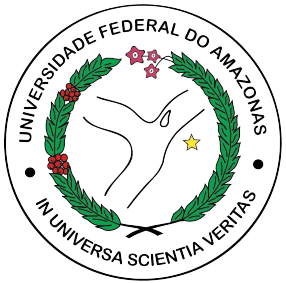
SigAPI AutoCraft: Um Método Otimizado e Generalista de Seleção de Características Para Detecção de *Malwares* Android

Laura C. Tschiedel, **Vanderson Rocha**,
Diego Kreutz, Hendrio Bragança,
Silvio E. Quincozes, Angelo G. D. Nogueira
e Joner Assolin

Universidade Federal do Pampa (UNIPAMPA)
Universidade Federal do Amazonas (UFAM)

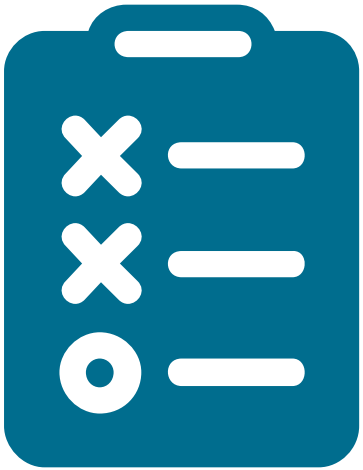


UFAM

unipampa

Universidade Federal do Pampa

Seleção de características



- O que é? É o processo de escolher um subconjunto de características mais relevantes.

- Objetivos:
  - Reduzir o tempo de treinamento
  - Melhorar o desempenho do modelo de classificação

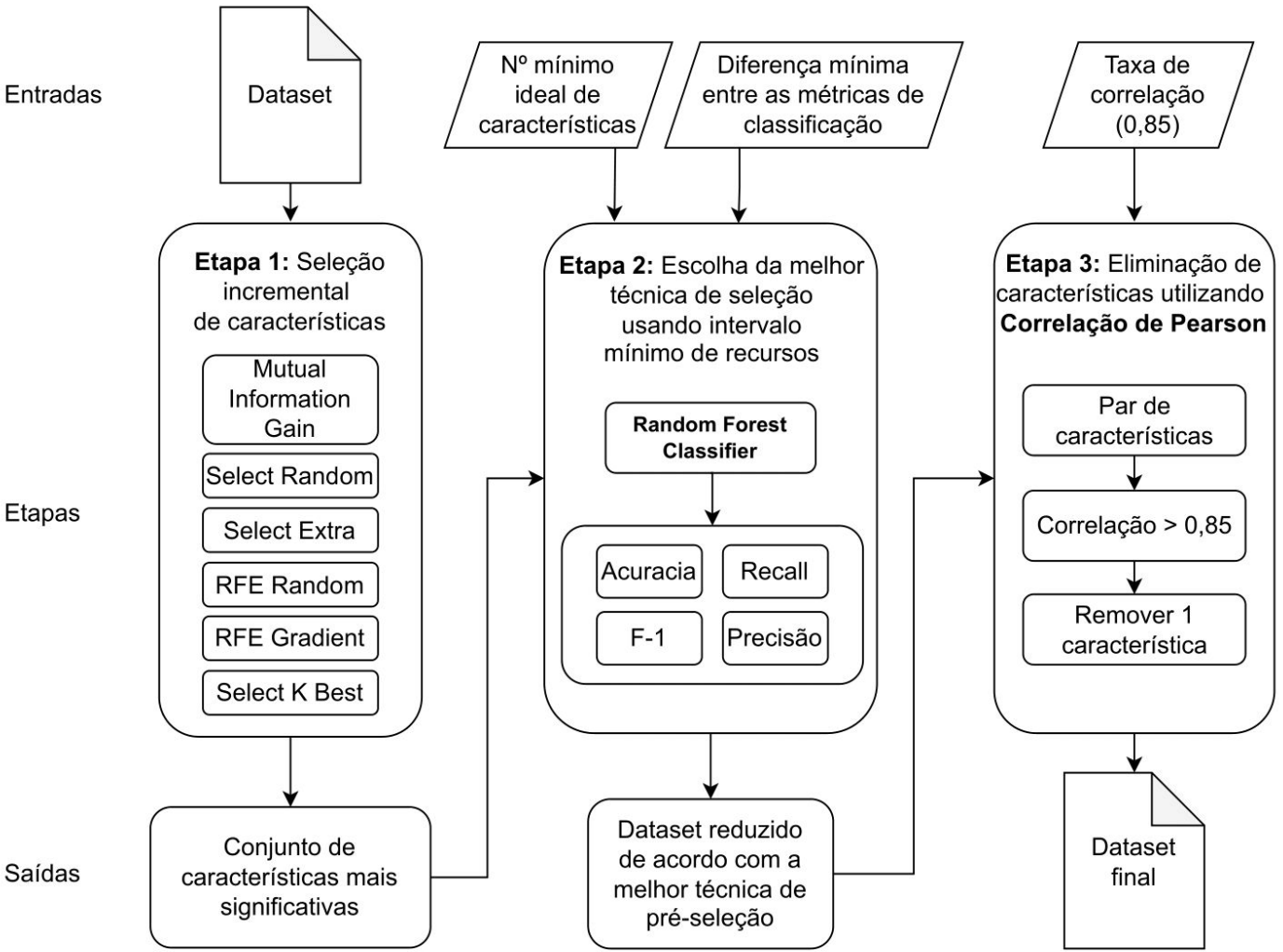
Seleção de características: Desafios

- Alta dimensionalidade
- Redundância
- Irrelevância
- Interpretação do modelo
- Ajuste excessivo de ruídos

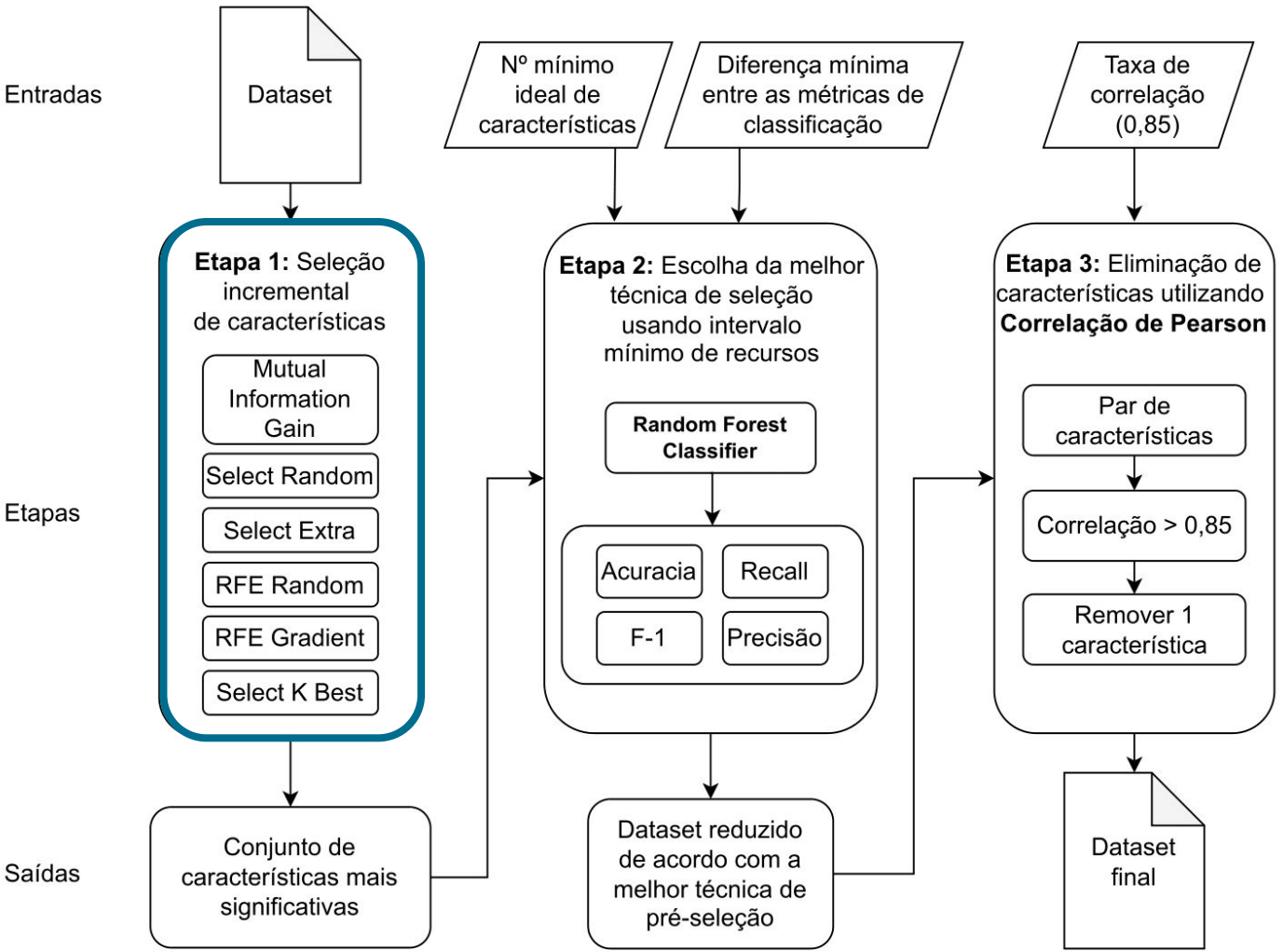
Seleção de características: Benefícios

- Economia de recursos
- Redução no tempo de treinamento
- Melhoria na performance

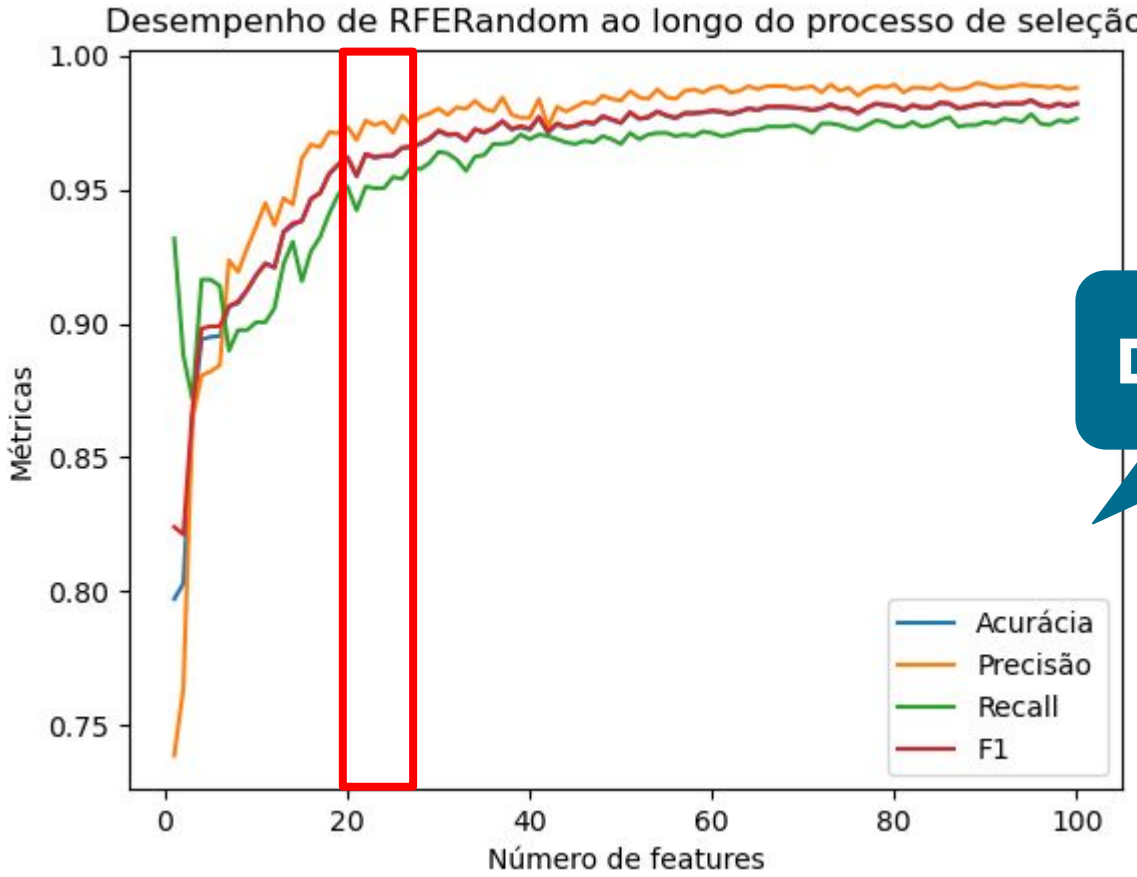
SigAPI


Entradas
Etapas
Saídas
Dataset
Nº mínimo ideal de características
Diferença mínima entre as métricas de classificação
Taxa de correlação (0,85)
Etapa 1: Seleção incremental de características
Mutual Information Gain
Select Random
Select Extra
RFE Random
RFE Gradient
Select K Best
Etapa 2: Escolha da melhor técnica de seleção usando intervalo mínimo de recursos
Random Forest Classifier
Acuracia
Recall
F-1
Precisão
Etapa 3: Eliminação de características utilizando Correlação de Pearson
Par de características
Correlação > 0,85
Remover 1 característica
Conjunto de características mais significativas
Dataset reduzido de acordo com a melhor técnica de pré-seleção
Dataset final

SigAPI

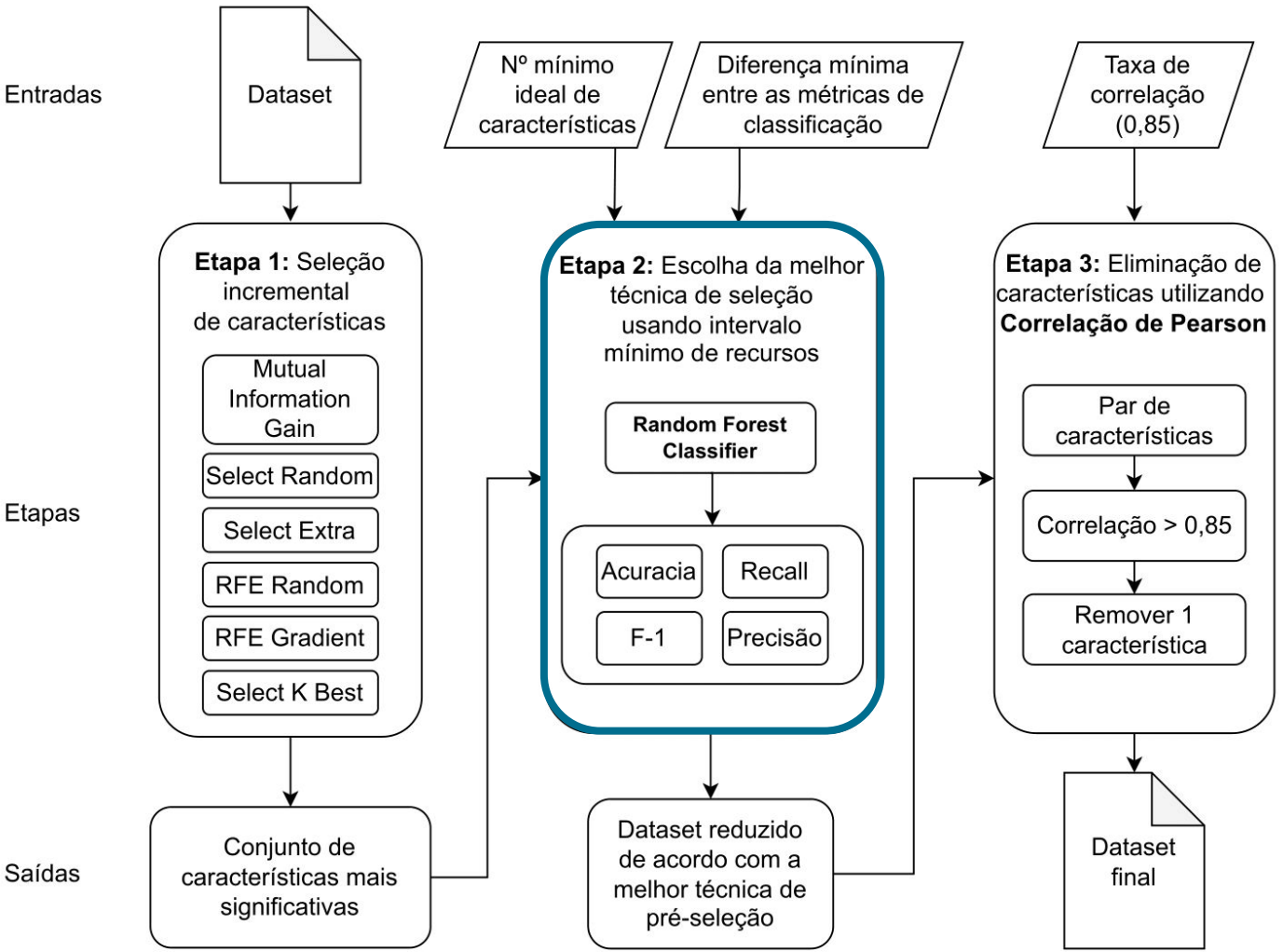

Entradas
Etapas
Saídas
Dataset
Nº mínimo ideal de características
Diferença mínima entre as métricas de classificação
Taxa de correlação (0,85)
Etapa 1: Seleção incremental de características
Mutual Information Gain
Select Random
Select Extra
RFE Random
RFE Gradient
Select K Best
Etapa 2: Escolha da melhor técnica de seleção usando intervalo mínimo de recursos
Random Forest Classifier
Acuracia
Recall
F-1
Precisão
Etapa 3: Eliminação de características utilizando Correlação de Pearson
Par de características
Correlação > 0,85
Remover 1 característica
Conjunto de características mais significativas
Dataset reduzido de acordo com a melhor técnica de pré-seleção
Dataset final

SigAPI

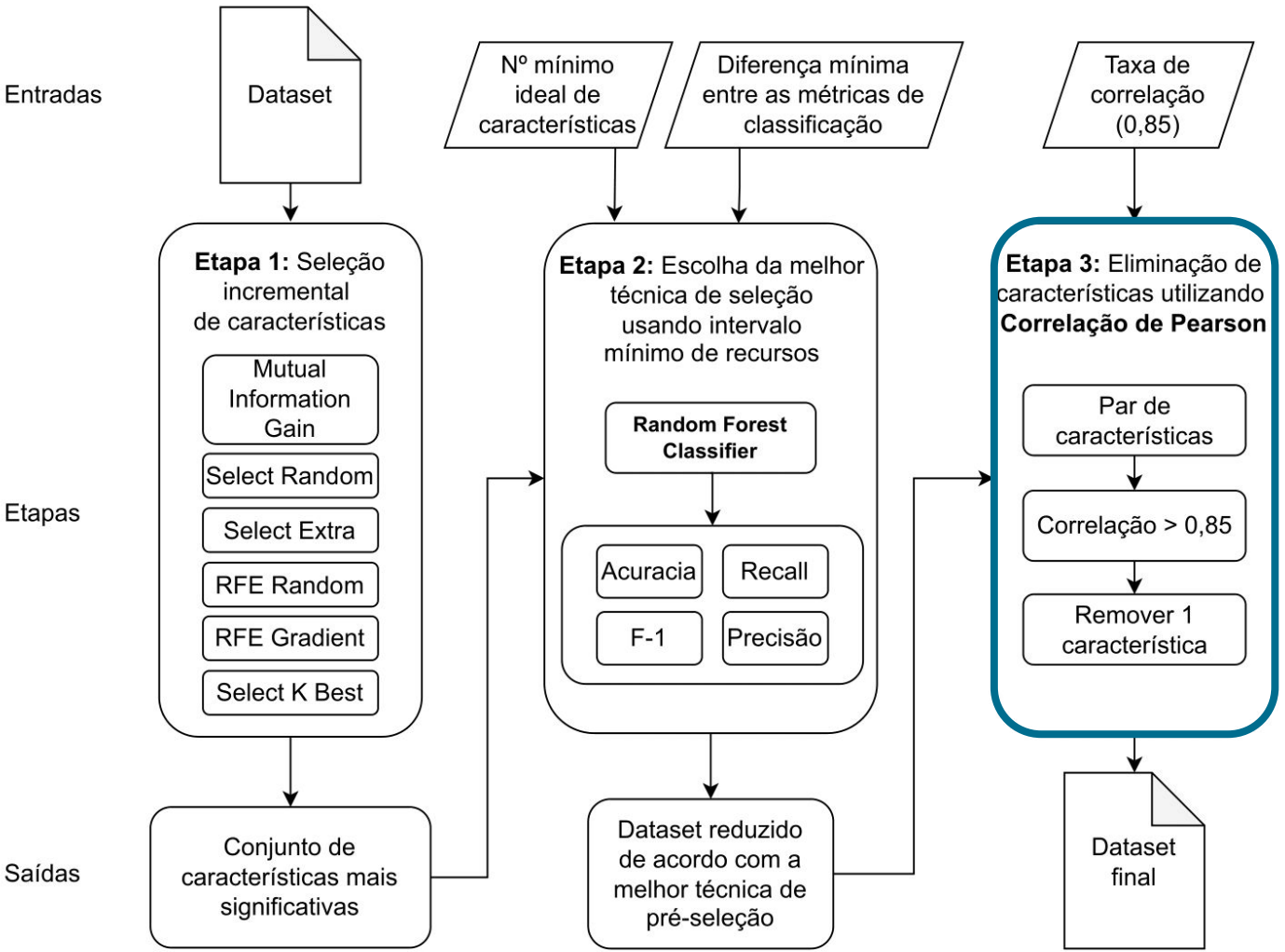


Drebin-215

SigAPI


Entradas
Etapas
Saídas
Dataset
Nº mínimo ideal de características
Diferença mínima entre as métricas de classificação
Taxa de correlação (0,85)
Etapa 1: Seleção incremental de características
Mutual Information Gain
Select Random
Select Extra
RFE Random
RFE Gradient
Select K Best
Etapa 2: Escolha da melhor técnica de seleção usando intervalo mínimo de recursos
Random Forest Classifier
Acuracia
Recall
F-1
Precisão
Etapa 3: Eliminação de características utilizando Correlação de Pearson
Par de características
Correlação > 0,85
Remover 1 característica
Conjunto de características mais significativas
Dataset reduzido de acordo com a melhor técnica de pré-seleção
Dataset final

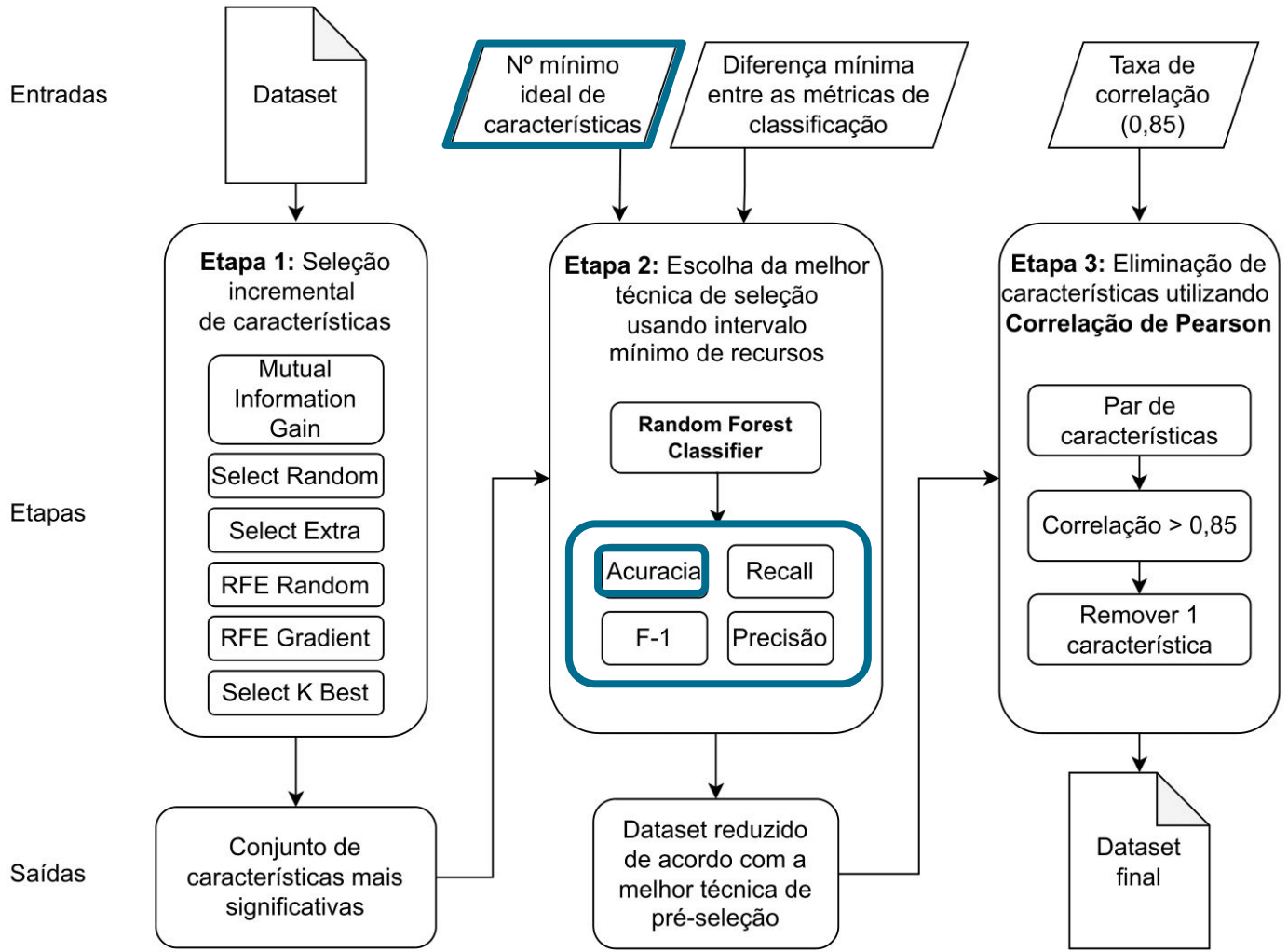
SigAPI


Entradas
Etapas
Saídas
Dataset
Nº mínimo ideal de características
Diferença mínima entre as métricas de classificação
Taxa de correlação (0,85)
Etapa 1: Seleção incremental de características
Mutual Information Gain
Select Random
Select Extra
RFE Random
RFE Gradient
Select K Best
Etapa 2: Escolha da melhor técnica de seleção usando intervalo mínimo de recursos
Random Forest Classifier
Acuracia
Recall
F-1
Precisão
Etapa 3: Eliminação de características utilizando Correlação de Pearson
Par de características
Correlação > 0,85
Remover 1 característica
Conjunto de características mais significativas
Dataset reduzido de acordo com a melhor técnica de pré-seleção
Dataset final

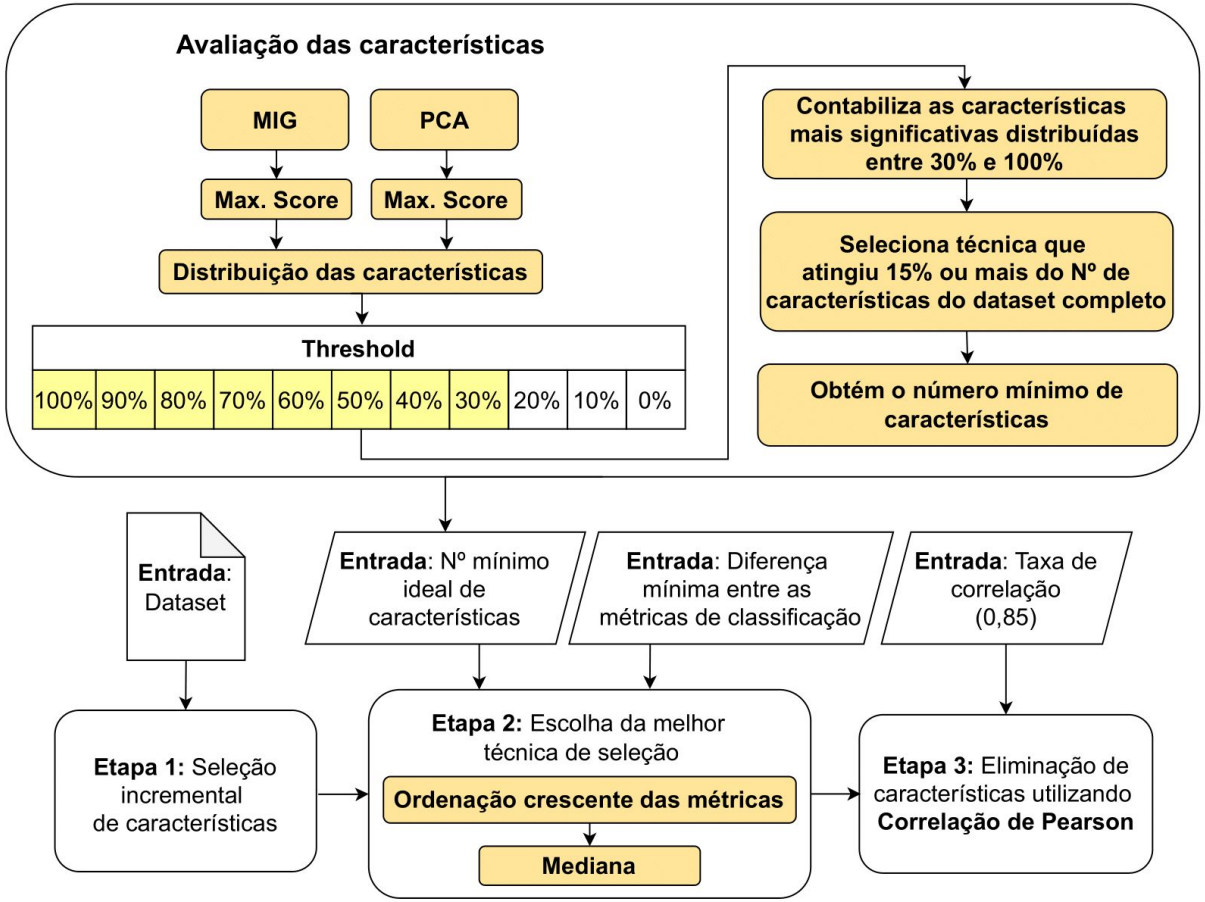
SigAPI: Desafios

- Definição manual do número mínimo de características
- Escolha da melhor técnica de pré-seleção

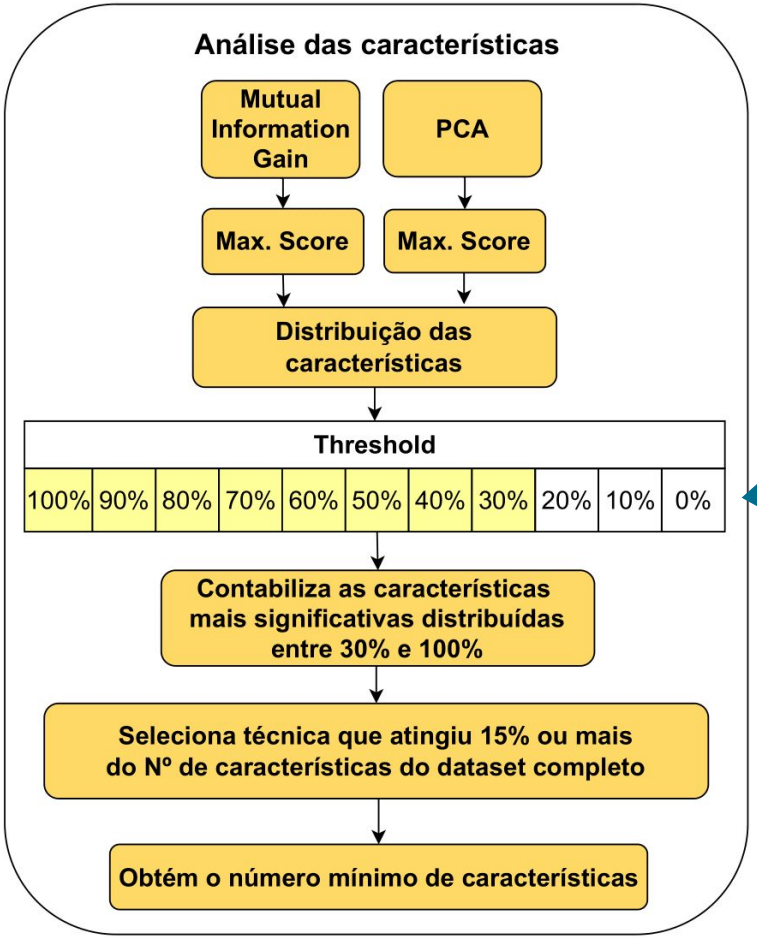
SigAPI


Entradas
Etapas
Saídas
Dataset
Nº mínimo ideal de características
Diferença mínima entre as métricas de classificação
Taxa de correlação (0,85)
Etapa 1: Seleção incremental de características
Mutual Information Gain
Select Random
Select Extra
RFE Random
RFE Gradient
Select K Best
Etapa 2: Escolha da melhor técnica de seleção usando intervalo mínimo de recursos
Random Forest Classifier
Acuracia
Recall
F-1
Precisão
Etapa 3: Eliminação de características utilizando Correlação de Pearson
Par de características
Correlação > 0,85
Remover 1 característica
Conjunto de características mais significativas
Dataset reduzido de acordo com a melhor técnica de pré-seleção
Dataset final

SigAPI AutoCraft



Número mínimo de características



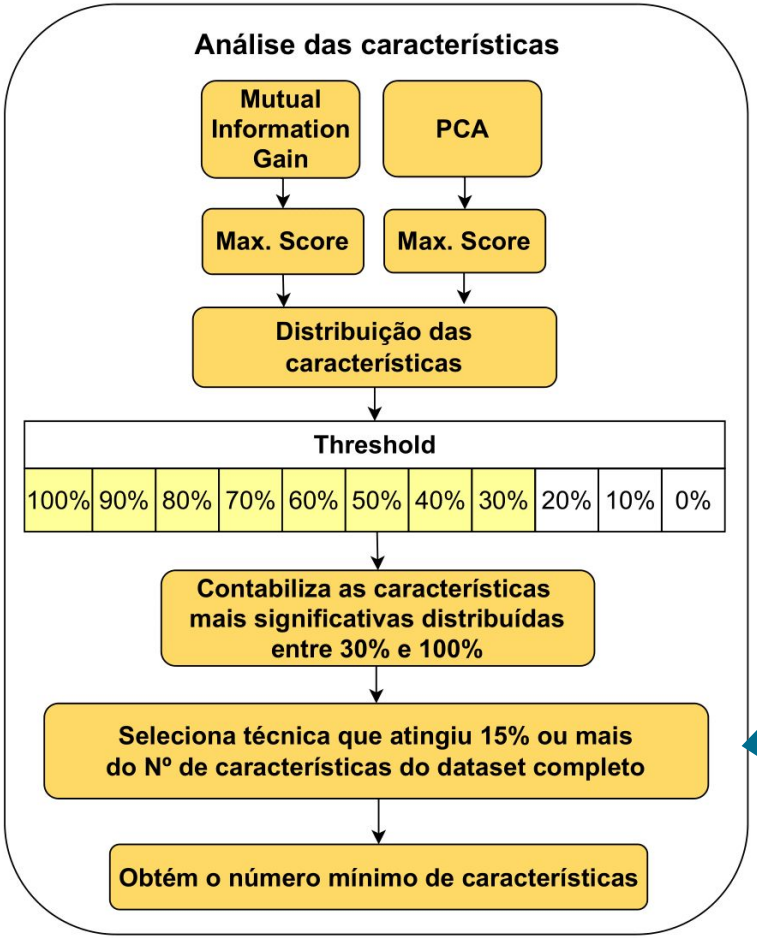
As características são agrupadas a cada threshold decremental de 10%.

Número mínimo de características

MAX. IG: 0,215	Threshold	100%	90%	80%	70%	60%	50%	40%	30%	20%	10%	0%
	Características	4	2	2	2	3	7	5	12	28	108	35

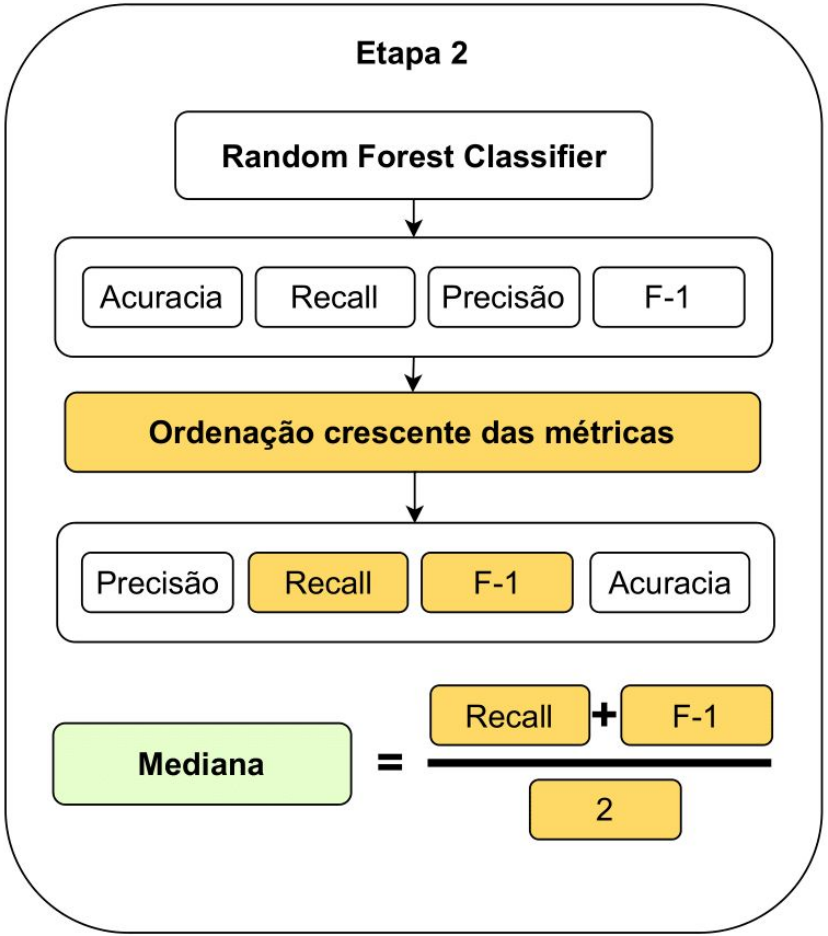
Dataset Drebin-215

Número mínimo de características



Se o MIG e o PCA não chegarem a 15% do nº total de características do dataset completo?

Técnica de seleção



Datasets

- 10 *datasets* distintos
- Domínio de detecção de *malware* Android
- Tipos de características
  - Chamadas de API
  - Permissões
- Balanceamento e redução
  - Random under sampler
  - Valor qui-quadrado

Datasets

Datasets Balanceados		Amostras				
Datasets	**Características**	**Benignas**		**Malwares**		**Total**
Androcrawl	81	10.170	50%	10.170	50%	20.340
Drebin-215	215	5.555	50%	5.555	50%	11.110
Adroit	166	3.418	50%	3.418	50%	6.836
Android Permissions	151	9.077	50%	9.077	50%	18.154
Kronodroid Real Device	286	36.755	50%	36.755	50%	73.510
Defensedroid Katz	200	5.222	50%	5.222	50%	10.444
Defensedroid Degree	200	5.222	50%	5.222	50%	10.444
Defensedroid Closeness	200	5.222	50%	5.222	50%	10.444
Defensedroid PRS	200	5.975	50%	5.975	50%	11.950
MH-100K	200	10.000	50%	10.000	50%	20.000

Resultados

Datasets	Datasets Completos		SigAPI Original		SigAPI AutoCraft	
	Características	Roc_AuC	Redução	RoC_AuC	Redução	RoC_AuC
Androcrawl	81	97.4%	97.5%	82.8%	76,6%	95.6%
Drebin-215	215	99.8%	99.5%	78.5%	85.2%	99.4%
Adroit	166	91.3%	99.3%	82.4%	89.8%	90.1%
Android Permissions	151	70.8%	98.6%	58.8%	76.9%	69.9%
Kronodroid Real Device	286	99.7%	99.3%	82.6%	87.5%	99.1%
Defensedroid Katz	200	98.8%	99%	79.5%	88%	96.1%
Defensedroid Degree	200	98.8%	99%	79.5%	88.5%	95.3%
Defensedroid Closeness	200	99.1%	99%	78.8%	88%	95.5%
Defensedroid PRS	200	96.9%	99%	78.1%	86%	95.5%
MH-100K	200	96.6%	99.5%	95.4%	97.5%	95.4%

Resultados

Datasets	Datasets Completos		SigAPI Original		SigAPI AutoCraft	
	Características	Roc_AuC	Redução	RoC_AuC	Redução	RoC_AuC
Androcrawl	81	97.4%	97.5%	82.8%	76,6%	95.6%
Drebin-215	215	99.8%	99.5%	78.5%	85.2%	99.4%
Adroit	166	91.3%	99.3%	82.4%	89.8%	90.1%
Android Permissions	151	70.8%	98.6%	58.8%	76.9%	69.9%
Kronodroid Real Device	286	99.7%	99.3%	82.6%	87.5%	99.1%
Defensedroid Katz	200	98.8%	99%	79.5%	88%	96.1%
Defensedroid Degree	200	98.8%	99%	79.5%	88.5%	95.3%
Defensedroid Closeness	200	99.1%	99%	78.8%	88%	95.5%
Defensedroid PRS	200	96.9%	99%	78.1%	86%	95.5%
MH-100K	200	96.6%	99.5%	95.4%	97.5%	95.4%

Resultados

Datasets	Datasets Completos		SigAPI Original		SigAPI AutoCraft	
	Características	Roc_AuC	Redução	RoC_AuC	Redução	RoC_AuC
Androcrawl	81	97.4%	97.5%	82.8%	76,6%	95.6%
Drebin-215	215	99.8%	99.5%	78.5%	85.2%	99.4%
Adroit	166	91.3%	99.3%	82.4%	89.8%	90.1%
Android Permissions	151	70.8%	98.6%	58.8%	76.9%	69.9%
Kronodroid Real Device	286	99.7%	99.3%	82.6%	87.5%	99.1%
Defensedroid Katz	200	98.8%	99%	79.5%	88%	96.1%
Defensedroid Degree	200	98.8%	99%	79.5%	88.5%	95.3%
Defensedroid Closeness	200	99.1%	99%	78.8%	88%	95.5%
Defensedroid PRS	200	96.9%	99%	78.1%	86%	95.5%
MH-100K	200	96.6%	99.5%	95.4%	97.5%	95.4%

Resultados

Datasets	Datasets Completos		SigAPI Original		SigAPI AutoCraft	
	Características	Roc_AuC	Redução	RoC_AuC	Redução	RoC_AuC
Androcrawl	81	97.4%	97.5%	82.8%	76,6%	95.6%
Drebin-215	215	99.8%	99.5%	78.5%	85.2%	99.4%
Adroit	166	91.3%	99.3%	82.4%	89.8%	90.1%
Android Permissions	151	70.8%	98.6%	58.8%	76.9%	69.9%
Kronodroid Real Device	286	99.7%	99.3%	82.6%	87.5%	99.1%
Defensedroid Katz	200	98.8%	99%	79.5%	88%	96.1%
Defensedroid Degree	200	98.8%	99%	79.5%	88.5%	95.3%
Defensedroid Closeness	200	99.1%	99%	78.8%	88%	95.5%
Defensedroid PRS	200	96.9%	99%	78.1%	86%	95.5%
MH-100K	200	96.6%	99.5%	95.4%	97.5%	95.4%

Resultados

Datasets	Datasets Completos		SigAPI Original		SigAPI AutoCraft	
	Características	Roc_AuC	Redução	RoC_AuC	Redução	RoC_AuC
Androcrawl	81	97.4%	97.5%	82.8%	76,6%	95.6%
Drebin-215	215	99.8%	99.5%	78.5%	85.2%	99.4%
Adroit	166	91.3%	99.3%	82.4%	89.8%	90.1%
Android Permissions	151	70.8%	98.6%	58.8%	76.9%	69.9%
Kronodroid Real Device	286	99.7%	99.3%	82.6%	87.5%	99.1%
Defensedroid Katz	200	98.8%	99%	79.5%	88%	96.1%
Defensedroid Degree	200	98.8%	99%	79.5%	88.5%	95.3%
Defensedroid Closeness	200	99.1%	99%	78.8%	88%	95.5%
Defensedroid PRS	200	96.9%	99%	78.1%	86%	95.5%
MH-100K	200	96.6%	99.5%	95.4%	97.5%	95.4%

SigAPI AutoCraft: Vantagens

- Capacidade de generalização independente do dataset
- Condição de parada dinâmica e automática
- Mantém uma taxa de redução elevada
- Mantém o resultado de predição do dataset reduzido próximo ao dataset completo
- Bom seletor de características

Repositório

- Exemplos de uso
- Documentação
- Scripts de demonstração
- Uso em docker
- Links

Demonstração

Obrigado!

UFAM

Universidade Federal do Pampa

FAPERGS

CAPES