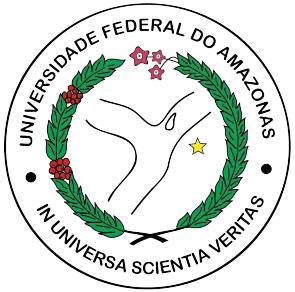
MH-FSF: um Framework para Reprodução, Experimentação e Avaliação de Métodos de Seleção de Características



UFAM

unipampa

Universidade Federal do Pampa

Vanderson Rocha, Hendrio Bragança, Diego Kreutz e Eduardo Feitosa

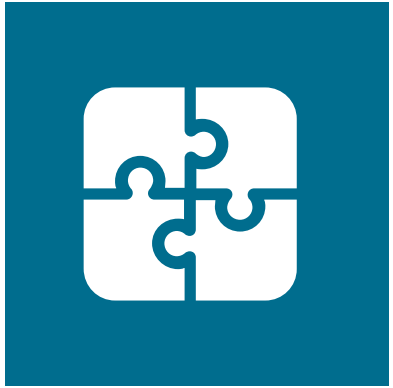
Universidade Federal do Amazonas (UFAM)
Universidade Federal do Pampa (UNIPAMPA)

Desafios

Adaptação
contínua
dos malwares

Desafios

Adaptação
contínua
dos malwares



Seleção
dinâmica
de características

Desafios

Adaptação
contínua
dos malwares

Seleção
dinâmica
de características

Volume
e diversidade
de dados

Desafios

Adaptação contínua dos malwares

Seleção dinâmica de características

Volume e diversidade de dados

Impacto na eficácia dos modelos

Limitações Atuais

Dependência de um único conjunto de dados;

Referência	# Métodos	Datasets
Sahin et al., 2023a	8	Próprio (APKPure, VirusShare)
Sahin et al., 2023b	11	Próprio (APKPure)
Islam et al., 2023	2	CICAndMal2020, Drebin, CICMaldroid2020
Salah et al., 2020	2	Drebin
Mahindru & Sangal, 2019	8	Próprio (Google Play)
Fatima et al., 2019	2	Próprio (IIT Kanpur)
Zhao et al., 2015	3	Drebin
Alomari et al., 2023	2	Kaggle

Limitações Atuais

Dificuldade na reprodutibilidade e verificação independente;

Referência	# Métodos	Datasets
Sahin et al., 2023a	8	Próprio (APKPure, VirusShare)
Sahin et al., 2023b	11	Próprio (APKPure)
Islam et al., 2023	2	CICAndMal2020, Drebin, CICMaldroid2020
Salah et al., 2020	2	Drebin
Mahindru & Sangal, 2019	8	Próprio (Google Play)
Fatima et al., 2019	2	Próprio (IIT Kanpur)
Zhao et al., 2015	3	Drebin
Alomari et al., 2023	2	Kaggle

Limitações Atuais

Comparação entre (poucos) métodos com datasets distintos.

Referência	# Métodos	Datasets
Sahin et al., 2023a	8	Próprio (APKPure, VirusShare)
Sahin et al., 2023b	11	Próprio (APKPure)
Islam et al., 2023	2	CICAndMal2020, Drebin, CICMaldroid2020
Salah et al., 2020	2	Drebin
Mahindru & Sangal, 2019	8	Próprio (Google Play)
Fatima et al., 2019	2	Próprio (IIT Kanpur)
Zhao et al., 2015	3	Drebin
Alomari et al., 2023	2	Kaggle

MH-FSF: Malware Hunter Features Selection Framework

MH-FSF: Objetivos

- Facilitar a incorporação de diversos métodos de seleção de características
- Permitir comparação direta entre diferentes abordagens de seleção de características, modelos de classificação e métricas
- Possibilitar o uso de técnicas mais eficazes e eficientes em diversas aplicações preditivas

Pipeline: Manipulação de dados

Dataset	Características		Amostas		
	#	Tipo	Maliciosas	Benignas	Total
ADROIT	166	P	3418	8058	11476
AndroCrawl	81	A (24) I (8) P (49)	10170	86562	96732
Android Permissions	183	P	20000	9999	29999
DefenseDroid PI	2938	P (1490) I (1448)	6000	5975	11975
DefensoDroid A (C)	4275	A	5254	5222	10476
DefensoDroid A (D)	6003				
DefensoDroid A (K)	6003				
DREBIN-215	215	A (73) P (113) O (6) I (23)	5555	9476	15036
KronoDroid R	246	P (146) A (100)	41382	36755	78137
KronoDroid E	268	P (145) A (123)	28745	35246	63991

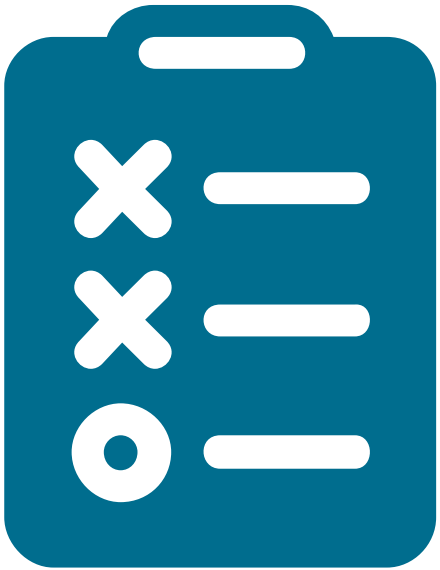
Pipeline: Manipulação de dados

Dataset	Características		Amostas		
	#	Tipo	Maliciosas	Benignas	Total
ADROIT	166	P	3418	8058	11476
AndroCrawl	81	A (24) I (8) P (49)	10170	86562	96732
Android Permissions	183	P	20000	9999	29999
DefenseDroid PI	2938	P (1490) I (1448)	6000	5975	11975
DefensoDroid A (C)	4275	A	5254	5222	10476
DefensoDroid A (D)	6003				
DefensoDroid A (K)	6003				
DREBIN-215	215	A (73) P (113) O (6) I (23)	5555	9476	15036
KronoDroid R	246	P (146) A (100)	41382	36755	78137
KronoDroid E	268	P (145) A (123)	28745	35246	63991

Pipeline: Manipulação de dados

Dataset	Características		Amostas		
	#	Tipo	Maliciosas	Benignas	Total
ADROIT	166	P	3418	8058	11476
AndroCrawl	81	A (24) I (8) P (49)	10170	86562	96732
Android Permissions	183	P	20000	9999	29999
DefenseDroid PI	2938	P (1490) I (1448)	6000	5975	11975
DefensoDroid A (C)	4275	A	5254	5222	10476
DefensoDroid A (D)	6003				
DefensoDroid A (K)	6003				
DREBIN-215	215	A (73) P (113) O (6) I (23)	5555	9476	15036
KronoDroid R	246	P (146) A (100)	41382	36755	78137
KronoDroid E	268	P (145) A (123)	28745	35246	63991

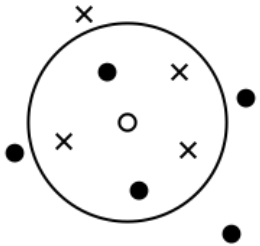
Pipeline: Seleção de características



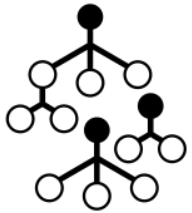
- Métodos clássicos
  - ABC
  - ANOVA
  - Qui-quadrado
  - IG
  - LASSO
  - LR
  - MAD
  - PCA
  - PCC
  - ReliefF
  - RFE
- Métodos específicos
  - JOWNDroid
  - MT
  - RFG
  - SemiDroid
  - SigAPI
  - SigPID

Pipeline: Treinamento e avaliação

- Treinamento de modelos



KNN
(agrupamento)



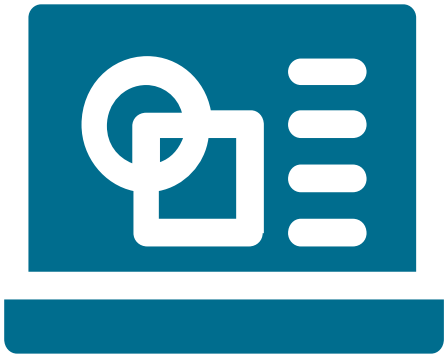
RF
(árvore)

SVM
(cluster)

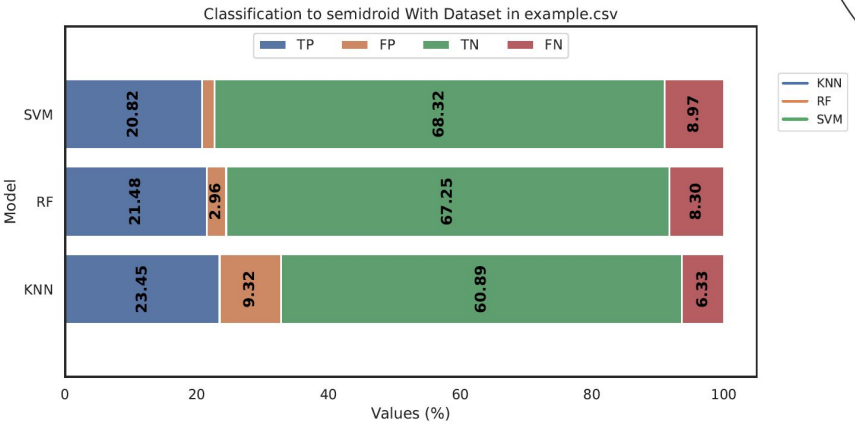
Pipeline: Treinamento e avaliação

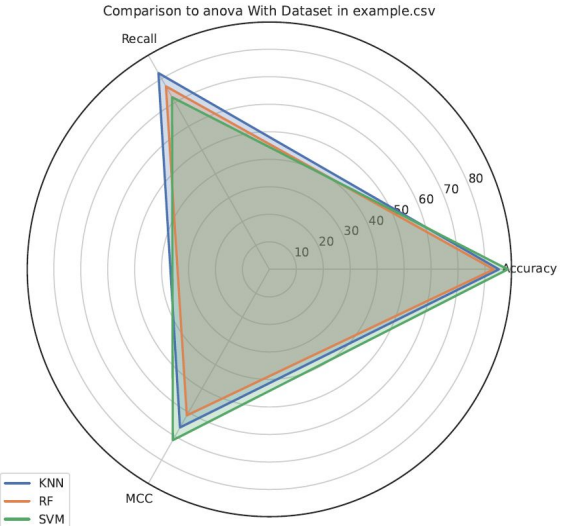
- Métricas de avaliação
  - Acurácia
  - Precisão
  - Recall
  - F1-Score
  - ROC-AuC
  - MCC
  - Matriz de Confusão

Pipeline: Visualização

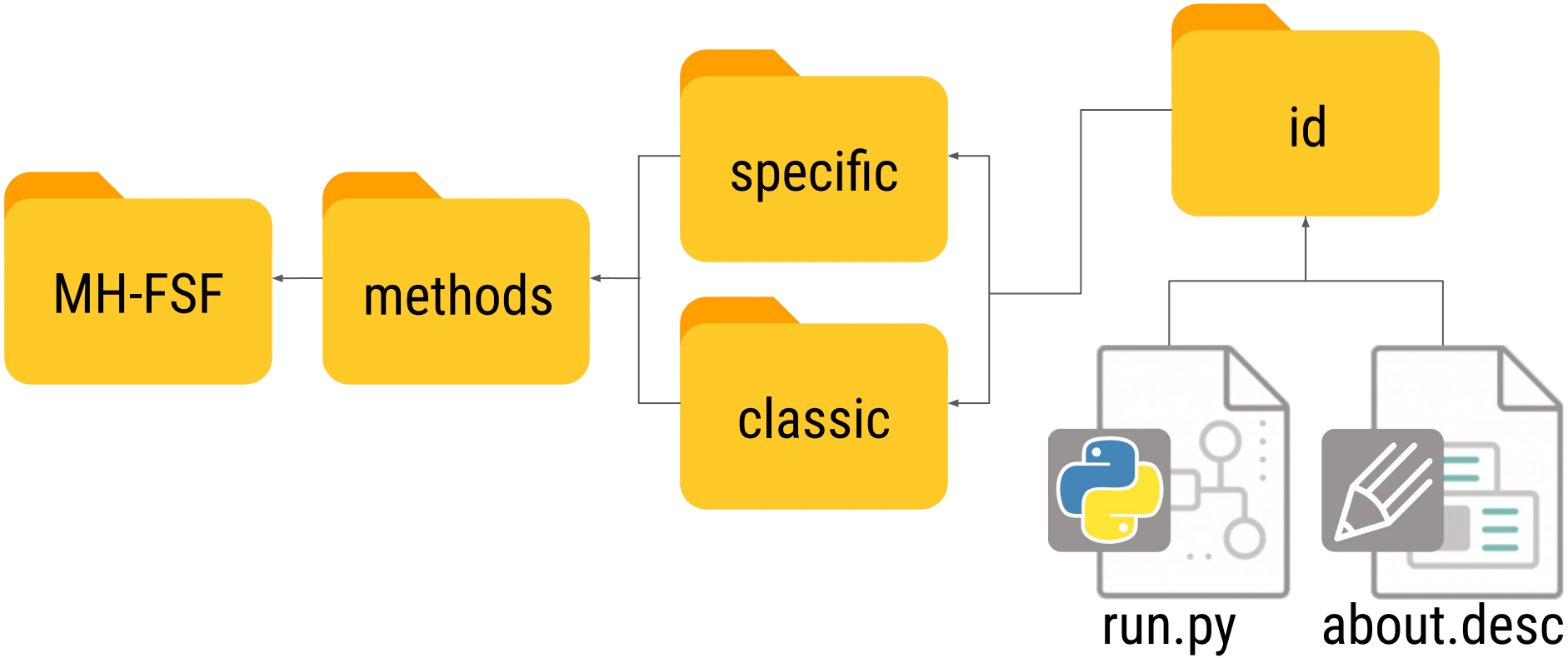


- Visualização
  - Gráfico de barras
  - Matriz de confusão
  - Gráfico de radar

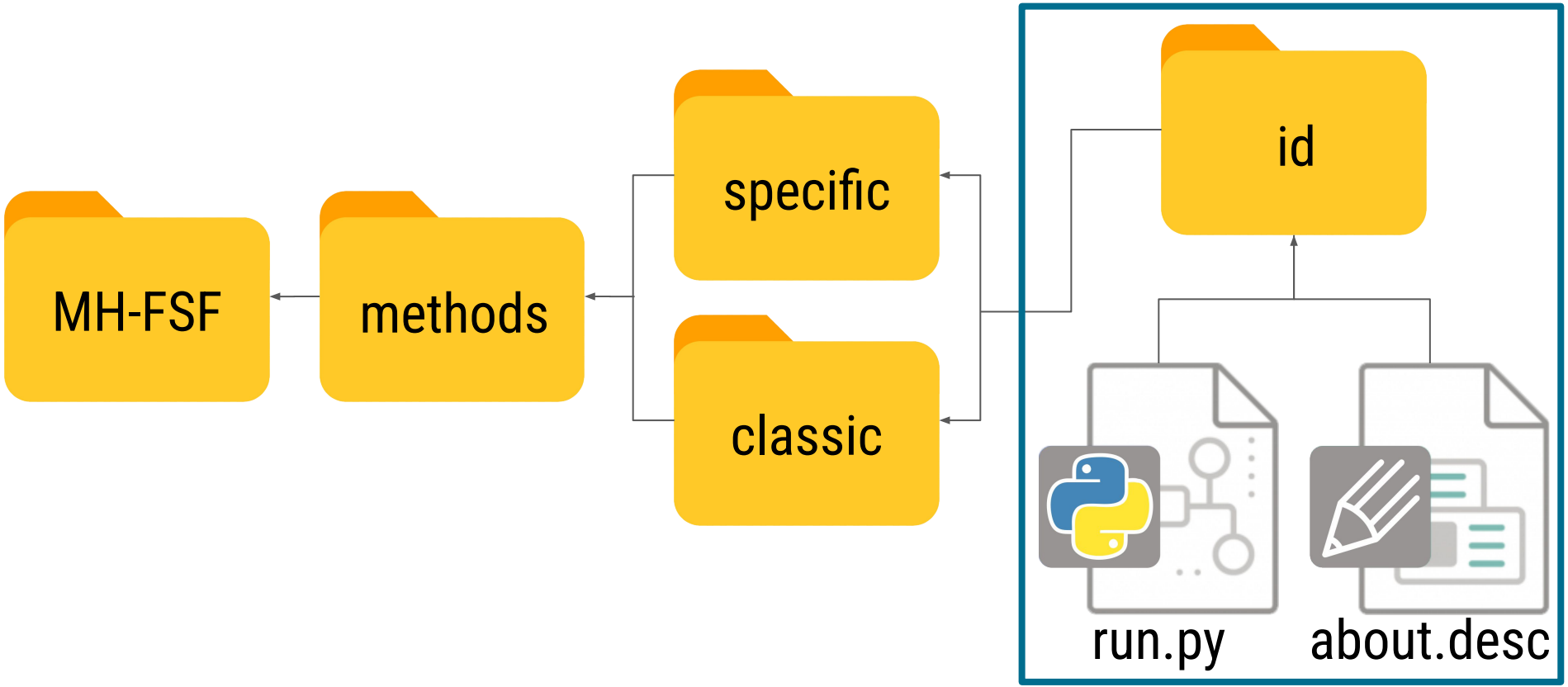




Extensibilidade de métodos



Extensibilidade de métodos



Extensibilidade de métodos

```
# Import libraries
def add_arguments(parser):
    parser = parser.add_argument_group('Arguments for Method')
# Others functions, if necessary
def run(args, path, dataset):
    . . .
    #To save reduced dataset
    method_id = 'use the same name as method directory'
    filename = f'{method_id}_{os.path.basename(path)}'
    output_file = os.path.join(args.output, filename)
    reduced_dataset.to_csv(output_file, index = False)
    return True
```

Extensibilidade de modelos

evaluation.py

```
available_ml_models = {
    'svm': svm.SVC(),
    'rf': RandomForestClassifier(random_state = 0),
    'knn': KNeighborsClassifier()
}
```

Resultados

Balanceados				Desbalanceados			
Método	**F1**	**Método**	**Recall**	**Método**			**Recall**
LASSO	0,90	LASSO	0,89	LASSO			0,91
RFE	0,90	RFE	0,89	RFE	0.90	RFE	0,90
SemiDroid	0,90	SemiDroid	0,89	SigAPI	0,90	SigAPI	0,90
JOWMDroid	0,90	JOWMDroid	0,89	MAD	0,90	PCC	0,90
...							
RFG	0,72	RFG	0,70	PCA	0,67	SigPID	0,67
ReliefF	0,71	ReliefF	0,68	SigPID	0,65	PCA	0,66
SigPID	0,71	SigPID	0,66	JOWMDroid	0,65	JOWMDroid	0,64
PCA	0,67	PCA	0,63	ReliefF	0,63	ReliefF	0,64

Métodos específicos

Resultados

Balanceados				Desbalanceados			
Método	**F1**	**Método**	**Recall**	**Método**	**F1**	**Método**	**Recall**
LASSO	0,90	LASSO	0,89	LASSO	0,91	LASSO	0,91
RFE	0,90	RFE	0,89	RFE	0.90	RFE	0,90
SemiDroid	0,90	SemiDroid	0,89	SigAPI	0,90	SigAPI	0,90
JOWMDroid	0,90	JOWMDroid	0,89	MAD	0,90	PCC	0,90
...							
RFG	0,72	RFG	0,70	PCA	0,67	SigPID	0,67
ReliefF	0,71	ReliefF	0,68	SigPID	0,65	PCA	0,66
SigPID	0,71	SigPID	0,66	JOWMDroid	0,65	JOWMDroid	0,64
PCA	0,67	PCA	0,63	ReliefF	0,63	ReliefF	0,64

Resultados

Balanceados				Desbalanceados			
Método	**F1**	**Método**	**Recall**	**Método**	**F1**	**Método**	**Recall**
LASSO	0,90	LASSO	0,89	LASSO	0,91	LASSO	0,91
RFE	0,90	RFE	0,89	RFE	0.90	RFE	0,90
SemiDroid	0,90	SemiDroid	0,89	SigAPI	0,90	SigAPI	0,90
JOWMDroid	0,90	JOWMDroid	0,89	MAD	0,90	PCC	0,90
...							
RFG	0,72	RFG	0,70	PCA	0,67	SigPID	0,67
ReliefF	0,71	ReliefF	0,68	SigPID	0,65	PCA	0,66
SigPID	0,71	SigPID	0,66	JOWMDroid	0,65	JOWMDroid	0,64
PCA	0,67	PCA	0,63	ReliefF	0,63	ReliefF	0,64

Resultados

Balanceados				Desbalanceados			
Método	**F1**	**Método**	**Recall**	**Método**	**F1**	**Método**	**Recall**
LASSO	0,90	LASSO	0,89	LASSO	0,91	LASSO	0,91
RFE	0,90	RFE	0,89	RFE	0.90	RFE	0,90
SemiDroid	0,90	SemiDroid	0,89	SigAPI	0,90	SigAPI	0,90
JOWMDroid	0,90	JOWMDroid	0,89	MAD	0,90	PCC	0,90
...							
RFG	0,72	RFG	0,70	PCA	0,67	SigPID	0,67
ReliefF	0,71	ReliefF	0,68	SigPID	0,65	PCA	0,66
SigPID	0,71	SigPID	0,66	JOWMDroid	0,65	JOWMDroid	0,64
PCA	0,67	PCA	0,63	ReliefF	0,63	ReliefF	0,64

Repositório

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

Demonstração

Parâmetro	Ajuda
--fs-types	Types of feature selection (FS) methods
--fs-methods	Run selected methods
--ml-models	ML model for evaluation
-d/–datasets	One or more datasets (csv files)
-th/--threshold	Percent of features to be selected
--parallelize	Parallel execution

Obrigado!

Vanderson Rocha

vanderson@ufam.edu.br

ppgi.ufam.edu.br

CAPES

MOTOROLA MOBILITY