Universidade Federal do Rio de Janeiro
Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia

Programa de Engenharia de Sistemas e Computação

COS868 - Probabilidade e Estatística para Aprendizado de Máquina

Profa. Dra. Rosa M. Leão (PESC/COPPE/UFRJ)

Projeto do Curso

Luiz Henrique Souza Caldas
email: lhscaldas@cos.ufrj.br

23 de dezembro de 2024

Conteúdo

1 Introdução

1.1 Objetivo

O objetivo deste trabalho é realizar uma análise de um conjunto de dados reais fornecidos por um provedor de Internet de médio porte, avaliando as taxas de upload e download de dispositivos domésticos, especificamente Smart-TVs e Chromecasts, com base na teoria aprendida em classe, destacando a importância de uma análise crítica dos resultados obtidos.

1.2 Análise Exploratória dos Dados

A análise exploratória foi realizada para compreender as características principais dos dados obtidos dos dispositivos Smart TV e Chromecast. Os resultados estão detalhados abaixo:

- **Primeiras linhas dos dados:**
  - **Smart TV:**

	device_id	date_hour	bytes_up	bytes_down
0	77209603	2021-11-22 15:23:00	132932.983607	2.818140e+06
1	77209603	2021-11-22 15:24:00	115770.491803	2.264410e+06
2	77209603	2021-11-22 15:25:00	114030.032787	2.309270e+06
3	77209603	2021-11-22 15:26:00	97170.622951	2.006544e+06
4	77209603	2021-11-22 15:27:00	39569.573770	8.061440e+05

  - **Chromecast:**

	device_id	date_hour	bytes_up	bytes_down
0	66161985	2021-09-06 00:01:00	2987.016393	49185.704918
1	66161985	2021-09-06 00:02:00	685.935484	328.258065
2	66161985	2021-09-06 00:03:00	4493.901639	37914.064516
3	66161985	2021-09-06 00:04:00	776.133333	229.200000
4	66161985	2021-09-06 00:05:00	3081.311475	51656.800000

- **Dimensões dos dados:**
  - Smart TV: (4417903, 4)
  - Chromecast: (1620529, 4)
- **Dados faltantes:**
  - Smart TV: Nenhum valor faltante em `device_id`, `date_hour`, `bytes_up`, `bytes_down`.
  - Chromecast: Nenhum valor faltante em `device_id`, `date_hour`, `bytes_up`, `bytes_down`.
- **Valores zero:**
  - Smart TV: `bytes_up` = 1.803.853, `bytes_down` = 1.978.337.

  - Chromecast: `bytes_up` = 6.057, `bytes_down` = 4.099.

- **Valores negativos:**
  - Smart TV: Nenhum valor negativo em `bytes_up` ou `bytes_down`.
  - Chromecast: Nenhum valor negativo em `bytes_up` ou `bytes_down`.

1.3 Pré-processamento

O pré-processamento foi realizado para preparar os dados dos dispositivos Smart TV e Chromecast para análises posteriores. As etapas realizadas são descritas a seguir:

- **Carregamento dos dados:** Os dados foram lidos a partir dos arquivos `dataset_smart-tv.csv` e `dataset_chromecast.csv`.
- **Correção de valores zero:** Como as colunas `bytes_up` e `bytes_down` apresentavam valores zero, foi aplicado um *shift* de +1 a todos os valores dessas colunas para evitar problemas no cálculo do logaritmo.
- **Reescalonamento dos dados:** Os valores das colunas `bytes_up` e `bytes_down` foram transformados para a escala logarítmica na base 10 (`log10`), devido à grande variação na ordem de grandeza desses valores.
- **Ordenação temporal:** Os dados foram ordenados pela coluna `date_hour` para garantir a consistência temporal nas análises subsequentes.
- **Salvamento dos dados processados:** Os datasets resultantes podem ser salvos como arquivos CSV (`smart_preprocessado.csv` e `chrome_preprocessado.csv`) para uso posterior.

Essa etapa garante que os dados estejam limpos, reescalonados e organizados, facilitando análises estatísticas e a geração de gráficos. Além disso, a transformação logarítmica reduz a influência de valores extremos, melhorando a interpretação dos resultados.

2 Estatísticas Gerais

Nesta seção, são apresentadas as estatísticas gerais dos dados coletados para os dispositivos Smart TV e Chromecast, sem considerar o horário em que os dados foram gerados. As análises incluem cálculos de medidas descritivas, como média, variância e desvio padrão, além de representações gráficas através de histogramas, boxplots e funções de distribuição empírica (ECDF).

2.1 Medidas Descritivas

As medidas descritivas para as taxas de upload e download (em escala logarítmica base 10) estão resumidas na Tabela 1. Para determinar o número adequado de bins para os histogramas, utilizamos a fórmula de Sturges:

$$k = 1 + \log_2(n), \tag{1}$$

onde n é o número total de amostras. Este método busca otimizar a visualização dos dados ao balancear granularidade e clareza.

Tabela 1: Medidas descritivas das taxas de upload e download.

Dispositivo	Tipo de Tráfego	Média	Desvio Padrão
Smart TV	Upload	2.16	2.03
Smart TV	Download	2.35	2.59
Chromecast	Upload	3.35	0.68
Chromecast	Download	3.80	1.29

O número de bins calculado para cada dispositivo, arredondando para cima, é o seguinte:

- **Smart TV:** $k = 24$ bins.
- **Chromecast:** $k = 22$ bins.

2.2 Visualizações Gráficas

Para compreender melhor a distribuição dos dados, utilizamos as seguintes representações gráficas:

- **Histogramas:** As distribuições das taxas de upload e download para cada dispositivo estão representadas nos histogramas da Figura 1.
- **Boxplots:** A Figura 2 mostra os boxplots comparando as taxas de upload e download entre Smart TV e Chromecast.
- **ECDF:** As funções de distribuição empírica, exibidas na Figura 3, demonstram a probabilidade acumulada para cada valor das taxas.

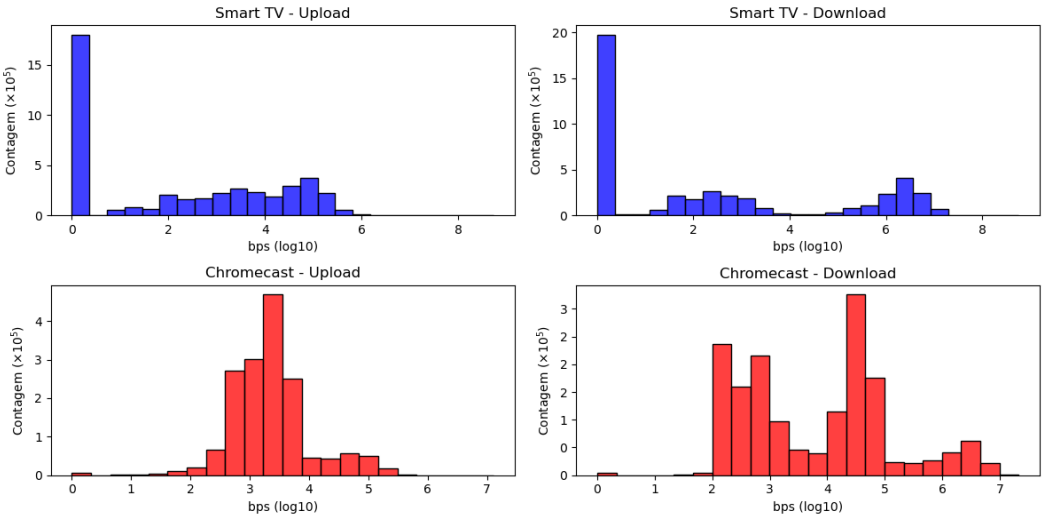

Figura 1: Histogramas das taxas de upload e download para Smart TV e Chromecast.

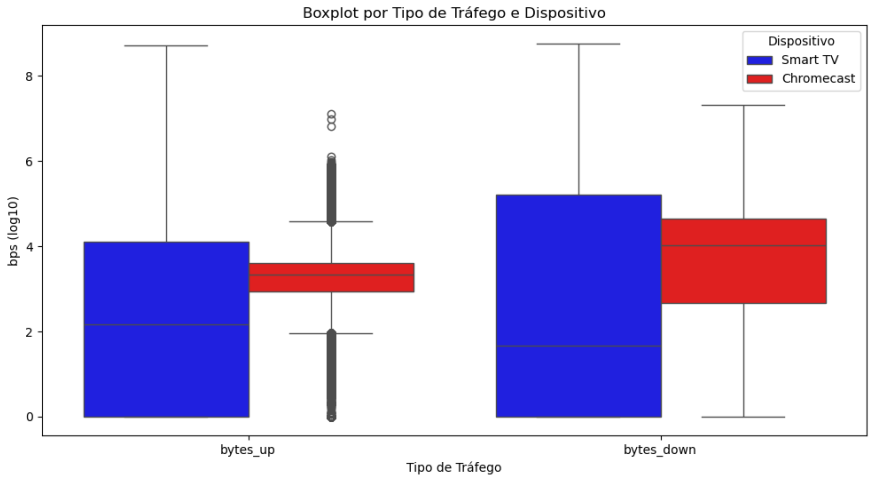

Figura 2: Boxplots das taxas de upload e download para Smart TV e Chromecast.

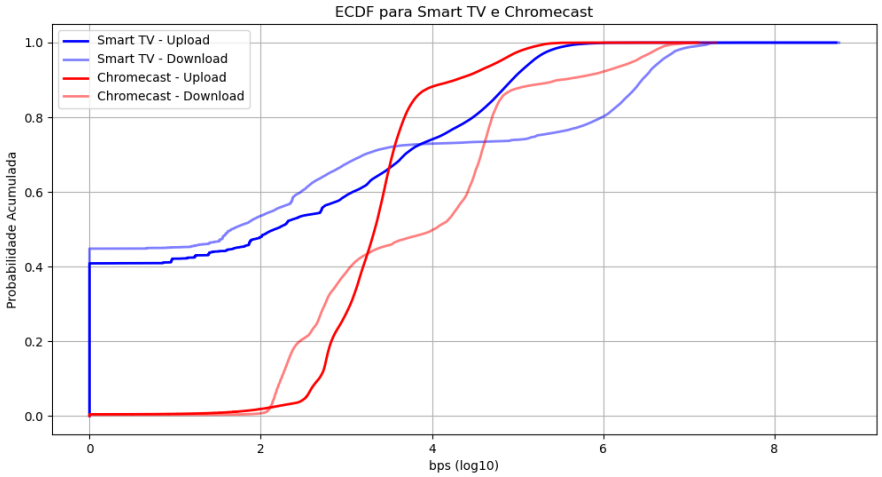

Figura 3: Funções de Distribuição Empírica (ECDF) das taxas de upload e download.

2.3 Análise dos Resultados

Os resultados indicam que as taxas de upload e download apresentam características distintas entre os dispositivos:

- **Smart TV:** As taxas exibem maior variação, refletida em desvios padrão mais elevados.
- **Chromecast:** As taxas são mais concentradas, indicando menor dispersão nos valores.

Essas diferenças podem ser atribuídas aos diferentes padrões de uso e capacidades técnicas dos dispositivos, como consumo contínuo em Smart TVs e interações esporádicas em Chromecasts.

3 Códigos

Os códigos utilizados em todas as etapas deste projeto estão disponíveis no repositório do GitHub: `https://github.com/lhscaldas/Projeto_Probabilidade_e_Estatistica`