UNIVERSIDADE FEDERAL
DO ESPÍRITO SANTO

Projeto Final

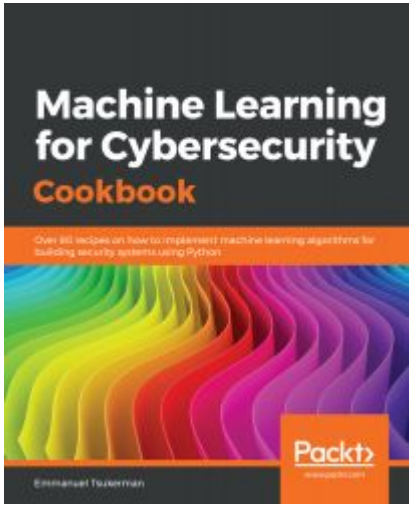
Igor Varejão

O tráfego de dados

- Com a internet das coisas cada vez mais presente no nosso dia a dia, mais dispositivos são introduzidos no ambiente onde vivemos e por consequência mais dados são transmitidos constantemente.
- Estudar como esses dados se comportam é de extrema importância para entender melhor o cenário atual

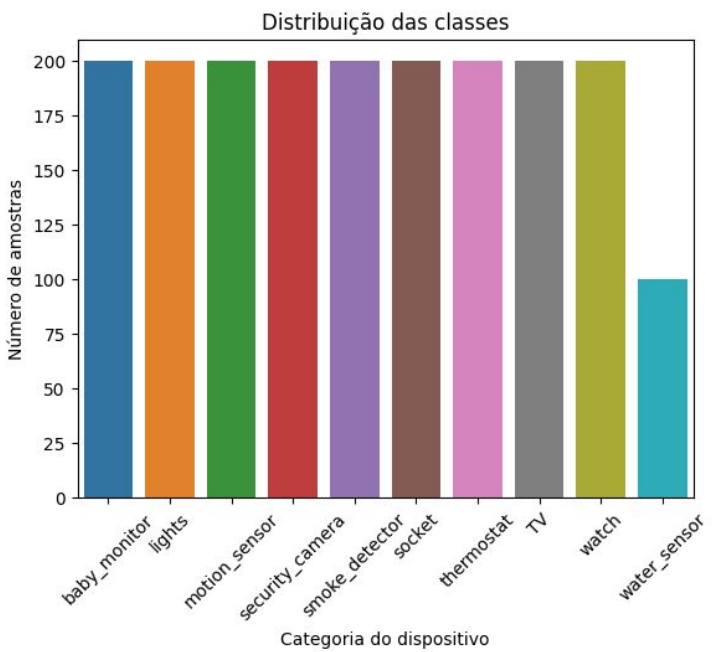
O dataset

- O dataset faz parte de um material do livro Machine Learning for Cybersecurity Cookbook
- É composto por amostras de pacotes coletadas de dispositivos de Internet das coisas.
- Cada amostra consiste em um período de monitoramento dos pacotes enviado por um determinado dispositivo, tal qual possui cerca de 297 informações como por exemplo:
  - *Número de pacotes mandado pelo cliente*
  - *Data de início / final da captura*
  - *…*



Características do dataset

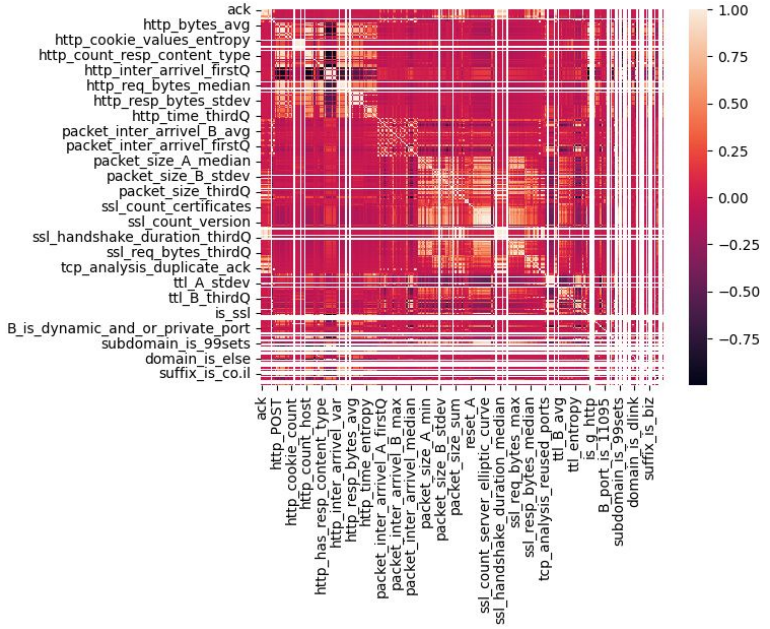
- No total, possui 1900 amostras, com 10 classes
- Faz a divisão das amostras em dois arquivos, treino e teste
- O arquivo de teste possui uma classe a mais do que o de treino, **water_sensor**. Logo havia duas possibilidades:
  a. *Eliminar as amostras dessa classe do dataset;*
  b. *Distribuir as amostras dessa classe para ambas os conjuntos;*



Pré Processamento

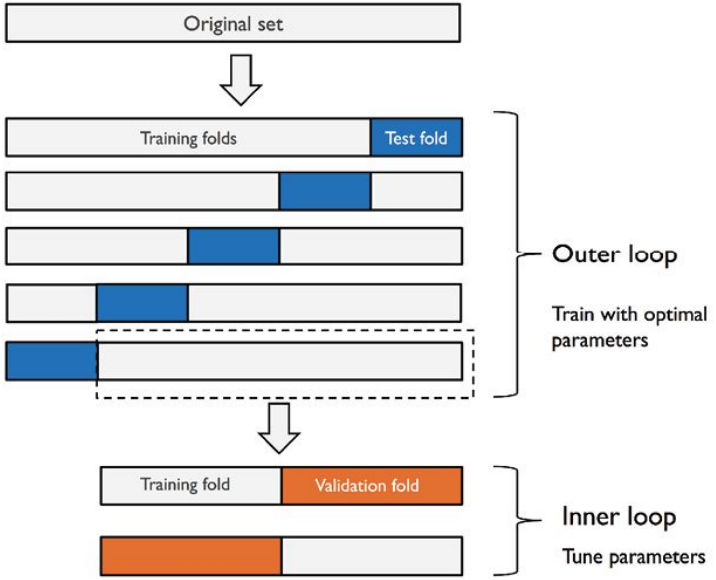
1. Remover os dados faltantes (Não aconteceu)
2. Remover características fortemente correlacionadas
   a. *Identificar os pares de características correlacionadas*
   b. *Remove a características do par com menor correlação com a classe*

O resultado deste pré-processamento foi a remoção de 166 características.



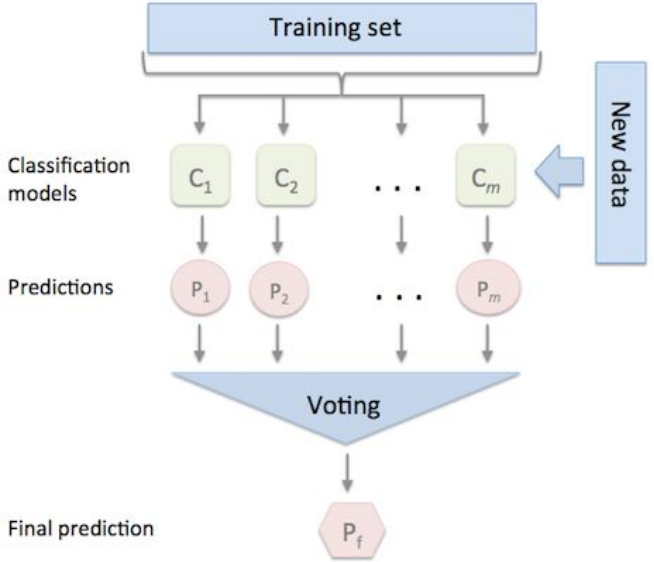
Treinamento

- Foram escolhidos 5 classificadores arbitrariamente:
  - *XGBClassifier*
  - *RandomForestClassifier*
  - *AdaBoostClassifier*
  - *GradientBoostingClassifier*
  - *BaggingClassifier*
- Validação cruzada aninhada
  - *Ciclo interno é feito uma busca em grade para encontrar os melhores parâmetros em cada classificador*



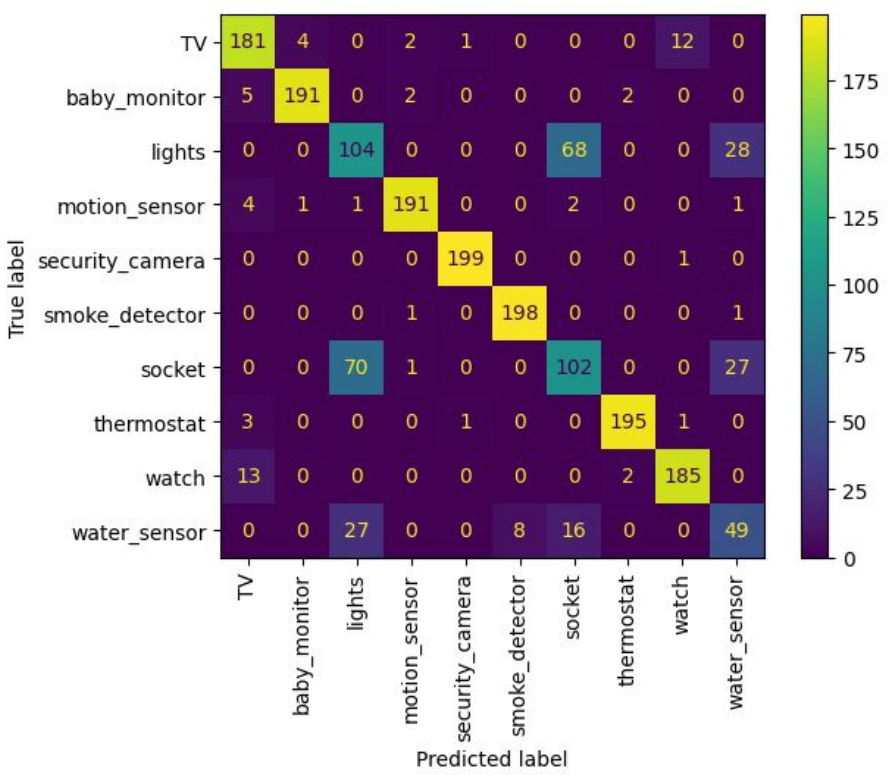
Teste

- A partir da escolha dos melhores parâmetros para cada cada classificador foi criado um classificador baseado em votos dos 5 classificadores.
- Este foi treinado e testado por uma validação cruzada simples com 5 folds e avaliado utilizando a métrica de **f1-score** e **acurácia**.



Performance

- A média das 5 execuções foram:
  - *f1-score: 0.825*
  - *acurácia: 0.844*



Mostrar os problemas encontrados

- Falta de datasets de tráfego de dispositivos de IoT, a maioria era focado para segurança, mais especificamente, identificar ataques DdoS (distributed denial-of-service).
- Os que existem não possuem um grande volume de dados, apesar de todas as tentativas o modelo chegou num teto de desempenho que poderia ser melhorado pelo maior número de dados.