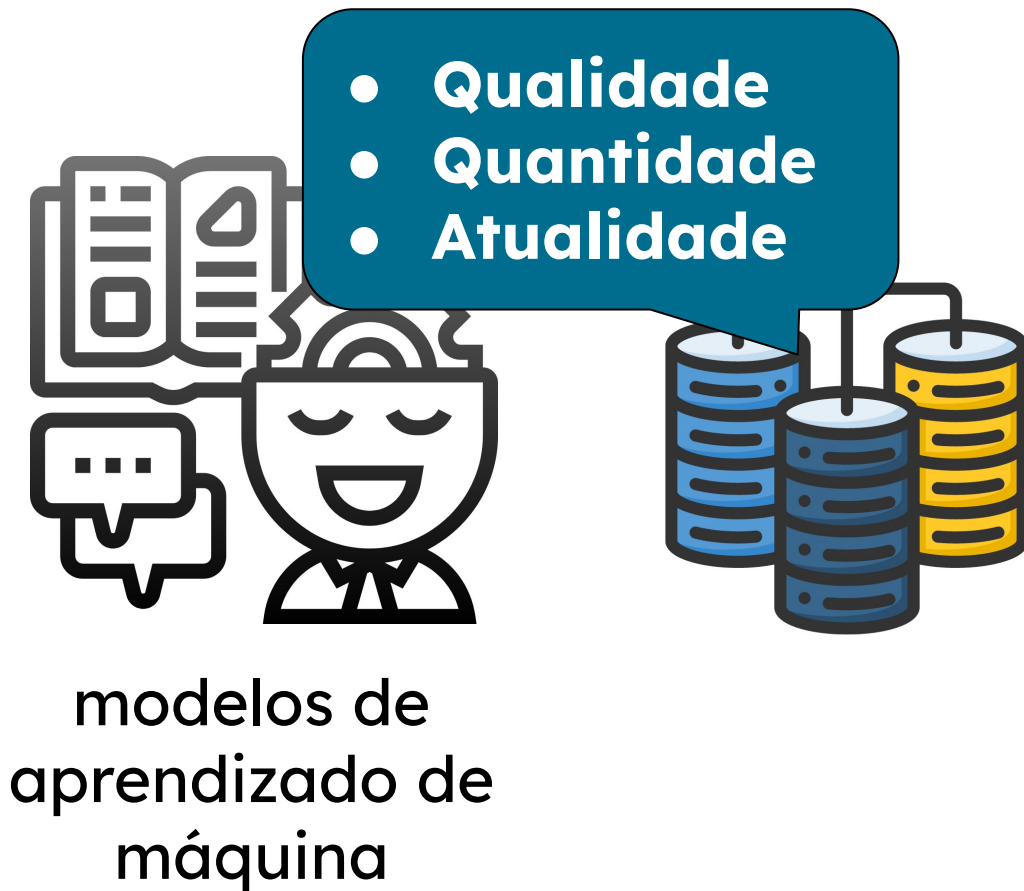
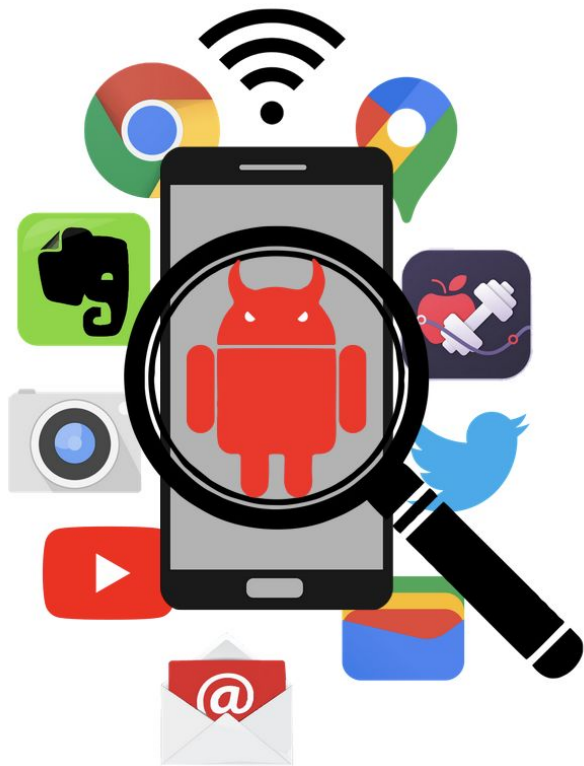
Motivação



modelos de
aprendizado de
máquina



Motivação



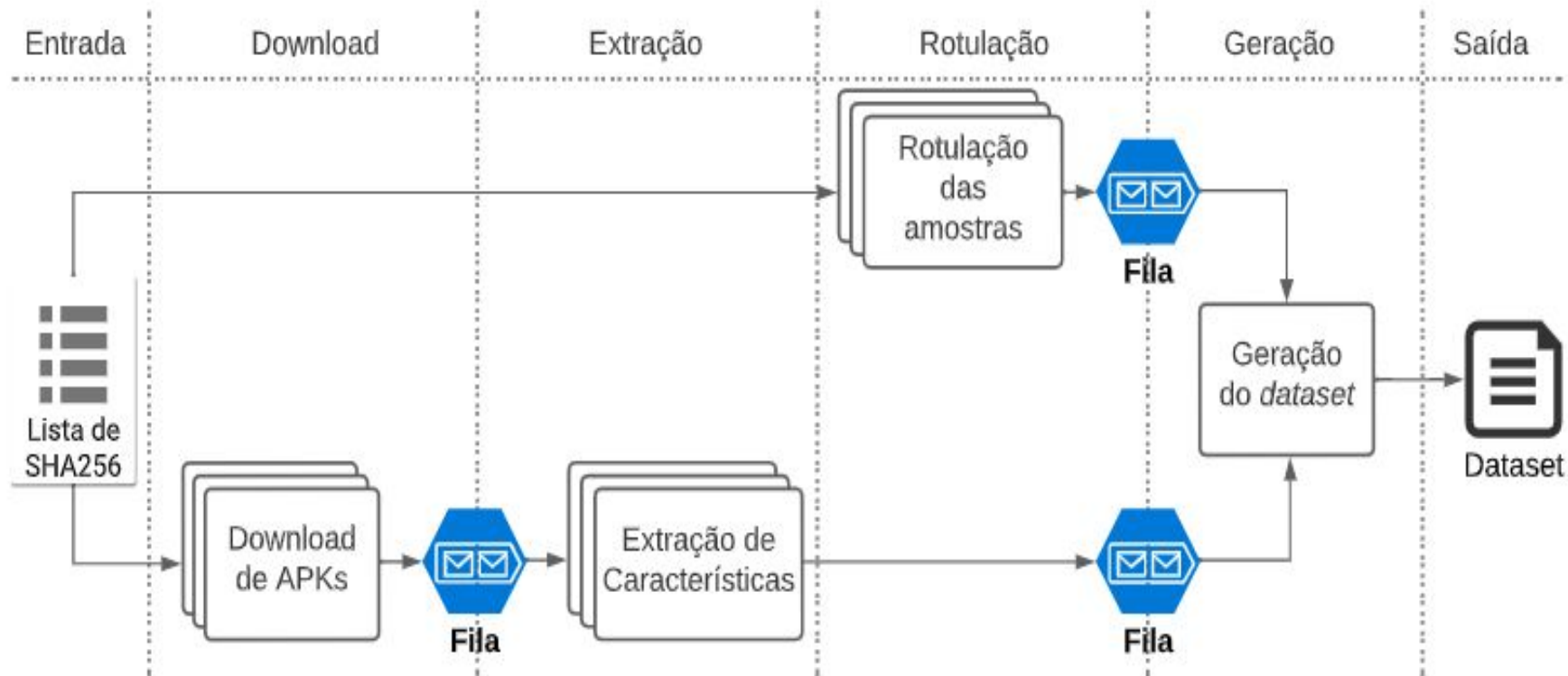
Motivação

80% dos projetos de IA falham por questões que envolvem dados

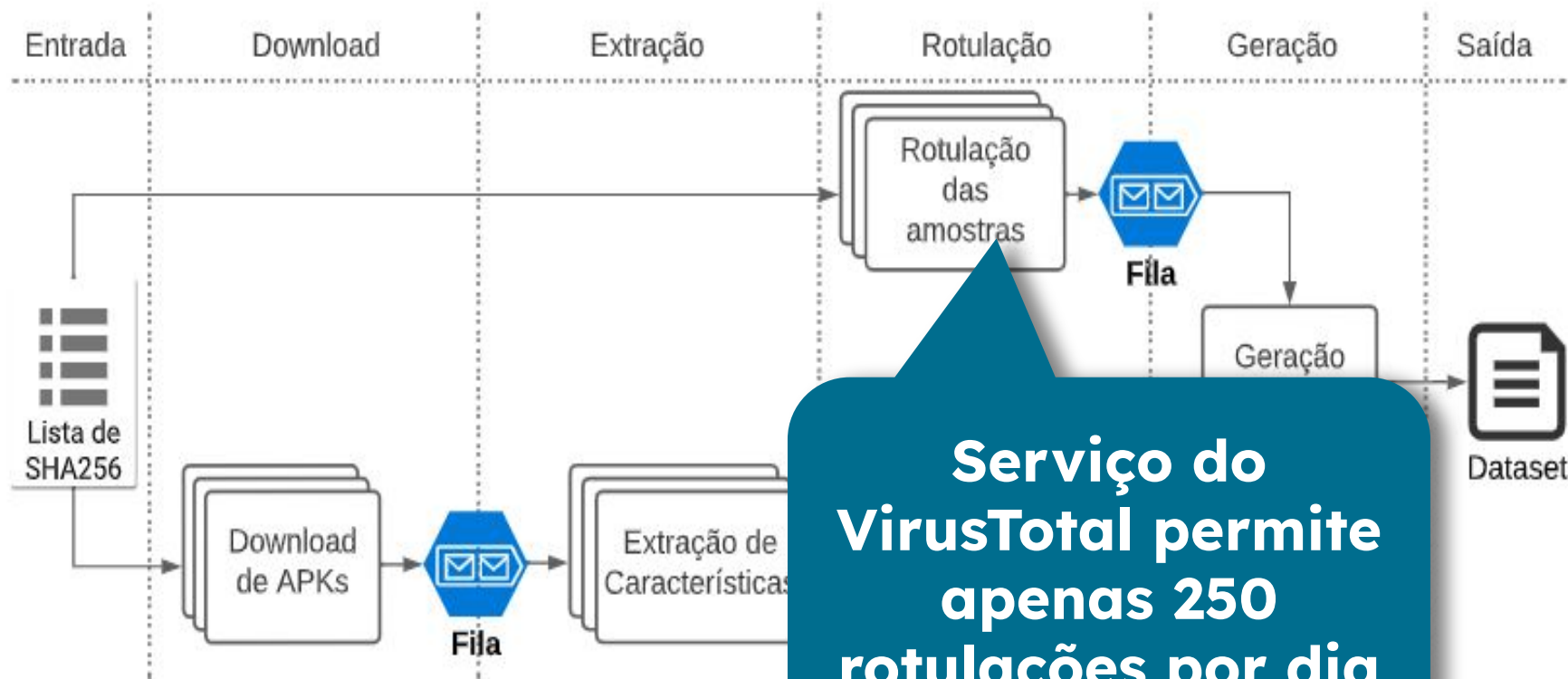


**Top 10 Reasons Why
AI Projects Fail**

Motivação (ADBuilder)

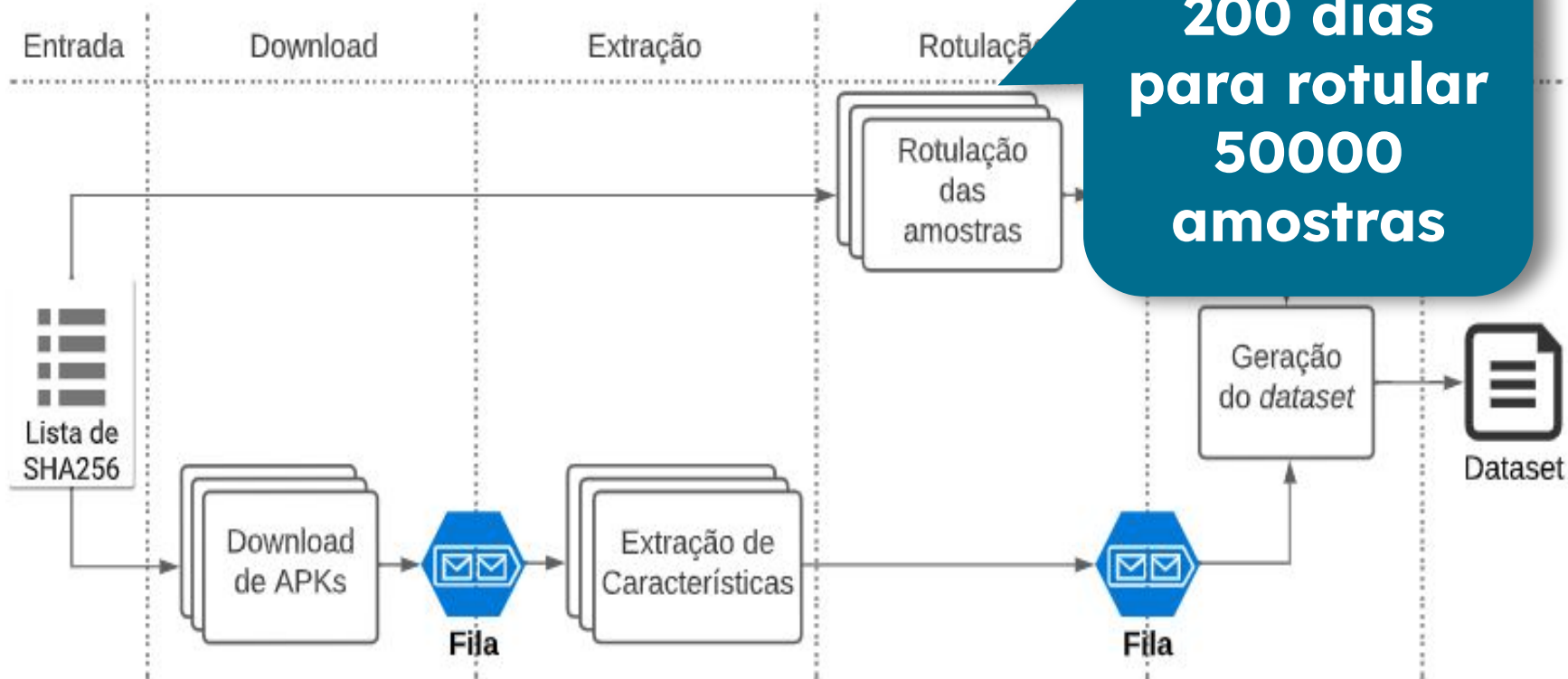


Motivação (ADBuilder)

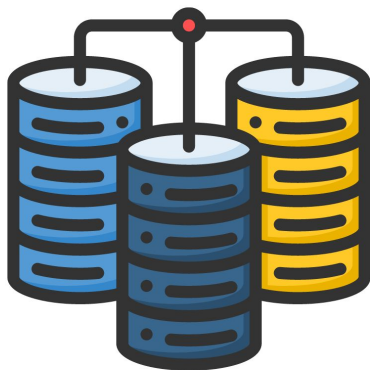


**Serviço do
VirusTotal permite
apenas 250
rotulações por dia**

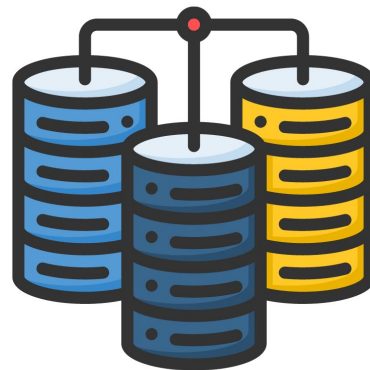
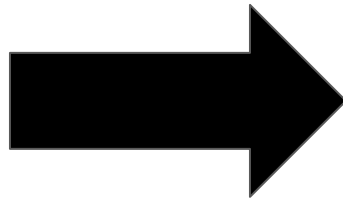
Motivação (ADBuilder)



Aumento de dados

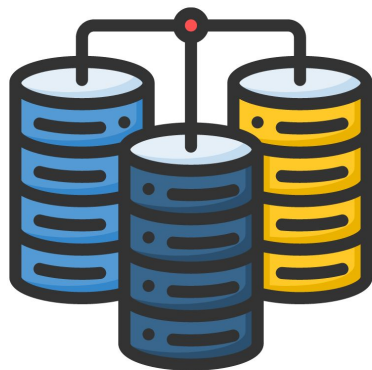


datasets reais



datasets sintéticos

Aumento de dados



datasets reais

Geração de dados
sintético a partir
de dados reais



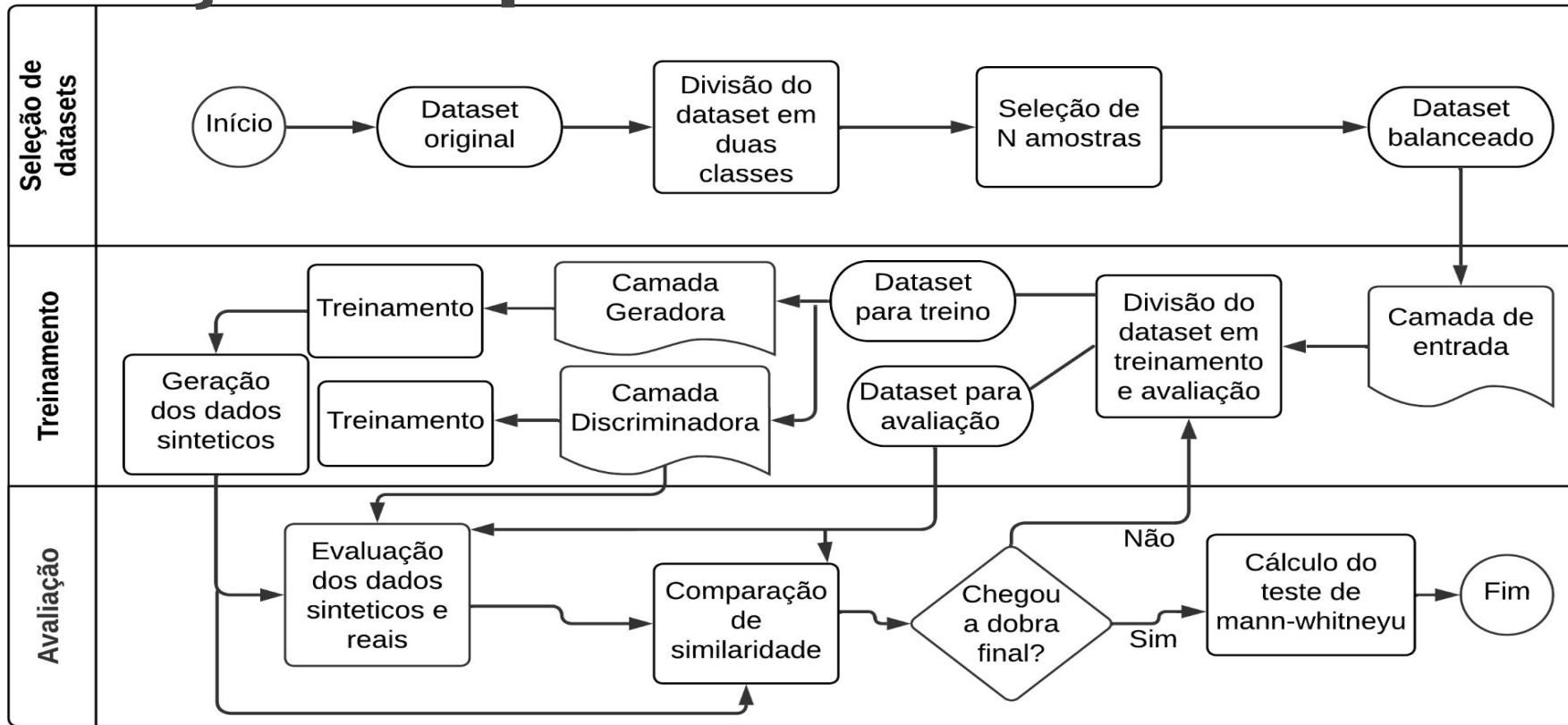
datasets sintéticos

Aumento de dados: benefícios

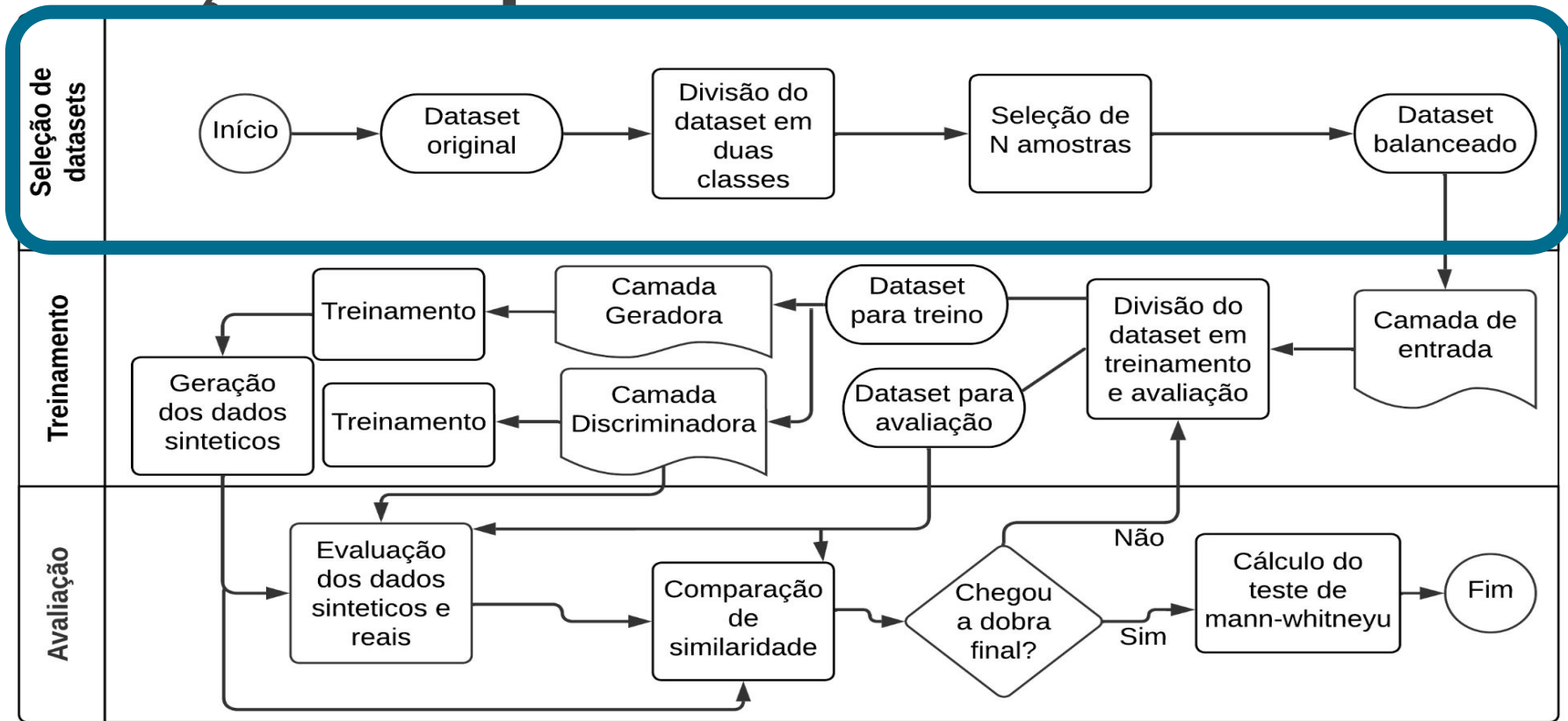


- **Performance aprimorada do modelo**
- **Evitar *overfitting***
- **Maior privacidade dos dados**

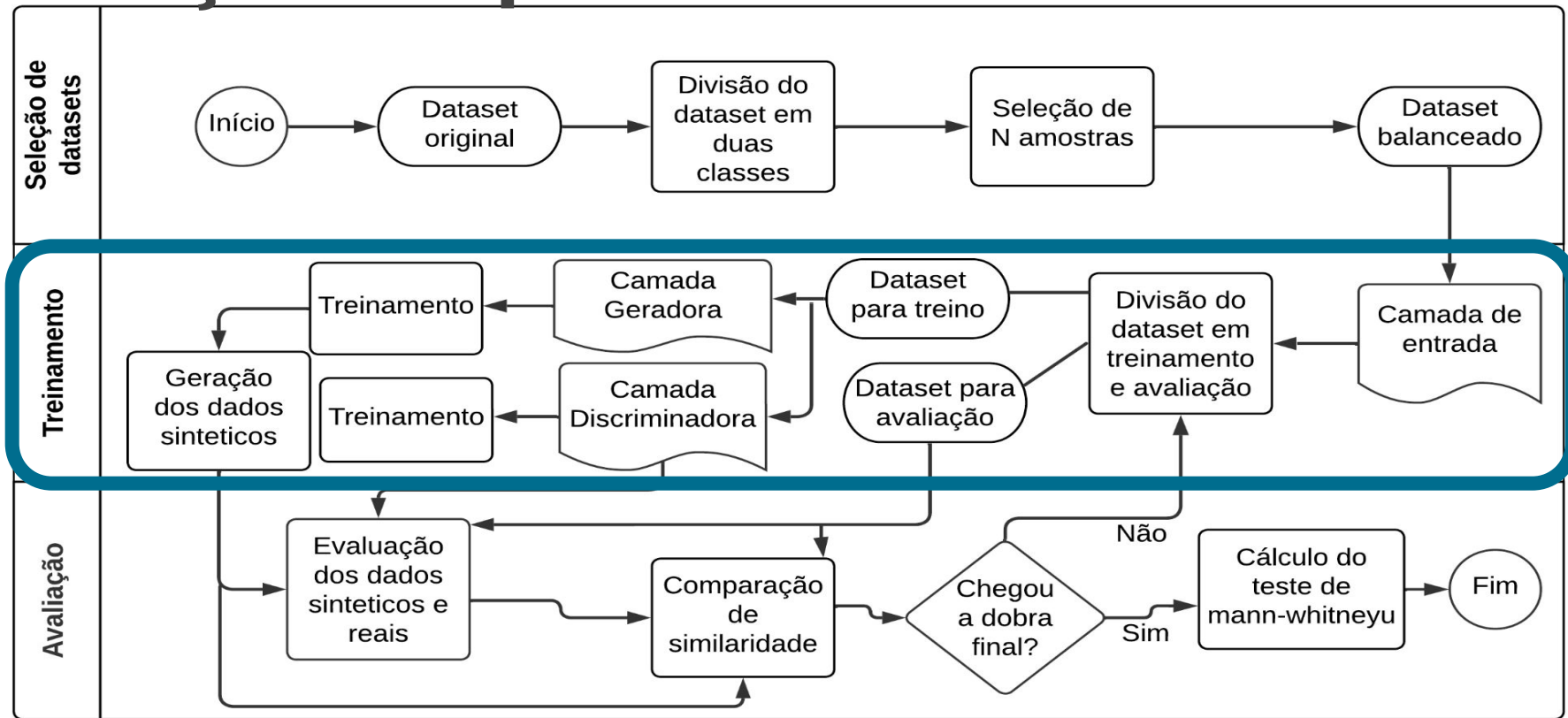
Solução Proposta



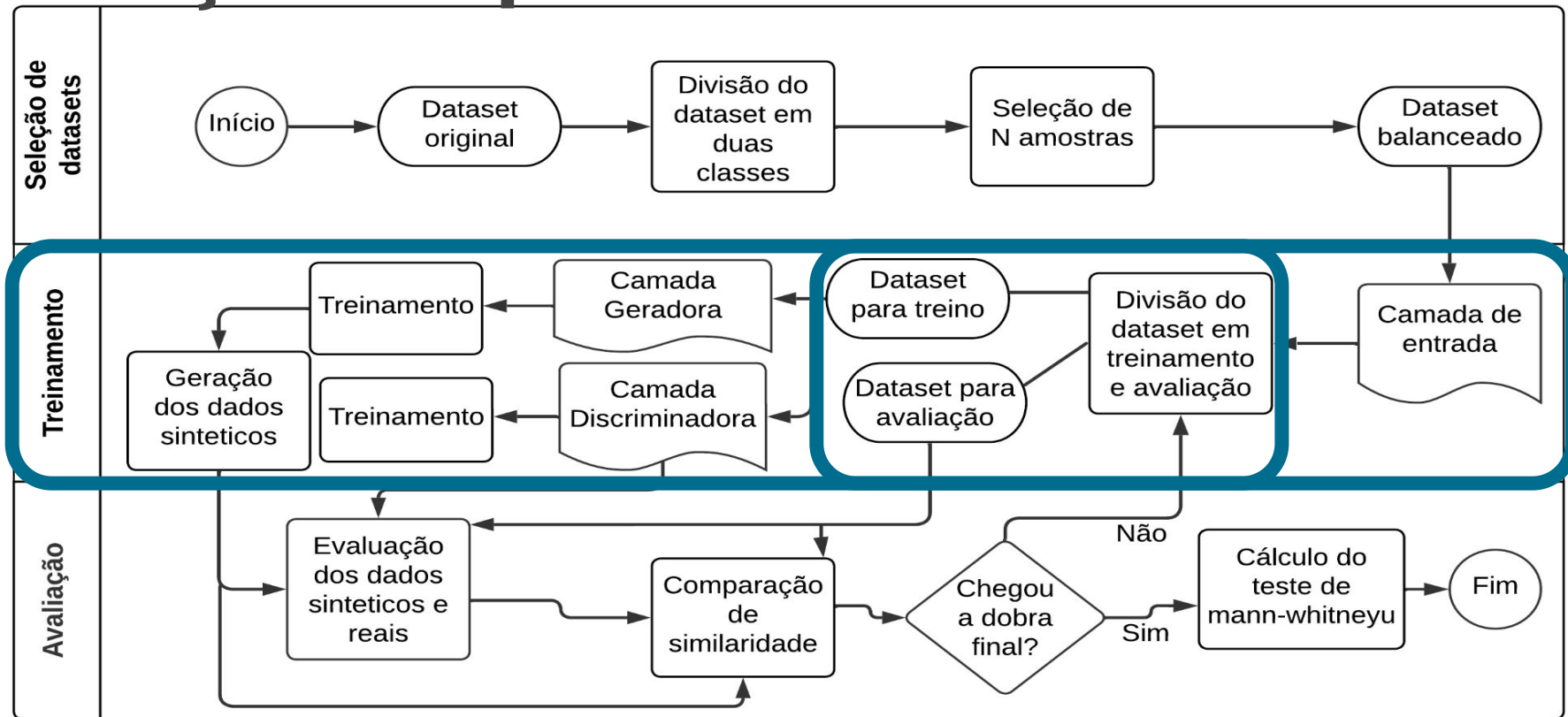
Solução Proposta



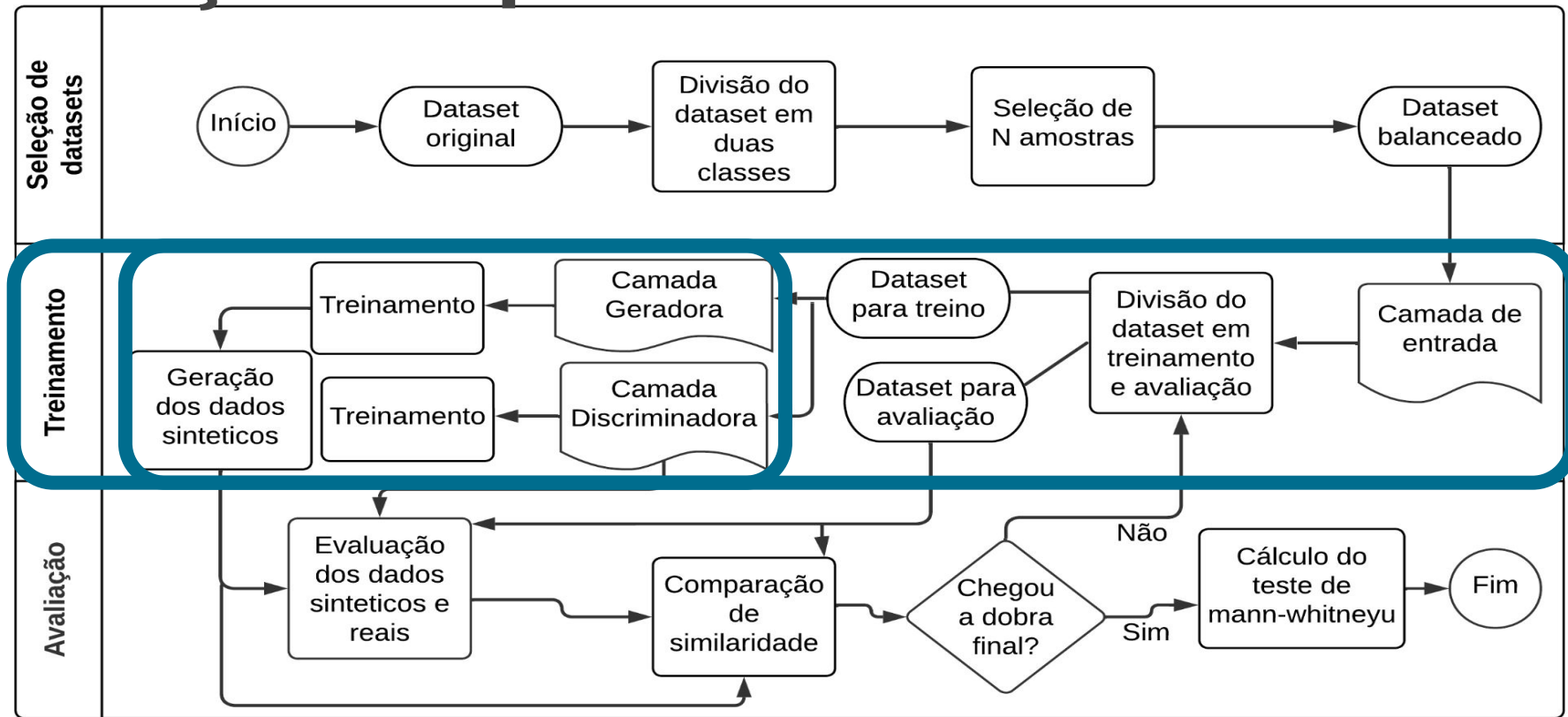
Solução Proposta



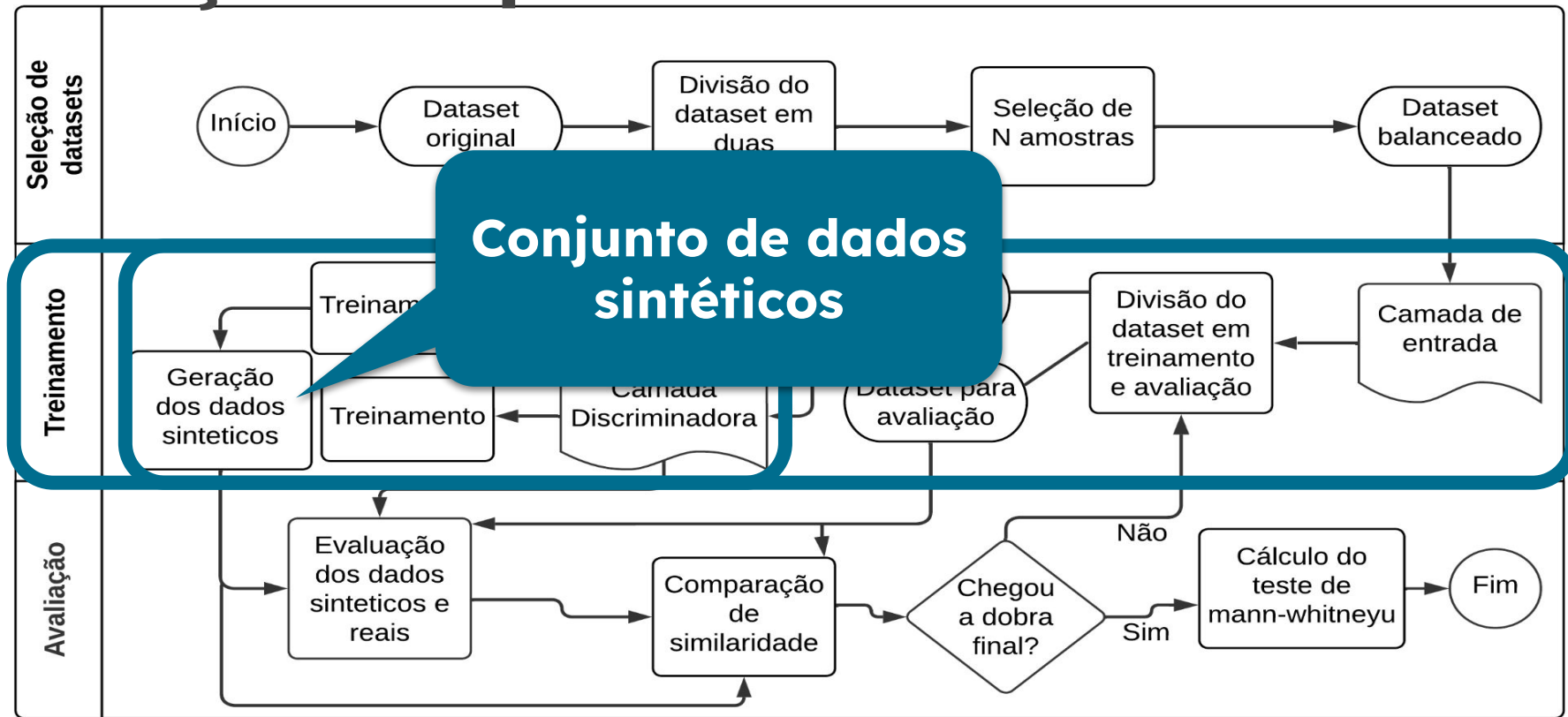
Solução Proposta



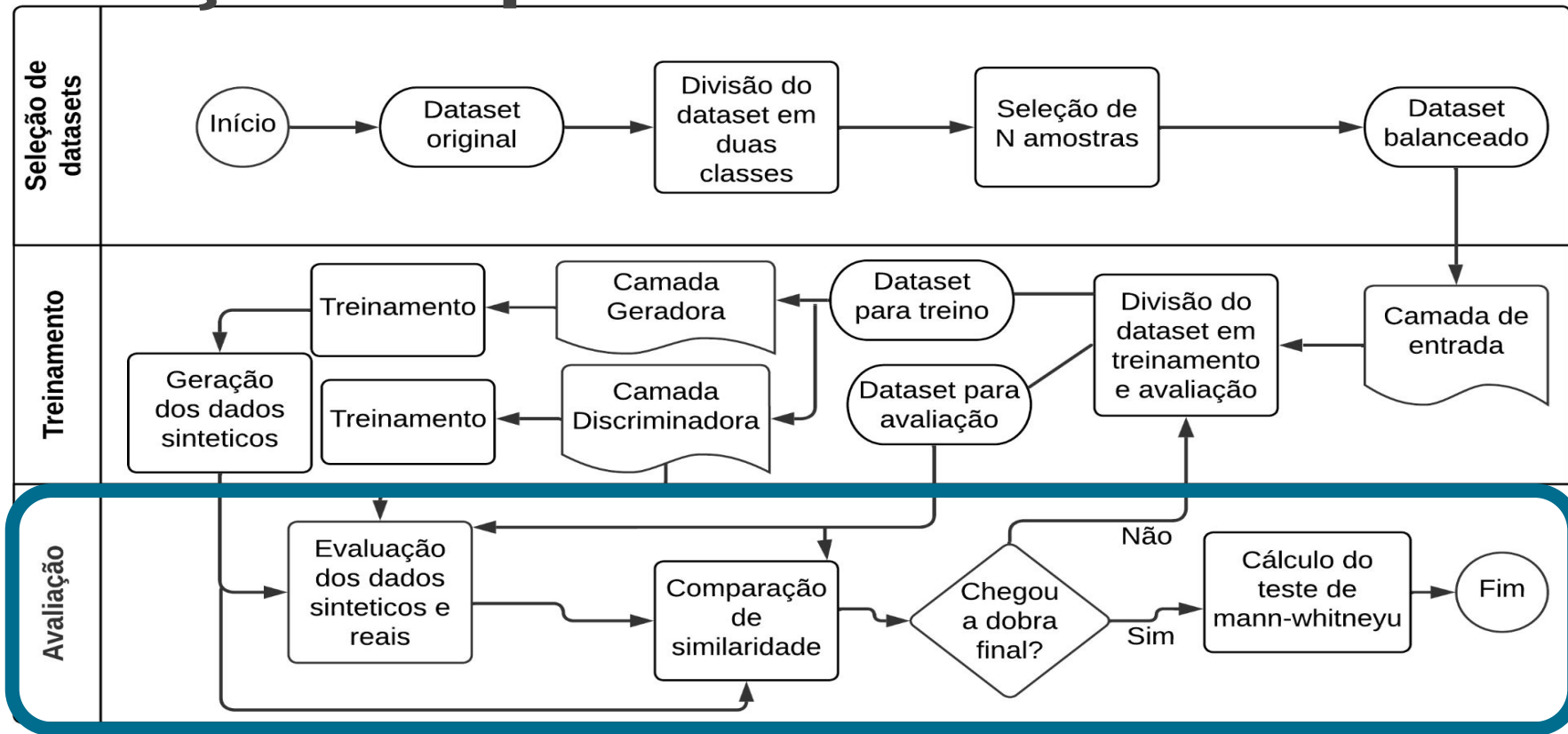
Solução Proposta



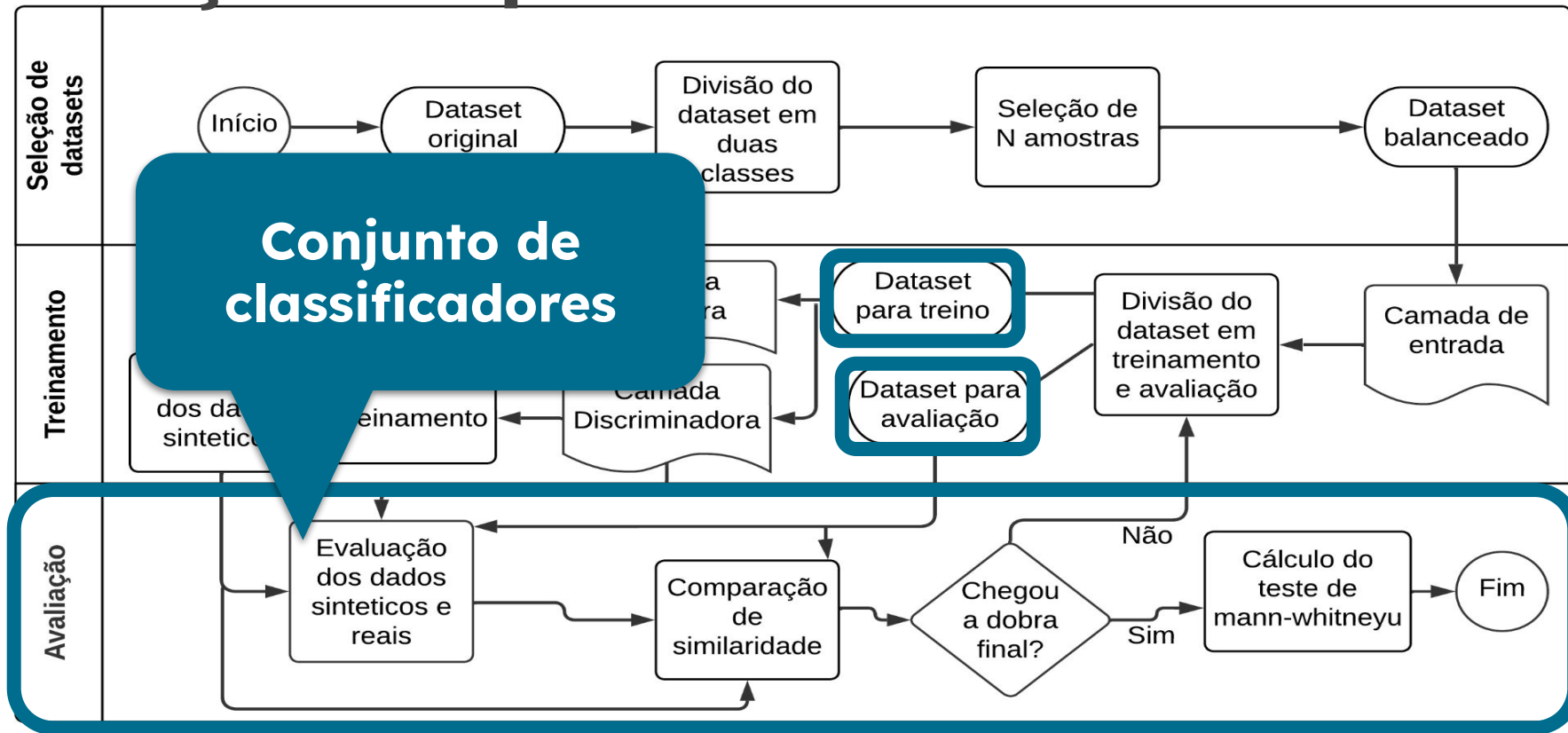
Solução Proposta



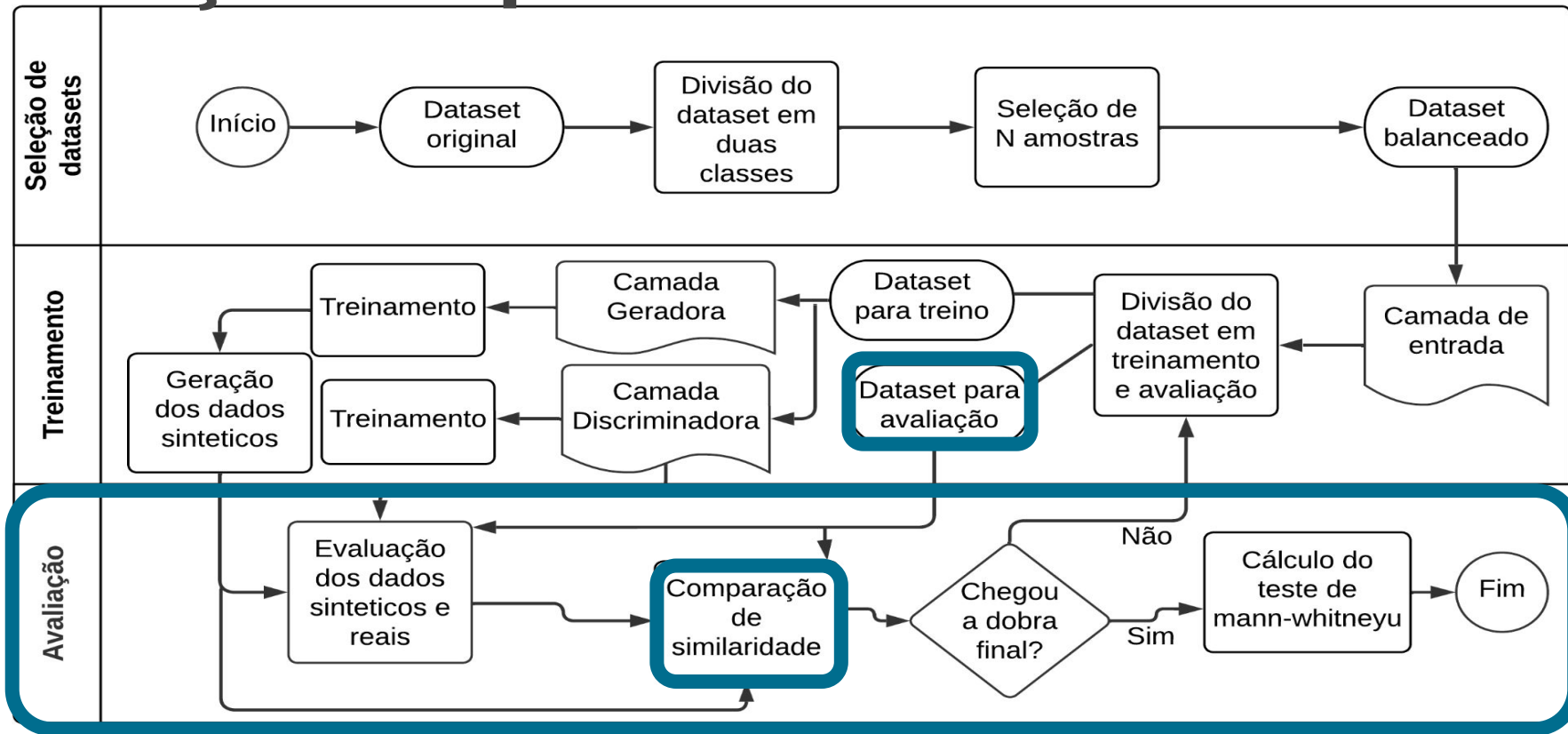
Solução Proposta



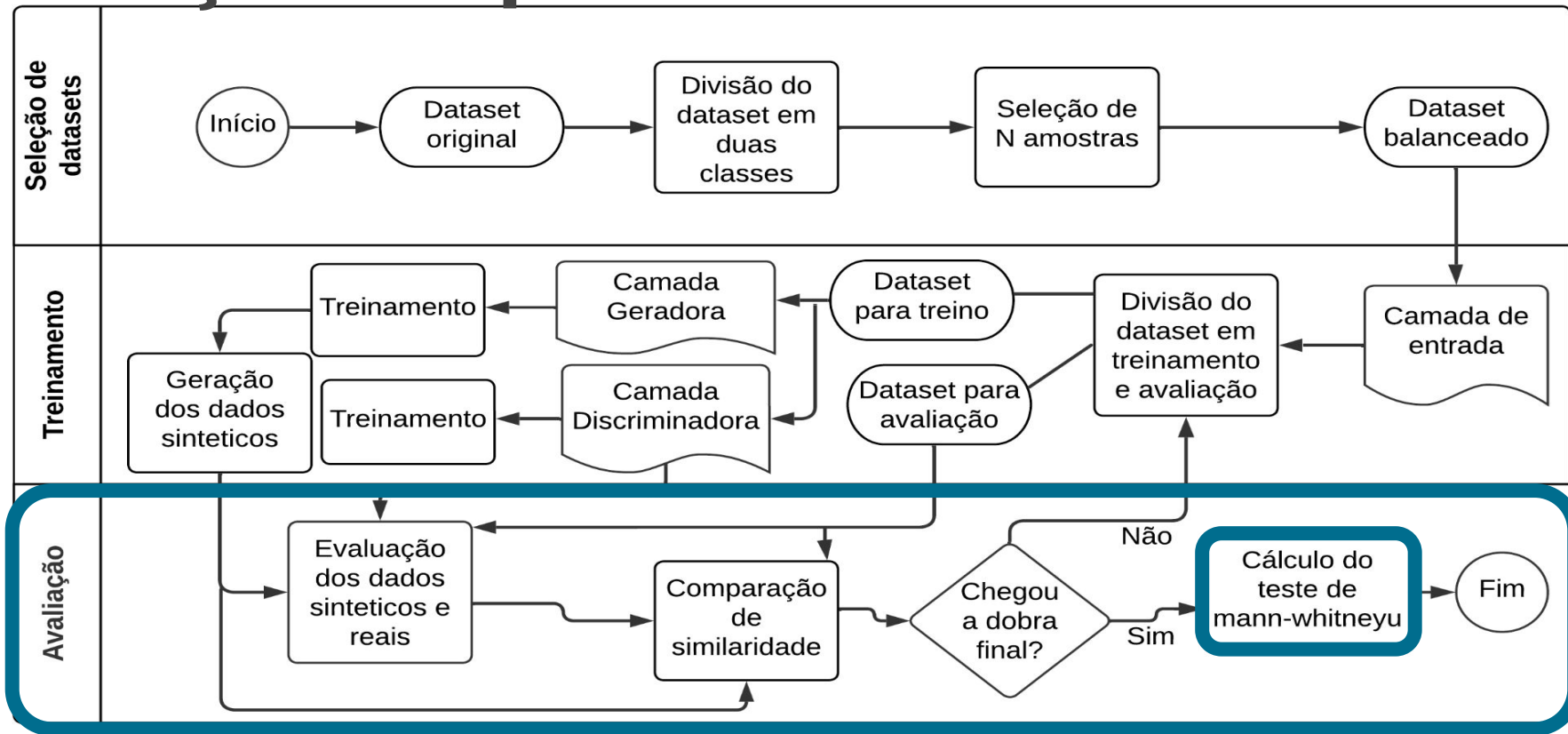
Solução Proposta



Solução Proposta



Solução Proposta



Ferramentas e tecnologias

Ambiente de testes:

- Ubuntu 22.04
- Python:
 - 3.8.10
 - 3.8.2
 - 3.10.12
- Docker 24.0.7 e 20.10.5

Bibliotecas python:

- Numpy 1.21.5
- Keras 2.9.0
- Tensorflow 2.9.1
- Pandas 1.4.4
- Scikit-learn 1.1.1
- Mlflow 2.12.1

Avaliação (Hiperparâmetros)

	Kronodroid	Drebin	Android P	Adroit
Dense Layer Sizes (d)	1024	1800	2048	1400
Dense Layer Sizes (g)	2048	3012	3012	1800
Dropout Decay Rate (d)	0,4	0,4	0,3	0,4
Dropout Decay Rate (g)	0,2	0,2	0,1	0,2
Initializer Deviation	0,004	0,003	0,001	0,01
Initializer Mean	0,1	0,1	0	0,1
Latent Dimension	128	128	128	150
Latent Standard Deviation	0,5	0,8	0,8	1
Number Epochs	500	300	460	450

Avaliação (Hiperparâmetros)

	Kronodroid	Drebin	Android P	Adroit
Dense Layer Sizes (d)	1024	1800	2048	1400
Dense Layer Sizes (g)	2048	3012	3012	1800
Dropout Decay Rate (d)	0,4	0,4	0,3	0,4
Dropout Decay Rate (g)	0,2	0,2	0,1	0,2
Initializer Deviation	0,004	0,003	0,001	0,01
Initializer Mean	0,1	0,1	0	0,1
Latent Dimension	128	128	128	150
Latent Standard Deviation	0,5	0,8	0,8	1
Number Epochs	500	300	460	450

Avaliação (Hiperparâmetros)

	Kronodroid	Drebin	Android P	Adroit
Dense Layer Sizes (d)	1024	1800	2048	1400
Dense Layer Sizes (g)	2048	3012	3012	1800
Dropout Decay Rate (d)	0,4	0,4	0,3	0,4
Dropout Decay Rate (g)	0,2	0,2	0,1	0,2
Initializer Deviation	0,004	0,003	0,001	0,01
Initializer Mean	0,1	0,1	0	0,1
Latent Dimension	128	128	128	150
Latent Standard Deviation	0,5	0,8	0,8	1
Number Epochs	500	300	460	450

Avaliação (conjuntos de dados)

<i>Dataset</i>	Características	Amostras		
		Malware	Benignos	Total
Adroit	166	50%	50%	6836
Drebin	215	50%	50%	11110
Kronodroid	286	50%	50%	20000
Android P	151	50%	50%	18154

Avaliação (Métricas de fidelidade)

<i>Dataset</i>	Positivo		Falso	
	Cosseno	Erro quadrático	Cosseno	Erro quadrático
Krondroid	0,69	0,06	0,74	0,07
Adroit	0,70	0,03	0,65	0,03
Drebin	0,37	0,12	0,50	0,16
Android P	0,22	0,03	0,51	0,03

Avaliação (Métricas de fidelidade)

<i>Dataset</i>	Positivo malware		Falso benigno	
	Cosseno	Erro quadrático	Cosseno	Erro quadrático
Krondroid	0,69	0,06	0,74	0,07
Adroit	0,70	0,03	0,65	0,03
Drebin	0,37	0,12	0,50	0,16
Android P	0,22	0,03	0,51	0,03

Avaliação (Métricas de fidelidade)

Dataset	P		Falso	
	Cosseno	Erro quadrático	Cosseno	Erro quadrático
Krondroid	0,69	0,06	0,74	0,07
Adroit	0,70	0,03	0,65	0,03
Drebin	0,37	0,12	0,50	0,16
Android P	0,22	0,03	0,51	0,03

Alta
similaridade
de cosseno

Avaliação (Métricas de fidelidade)

<i>Dataset</i>	Positivo		Falso	
	Cosseno	Cosseno	Cosseno	Erro quadrático
Krondroid	0,69		0,74	0,07
Adroit	0,70		0,65	0,03
Drebin	0,37	0,12	0,50	0,16
Android P	0,22	0,03	0,51	0,03

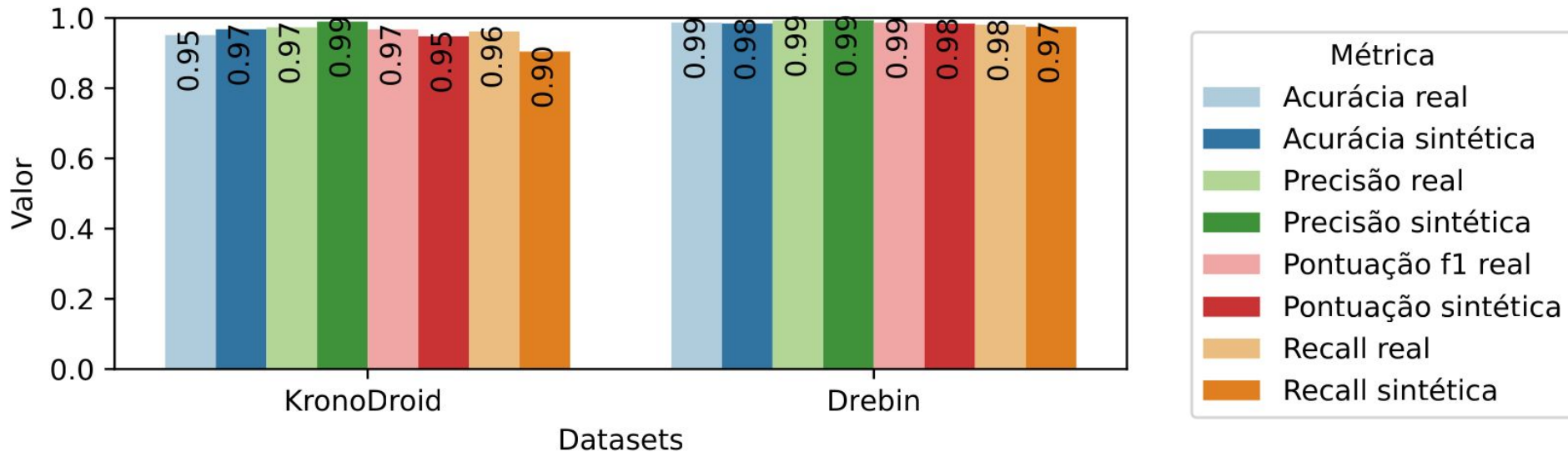
Valores baixos
de similaridade
de cosseno

Avaliação (Métricas de fidelidade)

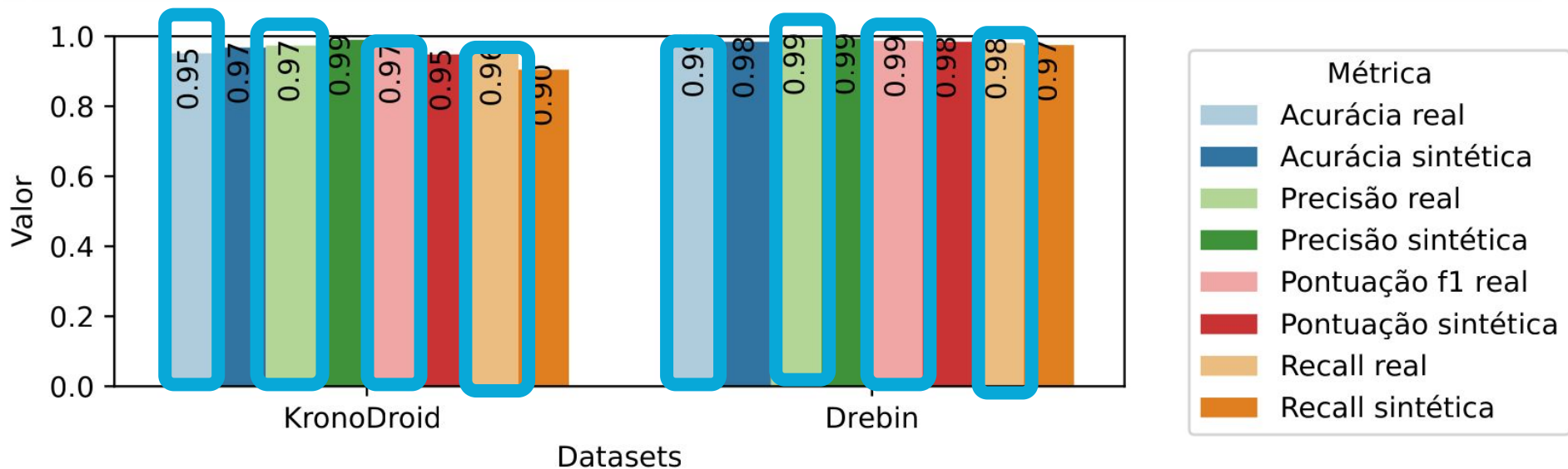
Dataset	Positivo			
	Cosseno	Erro quadrático	Cosseno	Erro quadrático
Krondroid	0,69	0,06	0,74	0,07
Adroit	0,70	0,03	0,65	0,03
Drebin	0,37	0,12	0,50	0,16
Android P	0,22	0,03	0,51	0,03

Erro quadrático próximo de 0

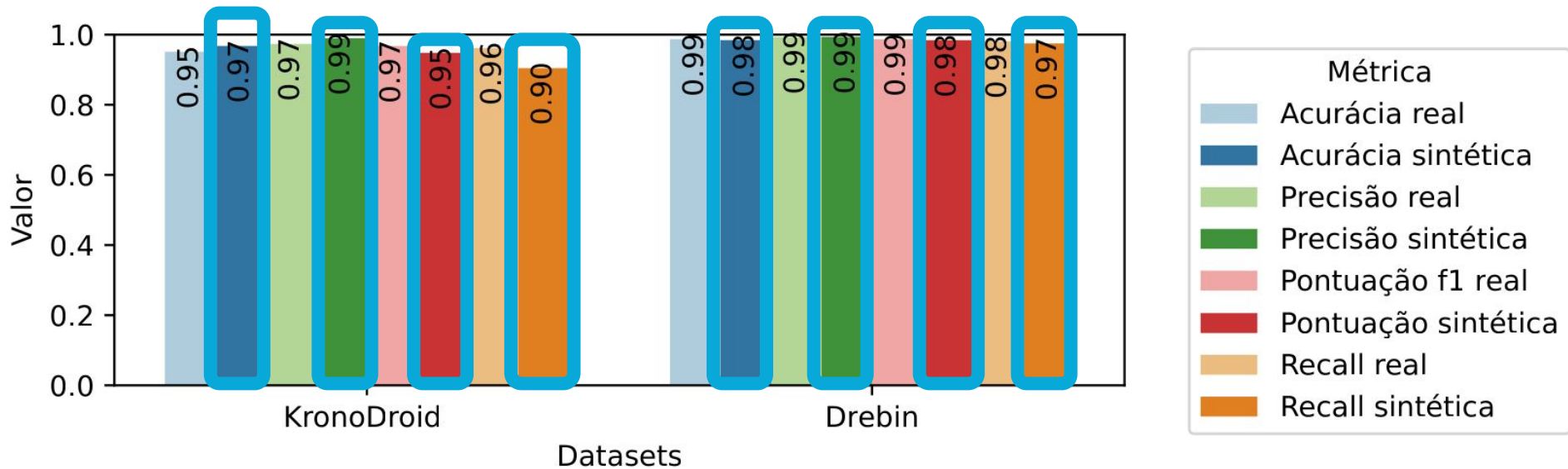
Avaliação (Métricas de utilidade)



Avaliação (Métricas de utilidade)

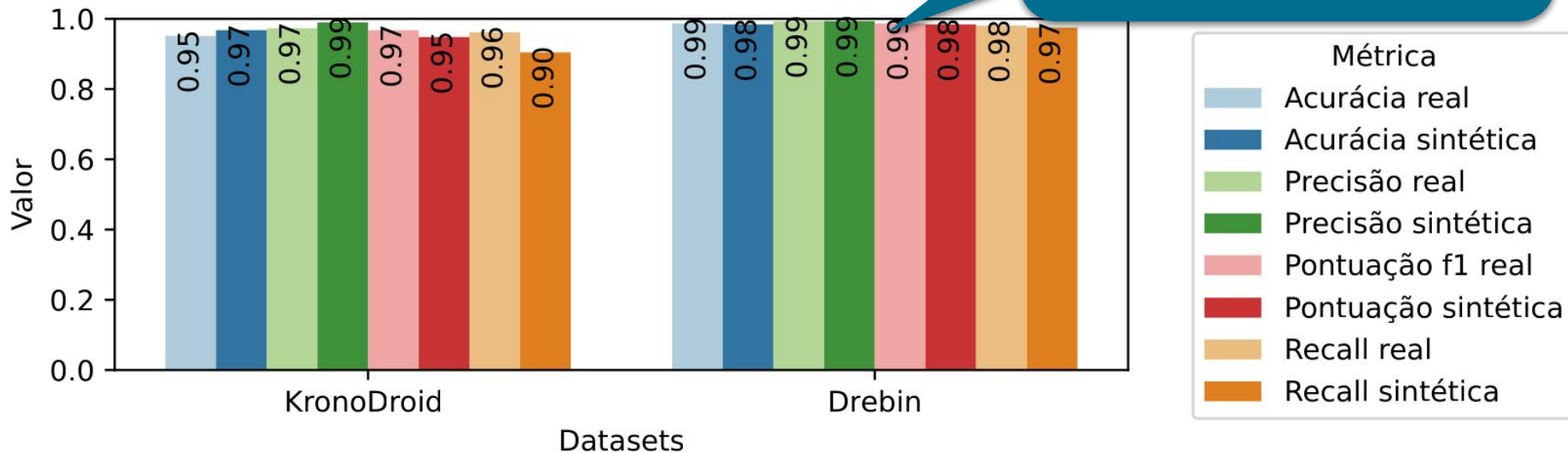


Avaliação (Métricas de utilidade)

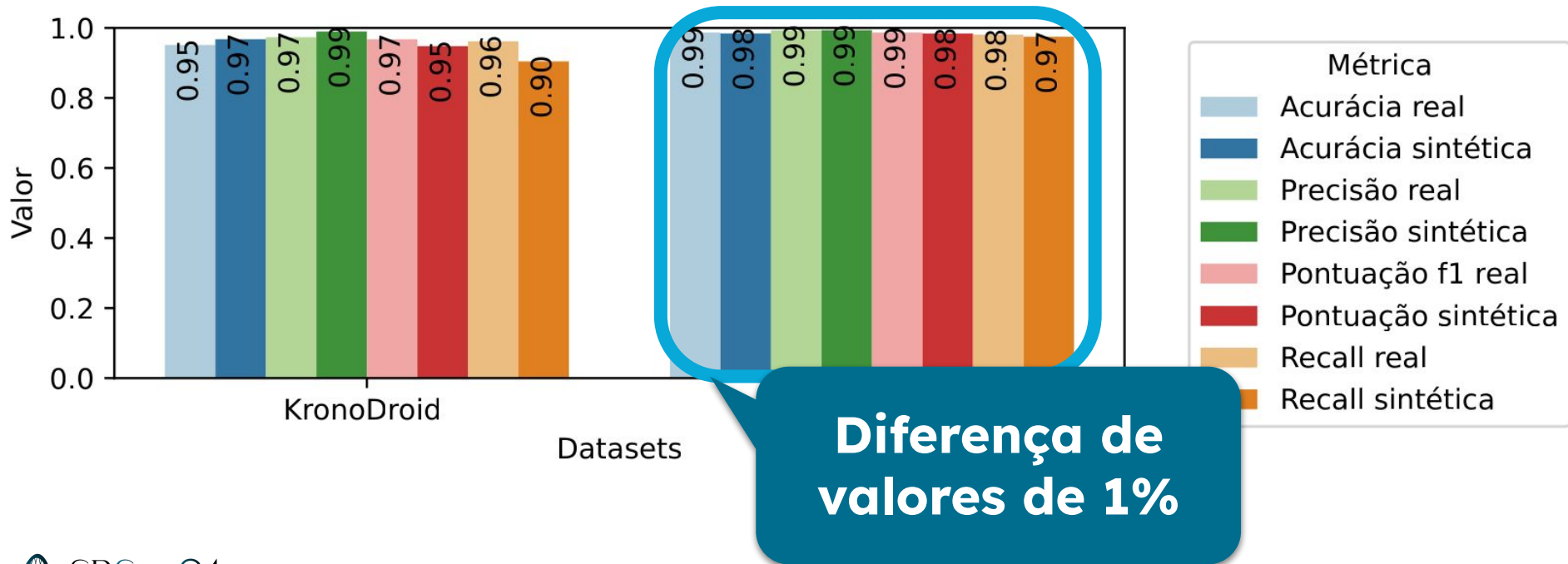


Avaliação (Métricas de utilidade)

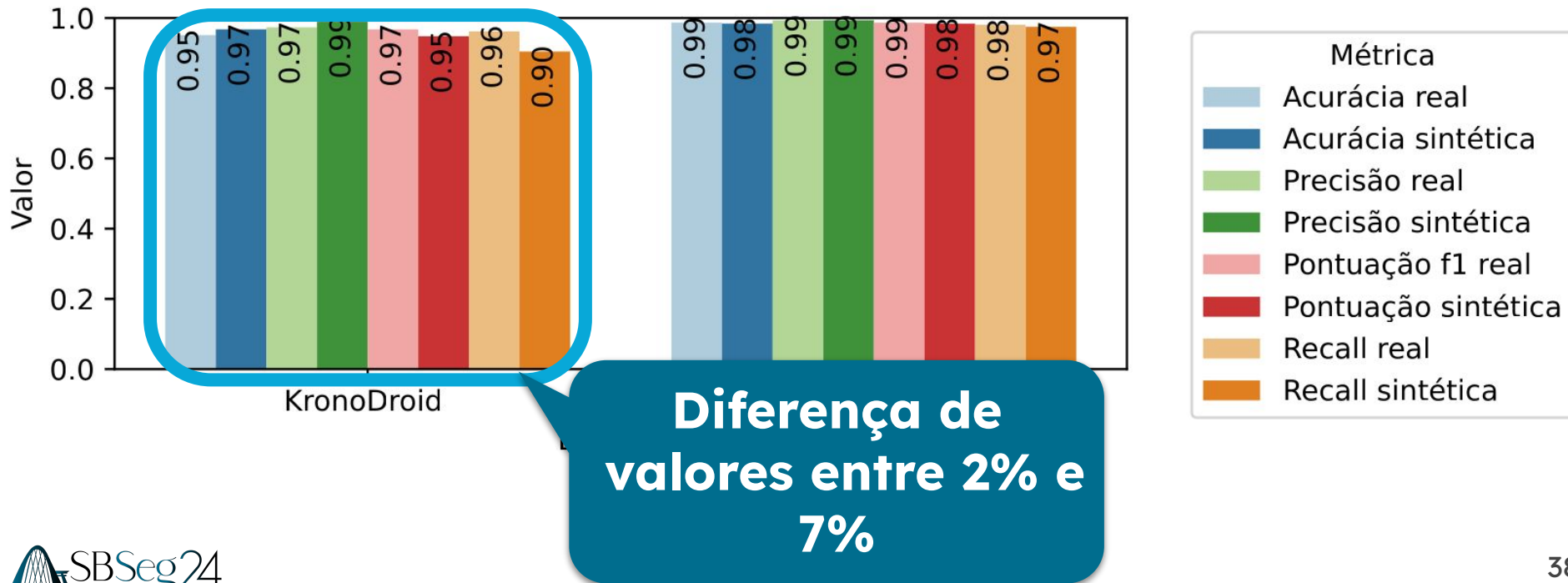
Random Forest



Avaliação (Métricas de utilidade)



Avaliação (Métricas de utilidade)



Avaliação (Métricas de utilidade)

<i>Dataset</i>	<i>P value</i>			
	Random Forest	Decision Tree	Perceptron	SGD Regressor
Krondroid	0,10	0,40	0,19	0,45
Adroit	0,23	0,07	0,05	0,10
Drebin	0,58	0,17	0,08	0,48
Android P	0,25	0,17	0,12	0,11

Avaliação (Métricas de utilidade)

**Todos os
classificadores estão
acima do limiar de
0,05**

	<i>P</i> value			
	Decision tree	Perceptron	SGD Regressor	
Krondroid	0,10	0,40	0,19	0,45
Adroit	0,23	0,07	0,05	0,10
Drebin	0,58	0,17	0,08	0,48
Android P	0,25	0,17	0,12	0,11

Trabalhos relacionados

Trabalho	Métricas	Dataset
Tanaka and Aranha, 2019	Recall, precisão, desvio padrão e distância euclidiana	2 Medicina e 1 Fraude
Mimura, 2020	Acurácia, recall e pontuação F1	1 Malware (Imagens)
Este trabalho	Acurácia, recall, precisão, pontuação F1, erro quadrático, similaridade de cosseno e valor de p	4 Malware (Android)

Considerações Finais

- Demonstramos a viabilidade de geração de dados para o contexto de detecção de Malwares Android
- Os datasets sintetizados são considerados fiéis e úteis
- A otimização dos hiperparâmetros é fundamental

Trabalhos futuros

- Avaliação de outras métricas:
 - Fidelidade
 - Utilidade
 - Privacidade
 - Eficiência computacional
- Uma análise comparativa de desempenho com ferramentas similares (gerais)

Obrigado!



Landing page: <https://malwaredatalab.github.io/>