

**آزمایشگاه تحلیل پیشرفته کلان داده (ابدال)**

شناسایی حملات در داده‌های شبکه با استفاده از روش‌های داده‌کاوی

درس: داده‌کاوی

استاد درس: دکتر حسین رحمانی

|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
| نسخه: | 1 |

فهرست مطالب

[1- مقدمه 1](#_Toc100947807)

[2- فایل ورودی 1](#_Toc100947808)

[3- توضیحات مجموعه‌داده 1](#_Toc100947809)

[3-1- حملات مبتنی بر بازتاب 2](#_Toc100947810)

[3-2- حملات مبتنی بر بهره‌برداری 2](#_Toc100947811)

[3-3- بررسی مجموعه‌ داده 3](#_Toc100947812)

[4- شرح مسئله 7](#_Toc100947813)

[4-1- گام اول 7](#_Toc100947814)

[4-2- گام دوم: پیش پردازش 7](#_Toc100947815)

[4-3- گام سوم: استفاده از روش‌های داده‌کاوی 8](#_Toc100947816)

[4-4- گام چهارم: آزمایش و تعیین معیارهای یادگیری مدل 8](#_Toc100947817)

[4-5- امتیازی 8](#_Toc100947818)

# مقدمه

امروزه اینترنت یکی از قسمت‌های اصلی جامعه را تشکیل می‌دهد. با توجه به فراگیر بودن اینترنت، دردسترس بودن آن یک امر ضروری به شمار می‌رود. از طرفی مهاجمان به دنبال از دسترس خارج کردن خدمات اینترنتی و سوءاستفاده از شرکت‌های خدمات اینترنتی هستند. اختلال و قطعی سرویس، ضرر‌های زیادی به شرکت وارد می‌کند تا جایی که 24 ساعت قطعی سرویس در یک شرکت بزرگ تجارت الکترونیک می‌تواند میلیون‌ها دلار ضرر به شرکت برساند. مهاجمان از ابزار‌ها و روش‌های مختلف جهت حمله به شبکه‌ها و زیرساخت‌های شرکت‌های ارائه کننده خدمات استفاده می‌کنند. به آن حملات، ناهنجاری در ترافیک شبکه نیز گفته می‌شود. به طور‌کلی، ناهنجار‌ی‌ها یا حملات، رویداد‌های شبکه هستند که از رفتار عادی مورد انتظار، منحرف می‌شوند و از نظر امنیتی مشکوک هستند. در سال‌های اخیر روش‌های محاسباتی بسیاری برای پیش‌بینی ناهنجاری در شبکه ایجاد شده است.

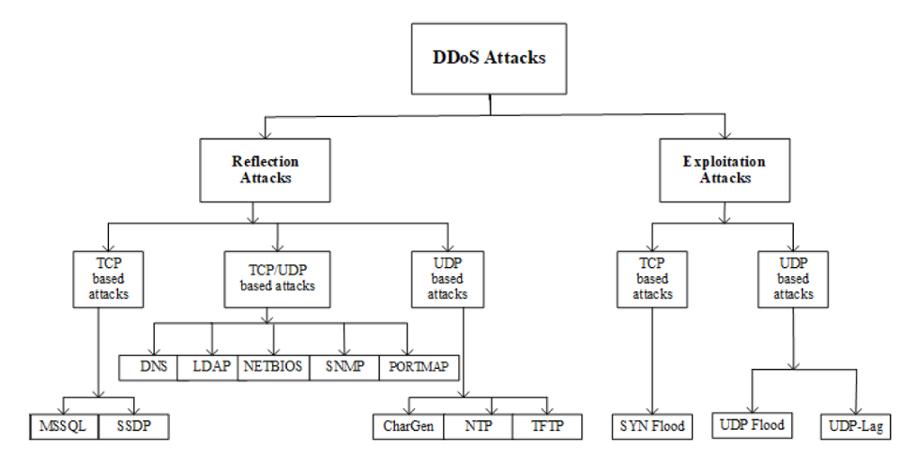
# فایل ورودی

لینک دانلود دیتاست:

<http://205.174.165.80/CICDataset/CICDDoS2019/Dataset/CSVs/CSV-03-11.zip>

# توضیحات مجموعه‌داده

مجموعه ‌داده‌هایی که ارائه شده است دامنه حملات محدود دارند و نیاز به شناسایی حملات جدید و ارائه طبقه‌بندی‌های جدید وجود دارد. از این رو، ما حملات جدیدی را می‌توانیم با استفاده از پروتکل‌های مبتنی‌بر TCP/UDP در لایه برنامه[[1]](#footnote-2) تجزیه و تحلیل کنیم و به یک طبقه‌بندی جدید پیشنهادی برسیم. همان‌طور که در شکل ۱ مشاهده می‌کنید حملات به دو دسته حملات مبتنی بر بازتاب و حملات مبتنی بر بهره‌برداری تقسیم می‌شوند.



شکل 1: دسته‌بندی حملات در مجموعه‌داده CICDDoS2019

## حملات مبتنی بر بازتاب

آن دسته از حملاتی هستند که در آن‌ها هویت مهاجم با استفاده از شخص ثالث قانونی پنهان می‌ماند. این حملات را می‌توان از طریق پروتکل‌های لایه کاربردی با استفاده از پروتکل‌های لایه انتقال، TCP،UDP یا ترکیبی از هر دو انجام داد. همان‌طور که در شکل ۱ مشا‌هده می‌کنید این دسته از حملات می‌تواند مبتنی بر TCP باشد که شامل MSSQL، SSDP است. دسته بعدی حملات مبتنی بر UDP‌ است که شامل CharGen، NTP و TFTP است و دسته آخر حملات با استفاده از هر دو پروتکل قابل انجام است که شامل DNS، LDAP، NETBIOS و SNMP است.

## حملات مبتنی بر بهره‌برداری

در این دسته از حملات بسته‌ها توسط مهاجمان به سرورهای بازتابنده ارسال می‌شوند که آدرس IP مبدأ روی آدرس IP قربانی هدف تنظیم شده‌ است تا قربانی را با بسته‌های پاسخ[[2]](#footnote-3) غرق کنند. این نوع حملات با توجه به پروتکل انتقال به دو دسته TCP و UDP تقسیم می‌شوند. حملات بهره برداری مبتنی بر TCP شامل SYN flood و حملات مبتنی بر UDP شامل UDP flood و UDP- Lag هستند. حمله UDP-Lag حمله‌ای است که ارتباط بین مشتری و سرور را مختل می‌کند. این حمله بیشتر در بازی‌های آنلاین مورد استفاده قرار می‌گیرد.

## بررسی مجموعه‌ داده

این مجموعه ‌داده شامل ۷ فایل با نام‌های LDAP، MSSQL، NetBIOS، Portmap، Syn، UDPLag، UDP با پسوند csv می‌باشد. همچنین این مجموعه داده شامل ۸۸ ویژگی به شرح زیر می‌باشد:

|  |  |
| --- | --- |
| نام ویژگی | شرح ویژگی |
| ACK Flag Count | تعداد بسته‌های دارای ACK |
| Active Max | حداکثر زمانی که یک جریان قبل از بیکار شدن فعال شده |
| Active Mean | میانگین زمانی که یک جریان قبل از بیکار شدن فعال شده |
| Active Min | حداقل زمانی که یک جریان قبل از بیکار شدن فعال شده |
| Active Std | انحراف معیار زمانی که یک جریان قبل از بیکار شدن فعال شده |
| Bwd IAT Max | حداکثر زمان بین دو بسته ارسال شده در جهت عقب[[3]](#footnote-4) |
| Bwd IAT Mean | میانگین زمان بین دو بسته ارسال شده در جهت عقب |
| Bwd IAT Min | حداقل زمان بین دو بسته ارسال شده در جهت عقب |
| Bwd IAT Std | زمان انحراف استاندارد بین دو بسته ارسال شده در جهت عقب |
| Bwd IAT Total | کل زمان بین دو بسته ارسال شده در جهت عقب |
| Bwd Packets/s | تعداد بسته های در جهت عقب در ثانیه |
| down/Up Ratio | نسبت دانلود و آپلود |
| ECE Flag Count | تعداد بسته‌های دارای پرچم ECE |
| FIN Flag Count | تعداد بسته‌های دارای پرچم FIN |
| Flow Bytes/s | تعداد بایت‌های جریان در ثانیه |
| Flow duration | مدت زمان جریان در میکروثانیه |
| Flow IAT Max | حداکثر زمان بین دو بسته ارسال شده در جریان |
| Flow IAT Mean | میانگین زمان بین دو بسته ارسال شده در جریان |
| Flow IAT Min | حداقل زمان بین دو بسته ارسال شده در جریان |
| Flow IAT Std | انحراف معیار زمان بین دو بسته ارسال شده در جریان |
| Average Packet Size | اندازه متوسط ​​بسته |
| Bwd Bulk Rate Avg | میانگین تعداد بایت‌ها نرخ انبوه[[4]](#footnote-5) |
| Bwd Header Length | کل بایت‌های استفاده شده برای سرآیندها در جهت عقب |
| Bwd Init Win bytes | تعداد کل بایت‌های ارسال شده در پنجره اولیه در جهت عقب |
| Bwd Packet Length Max | حداکثر اندازه بسته در جهت عقب |
| Bwd Packet Length Mean | اندازه متوسط ​​بسته در جهت عقب |
| Bwd Packet Length Min | حداقل اندازه بسته در جهت عقب |
| Bwd Packet Length Std | اندازه انحراف معیار بسته در جهت عقب |
| Bwd Segment Size Avg | میانگین تعداد بایت‌های نرخ انبوه در جهت عقب |
| Bwd PSH Flags | تعداد دفعاتی که پرچم PSH در بسته‌هایی که در جهت عقب حرکت می‌کنند تنظیم شده‌ است (0 برای UDP) |
| Bwd URG Flags | تعداد دفعاتی که پرچم URG در بسته‌هایی که در جهت عقب حرکت می‌کنند تنظیم شده‌است (0 برای UDP) |
| CWR Flag Count | تعداد بسته‌های دارای CWR |
| Destination IP | آدرس IP که بسته‌ها به آن ارسال شده‌ است |
| Destination Port | شماره پورتی که بسته‌ها به آن ارسال می‌شود |
| Fwd Act Data Pkts | تعداد بسته‌هایی با حداقل 1 بایت بار داده TCP در جهت جلو[[5]](#footnote-6) |
| Fwd Bulk Rate Avg | میانگین تعداد نرخ انبوه در جهت رو به جلو |
| Fwd Bytes/ Bulk Avg | میانگین تعداد بایت‌ها نرخ انبوه در جهت رو به جلو |
| Fwd Header Length | کل بایت‌های استفاده شده برای سرآینده‌ها در جهت جلو |
| Flow Packets/s | تعداد بسته‌های جریان در ثانیه |
| Fwd IAT Max | حداکثر زمان بین دو بسته ارسال شده در جهت جلو |
| Fwd IAT Mean | میانگین زمان بین دو بسته ارسال شده در جهت جلو |
| Fwd IAT Min | حداقل زمان بین دو بسته ارسال شده در جهت جلو |
| Fwd IAT Std | انحراف معیار زمان بین دو بسته ارسال شده در جهت جلو |
| Fwd IAT Total | کل زمان بین دو بسته ارسال شده در جهت جلو |
| FWD Packets/s | تعداد بسته‌های رو به جلو در ثانیه |
| Fwd PSH flags | تعداد دفعاتی که پرچم PSH در بسته هایی که در جهت جلو حرکت می‌کنند تنظیم شده است (0 برای UDP) |
| Fwd URG Flags | تعداد دفعاتی که پرچم URG در بسته هایی که در جهت جلو حرکت می‌کنند تنظیم شده است (0 برای UDP) |
| Idle Max | حداکثر زمانی که یک جریان قبل از فعال شدن بیکار شده |
| Idle Mean | میانگین زمانی که یک جریان قبل از فعال شدن بیکار شده |
| Idle Min | حداقل زمانی که یک جریان قبل از فعال شدن بیکار شده |
| Idle Std | انحراف معیار زمانی که یک جریان قبل از فعال شدن بیکار شده |
| Inbound | جهت ترافیک را مشخص می‌کند |
| Label | نوع بسته ها برای طبقه بندی |
| Protocol | نوع پروتکل را مشخص می‌کند |
| PSH Flag Count | تعداد بسته با پرچم PSH |
| RST Flag Count | تعداد بسته‌های دارای RST |
| SimillarHTTP | بسته‌هایی که شامل داده‌های http می‌شود |
| Source IP | آدرس IP فرستنده بسته‌ها |
| Source Port | شماره پورت فرستنده بسته‌ها |
| Timestamp | زمان انتقال بسته‌ها |
| Flow ID | ترکیبی از پورت مقصد IP مقصد، IP مبدا و پورت مبدا |
| Fwd Init Win bytes | تعداد کل بایت‌های ارسال شده در پنجره اولیه در جهت جلو |
| Fwd Packet Length Max | حداکثر اندازه بسته در جهت جلو |
| Fwd Packet | ​​میانگین اندازه بسته در جهت جلو |
| Fwd Packet Length Min | حداقل اندازه بسته در جهت جلو |
| Fwd Packet Length Std | انحراف معیار اندازه بسته در جهت جلو |
| Fwd Packet/Bulk Avg | میانگین تعداد بسته‌ها نرخ انبوه در جهت رو به جلو |
| Fwd Seg Size Min | حداقل اندازه بخش[[6]](#footnote-7) مشاهده شده در جهت جلو |
| Fwd Segment Size Avg | اندازه متوسط ​​در جهت جلو مشاهده شده است |
| Packet Length Max | حداکثر طول یک بسته |
| Packet Length Mean | میانگین طول یک بسته |
| Packet Length Min | حداقل طول یک بسته |
| Packet Length Std | انحراف معیار طول یک بسته |
| Packet Length Variance | واریانس طول یک بسته |
| Subflow Bwd Bytes | میانگین تعداد بایت‌ها در یک جریان فرعی در جهت عقب |
| Subflow Bwd Packets | میانگین تعداد بسته‌ها در یک زیر جریان در جهت عقب |
| Subflow Fwd Bytes | میانگین تعداد بایت‌ها در یک جریان فرعی در جهت جلو |
| Subflow Fwd Packets | میانگین تعداد بسته‌ها در یک زیر جریان در جهت رو به جلو |
| SYN Flag Count | تعداد بسته‌های دارای SYN |
| Total Bwd packets | مجموع بسته‌ها در جهت عقب |
| Total Fwd Packet | مجموع بسته‌ها در جهت جلو |
| Total Length of Bwd Packet | اندازه کل بسته در جهت عقب |
| Total Length of Fwd Packet | اندازه کل بسته در جهت جلو |
| URG Flag Count | تعداد بسته های دارای URG |

# شرح مسئله

## گام اول

یکپارچه سازی هفت مجموعه داده‌ و **نمونه برداری** از مجموعه داده کلی به دلیل حجم بالای مجموعه داده.

## گام دوم: پیش پردازش

1. سطر‌هایی که در آن مقادیر پرت وجود دارد حذف شود. سطر‌هایی که حاوی مقادیر NaN هستند حذف شود.
2. حذف ویژگی‌هایی که اهمیت زیادی ندارند و به آموزش کمک نمی‌کنند.
3. حذف ویژگی‌هایی که فقط حاوی مقدار صفر می‌باشد.
4. مجموعه داده برای تشخیص و طبقه بندی حملات DDoS به دو دسته تقسیم می شود. 'BENIGN' برچسب '0' و سایر حملات با '1' در مجموعه داده ایجاد شده برای شناسایی حمله به ترافیک شبکه برچسب گذاری شده است.
5. برای اینکه مقادیر موجود در مجموعه داده بر آموزش بیش از حد تأثیر نگذارد، فرآیند نرمال سازی برروی ویژگی‌ها اعمال شود.
6. تقسیم بندی داده‌های به داده‌های train و test.

## گام سوم: استفاده از روش‌های داده‌کاوی

استفاده از تمام روش‌های داده‌کاوی تدریس شده در کلاس برای تشخیص حملات مجاز است.

## گام چهارم: آزمایش و تعیین معیارهای یادگیری مدل

بدست آوردن ماتریس در‌هم‌ ریختگی (Confusion Matrix) و بدست آوردن ۴ معیار Accuracy، Precision، Recall و F-Score.

## امتیازی

به افرادی که الگوریتم پیشنهادیشان بیشترین دقت در معیارهای تعریف شده داشته باشد. نمره جایزه تعلق می‌گیرد. به افرادی که مهم‌ترین ویژگی‌ها در مجموعه داده‌ی ارائه شده را (با ذکر دلیل) معرفی کنند. نمره تعلق میگیرد.

**\*\*\* نکته بسیار مهم: به منظور تحویل تمرین، تمامی کدهای پیاده‌سازی خود را به همراه یک گزارش کامل و تحلیلی تحویل دهید. بدیهی است در صورت کامل نبودن فایل‌های ارسالی، نمره‌ای به دانشجو تعلق نخواهد گرفت.**

1. application layer [↑](#footnote-ref-2)
2. response packets [↑](#footnote-ref-3)
3. backward direction [↑](#footnote-ref-4)
4. bulk rate [↑](#footnote-ref-5)
5. forward direction [↑](#footnote-ref-6)
6. segment [↑](#footnote-ref-7)