APRESENTAÇÃO FINAL DATA SCIENCE SEGURANÇA

ARTUR RICARDO BIZON

INTRODUÇÃO

- O problema desta tarefa abordado é a classificação de malwares para android
- Foi utilizado o dataset CicMalDroid2020, para o treinamento e a classificação dos dados
- Para validar a aplicação com outros dados foi utilizado parte dos dados do dataset Androzoo, dos anos 2014, 2015, 2016, 2017 e2018

CICMALDROID2020

- Este dataset possui 11598 malwares mapeados em arquivos csv, estes arquivos possuem informações estáticas e dinâmicas sobre cada malware
- Para gerar os arquivos CSVs os autores utilizaram a aplicação de sandbox CopperDroid, esta aplicação é capaz de extrair tanto informações estáticas quando dinâmicas de arquivos apk

CLASSES PRESENTES NO DATASET

- Adware

Malware que consiste em exibir conteúdos de propaganda, mesmo quando o usuário tenta encerrar o aplicativo. Este tipo de malware além de exibir campanhas publicitárias indesejadas ainda pode roubar informaões senssíveis dos usuários

- Banking Malware

Malware especializado em coletar informações sensíveis de aplicações bancárias

CLASSES PRESENTES NO DATASET

- SMS Malware

Este malware faz uso da exploração do serviço de SMS para realizar ataques, enviar SMSs maliciosos, entre outros

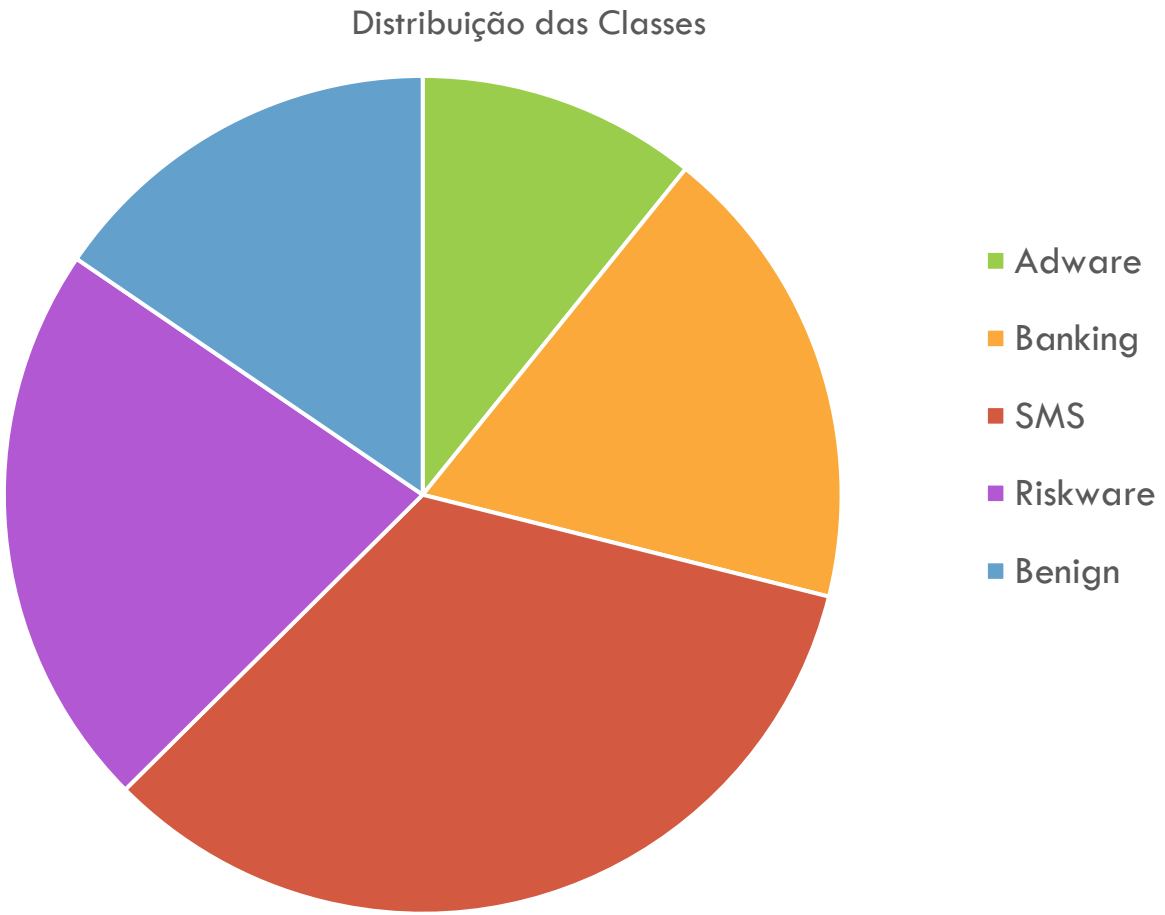
- Riskware

Aplicações legítimas que podem oferecer riscos aos usuários, caso um agente malicioso explore estas aplicações, este tipo de aplicação pode ser utilizado como porta de entrade de malwars, podendo ser, Adware, Ransoware, entre outros.

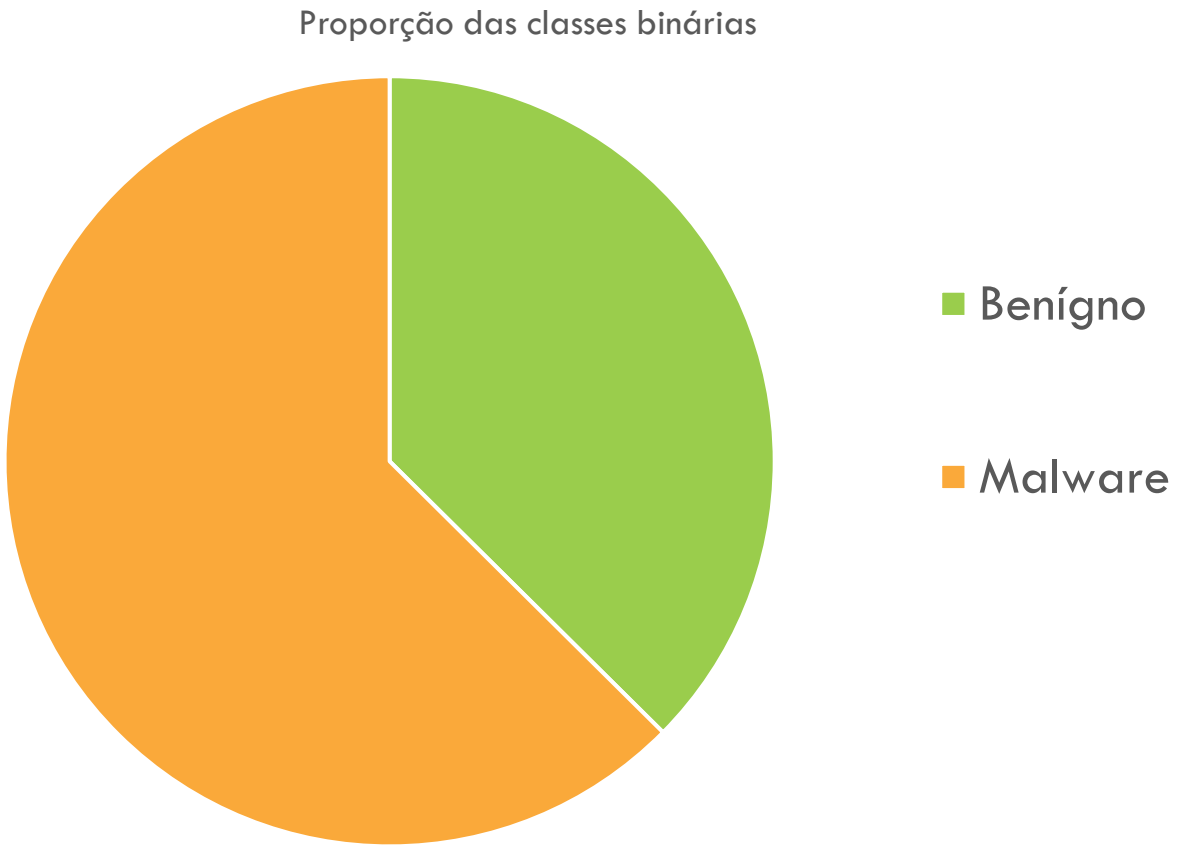
- Benign

Aplicações legítimas que não possuem intenções maliciosas

PROPORÇÃO DO DATASET



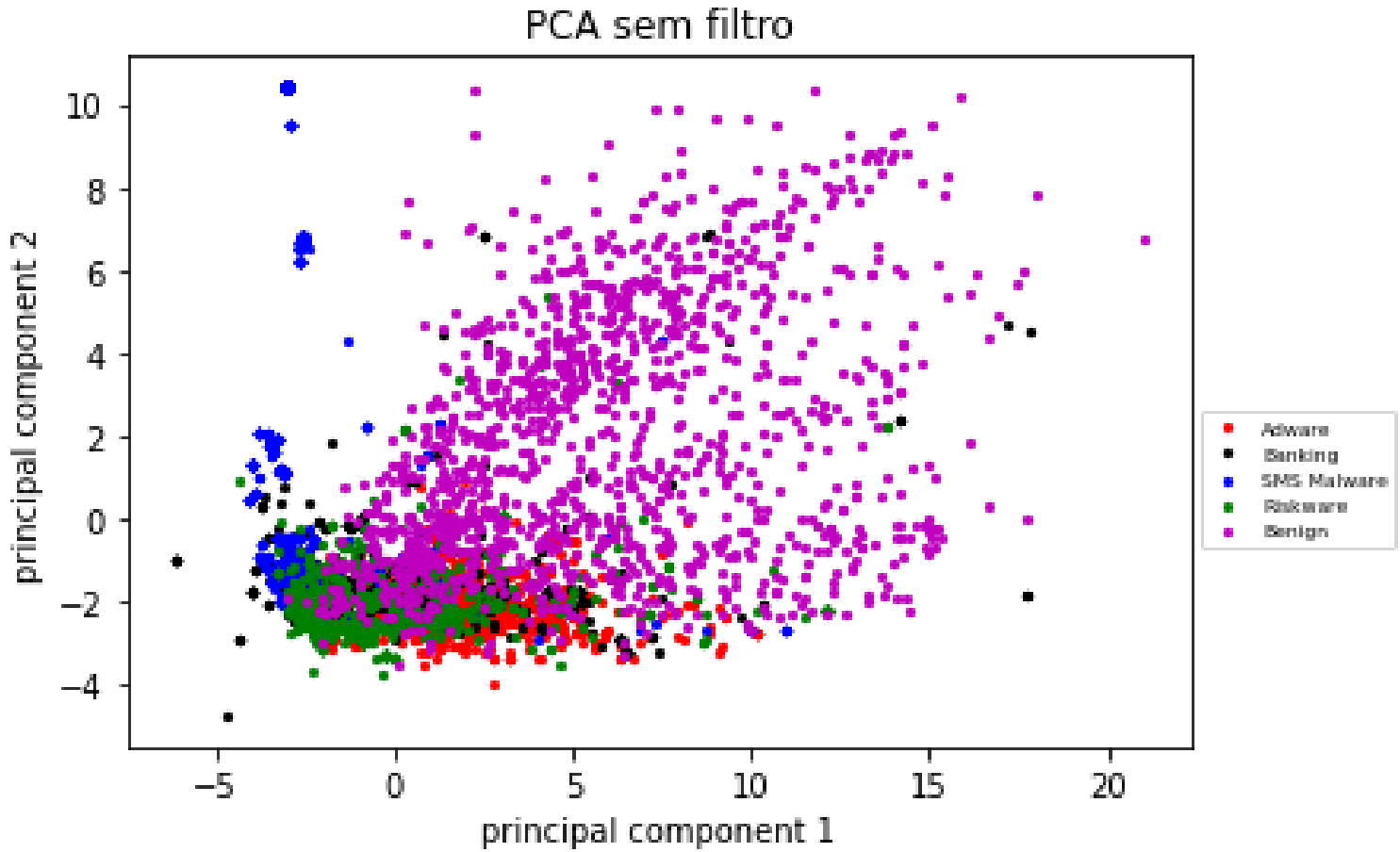
DISTRIBUIÇÃO DAS CLASSES



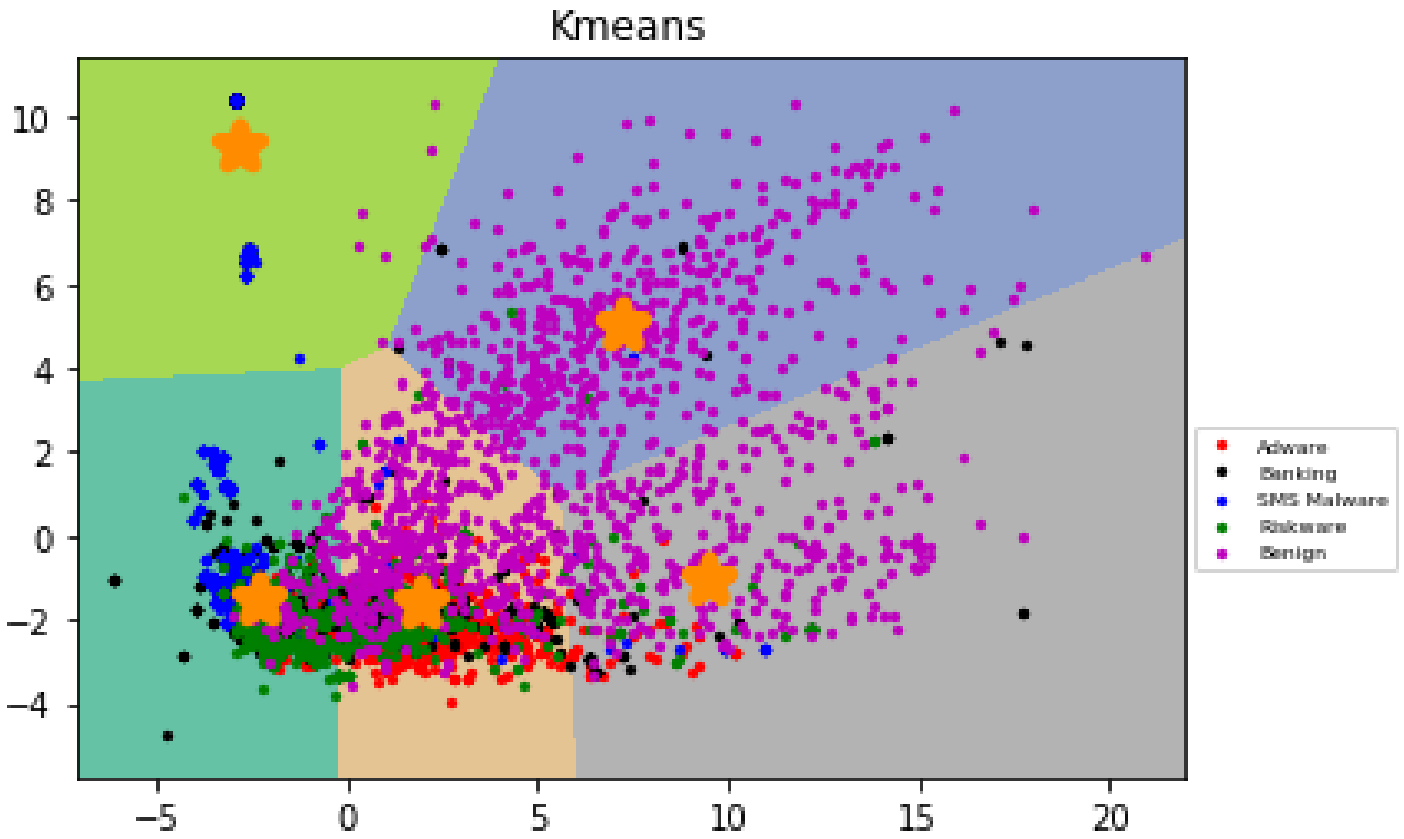
SELEÇÃO DE FEATURES

- Neste trabalho foram utilizadas apenas informações estáticas sobre os aplicativos
- Inicialmente o dataset possuía 50621 colunas no arquivo de informações estáticas, após uma filtragem por variância, restaram 101 colunas, sendo estas colunas de informações binárias (0 ou 1) ou informações numéricas

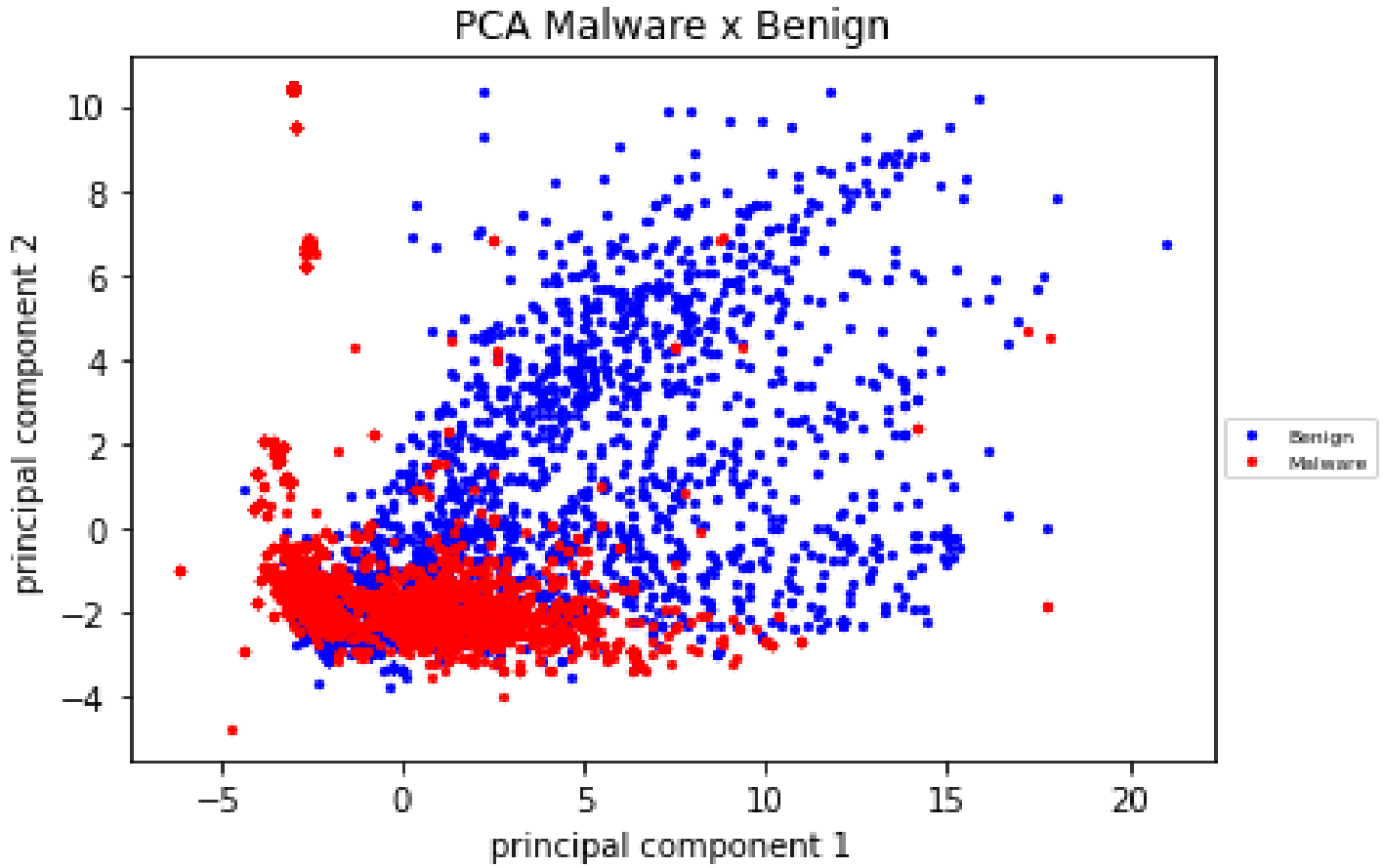
EXPLORAÇÃO DOS DADOS



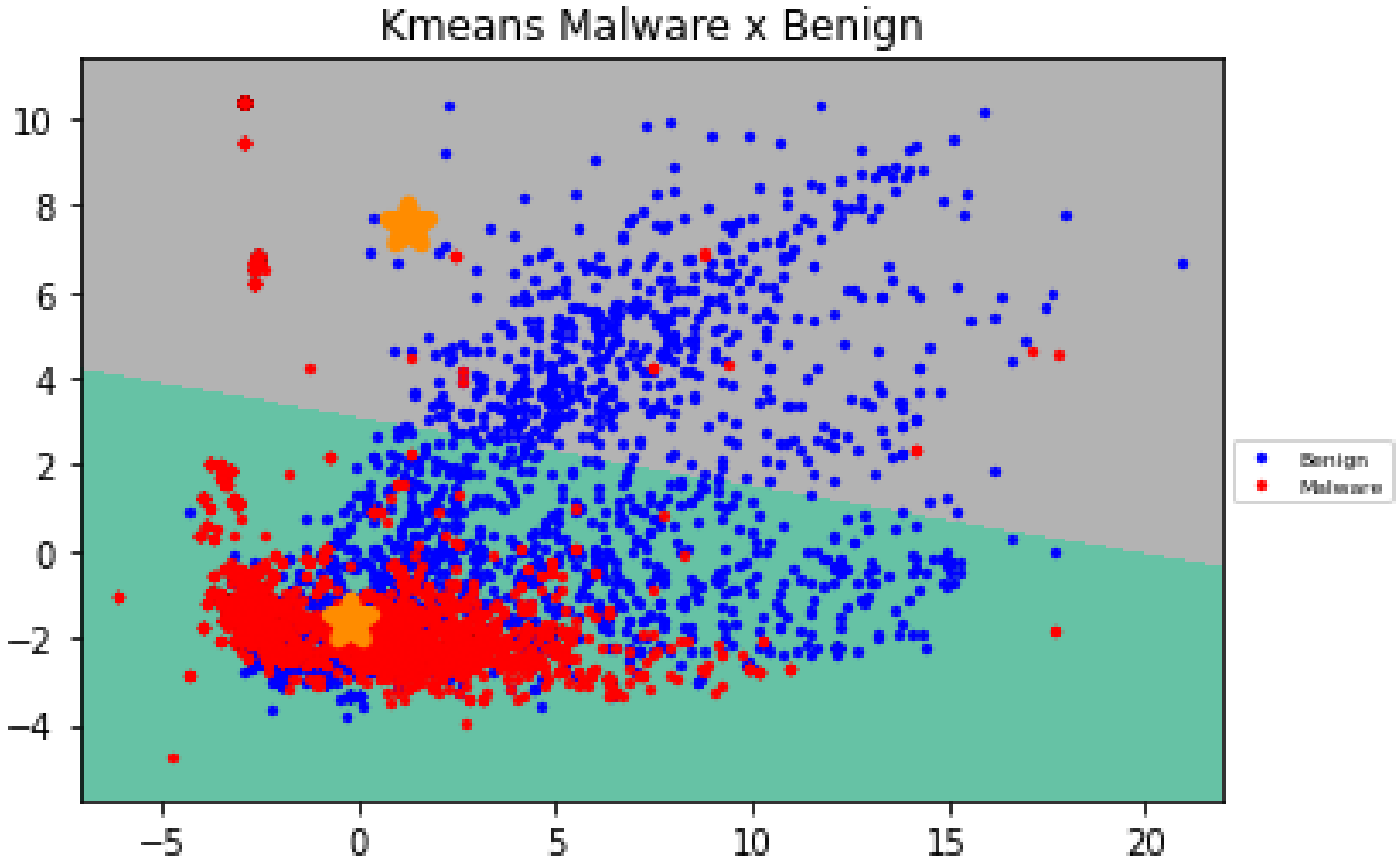
EXPLORAÇÃO DOS DADOS



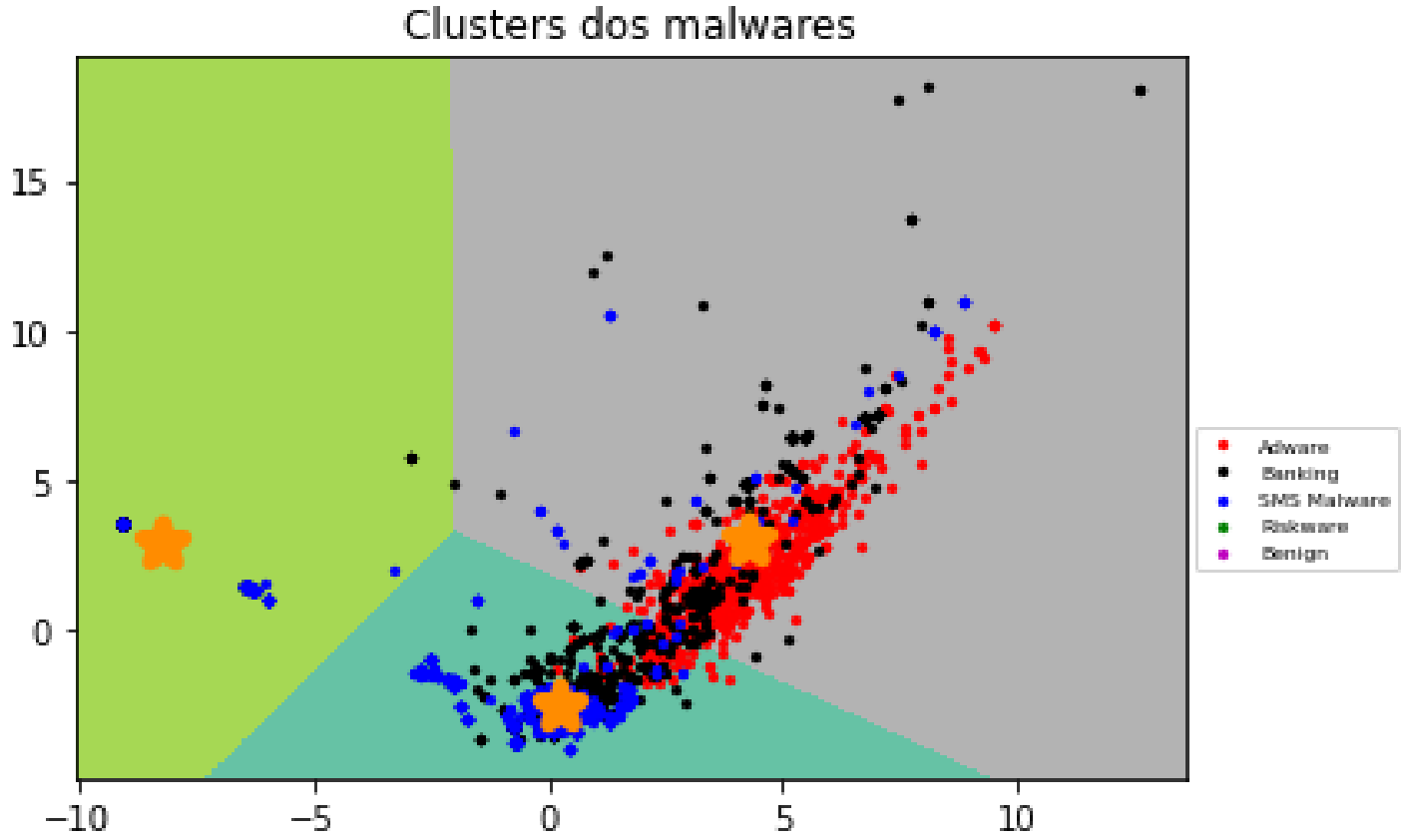
EXPLORAÇÃO DOS DADOS



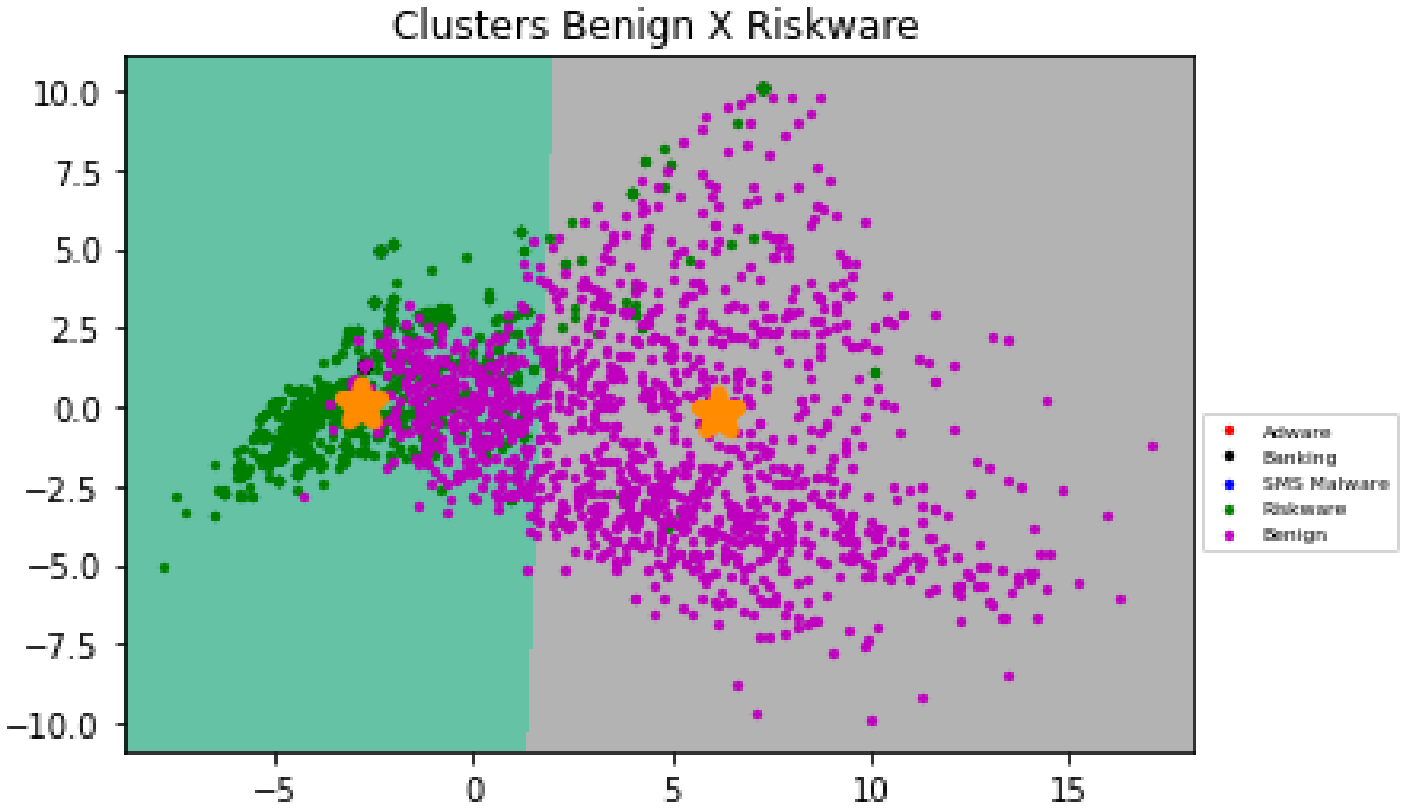
EXPLORAÇÃO DOS DADOS



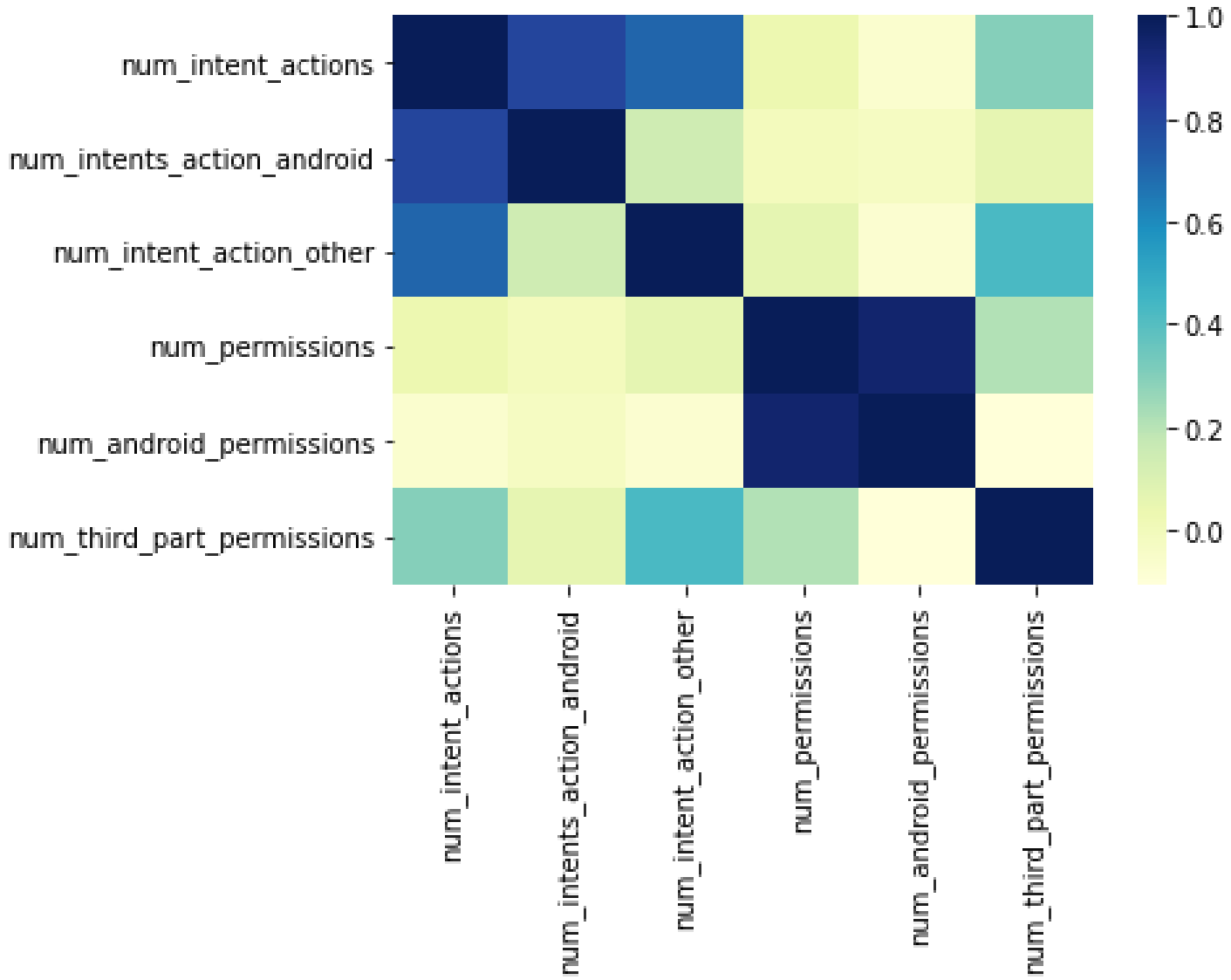
EXPLORAÇÃO DOS DADOS



EXPLORAÇÃO DOS DADOS



EXPLORAÇÃO DOS DADOS



GERAÇÃO DOS MODELOS MACHINE LEARNING

- Link colab: https://colab.research.google.com/drive/12X0erNQnuk4_KXu4Kx6qsGpawpDeXkOW?usp=sharing

VALIDAÇÃO DOS MODELOS

Classificador	Acurácia	Tempo Treino
RF 80X20	95	697ms
RF 50X50	95	473ms
KNN 80X20	95	5,45ms
KNN 50X50	94	5,28ms
MLP 80X20	96	1,5min
MLP 50X50	96	46seg

VALIDAÇÃO DOS MODELOS

- Resultados apresentados pelo artigo

Classificador	Acurácia	Tempo Treino
RF	98,32	não informado
PLSAE	98,52	23h

VALIDAÇÃO DOS MODELOS

- Testes com dados do dataset Androzoo

Classificador	Acurácia	Ano Amostras
RF 80X20	33	2014
RF 50X50	26	2014
KNN 80X20	66	2014
KNN 50X50	69	2014
MLP 80X20	23	2014
MLP 50X50	18	2014

VALIDAÇÃO DOS MODELOS

- Testes com dados do dataset Androzoo

Classificador	Acurácia	Ano Amostras
RF 80X20	31	2015
RF 50X50	28	2015
KNN 80X20	69	2015
KNN 50X50	72	2015
MLP 80X20	24	2015
MLP 50X50	16	2015

VALIDAÇÃO DOS MODELOS

- Testes com dados do dataset Androzoo

Classificador	Acurácia	Ano Amostras
RF 80X20	31	2016
RF 50X50	17	2016
KNN 80X20	65	2016
KNN 50X50	69	2016
MLP 80X20	20	2016
MLP 50X50	14	2016

VALIDAÇÃO DOS MODELOS

- Testes com dados do dataset Androzoo

Classificador	Acurácia	Ano Amostras
RF 80X20	25	2017
RF 50X50	18	2017
KNN 80X20	58	2017
KNN 50X50	64	2017
MLP 80X20	13	2017
MLP 50X50	8	2017

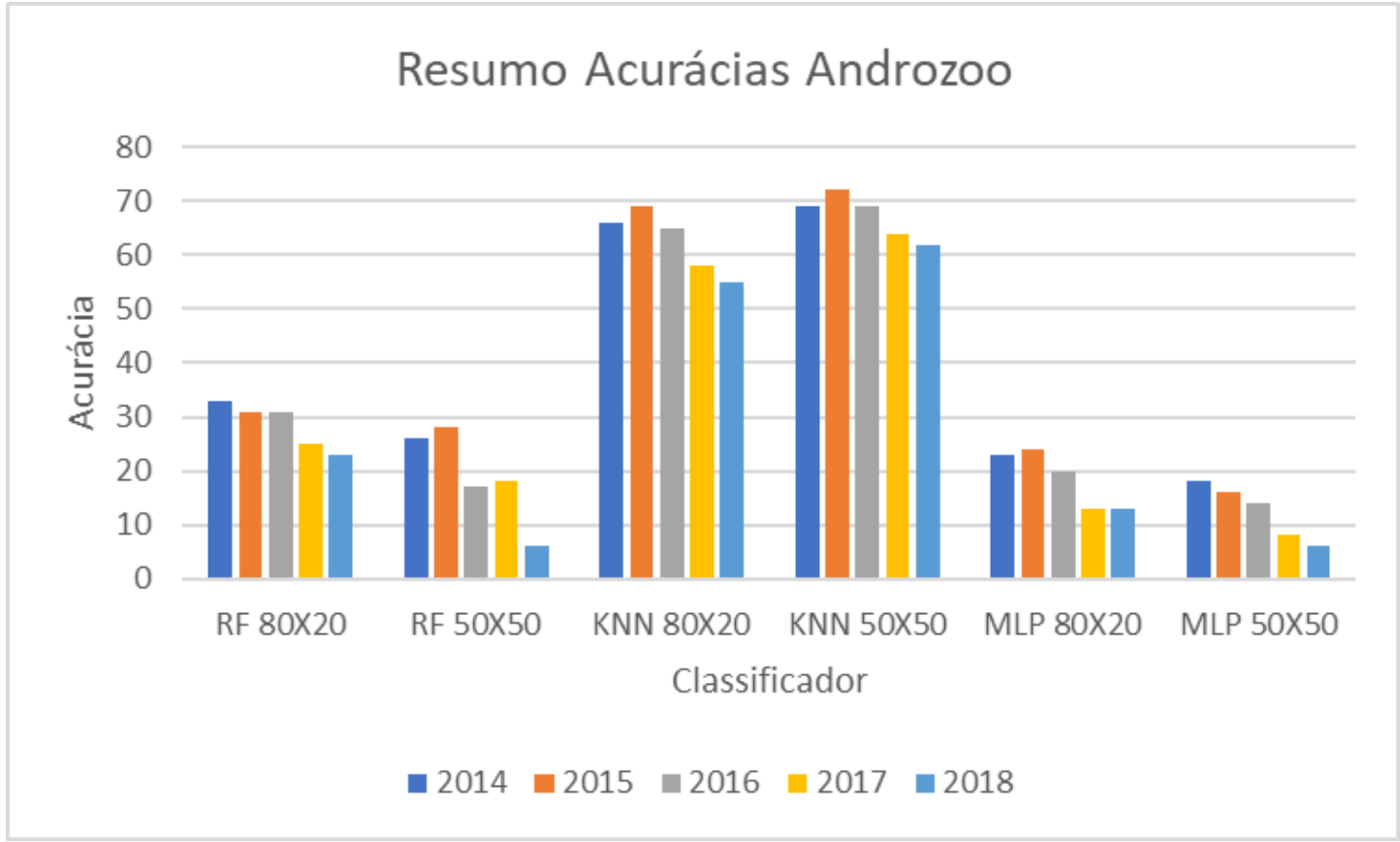
VALIDAÇÃO DOS MODELOS

- Testes com dados do dataset Androzoo

Classificador	Acurácia	Ano Amostras
RF 80X20	23	2018
RF 50X50	6	2018
KNN 80X20	55	2018
KNN 50X50	62	2018
MLP 80X20	13	2018
MLP 50X50	6	2018

VALIDAÇÃO DOS MODELOS

- Testes com dados do dataset Androzoo



VALIDAÇÃO DOS MODELOS

- https://colab.research.google.com/drive/1faRSTyliWD3bM310WjNaPR_iSHLx88G6?usp=sharing

APLICAÇÃO PRÁTICA

- https://colab.research.google.com/drive/1tf3_x9aXIJ-PiGVyrSxQyXI_uvQtsJx6?usp=sharing