UFAM

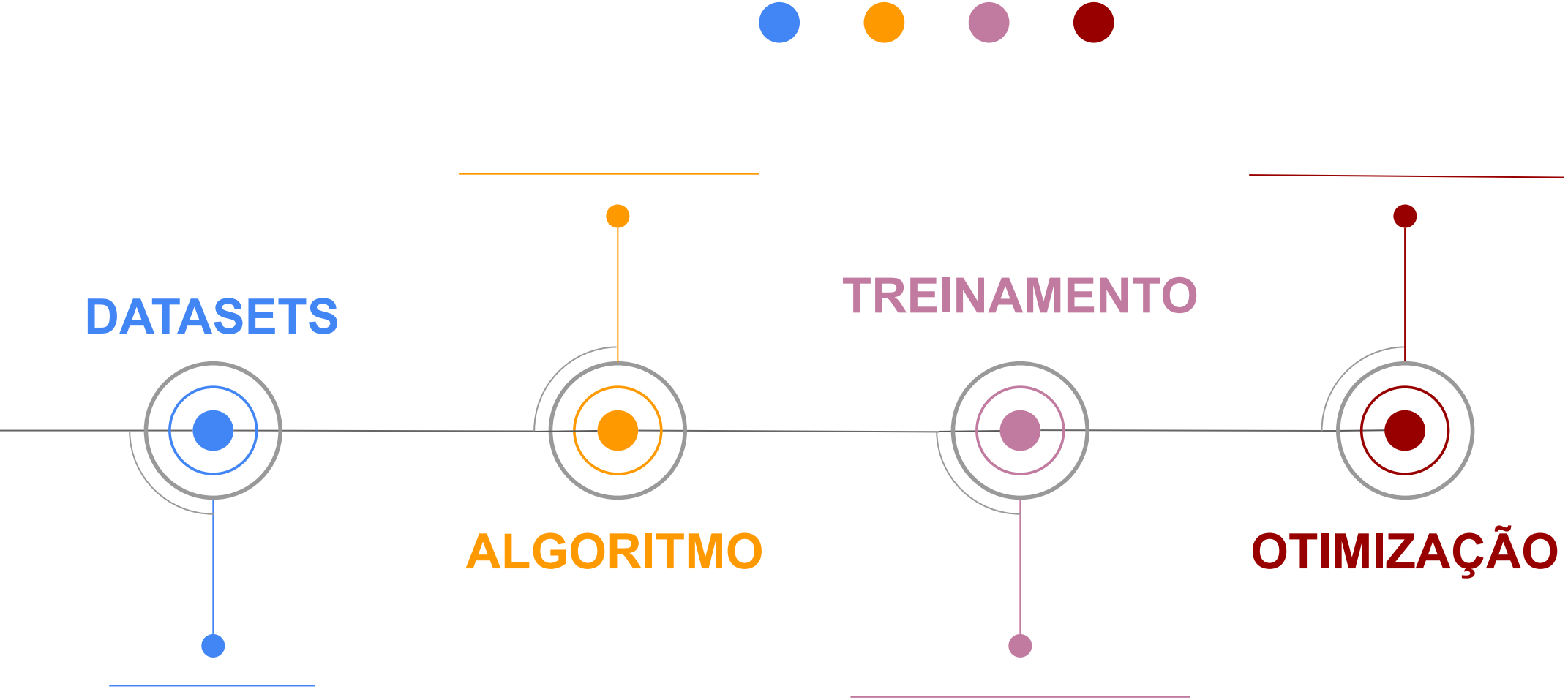
Análise do impacto de viés nos conjuntos de dados para detecção de Malwares Android

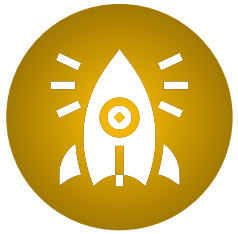
6º Workshop Regional de Segurança da Informação e de Sistemas Computacionais

WRSeg 2021

Lucas Vilanova, Renato Sayyed, Tainá Soares, Guilherme Siqueira, Gustavo Rodrigues, Eduardo Feitosa e Diego Kreutz

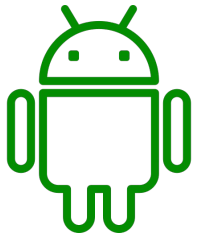
Aprendizado de Máquina





Motivação

Dados defasados



Aplicativos antigos

Motivação

Drebin-215 (2012)

READ_PHONE_STATE	SEND_SMS	WRITE_SMS	INTERNET	CHANGE_WIFI_STATE
1	1	1	0	0

DefenseDroid (2021)

READ_PHONE_STATE	SEND_SMS	WRITE_SMS	INTERNET	CHANGE_WIFI_STATE
1	1	1	1	1

New AbstractEmu malware roots Android devices, evades detection

By **Sergiu Gatlan**

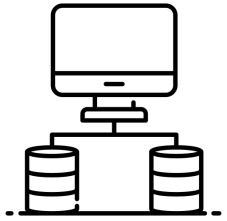
October 28, 2021 09:15 AM

Image: Jon Hunter

New Android malware can root infected devices to take complete control and silently tweak system settings, as well as evade detection using code abstraction and anti-emulation checks.

Será que um modelo configurado para um dataset defasado mantém o mesmo desempenho em predizer Malwares Android em datasets mais atuais?

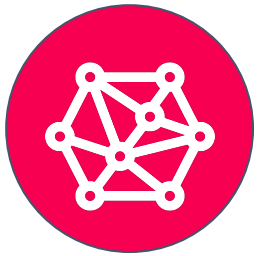
Objetivo



Desempenho de modelos preditivos em datasets de diferentes épocas

Seleção dos Conjuntos de Dados

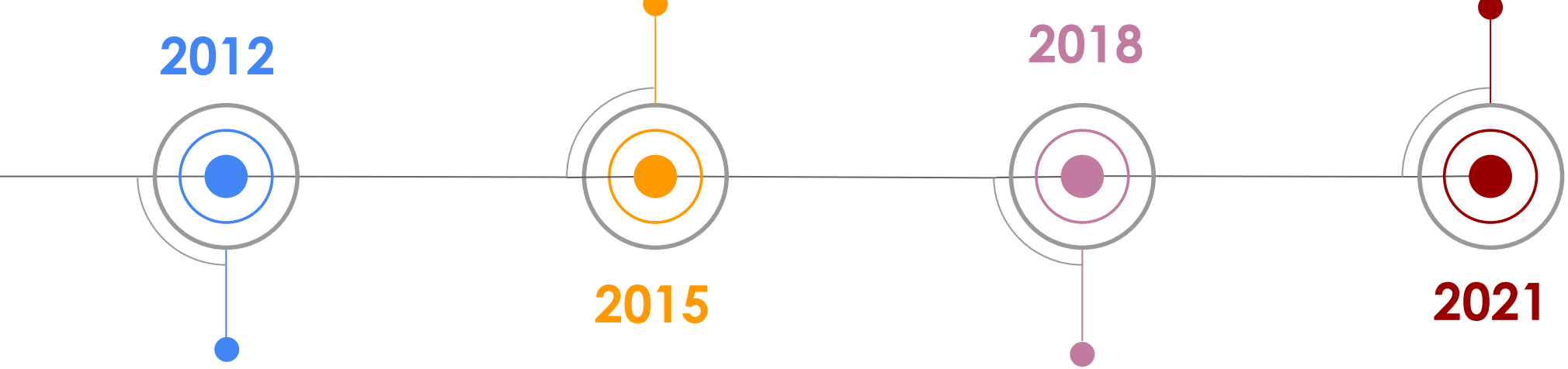
Tratamento dos Dados



Desenvolvimento do Modelo

Análise dos Resultados

Seleção dos Conjuntos de Dados



2012

Drebin-215

Exemplos: 15.036
Caracaterísticas: 215

2015

Androcrawl

Exemplos: 162.983
Características: 222

2018

AndroidMalwareNormal

Exemplos: 28.849
Características: 173

2021

DefenseDroid

Exemplos: 11.975
Características: 1.491

Tratamento dos Dados

Limpeza de dados

Seleção de Características

Limpeza de dados

Exemplos duplicados

Datasets	Quantidade de exemplos duplicados	% de exemplos duplicados
Drebin-215	7.777	51,72%
Androcrawl	869	0,53%
AndroidMalwareNormal	5.125	17,76%
DefenseDroid	4.250	35,49%

Exemplos duplicados

Datasets	Quantidade de exemplos duplicados	% de exemplos duplicados
Drebin-215	**7.777**	**51,72%**
Androcrawl	**869**	**0,53%**
AndroidMalwareNormal	**5.125**	**17,76%**
DefenseDroid	**4.250**	**35,49%**

Seleção de Características

Datasets	Quantidade de características	Quantidade de características após seleção
Drebin-215	215	113
Androcrawl	222	58
AndroidMalwareNormal	173	173
DefenseDroid	1.491	134

Seleção de Características

Datasets	Quantidade de características	Quantidade de características após seleção
Drebin-215	215	113
Androcrawl	222	58
AndroidMalwareNormal	173	173
DefenseDroid	1.491	134

Seleção de Características

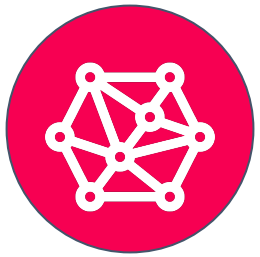
Datasets	Quantidade de características	Quantidade de características após seleção
Drebin-215	215	113
Androcrawl	222	58
AndroidMalwareNormal	173	173
DefenseDroid	1.491	134

Seleção de Características

Datasets	Quantidade de características	Quantidade de características após seleção
Drebin-215	215	113
Androcrawl	222	58
AndroidMalwareNormal	173	173
DefenseDroid	1.491	134

Seleção de Características

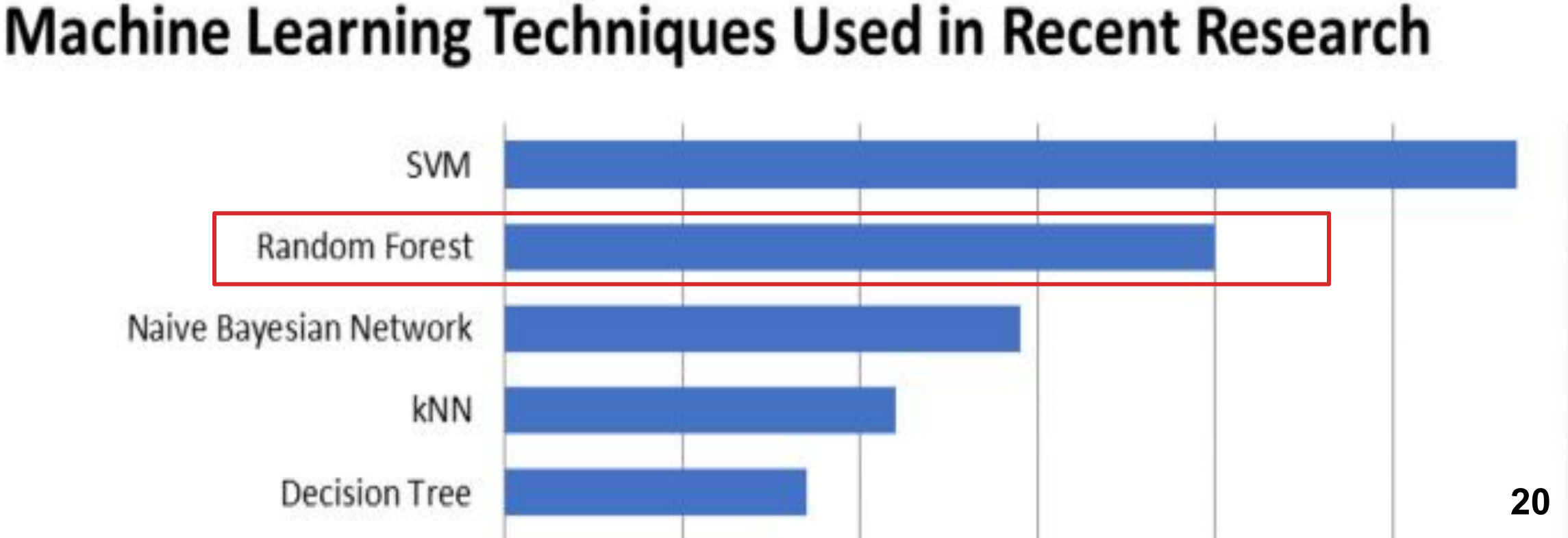
Datasets	Quantidade de características	Quantidade de características após seleção
Drebin-215	215	113
Androcrawl	222	58
AndroidMalwareNormal	173	173
DefenseDroid	1.491	134



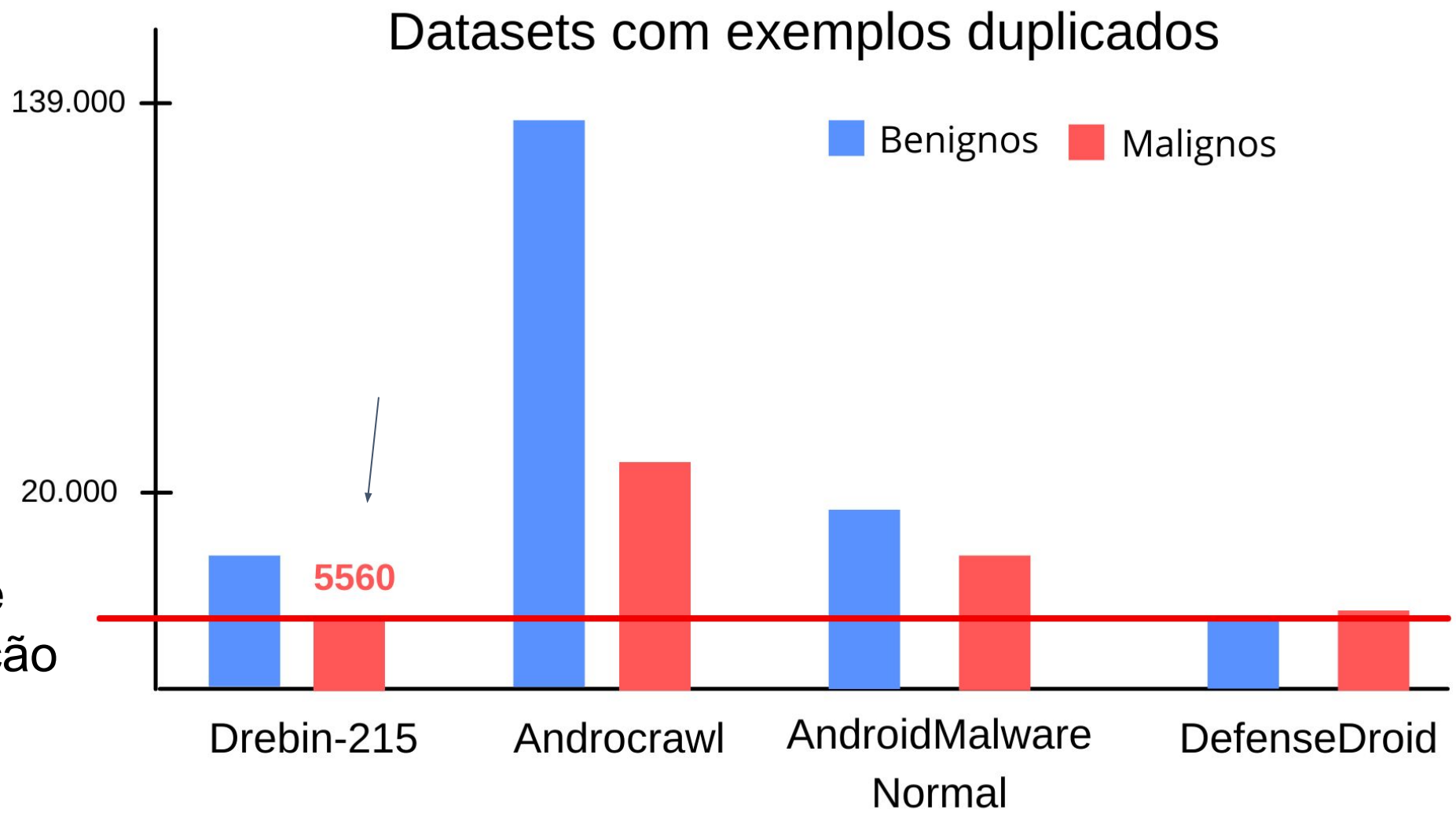
Desenvolvimento do Modelo

Algoritmo RandomForest

Link do paper:

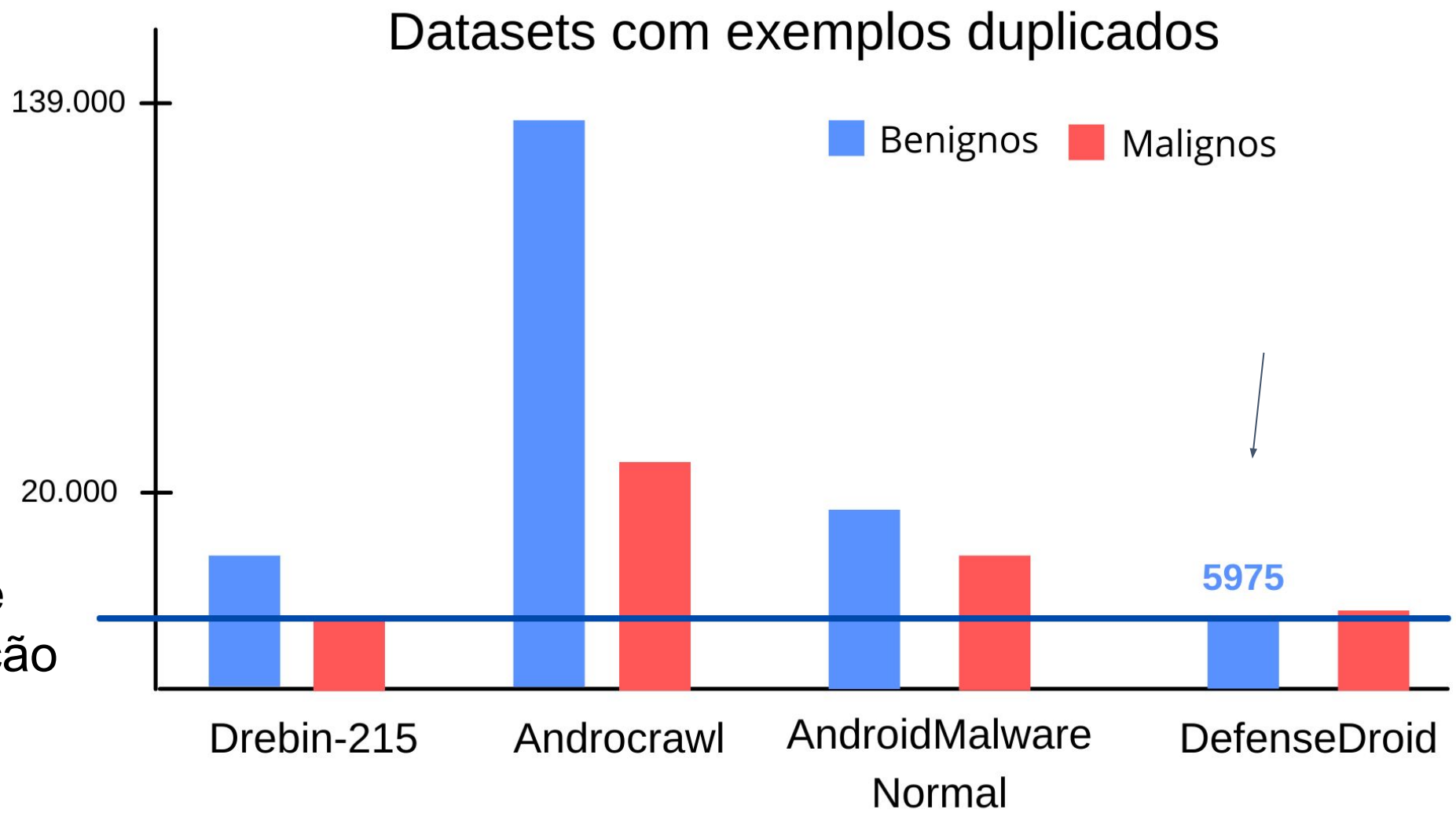


Normalização dos dados

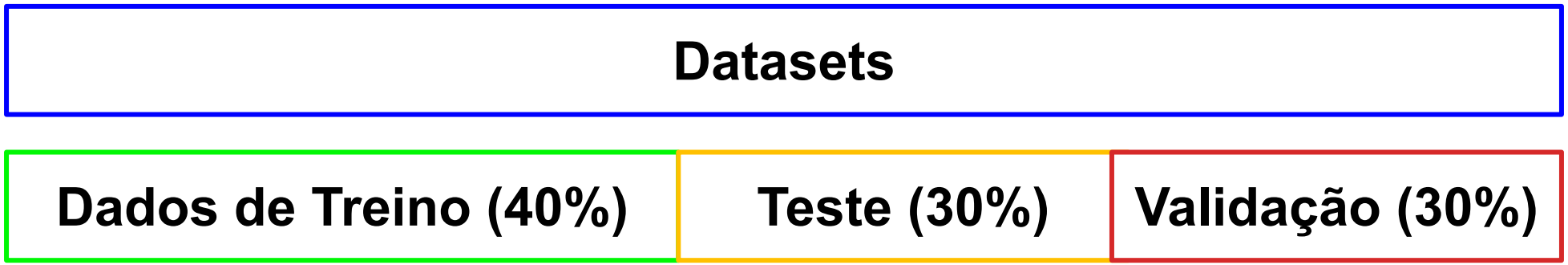


Linha de normalização

Normalização dos dados



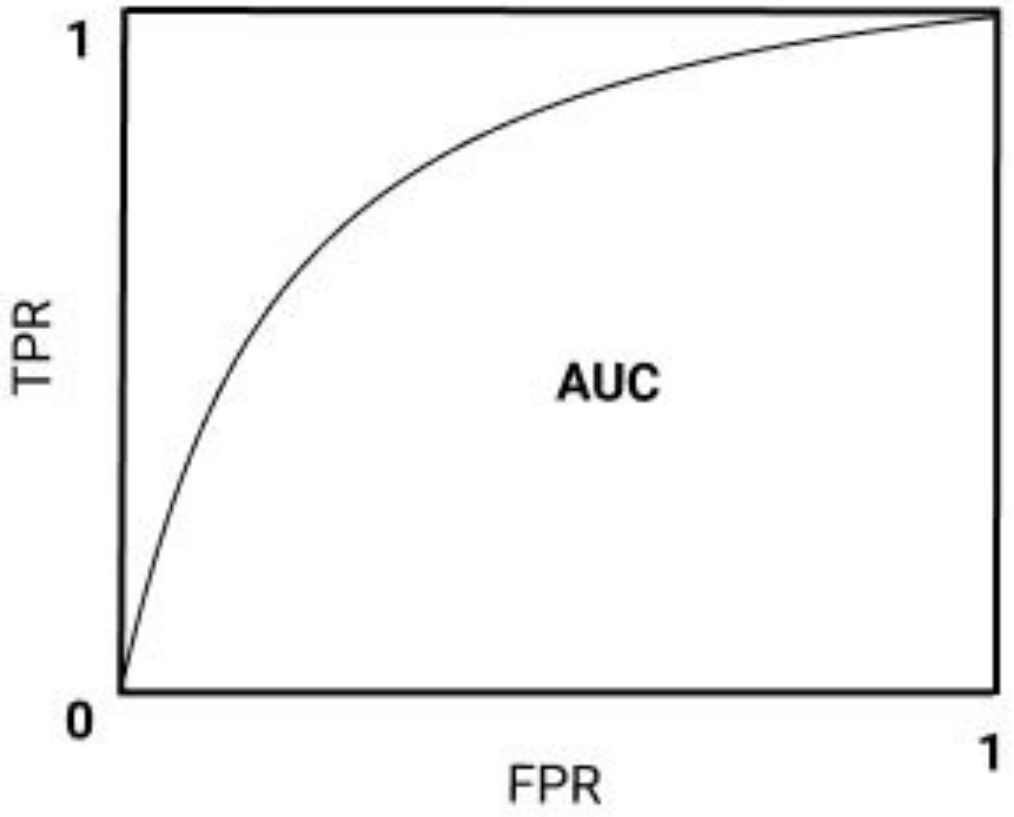
Repartições dos datasets



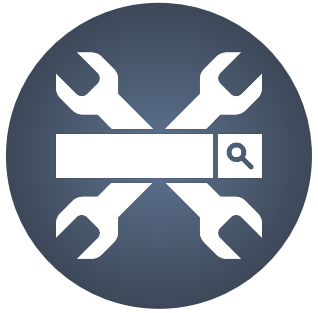
Dataset base do modelo >> **Drebin-215 (2012)**

Métrica ROC-AUC

Taxa de
Verdadeiros Positivos



Taxa de
Falsos Positivos



Otimização do Classificador RandomForest

Configurações de Hiperparâmetros

Hiperparâmetro	Padrão	Intervalo de busca	Busca aleatória
n_estimators	100	200 : 2000	1800
min_samples_split	2	2, 5 e 10	2
min_samples_leaf	1	1, 2 e 4	1
max_features	auto	auto e sqrt	auto
max_depth	None	10 : 110 e None	20
bootstrap	True	True e False	False

Configurações de Hiperparâmetros

Hiperparâmetro	Padrão	Intervalo de busca	Busca aleatória
n_estimators	100	200 : 2000	1800
min_samples_split	2	2, 5 e 10	2
min_samples_leaf	1	1, 2 e 4	1
max_features	auto	auto e sqrt	auto
max_depth	None	10 : 110 e None	20
bootstrap	True	True e False	False

Configurações de Hiperparâmetros

Hiperparâmetro	Padrão	Intervalo de busca	Busca aleatória
n_estimators	100	200 : 2000	1800
min_samples_split	2	2, 5 e 10	2
min_samples_leaf	1	1, 2 e 4	1
max_features	auto	auto e sqrt	auto
max_depth	None	10 : 110 e None	20
bootstrap	True	True e False	False

Configurações de Hiperparâmetros

Hiperparâmetro	Padrão	Intervalo de busca	Busca aleatória
n_estimators	100	200 : 2000	1800
min_samples_split	2	2, 5 e 10	2
min_samples_leaf	1	1, 2 e 4	1
max_features	auto	auto e sqrt	auto
max_depth	None	10 : 110 e None	20
bootstrap	True	True e False	False

Resultados

Resultados

Dataset	ROC-AUC	
	Com duplicata	Sem duplicata
Drebin-215 (2012)	94,37%	89,58%
Androcrawl (2014/2015)	55,80%	54,39%
AndroidMalwareNormal (2018)	55,80%	50,47%
DefenseDroid (2021)	84,61%	84,85%

Resultados

Dataset	ROC-AUC	
	Com duplicata	Sem duplicata
Drebin-215 (2012)	94,37%	89,58%
Androcrawl (2014/2015)	55,80%	54,39%
AndroidMalwareNormal (2018)	55,80%	50,47%
DefenseDroid (2021)	84,61%	84,85%

Resultados

Dataset	ROC-AUC	
	Com duplicata	Sem duplicata
Drebin-215 (2012)	94,37%	89,58%
Androcrawl (2014/2015)	55,80%	54,39%
AndroidMalwareNormal (2018)	55,80%	50,47%
DefenseDroid (2021)	84,61%	84,85%

Resultados

Dataset	ROC-AUC	
	Com duplicata	Sem duplicata
Drebin-215 (2012)	94,37%	89,58%
Androcrawl (2014/2015)	55,80%	54,39%
AndroidMalwareNormal (2018)	55,80%	50,47%
DefenseDroid (2021)	84,61%	84,85%

Resultados

Dataset	ROC-AUC	
	Com duplicata	Sem duplicata
Drebin-215 (2012)	94,37%	89,58%
Androcrawl (2014/2015)	55,80%	54,39%
AndroidMalwareNormal (2018)	55,80%	50,47%
DefenseDroid (2021)	84,61%	84,85%

Resultados

Dataset	ROC-AUC	
	Com duplicata	Sem duplicata
Drebin-215 (2012)	94,37%	89,58%
Androcrawl (2014/2015)	55,80%	54,39%
AndroidMalwareNormal (2018)	55,80%	50,47%
DefenseDroid (2021)	84,61%	84,85%

Trabalhos futuros

01 Criação de um dataset com dados atuais

02 Investigar extensivamente o enviesamento causado pelos dados de datasets existentes

03 Investigar questões relacionadas a rastreabilidade das amostras dos datasets

Obrigado!

Repositório GitHub

lucasvilanova.aluno@unipampa.edu.br

UFAM

unipampa

Universidade Federal do Pampa

Slides de Backup

Valores inválidos

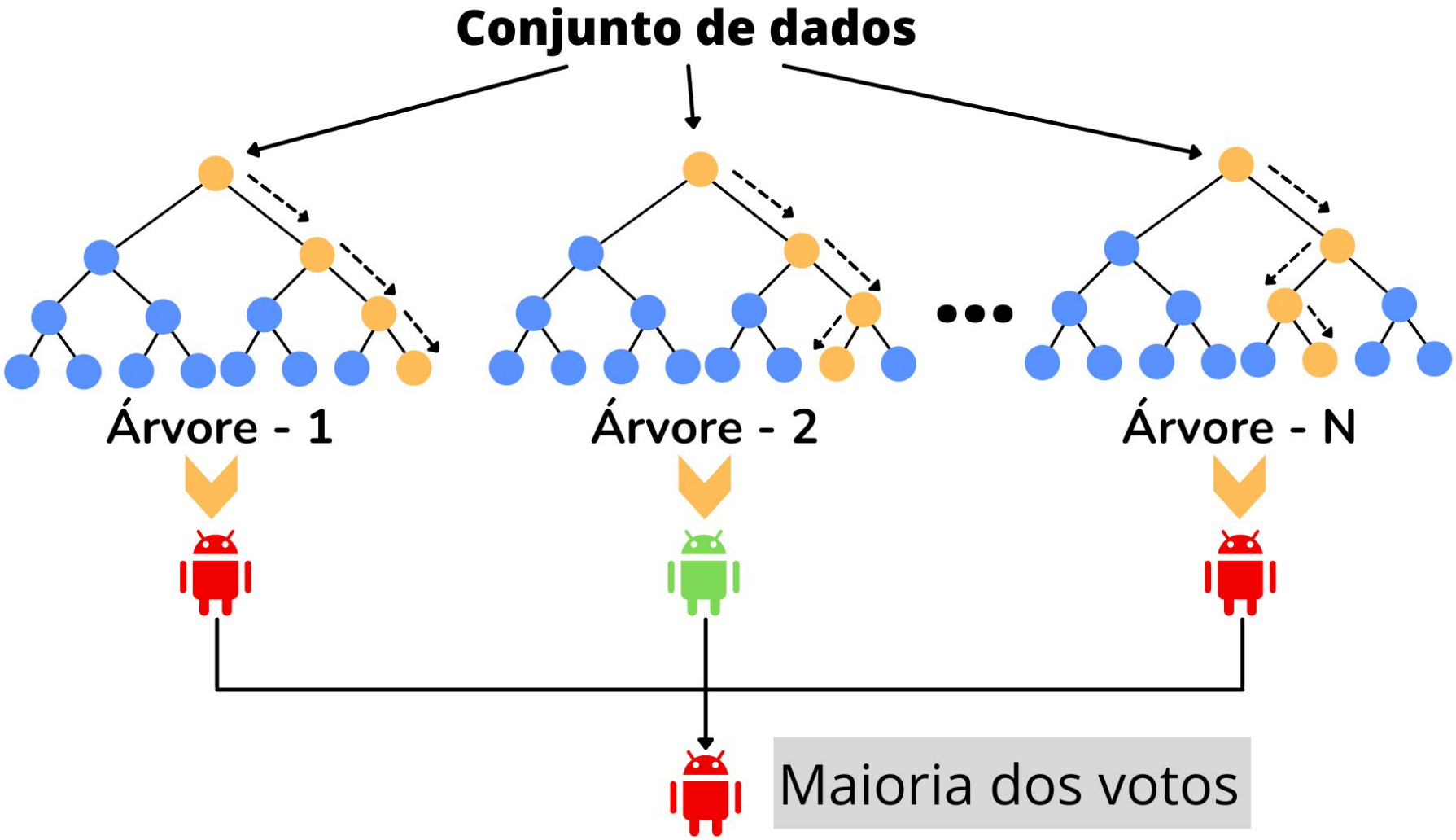
Dataset	Permissões	Exemplos com valores “?”
Androcrawl	**RECEIVE, MESSAGE, SEND**	**10.262**

Transformação de Variáveis

Objeto >> **INT**, **Boolean**

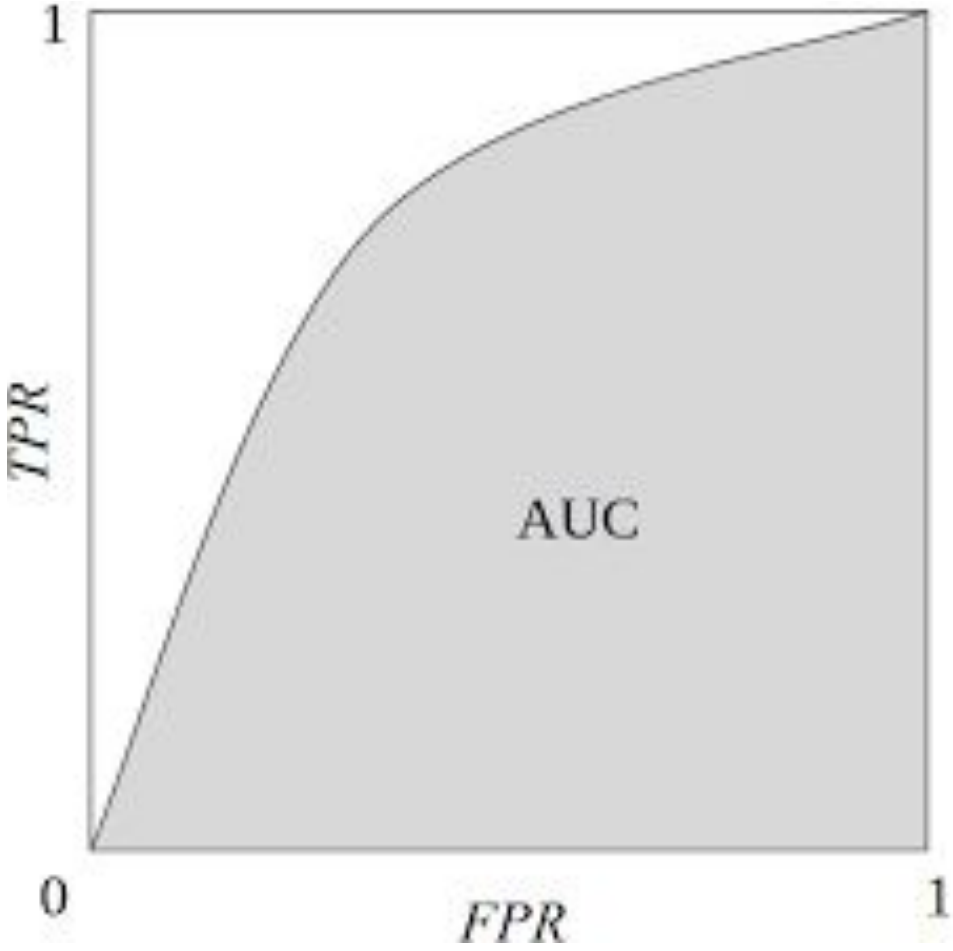
Boolean >> **INT (0 e 1)**

Algoritmo RandomForest



Métrica ROC-AUC

TPR = Sensibilidade



FPR = 1 - Especificidade

Link da lista das permissões manifestas do Android

https://developer.android.com/reference/android/Manifest.permission